

Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

William Campbell

Date

Prediction Impact Curve: A New Graphical Approach Integrating Intervention Effects in
the Evaluation of Prediction Model Utility

By

William Campbell
Master of Public Health

Epidemiology

A. Cecile J.W. Janssens
Committee Chair

Prediction Impact Curve: A New Graphical Approach Integrating Intervention Effects in
the Evaluation of Prediction Model Utility

By

William Campbell

B.A., Emory University, 2011

Thesis Committee Chair: A. Cecile J.W. Janssens

An abstract of
a thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Public Health
in Epidemiology
2014

Abstract

Prediction Impact Curve: A New Graphical Approach Integrating Intervention Effects in the Evaluation of Prediction Model Utility

By William Campbell

B.A., Emory University, 2011

Traditional measures of model performance generally address discrimination and calibration, while novel measures focus on the potential for risk models to change medical decisions. This document first provides a review of current traditional and novel model performance measures. Then, we propose a graphical approach, the prediction impact curve, which evaluates the performance of risk models in terms of their expected preventive effect in the population. Using simulated data and estimates from the literature, we illustrate how the prediction impact curve is used to estimate the expected reduction in events when using a risk model to assign individuals to a preventive intervention and how to compare nested risk models. We apply the prediction impact curve to the Atherosclerosis Risk in Communities (ARIC) Study to illustrate its application toward primary prevention of coronary heart disease. We estimated that if the ARIC cohort received statin intervention at baseline, 5% of events were expected to be prevented when evaluated at a cut-off threshold of 20% predicted risk. Additionally, we estimated that an average of 15% of events were expected to be prevented when considering performance across all possible thresholds. We conclude that the prediction impact curve is a useful and intuitive graphical approach for assessing the expected performance of risk models and is most beneficial when considered alongside existing measures of model performance.

Prediction Impact Curve: A New Graphical Approach Integrating Intervention Effects in
the Evaluation of Prediction Model Utility

By

William Campbell

B.A., Emory University, 2011

Thesis Committee Chair: A. Cecile J.W. Janssens

A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Public Health
in Epidemiology
2014

ACKNOWLEDGEMENTS

I would like to thank my advisor, Dr. A. Cecile J.W. Janssens, for her time, support, and guidance. I also thank Suman Kundu and Rachel Kalf for their helpful advice.

TABLE OF CONTENTS

Chapter I: Literature Review	1
Traditional Measures of Model Performance	1
Overall performance measures and goodness-of-fit	1
Calibration	2
Discrimination	3
Limitations of AUC	3
Novel Measures of Model Performance	4
Reclassification measures	4
Net reclassification improvement	5
Integrated discrimination improvement	6
Limitations of NRI and IDI	7
Traditional decision analysis	8
Decision curve analysis and net benefit	9
Limitations of decision curve analysis	10
Chapter II: Manuscript	11
Abstract	12
Introduction	13
Materials and Methods	14
Results	18
Discussion	21
Funding and Acknowledgements	23
Chapter III: Discussion and Future Direction	24
Appendix	26
Figure Legend	26
References	27
Figures and Tables	29
IRB Exemption Letter	34

CHAPTER I: LITERATURE REVIEW

Risk models are used to predict individual risk of disease and hold great potential for clinical decision making. Current epidemiologic research is continually identifying new predictive markers and proposing their integration into existing risk models. The current challenge is evaluating the utility of a new marker when added to a risk model with established predictors because strong association between the novel marker and outcome does not imply improved performance according to traditional measures. As a result, many researchers believe that traditional measurements of model performance are insufficient in capturing the utility of an additional risk marker (1, 2).

TRADITIONAL MEASURES OF MODEL PERFORMANCE

Overall performance measures and goodness-of-fit

Overall measures of fit are fundamental in assessing model performance and are generally implemented during the model selection process. Goodness-of-fit approaches are used to measure the distance between predicted outcomes and observed outcomes, with better models having smaller distances between predicted and observed values (3).

For ordinary least squares (OLS) linear regression, R^2 is the most common overall performance measure, which is the square of the Pearson correlation coefficient.

Variations of R^2 have been developed for nonlinear regression and generally fall under the umbrella of “pseudo R^2 ” measures. Pseudo R^2 measurements can be approached in different ways depending on the particular modeling strategy (4, 5).

Common pseudo R^2 measures are often interpreted in terms of the “proportion of

explained variation” and as such, pseudo R^2 measurements retain much of their interpretability when compared to R^2 for OLS (6).

A test of fit can be achieved using deviance values, which is performed on the base model and compared to a more complex model (e.g., saturated or fully parameterized model). Another common test of fit is the Pearson χ^2 test, which calculates the probability of observing a given distance between observed and expected values. Overall tests of fit, however, are unlikely to detect small disagreements between the fitted model and observed data and should be regarded as preliminary screening tools to reject grossly inadequate models (7).

Calibration

Calibration measures quantify the agreement between predicted probabilities and observed outcomes (8). The main difference between goodness-of-fit and calibration is that the former is employed during the model selection process while the latter is an assessment of the final model. Calibration can be viewed graphically by plotting predictions on the x -axis and observed outcomes on the y -axis. “Calibration-in-the-large” refers to the intercept, a , of the calibration curve, which indicates predictions that are systematically too low or too high. The calibration slope, b , is equal to 1 for a perfectly calibrated model and a calibration slope less than 1 suggests an overestimation of model coefficients (9). A common method of calibration for binary outcomes is the Hosmer-Lemeshow (H-L) statistic, which groups deciles of predictions and compares the mean observed outcome against the mean predicted probability for each group (10). This grouping strategy, however, is arbitrary and can be imprecise (11).

Discrimination

An effective risk model will discriminate between those with and without the outcome. The most frequently used discrimination measure is the concordance statistic, which is equal to the area under the receiver operating characteristic (ROC) curve (AUC). The receiver operating characteristic curve plots sensitivity over '1-specificity' for every possible cut-off risk threshold (12). The number of risk thresholds corresponds to the number of unique combinations of predictors in the model that yield unique predicted risk scores. For each unique risk score, the sensitivity is estimated as the proportion of all events that are correctly classified by the model and '1-specificity' is the proportion of nonevents that are incorrectly classified as events. Therefore, the ROC can be seen as a graphical representation of the tradeoff between true and false positive classifications when considering a risk model over all possible thresholds.

