Published in final edited form as:

Med Decis Making. 2017 October ; 37(7): 790–801. doi:10.1177/0272989X17704855.

Understanding the value of individualized information: the impact of poor calibration or discrimination in outcome prediction models

Natalia Olchanski, MS¹, Joshua T. Cohen, PhD¹, Peter J. Neumann, ScD¹, John B. Wong, MD^{1,2}, and David Kent, MD, MSc¹

¹Institute for Clinical Research and Health Policy Studies, Tufts Medical Center, Boston, MA

²Division of Clinical Decision Making, Tufts Medical Center, Boston, MA

Abstract

Purpose—Risk prediction models allow incorporation of individualized risk and clinical effectiveness information to identify patients for whom therapy is most appropriate and cost-effective. This approach has the potential to identify inefficient (or harmful) care in subgroups at different risks even when overall results appear favorable. We explore the value of personalized risk information, and factors that influence it.

Methods—Using an expected value of individualized care (EVIC) framework which monetizes the value of customizing care, we developed a general approach to calculate individualized incremental cost effectiveness ratios (ICERs) as a function of individual outcome risk. For a case study (tPA vs. streptokinase to treat possible myocardial infarction), we used simulation to explore how EVIC is influenced by population outcome prevalence, model discrimination (c-statistic) and calibration, and willingness-to-pay (WTP) thresholds.

Results—In our simulations, for well calibrated models, which do not over or underestimate predicted vs observed event risk, EVIC ranged from 0 to \$700 per person, with better discrimination (higher c-statistic values) yielding progressively higher EVIC values. For miscalibrated models EVIC ranged from −\$600 to \$600 in different simulated scenarios. The EVIC values decreased as discrimination improved from a c-statistic of 0.5 to 0.6, before becoming positive as the c-statistic reached values of ~0.8.

Conclusions—Individualizing treatment decisions using risk may produce substantial value, but also has the potential for net harm. Good model calibration ensures a non-negative EVIC. Improvements in discrimination generally increase EVIC, but when models are miscalibrated, greater discriminating power can paradoxically reduce the EVIC under some circumstances.

Introduction

It has become more widely appreciated that patients enrolled in clinical trials may vary greatly in their risk of the outcome of interest and that, because of this, the average trial results (especially when expressed as an absolute risk reduction) typically do not apply to all enrolled patients[1–5]. Thus, the optimal clinical approach can vary across patients, particularly when a therapy is associated with some treatment-related harm or substantial cost. Hence, recommendations tailored to the individual, based in particular on individual

patient risks, hold potential for improving care compared to “one-size-fits-all” recommendations (e.g., “treat all” or “treat none”). It is unclear how much of an improvement in specific treatment decisions can be made with currently available information incorporating individualized risk estimates. Key outstanding questions include: (1) how great are the incremental benefits of individualizing care compared to a treat all or treat none approach, (2) under what conditions are these benefits greatest, and (3) when might targeted strategies result in net harm?

This study uses simulation to explore how the underlying characteristics of a decision scenario can influence the additional benefits to be derived from an individualized (risk-based) approach versus a “one-size-fits-all” approach, based on calculating the expected value of individualized care (EVIC)[6]. In particular, we explore how changes in the clinical risk prediction model performance (such as predictive model discrimination and calibration) might influence the EVIC. At the same time we assess the impact of different levels of the willingness to pay and the average population risk. Our goal is to understand the decisional context for which an individualized tailored risk-based approach is most likely to add value, as well as when it might lead to harm.

Theory

Expected Value of Individualized Care

In this study we use a previously developed framework known as the Expected Value of Individualized Care (EVIC)[6–8]. EVIC is the difference between the expected value of the best population-wide strategy – i.e., either treating every population member or treating no one – and the expected value of selecting treatment for each population member individually.

Identifying and valuing the best population-wide strategy—By convention, we designate the expected value of treating no one to be zero. The incremental value of treating everybody is therefore the monetized treatment benefit for the population (where benefits are monetized based on the “willingness to pay” (WTP) threshold) minus treatment costs, averaged over the entire population. If this incremental net benefit value is positive, treating everyone is the optimal population-wide strategy. Otherwise, treating nobody (value of zero) is the optimal population-wide strategy.

The analysis in this paper estimates each individual’s treatment benefit by assuming it is proportional to the product of three quantities: (1) the baseline risk of the adverse event for individual i in the absence of treatment, designated $R_{i,baseline}$; (2) the risk reduction conferred by treatment, designated $(1 - RR_{tx})$, where RR_{tx} is the relative risk of the adverse event with treatment; and (3) the QALY gain associated with averting the adverse outcome, designated $QALY$. Hence, the expected QALY gain from treatment of individual i is $R_{i,baseline} \times (1 - RR_{tx}) \times QALY$. Note that this formulation assumes that both RR_{tx} and $QALY$ are the same for all members of the population; only baseline outcome risk ($R_{i,baseline}$) differs across individuals. This is a simplification that does not change the overall results of the framework presented. Each individual’s outcome in the absence of

treatment is not known ex ante, but the baseline outcome risk can be estimated using a predictive model on the basis of baseline characteristics of the individual.

Multiplying the expected QALY gain by WTP (the willingness to pay for one QALY) yields the expected monetized treatment benefit. If this benefit exceeds the treatment cost on average, then treating everyone is the optimal population-wide strategy, and its expected per-person value is $WTP \times \bar{R}_{baseline} \times (1 - RR_{tx}) \times QALY - Cost_{TX}$. Here, $\bar{R}_{baseline}$ is the average population baseline risk and $Cost_{TX}$ is the per-patient treatment cost. Otherwise, the optimal population strategy is treating nobody and its monetized value is zero, by convention.

Valuing individualized care—We again estimate benefits minus costs for all members of the population. In this case, however, individuals receive treatment only if their monetized benefits exceed costs. For individuals receiving treatment, treatment value is $WTP \times R_{i,baseline} \times (1 - RR_{tx}) \times QALY - Cost_{TX}$, exactly as above; for others (who remain untreated), it is zero.

Assuming that the individualized decisions are based on accurate estimates, EVIC cannot be negative; that is, individualized care must be at least as good as the best population-wide strategy, because it only eliminates treatments in those patients for whom treatment has a negative value. At worst, EVIC is zero when the optimal decision does not vary across individuals and thus decision-making is the same for both the individualized and “one-size-fits-all” approaches.

Perfect risk prediction

Consider a hypothetical clinical prediction model that perfectly predicts which population members will experience an adverse outcome in the absence of treatment. Under this condition, only that group which would benefit from treatment would receive it (assuming the therapy provides overall benefit to them, and they would not have a good outcome regardless of therapy).

If the optimal population-wide strategy is to treat everyone, then EVIC is the difference between the value of targeting individuals who will experience the adverse outcome in the absence of treatment, and the value of treating everyone (including those who will not benefit from treatment or do well even if untreated). In the absence of treatment harm, the treatment benefit is the same for both strategies since both strategies treat all members of the population who could benefit. The costs differ, however. For the targeted strategy, costs are $N \times \bar{R}_{baseline} \times Cost_{TX}$. For the “treat all” strategy, costs are $N \times Cost_{TX}$. EVIC is the difference between these two values, or $(1 - \bar{R}_{baseline})N \times Cost_{TX}$ and is maximized when $\bar{R}_{baseline}$ is small, since a smaller number of patients would need to be treated with individualized care.

If the optimal population-wide strategy is “treat none” (value of zero by convention), then EVIC is the treatment benefits minus the treatment costs. That is, EVIC is $N \times \bar{R}_{baseline} \times [WTP \times QALY \times (1 - RR_{tx}) - Cost_{TX}]$. Note that as $\bar{R}_{baseline}$ increases, EVIC also grows, since we would be changing the treatment strategy on progressively more individuals. This

relationship may not be monotonic because at some value of $\bar{R}_{baseline}$ the optimal population strategy can switch between “treat all” and “treat none”.