AUC is the integral of the ROC and can be directly estimated from the data. The value of AUC is equal to the probability that the model will assign a higher risk score to a randomly chosen event than to a randomly chosen nonevent (13). However, in clinical settings, individuals are never presented in pairs and it can be argued that AUC is not clinically relevant (14).

Limitations of AUC

Current research often judges the performance of risk models solely based on AUC, however, this practice is discouraged due to inherent limitations of the measurement. As a rank statistic, AUC is not a function of the actual predicted probabilities. AUC

describes how well a model can rank order individuals, however, the probabilities themselves may not be useful (15). For example, a model that predicts a risk of 0.51 for all events and 0.50 for all nonevents would achieve perfect discrimination even though the probabilities themselves are not indicative of true underlying risk. Another limitation of AUC is that the distribution of risk in the population is not taken into account. That is, AUC does not consider the distance between predicted risks of ranked individuals. Therefore, a model that can significantly separate low vs. high-risk individuals does not yield a higher AUC than a model that can barely separate them.

The limitations of AUC are most clear in the context of model selection or when contemplating the integration of a novel risk factor into an existing model. The main criticism of AUC is that it is unresponsive to strong predictors when compared to likelihood- or deviance-based measures of fit. A novel predictor with a large effect size, for example, generally does not cause a large increase in AUC when added to an existing model with established predictors. Cook demonstrates this characteristic using the Framingham risk model for cardiovascular disease (CVD) as an illustration (15). Based on likelihood-based ratios, systolic blood pressure (SBP) was found to be the second strongest predictor of CVD after age. However, when constructing models with and without SBP, the AUC of the model only increased 1% (from 73% to 74%). Because of the relative unresponsiveness of AUC, it is poorly suited for determining inclusion of individual predictors.

NOVEL MEASURES OF MODEL PERFORMANCE

Reclassification measures

Limitations of AUC have caused investigators to turn to novel measures of model performance, such as reclassification measures. Reclassification measures are a response to the fact that AUC, and other traditional measures, do not reflect the model's ability to change clinical decisions. For example, a novel predictor may cause no change in AUC given existing predictors, but it may modify the predicted risks of certain individuals in such a way that classifies them into different treatment groups. Thus, reclassification measures attempt to measure the importance of a predictor in terms of how the updated model changes treatment decisions.

The crudest forms of reclassification measures can be calculated from classification tables. After grouping predicted risk scores into clinically meaningful categories, the investigator can calculate the percentage of individuals that change categories due to updating the model. Measuring the crude change in classification, however, is insufficient because it is ambiguous whether the direction of movement is appropriate given the individual's case status.

Net reclassification improvement

Net reclassification improvement (NRI), proposed by Pencina *et al.*, considers separately individuals with and without the outcome of interest (16). NRI, like other reclassification measures, can be viewed as a modification of discrimination measures. NRI requires two nested models in which one or more predictors have been added. NRI is evaluated at a specific risk threshold and requires the specification of clinically meaningful risk groups based on absolute predicted risk. "Upward" movement is defined as a change to a higher risk category based on the updated model and "downward" movement is defined as a

change to a lower risk category. As such, upward movement for events indicates improved classification and downward movement implies worse classification (the opposite is true for nonevents). When using sample data to estimate probabilities, NRI can be expressed:

$$NRI = (\hat{p}_{up,events} - \hat{p}_{down,events}) - (\hat{p}_{up,nonevents} - \hat{p}_{down,nonevents}) \quad (1)$$

Where the estimated probability, \hat{p} , equals the number of individuals moving in the specified direction divided by the total number of events or nonevents, as appropriate. NRI can be interpreted as the net sum of improved classification due to updating the model.

Integrated discrimination improvement

The integrated discrimination improvement (IDI) is an extension of NRI that considers discrimination across all possible risk thresholds (16). If IS denotes the integral of sensitivity over all possible thresholds and IP denotes the integral of ‘1-specificity’ then IDI is defined as follows (see (16) for derivation):

$$IDI = (IS_{new} - IS_{old}) - (IP_{new} - IP_{old}) \quad (2)$$

Where ‘new’ signifies the updated model and ‘old’ signifies the original model. Since an integral over the interval (0, 1) signifies an average, IDI can be interpreted as the difference between increased average sensitivity and increased average ‘1-specificity’ for

the original and updated models. IDI has been shown to be equivalent to the difference in discrimination slopes of two models, and to the difference in Pearson R^2 values (2).

The popularity of NRI and IDI is due in part because the metrics are more responsive when evaluating the impact of novel predictors. This becomes more clear when IDI is examined alongside change in AUC. Change in AUC and IDI can both be viewed as average sensitivities adjusted for the undesirable increase in ‘1-specificity’ that occurs during classification of individuals. AUC adjusts for ‘1-specificity’ by weighting the sensitivities at each risk threshold with the corresponding derivatives of ‘1-specificity’. IDI, however, adjusts for ‘1-specificity’ by means of subtraction, which accounts for the increased responsiveness of IDI when compared to change in AUC (11).

Limitations of NRI and IDI

Net reclassification improvement and integrated discrimination improvement were quickly adopted, but more research is needed to evaluate and refine the measurements. Pepe points out that NRI and IDI do not measure the size of movement between categories (14). A few large upward movements, for example, are indistinguishable from a medium number of small upward movements. A scatterplot of ‘new’ versus ‘old’ predictions, however, can reveal these characteristics. The author also notes that certain classifications may be more important than others. Being correctly classified into a high-risk category, for example, may be of more importance than being classified into a medium-risk category. NRI, however, treats all reclassifications equivalently if movement occurs in the correct direction.

A second limitation of reclassification measures is that “correct” classification is determined based on observed case status as opposed to true underlying risk (which is impossible to measure). In theory, the correctness of classification for an individual should be based on movement towards a category that better reflects his or her true underlying risk (17). NRI and IDI, however, define correct classification as movement toward a category that better reflects the individual’s observed outcome. Therefore, “correct” classification, as proposed by Pencina, *et al.*, may in reality be incorrect when considering true underlying risk.

A third limitation of reclassification measures is that improved classification does not necessarily imply improvement in model performance when compared to traditional measures (14). The amount of reclassification is determined in part by the choice in risk categories as well as correlation between risk predictors (18). As mentioned earlier, reclassification measures are often used because of their responsiveness to changes in the base model, however, this responsiveness also has the potential to yield overly optimistic results compared to other measures (17). Lastly, reclassification measures depend on the existence of clinically meaningful risk categories, which are rarely available or agreed upon.