Imperfect risk prediction

We are interested in EVIC in more typical circumstances in which risk prediction is imperfect. Below we consider briefly how EVIC might change when model discrimination is imperfect and also when a model is miscalibrated.

Imperfect Discrimination—“Risk” refers to the *ex ante* probability that a particular member of a population will ultimately experience an adverse outcome. Discrimination refers to a model’s ability to assign higher risks to subjects who will ultimately experience the adverse outcome than to subjects who will not.

Formally, discrimination is measured in terms of the c-statistic, which represents the proportion of all possible discordant subject pairs (i.e., pairs in which exactly one subject will experience the adverse outcome) that the model correctly orders by risk. That is, in a correctly ordered pair, the model predicts a higher risk for the subject who will experience the adverse outcome than it predicts for the subject who will not.

A model with perfect discrimination will always assign a higher risk to the subject who will experience the adverse outcome, so a model with perfect discrimination has a c-statistic value of 1.0 (100% of the pairs are correctly ordered). A model with no discriminatory power effectively “flips a coin” and hence has a 50% chance of correctly assigning a higher risk to the subject who will ultimately experience the adverse outcome. Therefore, a model with no discriminatory power has a c-statistic of 0.5 (50%). Note that the c-statistic says nothing about the relative magnitude of the risks assigned by the model to individuals who do and do not experience the adverse outcome. Only the order matters.

Figure 1 illustrates empirically derived risk distributions corresponding to c-statistic values ranging from 0.5 (model predictions are no better than chance) to 1.0 (model achieves perfect discrimination). At c-statistic values modestly exceeding 0.5 (e.g., c-statistic = 0.6), modeled risk estimates for population members cluster around the value of $\bar{R}_{baseline}$ (risk values for some individuals are slightly above $\bar{R}_{baseline}$, while others are a bit below). That is, models with low c-statistic values typically predict that all subjects face similar risks, and these risks are roughly equal to the prevalence of the adverse outcome in the population. As the c-statistic increases, the resulting risk heterogeneity across the population increases because of improvements in the model’s ability to discriminate high from low risk individuals. At c-statistic values closer to 1.0 (e.g., c-statistic = 0.8 or 0.9), the model-predicted risk for a portion of the population is large (eventually approaching 1.0 as the c-statistic grows), while for the remainder of the population, the model-predicted risk is small (approaching zero as the c-statistic grows). That is, at high c-statistic values, the model distinguishes individuals very likely to experience the adverse outcome from those very likely not to.

In this paper we use simulations to explore the relationship between the predictive model’s c-statistic and the EVIC.

Imperfect calibration—Calibration measures the extent to which model-predicted prevalence approximates the actual outcome prevalence, typically across some subgrouping specification (e.g., risk deciles). When calibration is perfect, the model is unbiased, and treatment decisions made for each member of the population individually are at least as good as the population-wide optimal treatment, regardless of how good or poor model discrimination is[9]. Hence, in these circumstances, EVIC is non-negative.

Often, risk models do not accurately predict risk for individuals outside the population used to construct the model. They may overestimate risk across quantiles, underestimate risk across quantiles, be “overfit” (under-estimate in low risk and over-estimate in high risk patients), be “underfit” (over-estimate low risk, under-estimate high risk) or have various other patterns. The risk prediction model in these situations can be said to have poor “external validity”, or to be “miscalibrated” or “biased”.

A miscalibrated model can incorrectly classify members of the population with regard to whether their expected benefit from treatment exceeds the treatment costs because the expected benefits are inaccurate. Misclassification can, in turn, produce erroneous treatment recommendations. Because the model recommendation might be erroneous, model-based recommendations for individuals can be worse than the optimal population-wide treatment recommendation. As a result, for misclassified models, the EVIC can be negative.

Finally, miscalibration is a potential problem not just when risk models are used to individualize treatment decisions, but anytime external evidence is applied to a population. Indeed, whenever applying clinical trial results to a population in actual practice, the degree of bias in the effect estimates is generally unknown, and thus population-wide recommendations can also yield sub-optimal population-wide decisions.

Illustrative Clinical Example

Methods

We illustrate the concepts described above through a case study based on the Global Utilization of Streptokinase and Tissue Plasminogen Activator for Occluded Coronary Arteries (GUSTO) trial, a large multicenter randomized clinical trial. The GUSTO trial showed that a newer more expensive thrombolytic agent, tissue plasminogen activator (tPA), conferred a survival benefit compared to the contemporaneous standard agent, streptokinase, for the treatment of acute myocardial infarction (MI)[10].

Assumptions—We based outcome rate and cost estimates for our simulations on the GUSTO trial[10]. We defined the adverse outcome event to be 30-day mortality. We assumed the average population risk ($\bar{R}_{baseline}$) for this outcome is 7%, that tPA achieves a 15% relative risk reduction ($RR_{tx} = 0.85$) for mortality compared with streptokinase, and that the total cost difference for patients treated with tPA vs. streptokinase ($Cost_{TX}$) is \$2000. This difference incorporates the short term direct medical costs. We assumed that the QALY gain due to averting death at 30 days is 12 additional QALYs, an estimate corresponding to average survival duration for acute MI survivors [11–13], but that there is no additional difference in QALYs due to treatment choice. Further, to simplify interpretation, we

excluded future health care costs for MI survivors, beyond the costs of tPA and streptokinase treatment. We investigated the impact of altering some of these assumptions, using ranges detailed in Table 1.

Identifying the optimal population-wide strategy—In addition to a treatment's effectiveness and cost, the optimal population-wide strategy (“treat all” or “treat none”) depended on several factors. These factors were the expected prevalence of the adverse outcome (in this example, the 30 day mortality rate $\bar{R}_{baseline}$), and the monetized value of each QALY gained (i.e., the decision maker's WTP threshold). Specifically, *Net Benefit* = $WTP \times \bar{R}_{baseline} \times (1 - RR_{Tx}) \times QALY - Cost_{Tx}$. If $R_{Baseline} = 0$, treatment can have no benefit, and net benefits cannot exceed zero (if treatment incurs any cost, net benefits will be negative). As $R_{Baseline}$ increases, net benefits increase with a slope of $WTP \times (1 - RR_{Tx}) \times QALY$ (see Figure 1). For $\bar{R}_{baseline} = \frac{Cost_{Tx}}{WTP \times (1 - RR_{Tx}) \times \Delta QALY}$, net benefit is zero.

Generating baseline risk values for each simulated population member—We assigned each member of our simulated population a baseline risk value drawn from a beta probability distribution. We created a series of baseline risk probability distributions, each corresponding to assumed values for (1) the population adverse event prevalence ($R_{Baseline} = 5\%, 7\%, 10\%, \text{ or } 20\%$), and (2) the c-statistic (values ranging from 0.5 - no discriminatory power - to 0.9 - high discriminatory power).