Traditional decision analysis

Decision analysis addresses the fact that AUC, sensitivity, and specificity do not directly measure the clinical value of the model. AUC is concerned with predictive accuracy of the model and is weighted against the consequence of false positive classification. AUC does not account for additional harms, however, such as false negative classification and

cost of the intervention. Traditional decision analysis generally cannot be accomplished without gathering additional data, such as cost of the intervention or quality adjusted life-years saved (19). Typically, decision analysis requires a binary outcome (or dichotomization) in order to calculate rates of negative and positive classifications.

Decision curve analysis and net benefit

Decision curve analysis was developed by Vickers, *et al.* in order to quantify model performance in clinically appropriate terms (19). Decision curve analysis compensates for the limitations of traditional decision analysis by using the theoretical relationship between benefits and harms rather than collecting additional data. The relationship between benefits and harms requires the specification of a threshold of predicted risk (p_t) at which the benefits of treatment are equal to the harms from the perspective of the analyst or clinician.

The net benefit of using a risk model is calculated as follows (see (19) for full explanation):

$$Net\ Benefit = TPR - FPR \left(\frac{p_t}{1-p_t} \right) \quad (3)$$

where TPR denotes the proportion of true positives, FPR denotes the proportion of false positives, and p_t signifies the threshold of evaluation. FPR is subtracted from TPR and weighted by the relative harm of a false positive and false negative classification. To construct a decision curve, p_t is first specified at a value of predicted risk where the benefits and harms of intervention are perceived to be equal. Second, net benefit is

calculated and recorded. Third, the analyst varies p_t and records resulting values of net benefit. Lastly, the analyst plots values of net benefit over p_t and compares the curve to hypothetical situations in which all or none of the patients are treated. The resulting graph shows the estimated values of net benefit across multiple values of p_t .

Limitations of decision curve analysis

Decision curve analysis has not been widely adopted in prediction research due in part to its inherent limitations and assumptions. First, determining p_t requires a comprehensive understanding of the benefits and harms of the intervention. Second, decision curve analysis assumes that the relationship between benefits and harms is approximately equal for all individuals. And third, predicted probabilities are assumed to be independent of p_t (19).

CHAPTER II: MANUSCRIPT

Prediction Impact Curve: A New Graphical Approach Integrating Intervention Effects in the Evaluation of Prediction Model Utility.

William Campbell¹, Andrea Ganna², A. Cecile J.W. Janssens^{1*}, Erik Ingelsson³

¹ Department of Epidemiology, Rollins School of Public Health, Emory University, Atlanta GA, USA.

² Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden;

³ Department of Medical Sciences, Molecular Epidemiology and Science for Life Laboratory, Uppsala University, Uppsala, Sweden;

* Correspondence to

Professor A. Cecile J. W. Janssens, Department of Epidemiology, Rollins School of Public Health, Emory University, 1518 Clifton Road NE, Atlanta, Georgia 30322, USA.

E-mail: cecile.janssens@emory.edu, Telephone: +1 404 727 6307, Fax: +1 404 727 8737

Abbreviations:

ROC: Receiver operating characteristic curve

AUC: Area under the receiver operating characteristic curve

NRI: Net reclassification improvement

PIC: Prediction impact curve

AUPIC: Area under the prediction impact curve

ABSTRACT

Traditional measures of model performance generally address discrimination and calibration, while novel measures focus on the potential for risk models to change medical decisions. We propose a graphical approach, the prediction impact curve, which evaluates the performance of risk models in terms of their expected preventive effect in the population. Using simulated data and estimates from the literature, we illustrate how the prediction impact curve is used to estimate the expected reduction in events when using a risk model to assign individuals to a preventive intervention and how to compare nested risk models. We apply the prediction impact curve to the Atherosclerosis Risk in Communities (ARIC) Study to illustrate its application toward primary prevention of coronary heart disease. We estimated that if the ARIC cohort received statin intervention at baseline, 5% of events were expected to be prevented when evaluated at a cut-off threshold of 20% predicted risk. Additionally, we estimated that an average of 15% of events were expected to be prevented when considering performance across all possible thresholds. We conclude that the prediction impact curve is a useful and intuitive graphical approach for assessing the expected performance of risk models and is most beneficial when considered alongside existing measures of model performance.

INTRODUCTION

The performance of risk models is evaluated in terms of clinical validity and clinical utility. To assess the clinical validity and utility of risk models, several traditional and novel measurements are used (see (20) for a review). The most frequently used measure of clinical validity is the area under the receiver operating characteristic (ROC) curve (AUC or c-statistic) (12). AUC quantifies the ability of a risk model to discriminate between individuals who will or will not manifest the outcome of interest, and the increase in AUC between two nested models indicates the improvement in discrimination offered by additional predictors. Although widely used, the measure is criticized for unresponsiveness when used to detect the added value of major risk factors (1).

Reclassification measures were proposed in order to quantify the influence of an updated model on treatment decisions (16). A commonly reported measure of reclassification, net reclassification improvement (NRI), considers reclassification separately for individuals with and without the outcome. NRI requires the formation of clinically meaningful risk categories based on absolute predicted risk and “upward” movement is defined as a change to a higher risk category and “downward” movement is defined as a change to a lower risk category. As such, upward movement for events indicates improved classification and downward movement implies worse classification (the opposite is true for nonevents). NRI can then be interpreted as the net sum of improved classification due to updating the model.

AUC and NRI are widely reported summary statistics, although these measures can lack intuitive interpretations. We propose a new curve, the prediction impact curve

(PIC), which is intended as an intuitive graph of the expected preventive effect that results from using a risk model to assign at-risk individuals to an intervention.

In this paper, we illustrate how the PIC is constructed, demonstrate its interpretation, and propose to calculate the area underneath as a summary statistic. Second, we investigate the properties of the area under the prediction impact curve (AUPIC) in relation to its determinants. Lastly, we apply the PIC to data from the Atherosclerosis Risk in Communities (ARIC) Study and illustrate how this graphical approach can be used to calculate the expected reduction in coronary heart disease (CHD) events when using a risk model to assign statin treatment at baseline.

MATERIALS AND METHODS

Constructing the prediction impact curve

The prediction impact curve plots the size of the risk group against event reduction, which is the percentage of events expected to be prevented when the intervention is given to the risk group (**Figure 1**). Event reduction is obtained for every possible risk group size, which is the percentage of total individuals with predicted risks that are higher than the risk threshold. The smallest increment in risk group size is achieved by adding the next ranked individual to the risk group (individuals are ranked by decreasing predicted risk). Three parameters must be considered in order to construct the prediction impact curve: the sensitivity of the risk model, the preventive fraction (PF) of the intervention, and the event incidence in the population.

First, the sensitivity for a given threshold is the proportion of events that the risk model correctly classifies as being in the risk group. When the entire risk group receives

the intervention, the sensitivity is also interpreted as the proportion of events that is assigned to the intervention.

Second, the preventive fraction is the proportion of events expected to be prevented by the intervention (21). PF is estimated as $1-RR$, where RR is the risk ratio obtained from, e.g., randomized controlled trials (RCT) that investigated the intervention effect. Note that the choice of RCT implicitly dictates which treatment is assumed for the individuals that are not selected for the risk group, which may be alternative intervention, usual care, or no intervention. Event reduction is the product of PF and sensitivity for a given risk group size. PF is assumed to be independent of risk and remains constant across risk group sizes.

Third, event incidence affects the PIC because it limits event reduction when the size of the risk group is smaller than the percentage of individuals that will develop the disease. Even a perfect model can only reach maximum event reduction when the percentage of individuals assigned to the intervention is equal to or larger than the event incidence (upper boundary in **Figure 1**). In **Figure 1**, for example, when the size of the risk group was 10%, the perfect model could only achieve 10% event reduction even though the PF was 0.20.

Quantifying the area under the prediction impact curve

The area under the prediction impact curve can be estimated from existing data by averaging the sensitivities for all possible risk group sizes and multiplying by the preventive fraction.

$$AUPIC = \frac{\sum Se}{n} * PF \quad (4)$$

Where Se denotes the sensitivity for a given risk group size, PF denotes the preventive fraction associated with the intervention, and the denominator is represented by n because the total number of risk group sizes is equal to the number of individuals in the population.

It is necessary to consider the theoretical minimum and maximum AUPIC for a given scenario. The minimum AUPIC is the area under the PIC had no risk model been used (AUC=0.5, diagonal line in **Figure 1**) and is equal to PF divided by 2. The maximum AUPIC is the area under the PIC corresponding to a perfect model and is determined by event incidence and PF:

$$AUPIC_{max} = PF - \left[\frac{Incidence * PF}{2} \right] \quad (5)$$

Simulated data

To investigate the properties of the prediction impact curve, we used a simulation method that allowed us to systematically vary all relevant parameters. To construct simulated datasets, we adopted a modeling procedure that is described in detail elsewhere (22), and for which the function is available in the R package *PredictABEL* (23). This procedure was originally created for simulating a dataset containing individual genotype data, but can be used to obtain risk data of any kind. It requires the specification of four parameters: frequencies and ORs of the predictors, population size, and event incidence. In order to obtain the desired AUC, we varied OR and frequencies of the predictors; that

is, we added as many predictors to the risk model until the AUC reached the predefined value.

Analyses

To demonstrate the interpretation of prediction impact curves, a scenario was simulated for a population of 100,000 individuals in which event incidence was specified at 20% and the hypothetical intervention had a PF of 0.20. Prediction impact curves corresponded to two risk models: one with AUC of 0.65 and another with improved AUC of 0.75. The curves were interpreted and compared using various approaches.

To further investigate the effects of event incidence, PF, and AUC on the prediction impact curve, four scenarios (*a*, *b*, *c*, and *d*) were simulated, using populations of 100,000 individuals, that systematically varied the parameters of interest. Scenarios *a* and *b* considered a hypothetical intervention with PF of 0.20 and discordant event incidences (10% and 40%, respectively), whereas scenarios *c* and *d* considered an intervention with PF of 0.60 and discordant event incidences (10% and 40%, respectively). Four prediction impact curves were plotted for each scenario, which corresponded to risk models with varying AUC (0.60-0.90). The AUPIC was calculated for all prediction impact curves and trends were reported.

Illustration for prevention of coronary heart disease

In order to illustrate a practical application, we applied the prediction impact curve to data originating from the Atherosclerosis Risk in Communities (ARIC) Study. The ARIC Study is a prospective study of cardiovascular disease in a cohort of 15,792

individuals sampled from four U.S. communities in 1987-1989. Follow-up was conducted through 1998, for a median of 10.2 years. The sample consisted of 45-64 year-old men and women who underwent three follow-up examinations in 1990-1992, 1993-1995, and 1996-1995.

A 10-year risk model for CHD was derived from these data by Chambless, *et al.*, and included the following main predictors: age, total cholesterol, HDL cholesterol, blood pressure, and smoking (24). Predicted risks for individuals were obtained from sex- and race-specific Cox regression models. A CHD event was defined in detail by Chambless, *et al.* and can be simplified as an individual that experienced myocardial infarction or CHD related death (see (24) for a full review of event ascertainment).

All individuals with missing outcome or predicted risk were excluded, which included those with missing values for predictors and those with preexisting CHD at baseline. Our illustration concerned the potential effect of statin treatment on the ARIC cohort if prescribed at baseline, which has a PF of 0.20 (RR=0.80) according to clinical trials (25). Because of this, we further excluded individuals that used statins prior to baseline or follow-up visits. Outcomes and predicted risks from the resulting population were used to construct the PIC and calculate the AUPIC. Due to the unavailability of an established cut-off threshold for this model, we chose an arbitrary threshold of 20% predicted risk at which to interpret the curve.

RESULTS

Interpreting the prediction impact curve

The prediction impact curve indicates the event reduction that is achieved when a certain percentage of the population receives the intervention, and vice versa. **Figure 2** reflects a simulated scenario in which event incidence was specified at 20% and the hypothetical intervention had a PF of 0.20. First consider the PIC corresponding to the risk model with AUC of 0.65 (solid curve). **Figure 2** shows that when the size of the risk group comprised 20% of the population, the expected event reduction was 7%. When no risk model was used, meaning that a random 20% of individuals received the intervention, the expected event reduction was 4% and when a perfect model was used, the expected event reduction was 20%. Thus, using the risk model reduced 3% more events than using no model. Alternatively, if the aim were to reduce 7% of events, 20% of the population needed treatment when using the model compared to 34% when no model was used.