In summary, a different risk distribution corresponds to each combination of values for $\bar{R}_{baseline}$ and the c-statistic, and can be modeled for a given population and predictive model, as previously shown in an analysis of 32 large clinical trials by Kent et al[14]. We described each distribution as a beta distribution with parameters that depend on the assumed values for the c-statistic and population prevalence, $\bar{R}_{baseline}$, in our simulations. We estimated the beta distribution parameters from empirical relationships across multiple clinical trials, derived previously[14]. To characterize the beta distribution, we first estimated the alpha parameter using ordinary least squares regression applied to alpha values derived for specific beta distributions fitted to each clinical trial in a set of 25 large randomized clinical trials (including ACCORD[15], AFFIRM[16], ALLHAT[17, 18], AMIS[19], BARI[20], BEST[21], BHAT[22], CAST[23], CPPT[24], DCCT[25], DIG[26], DPP[27], ENRICH[28], FAVORIT[29], HALTC[30], HDFP[31], HEMO[32], MRFIT[33], MTOPS[34], OAT[35], PEACE[36], ROC[37, 38], SHEP[39], SOLVD[40, 41], TIMI II[42].) The resulting relationship was $\ln \alpha = 9.35 - 12.15c$, where c is the c-statistic. The R^2 for this regression was 0.90. After estimating α , we estimated β from the relationship $\frac{\alpha}{\alpha + \beta} = p$, where p is the population prevalence of the adverse outcome (equal to $\bar{R}_{baseline}$). Because the clinical trials in our sample had c-statistic values ranging from 0.58 to 0.75, our characterization of the population baseline risk distribution is least uncertain in this range.

We simulated a cohort of 5,000 individuals for each scenario (i.e., for each combination of c-statistic and $\bar{R}_{baseline}$ values). For each scenario, we reported the EVIC and the population distribution of cost-effectiveness ratios. By comparing scenarios, we described how the cost-effectiveness distribution changes in response to changes in the assumed discriminatory power of the risk model.

EVIC when risk estimates are well calibrated—To calculate EVIC, we first identified the optimal “one-size-fits-all” population strategy (treat all or treat none) for each scenario. We then identified the optimal strategy for each population member based on his or her baseline 30-day mortality risk. Specifically, we identified tPA to be the optimal treatment for individual i if $WTP \times \bar{R}_{i,baseline} \times (1 - RR_{tx}) \times QALY > Cost_{Tx}$, where $\bar{R}_{i,baseline}$ is the baseline 30-day mortality risk for that individual, $(1 - RR_{tx})$ is the relative risk conferred by tPA, and $Cost_{Tx}$ is the incremental cost of tPA compared to the streptokinase treatment. Finally, we estimated EVIC by summing the monetized gains accrued for all population members for whom the individualized treatment differed from the optimal population-wide strategy, and averaging them across the population. For example, if the optimal population-wide strategy was “treat none”, the EVIC was the sum of the monetized net gains accrued for all population members who, based on an individualized care, switched to treatment.

EVIC when risk estimates are miscalibrated—For this scenario, we first characterized the “true” baseline risk distribution using the approach described above for the perfect calibration scenario. We then calculated a reported baseline risk value for each member of the simulated cohort – the value that would be observed in the “real world” – by multiplying that individual’s simulated “true” baseline risk value by a miscalibration factor that ranged from 0.5 (underestimate) to 2.0 (overestimate). Next, we identified a putative optimal treatment for each individual by plugging that individual’s reported baseline risk value into the benefit expression ($WTP \times \bar{R}_{i,baseline} \times (1 - RR_{tx}) \times QALY$) and comparing it to tPA’s incremental cost. EVIC was calculated based on the “true” underlying costs and benefits of the selected treatment for each individual.

Results

Mortality risk and cost-effectiveness of population-wide treatment—Figure 2 illustrates the relationship between the net benefit of treating all patients with MI with tPA (vertical axis) when assumptions about 30-day mortality (horizontal axis) are varied. As expected, the net benefit is negative when the 30-day mortality is low; the net benefit is greatest when the 30-day mortality is highest. As illustrated in Figure 2, if a QALY is worth \$100,000, net benefits are zero when population 30 day mortality is assumed to be 1.1% (i.e., $P_{100k} = 1.1\%$). The corresponding break-even points for QALY values of \$50,000 and \$100,000 are $P_{50k} = 2.2\%$, and $P_{20k} = 5.6\%$, respectively. When the population-wide mortality risk falls below these thresholds, the optimal population-wide strategy is to treat none (i.e. use streptokinase in all patients); when the average mortality risk is higher, the optimal population-wide strategy is to treat all.