The PIC can also be used to quantify the change in event reduction that results from adding predictors to an existing model. **Figure 2** shows that increasing the AUC from 0.65 to 0.75 led to an expected 2% increase in event reduction (from 7% to 9%) when the size of the risk group was 20%. Alternatively, if the aim were to reduce 7% of events, 7% fewer people (20% versus 13%) needed treatment using the updated model in order to achieve the same event reduction.

The PIC in relation to its determinants

Figure 3 demonstrates how the PIC varies with event incidence, PF, and AUC of the model. Three trends were apparent after systematically varying each parameter. First, event incidence inhibited the PIC and restricted the area underneath because prevention is suboptimal when the size of the risk group is lower than the incidence (**Figures 3a** and **3c**

versus **Figures 3b** and **3d**). Second, PF determined the absolute event reduction but did not impact the overall shape of the prediction impact curve (**Figures 3a** and **3b** versus **Figures 3c** and **3d**). Third, risk models with higher AUC obtained higher values of event reduction across the entire interval (0%-100%).

The AUPIC in relation to its determinants

Table 1 presents the areas under the prediction impact curves for the four scenarios presented in **Figure 3**. Comparing between and within scenarios reveals certain trends regarding the AUPIC and its theoretical maximum and minimum. The minimum AUPIC for a given intervention was based solely on PF and remained constant regardless of event incidence or AUC of the model. The maximum AUPIC was constant for a given combination of incidence and PF, regardless of the AUC of the risk model.

When all other parameters were held constant, three conclusions were drawn regarding the AUPIC in relation to each of its determinants. First, the AUPIC decreased as the event incidence increased. Second, the AUPIC was larger for interventions with a larger preventive effect, or PF. Third, the AUPIC increased as the AUC of the model increased, reflecting improved sensitivity over the entire interval (0%-100%).

Illustration for prevention of coronary heart disease

After exclusions (n=5,729), the final dataset contained outcomes and predicted risks for 10,063 individuals from the ARIC cohort for a mean follow-up time of 10.9 years. The 10-year risk model had an AUC of 0.77 when applied to the final dataset. **Figure 4** summarizes event reduction across all risk group sizes (0%-100%) for the 10-year risk

model. By the end of 10.9 years CHD incidence was 5%, of which the model assigned 24% to the risk group at baseline (sensitivity=24%) based on a threshold of 20% predicted risk. Choosing a cut-off threshold of 20% predicted risk resulted in a risk group that comprised about 7% of the cohort. When considering a statin intervention (PF=0.20) at baseline, 5% of events were expected to be prevented using the model compared to 1% when no model was used.

Alternatively, if the aim were to reduce 15% of events in the population, an expected 36% of the population would need to be treated using the model compared to 75% when no model was used. When considering performance across the entire interval, average event reduction was 15% (AUPIC=0.15), which suggests that using the model has the potential to prevent an additional 5% of events on average compared to using no model (AUPIC_{min}=0.10).

DISCUSSION

The prediction impact curve quantifies the expected reduction in events due to an intervention program. Similar to ROC, the prediction impact curve allows for calculations across multiple thresholds and, thus, allows for the calculation of a summary statistic. In contrast to existing measures of model performance, the PIC considers preventive effects of an intervention when examining a model's utility.

The PIC has several limitations. First, we assumed the PF of the intervention to be independent of risk. Second, the PIC is essentially a form of sensitivity analysis and does not consider possible harms of the intervention. Therefore, the PIC is not designed

to weigh benefits and harms of using a risk model, but rather to describe the sensitivity of the model in clinically relevant terms.

The PIC offers information about risk models that may not be immediately apparent when using traditional measures, such as AUC and NRI. The PIC adds interpretability to change in AUC and offers added information regarding the role of event incidence when assigning interventions. Additional research, however, is needed in regards to the statistical properties of the PIC.

In conclusion, the prediction impact curve is an exceptionally intuitive form of sensitivity analysis based on the ultimate goal of risk models in public health. The additional information provided by the PIC may prove to be useful for researchers and clinicians when evaluating the performance of risk models and novel predictors.

FUNDING

This work was supported by The European Community's Seventh Framework Programme (FP7/2007-2013), ENGAGE Consortium, grant agreement HEALTH-F4-2007-201413, the Swedish Research Council (project grant no. 2012-1397), the Swedish Heart-Lung Foundation (project grant no. 20120197), the National Cancer Institute at the National Institutes of Health (grant number HHSN261201200425P) and the Vidi grant from the Netherlands Organisation for Scientific Research.

ACKNOWLEDGEMENTS

The authors thank Dr. Marie Really, Dr. Arvid Sjölander, Rachel Kalf and Suman Kundu for their helpful advice.

Conflicts of interest: None declared.

CHAPTER III: DISCUSSION AND FUTURE DIRECTION

The prediction impact curve addresses the current needs of prediction research by quantifying model performance in clinically relevant terms. Unlike existing measurements of model performance, the PIC integrates the effect of an intervention by predicting the expected reduction in events due to its implementation at baseline.

The PIC is likely to receive criticism for not incorporating harms in its calculation. This is a valid point because, given a situation without harms, it is most beneficial to treat all individuals in the population. In essence, the PIC describes the sensitivity of the model as a proportion of the preventive fraction of the intervention. As such, the PIC has similar limitations as traditional sensitivity analysis and does not pretend to weigh benefits and harms of the intervention. Since the values on the prediction impact curve correspond to specific risk thresholds, they can be directly compared to other measurements that assess harms of the model at one or more thresholds.

Among the main advantages of the PIC are its intuitive approach and straightforward interpretation. The PIC and AUPIC may prove to be useful for researchers, clinicians, and students to conceptualize changes in sensitivity and the expected impact of such changes on the population. Furthermore, the PIC could help bridge the gap between prediction research and the adoption of risk models in clinical practice.

The concept behind the PIC could potentially be applied at a broader level. For example, plotting '1-specificity' as a proportion of total cost of the intervention could estimate the proportion of total cost that goes towards unnecessary treatment. The PIC could potentially be the first of a broader family of measurements that examine

sensitivity and specificity as proportions of clinically relevant factors. Future research should focus on developing measures of model performance that are intuitive and applicable to clinical settings.

APPENDIX

FIGURE LEGEND

Figure 1. Prediction impact curve for one risk model.

Legend: The plot represents a scenario in which event incidence was specified at 20% and the hypothetical intervention had a PF of 0.20. Prediction impact curve corresponds to a risk model with AUC of 0.65.

Figure 2. Prediction impact curve for two nested risk models with AUC of 0.65 and 0.75.

Legend: The plot represents a scenario in which event incidence was specified at 20% and the hypothetical intervention had a PF of 0.20. Prediction impact curves correspond to risk models with AUC of 0.65 and 0.75.

Figure 3. Prediction impact curves for varying event incidence, PF, and AUC.

Legend: The plots represent four different scenarios with varying event incidence and PF. Prediction impact curves correspond to risk models with varying AUC (0.60-0.90).

Figure 4. Prediction impact curve for 10-year CHD risk model derived from Artherosclerosis Risk in Communities Study data.

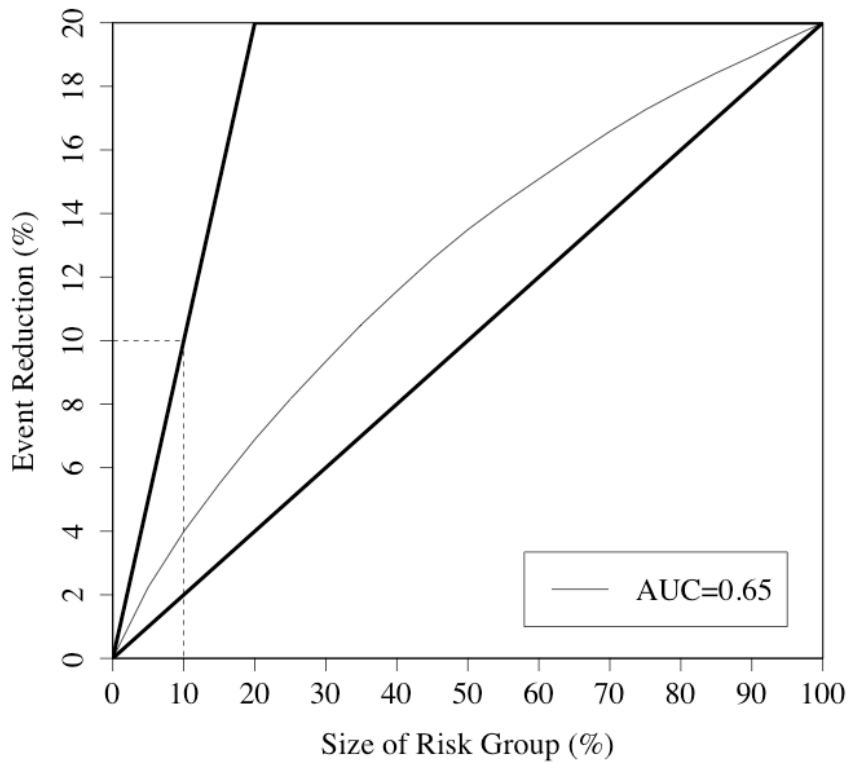
Legend: The plot represents a scenario in which CHD incidence was 5% and the statin intervention had a PF of 0.20. Prediction impact curve corresponds to the 10-year CHD risk model derived from the ARIC data.

REFERENCES

1. Cook, N.R., *Use and misuse of the receiver operating characteristic curve in risk prediction*. *Circulation*, 2007. **115**(7): p. 928-35.
2. Pepe, M.S., et al., *Integrating the predictiveness of a marker with its performance as a classifier*. *Am J Epidemiol*, 2008. **167**(3): p. 362-8.
3. Legates, D.R., *Evaluating the use of "goodness - of - fit" measures in hydrologic and hydroclimatic model validation*. Vol. 35. 1999, Water Resources Research.
4. Nagelkerke, N.J.D., *A Note on a General Definition of the Coefficient of Determination*. *Biometrika*, 1991. **78**(Sep.): p. 691-692.
5. Menard, S., *Coefficients of Determination for Multiple Logistic Regression Analysis*. *The American Statistician*, 2000. **54**: p. 17-24.
6. Magee, L., *R2 Measures Based on Wald and Likelihood Ratio Joint Significance Tests*. *The American Statistician*, 1990. **44**(No. 3): p. 250-253.
7. Rothman, K.J., S. Greenland, and T.L. Lash, *Modern epidemiology*. 3rd ed. 2008, Philadelphia: Wolters Kluwer Health/Lippincott Williams & Wilkins. x, 758 p.
8. Habbema, J.D. and J. Hilden, *The measurement of performance in probabilistic diagnosis. IV. Utility considerations in therapeutics and prognostics*. *Methods Inf Med*, 1981. **20**(2): p. 80-96.
9. Vach, W., *Calibration of clinical prediction rules does not just assess bias*. *J Clin Epidemiol*, 2013. **66**(11): p. 1296-301.
10. Lemeshow, S. and D.W. Hosmer, Jr., *A review of goodness of fit statistics for use in the development of logistic regression models*. *Am J Epidemiol*, 1982. **115**(1): p. 92-106.
11. Steyerberg, E.W., et al., *Assessing the performance of prediction models: a framework for traditional and novel measures*. *Epidemiology*, 2010. **21**(1): p. 128-38.
12. Hanley, J.A. and B.J. McNeil, *The meaning and use of the area under a receiver operating characteristic (ROC) curve*. *Radiology*, 1982. **143**(1): p. 29-36.
13. Harrell, F.E., Jr., et al., *Evaluating the yield of medical tests*. *JAMA*, 1982. **247**(18): p. 2543-6.
14. Pepe, M.S., Z. Feng, and J.W. Gu, *Comments on 'Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond' by M. J. Pencina et al., Statistics in Medicine (DOI: 10.1002/sim.2929)*. *Stat Med*, 2008. **27**(2): p. 173-81.
15. Cook, N.R., *Statistical evaluation of prognostic versus diagnostic models: beyond the ROC curve*. *Clin Chem*, 2008. **54**(1): p. 17-23.
16. Pencina, M.J., et al., *Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond*. *Stat Med*, 2008. **27**(2): p. 157-72; discussion 207-12.
17. Pepe, M.S., *Problems with risk reclassification methods for evaluating prediction models*. *Am J Epidemiol*, 2011. **173**(11): p. 1327-35.
18. Janes, H., et al., *Measuring the performance of markers for guiding treatment decisions*. *Ann Intern Med*, 2011. **154**(4): p. 253-9.