For example, if the QALY value (WTP) = \$50,000/QALY, the EVIC for a model with a c-statistic of 0.8 is \$200 per population member when $\bar{R}_{baseline} = 20\%$. The per-person EVIC of the same model increases to \$500 in a population with a $\bar{R}_{baseline} = 5\%$ because $\bar{R}_{baseline} = 5\%$ is closer to the break-even risk value of 2% (Figure 2). Similarly, when $WTP = \$20,000/QALY$, the break-even average 30-day mortality is approximately 6%, and the EVIC is in fact highest when $\bar{R}_{baseline}$ is between 5% and 7%.

Holding $\bar{R}_{baseline}$ constant and instead changing the assumed monetized value of a QALY (WTP) likewise influences EVIC for analogous reasons. As WTP becomes smaller (e.g., reducing it from \$100,000/QALY to \$20,000/QALY), EVIC increases from \$350 per subject to \$700 per subject. Note that in both scenarios, the optimal population-wide strategy is to treat all with tPA. But as the value for WTP falls, the less expensive streptokinase therapy becomes the optimal strategy for a growing number of population members because individualized care accrues benefits for this growing number of individuals (by switching these patients from tPA to streptokinase) so that EVIC increases.

Changes in individualized ICER with changes in discrimination—Figure 3 illustrates individualized cost-effectiveness value distributions corresponding to different assumptions regarding the risk model's discriminatory power. Because a well-calibrated model with low discriminatory power (c statistic) yields similar baseline risks for all members of the population, individualized cost-effectiveness estimates are likewise similar for all members of the population and the mean and median cost-effectiveness estimates do not substantially differ. As discriminatory power improves, predicted risk varies across population members and so do the individualized ICERs. The individualized cost-effectiveness of tPA for some patients (i.e. those at low risk for mortality) becomes highly unfavorable (very large) as the incremental benefit in the cost-effectiveness ratio denominator becomes small; for others (i.e. those at high risk), tPA's cost-effectiveness ratio becomes more favorable (relatively small) as the incremental benefit in the ratio's denominator grows. Because the resulting distribution of cost-effectiveness ratios is skewed, the mean and median diverge, making tPA less worthwhile in the typical patient.

EVIC when risk estimates are perfectly calibrated—Figure 4 illustrates our EVIC calculations for the simulated cohorts. The EVIC ranges from 0 to \$700 per person, with better discrimination (higher c-statistic values) yielding higher EVIC values. The association between discriminatory power and EVIC persists across different assumed values for baseline average 30-day mortality ($\bar{R}_{baseline}$) and the monetized value for a QALY (WTP).

EVIC when risk estimates are miscalibrated—Figure 5 illustrates our findings for scenarios in which the prediction models produce miscalibrated individual risk estimates. Under the assumptions in our study, EVIC ranged from -\$600 to \$700 per person, reflecting lower value for individualized care compared to optimal population-wide decisions in some circumstances. The most striking feature of these results is evident in the plots on the left (miscalibration factor of 0.5, meaning that the model-reported risk underestimates each individual true 30-day mortality probability by one-half). In those plots, improvements in discrimination were non-monotonically related to changes in EVIC. Increasing discriminatory power lowered EVIC when discrimination was low (i.e. as the c-statistic improved from 0.5 to between 0.6 and 0.7) but then increased EVIC when the c-statistic was higher (i.e. as the c-statistic improves to between 0.8 and 0.9).