19. Vickers, A.J. and E.B. Elkin, *Decision curve analysis: a novel method for evaluating prediction models*. Med Decis Making, 2006. **26**(6): p. 565-74.
20. Steyerberg, E.W., *Clinical prediction models : a practical approach to development, validation, and updating*. Statistics for biology and health. 2009, New York, NY: Springer. xxviii, 497 p.
21. Miettinen, O.S., *Proportion of disease caused or prevented by a given exposure, trait or intervention*. Am J Epidemiol, 1974. **99**(5): p. 325-32.
22. Janssens, A.C., et al., *Predictive testing for complex diseases using multiple genes: fact or fiction?* Genet Med, 2006. **8**(7): p. 395-400.
23. Kundu, S., et al., *PredictABEL: an R package for the assessment of risk prediction models*. Eur J Epidemiol, 2011. **26**(4): p. 261-4.
24. Chambless, L.E., et al., *Coronary heart disease risk prediction in the Atherosclerosis Risk in Communities (ARIC) study*. J Clin Epidemiol, 2003. **56**(9): p. 880-90.
25. Baigent, C., et al., *Efficacy and safety of cholesterol-lowering treatment: prospective meta-analysis of data from 90,056 participants in 14 randomised trials of statins*. Lancet, 2005. **366**(9493): p. 1267-78.

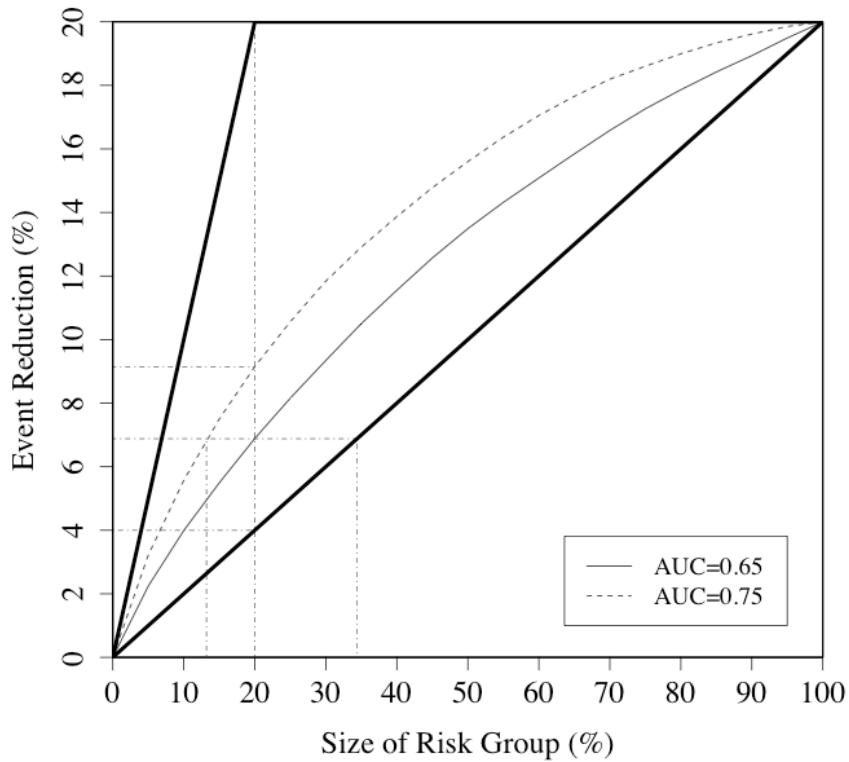
Figure 1. Prediction impact curve for one risk model.



AUC, Area under the receiver operating characteristic curve;
 PF, preventive fraction

Legend: The plot represents a scenario in which event incidence was specified at 20% and the hypothetical intervention had a PF of 0.20. Prediction impact curve corresponds to a risk model with AUC of 0.65.

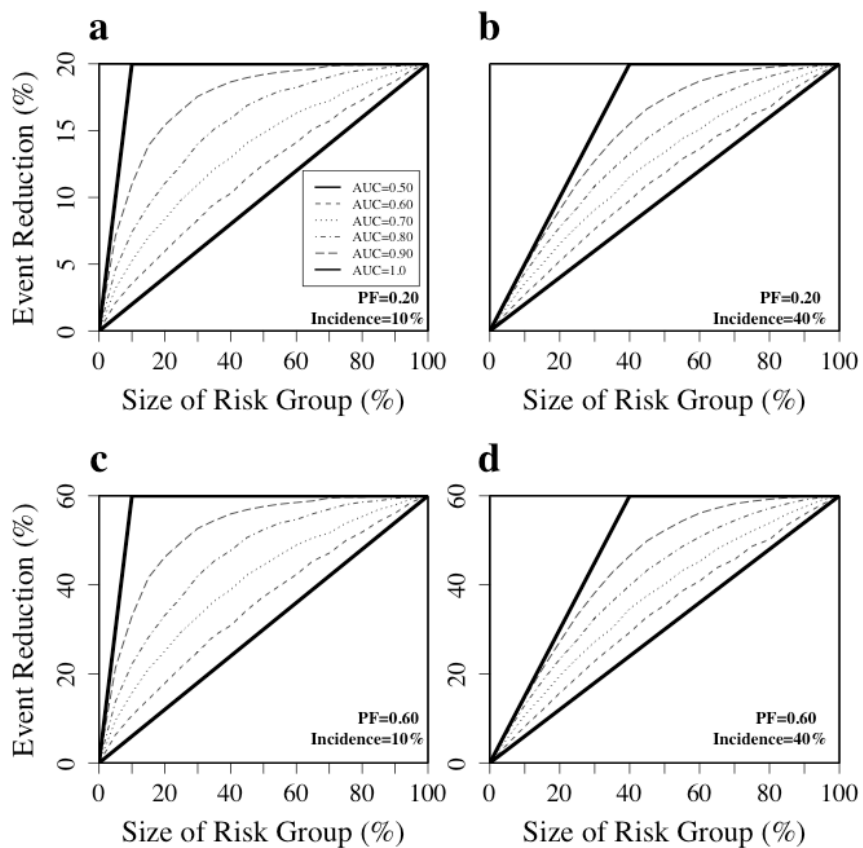
Figure 2. Prediction impact curve for two nested risk models with AUC of 0.65 and 0.75.



AUC, Area under the receiver operating characteristic curve; PF, Preventive fraction

Legend: The plot represents a scenario in which event incidence was specified at 20% and the hypothetical intervention had a PF of 0.20. Prediction impact curves correspond to risk models with AUC of 0.65 and 0.75.