Discussion

Cost-effectiveness analyses are typically performed at a population level, yet most medical decisions are made, and most resources are allocated, one patient at a time. Therefore, cost-

effectiveness analyses are generally at a level that is incongruent with medical decision making, and individual cost-effectiveness estimates can vary substantially from the overall average cost-effectiveness[13, 43, 44]. In the simulations reported we examine the value of individualizing information based on estimates of individual patient risk. Our results suggest that individualizing treatment decisions may yield substantial value, but also has the potential for harm. We have verified what others have observed for well calibrated models: EVIC increases as discriminatory power increases[9]. As model discrimination improves, risk heterogeneity increases, and consequently there is also greater heterogeneity in individual ICER estimates. Because these individual ICER estimates may diverge substantially from the population average (and the ICER for the typical patient may be quite different from the population average), the optimized individualized strategy may diverge substantially from the optimized strategy based on a population average. Our simulations, based on a simplified version of a classic cost-effectiveness analysis, suggest that moderate improvements in discrimination can add substantial value, equivalent to several hundred dollars on a per person basis in some scenarios.

When models are well calibrated, individualized risk based decisions are, at worst, non-harmful. This is consistent with other work using a decision analytic framework[9]. We also found that well calibrated risk models are of greatest value to decisions when the overall prevalence places a population-wide strategy near the decision threshold and when model discrimination is good. When the mortality risk yields an incremental cost-effectiveness ratio that is far from the WTP decision threshold, the optimal population-wide and individualized strategies will be the same and EVIC will be near-zero, except when model discrimination is very good.

The EVIC is higher when the population baseline risk estimate ($\bar{R}_{baseline}$) is closer to the “break-even” value that makes the expected net benefit or incremental monetized value of tPA close to zero. In this circumstance, there is no clear advantage to either the “treat all with tPA” or “treat none with tPA” strategy (see Figure 1). This balance implies that the break-even risk value is near the middle of the population risk distribution, and that the optimized individualized strategy will differ from the best population-wide option (either treat all with tPA or treat none with tPA) for many. Hence, individualized care will result in a substantial number of treatment switches, and because gains from individualized care accrue only in these switched cases, EVIC can grow larger as discrimination improves and the risk distribution broadens.

Our analysis goes beyond past work by carefully examining what happens when models are miscalibrated. That scenario is important because in the real world, models are typically developed on one population and applied to others.

When models are miscalibrated, EVIC can be negative and improvements in model discrimination can paradoxically decrease the value of individualizing care in some circumstances. We explored the impact of individualization based on miscalibrated predictive models, which can lead to treatment assignment “mistakes”. When the population average ICER is near the WTP threshold, decision making is more sensitive to miscalibration and these treatment assignment mistakes are more common (making EVIC

more negative). This can be explained by looking at the distribution of individual ICERs (Figure 3). The effect of model miscalibration in our simulation was to shift the ICERs of some individuals across the WTP threshold, leading to the wrong decision for those patients, thus lowering the overall EVIC. When the WTP threshold is not close to the population ICER, the change in individual ICERs does not affect the treatment decision, and there is no effect on the overall EVIC. Interestingly, a poor c-statistic can sometimes protect from misclassification due to bias, which becomes revealed only when the c-statistic improves; thus, greater discriminating power can paradoxically reduce the EVIC under some circumstances.

These simulations were based on several simplifying assumptions. Most importantly, we assumed the following parameters were independent of baseline risk: treatment effect (measured as relative risk reduction); treatment costs; and QALYs gained conditioned on outcome. It is likely that long term costs and life expectancy or QALYs are correlated with the short-term outcome risk. For example, individuals at higher risk of a cardiovascular death are also more likely to be older and have more severe heart attacks and other risk factors or health conditions that would impact their long-term life expectancy. Including the correlations between short-term risk and long-term benefit in the decision model is likely to attenuate the heterogeneity of individualized benefits across the population (and diminish EVIC), although it may be difficult to generalize whether this would be true across other conditions. A further complicating factor not included in this analysis is heterogeneity of treatment-related harm (i.e., intracranial bleeding)[12], especially its interplay with the heterogeneity of the individual outcome risk and benefit from treatment. These issues underscore the heavy data and analytic burdens of more individualized cost effectiveness studies.