Figure 3. Prediction impact curves for varying event incidence, PF, and AUC.



AUC, Area under the receiver operating characteristic curve; PF, Preventive fraction

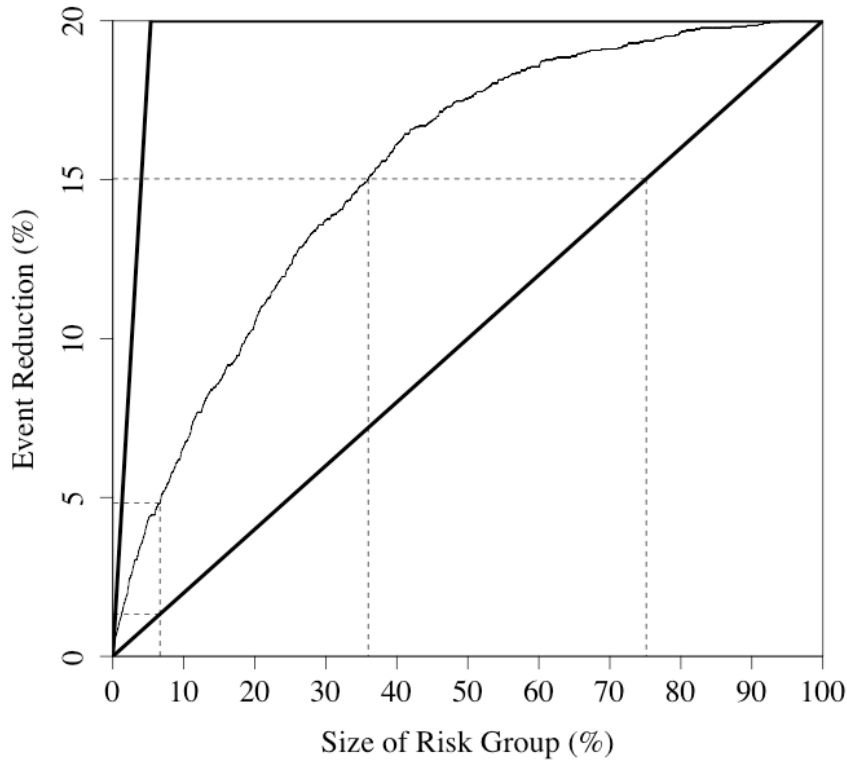
Legend: The plots represent four different scenarios with varying event incidence and PF. Prediction impact curves correspond to risk models with varying AUC (0.60-0.90).

Table 1. Area under the prediction impact curve for varying event incidence, PF, and AUC.

Incidence	Min. AUPIC	Max. AUPIC	AUC	AUPIC
PF=0.20				
10%	0.10	0.19	0.60	0.12
	0.10	0.19	0.70	0.14
	0.10	0.19	0.80	0.15
	0.10	0.19	0.90	0.17
40%	0.10	0.16	0.60	0.11
	0.10	0.16	0.70	0.12
	0.10	0.16	0.80	0.14
	0.10	0.16	0.90	0.15
PF=0.60				
10%	0.30	0.57	0.60	0.35
	0.30	0.57	0.70	0.41
	0.30	0.57	0.80	0.46
	0.30	0.57	0.90	0.52
40%	0.30	0.48	0.60	0.34
	0.30	0.48	0.70	0.37
	0.30	0.48	0.80	0.41
	0.30	0.48	0.90	0.44

AUC, Area under the receiver operating characteristic curve; AUPIC, Area under the prediction impact curve; PF, Preventive fraction

Figure 4. Prediction impact curve for 10-year CHD risk model derived from Artherosclerosis Risk in Communities Study data.



AUC, Area under the receiver operating characteristic curve; CHD, Coronary heart disease; PF, preventive fraction, ARIC, Artherosclerosis Risk in Communities Study

Legend: The plot represents a scenario in which CHD incidence was 5% and the statin intervention had a PF of 0.20. Prediction impact curve corresponds to the 10-year CHD risk model derived from the ARIC data.



EMORY
UNIVERSITY

Institutional Review Board

April 7, 2014

William Campbell,
Principal Investigator
Athletics and Recreation

RE: **Exemption of Human Subjects Research**

IRB00072937

Prediction impact curve: a new graphical approach integrating intervention effects in the evaluation of prediction model utility.

Dear Mr. Campbell:

Thank you for submitting an application to the Emory IRB for the above-referenced project. Based on the information you have provided, we have determined on 4/7/2014 that although it is human subjects research, it is exempt from further IRB review and approval.

This determination is good indefinitely unless substantive revisions to the study design (e.g., population or type of data to be obtained) occur which alter our analysis. Please consult the Emory IRB for clarification in case of such a change. Exempt projects do not require continuing renewal applications.

This project meets the criteria for exemption under 45 CFR 46.101(b)(4). Specifically, you will develop and test a new metric, the Prediction Impact Curve, for measuring the performance of disease prediction models. You have requested use of data from the ARIC Study (atherosclerosis) made publicly available to researchers through NLHBI in order to test this model and demonstrate its applications. The data is coded and you have indicated that you will not have access to the key(s) that links data to individual identifiers nor seek to determine individual identities. The following is associated with this approval:

- Protocol PIC, 4/5/2014

Please note that the Belmont Report principles apply to this research: respect for persons, beneficence, and justice. You should use the informed consent materials reviewed by the IRB unless a waiver of consent was granted. Similarly, if HIPAA applies to this project, you should

4/8/2014

<https://eresearch.emory.edu/Emory/Doc/0/T44T1NFCRQ8KB6EQLLNU9M9F78/fromString.html>



use the HIPAA patient authorization and revocation materials reviewed by the IRB unless a waiver was granted. CITI certification is required of all personnel conducting this research.

Unanticipated problems involving risk to subjects or others or violations of the HIPAA Privacy Rule must be reported promptly to the Emory IRB and the sponsoring agency (if any).

In future correspondence about this matter, please refer to the study ID shown above. Thank you.

Sincerely,

Regina Drake, M.Div, CIP
Senior Research Protocol Analyst
This letter has been digitally signed

Emory University
1599 Clifton Road, 5th Floor - Atlanta, Georgia 30322
Tel: 404.712.0720  - Fax: 404.727.1358  - Email: irb@emory.edu - Web: <http://www.irb.emory.edu/>
An equal opportunity, affirmative action university