Conclusions

Individualized estimates can help determine treatment course. However, predictive models should be used with care and an understanding of the model's applicability. First, if the treatment's average cost-effectiveness is far from any important decision threshold, there is a high probability that the optimal population-wide strategy is the right course for any individual. When the net benefit of a population wide strategy is near zero, individualized information may be very helpful in assigning the best treatment strategy. Second, if a predictive model has poor discriminating power, it may be of very low value. Finally, poorly calibrated models may provide misleading predictions and cause patient harm, reducing EVIC. These results emphasize the importance of paying special attention to the decisional context in which models will be deployed and to model calibration, an often neglected aspect of predictive model validation. Caution is needed when transporting risk models to new settings. In general, models developed on local samples are anticipated to be better calibrated than models developed on "different but related" samples. Whenever possible, risk models should be updated and recalibrated to ensure the accuracy of the risk information they yield. While the potential for miscalibration is ubiquitous, more research is needed to determine those situations in which risk models are especially likely to be vulnerable to miscalibration.

Acknowledgments

Financial support for this study was provided entirely by a grant from the National Institutes of Health (grant U01 NS086294). The funding agreement ensured the authors' independence in designing the study, interpreting the data, writing, and publishing the report.

The authors would like to acknowledge Issa Dahabreh, MD, MS, for helpful discussions assisting in the derivation of parameter estimates for risk distributions in our simulations.

This article was prepared using research materials from Action to Control Cardiovascular Risk in Diabetes (ACCORD), Atrial Fibrillation Follow-Up Investigation of Rhythm Management (AFFIRM), Antihypertensive and Lipid-Lowering Treatment to Prevent Heart Attack Trial (ALLHAT), Aspirin-Myocardial Infarction Study (AMIS), Bypass Angioplasty Revascularization Investigation (BARI), Beta-Blocker Evaluation in Survival Trial (BEST), Beta-Blocker Heart Attack Trial (BHAT), Cardiac Arrhythmia Suppression Trial (CAST), Digitalis Investigation Group (DIG), Enhancing Recovery in Coronary Heart Disease Patients (ENRICH), Hypertension Detection and Follow-Up Program (HDFP), Lipid Research Clinics (LRC) Coronary Primary Prevention Trial (CPPT), Multiple Risk Factor Intervention Trial for the Prevention of Coronary Heart Disease (MRFIT), Occluded Artery Trial (OAT), Prevention of Events With Angiotensin-Converting Enzyme Inhibitor Therapy (PEACE), Resuscitation Outcomes Consortium (ROC) Hypertonic Saline Trial Shock Study (HS) and Traumatic Brain Injury Study (TBI), Systolic Hypertension in the Elderly Program (SHEP), Studies of Left Ventricular Dysfunction (SOLVD), and Thrombolysis in Myocardial Ischemia Trial II (TIMI II) obtained from the National Heart, Lung, and Blood Institute Biologic Specimen and Data Repository Information Coordinating Center and does not necessarily reflect the opinions or views of the study investigators or the National Heart, Lung, and Blood Institute.

The Acute Renal Failure Trial Network (ATN), Diabetes Control and Complications Trial (DCCT), Diabetes Prevention Program (DPP), Folic Acid for Vascular Outcome Reduction in Transplantation Trial (FAVORIT), The Hepatitis C Antiviral Long-Term Treatment Against Cirrhosis (HALT-C), Hemodialysis Study (HEMO), and the Medical Therapy of Prostatic Symptoms (MTOPS) were conducted by study Investigators and supported by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). The data from the trials reported here were supplied by the NIDDK Central Repositories. This manuscript was not prepared in collaboration with Investigators of these studies and does not necessarily reflect the opinions or views of Investigators, the NIDDK Central Repositories, or the NIDDK.

References

1. Rothwell PM. Can overall results of clinical trials be applied to all patients? *Lancet*. 1995; 345(8965):1616–9. [PubMed: 7783541]
2. Kent DM, Hayward RA. Limitations of applying summary results of clinical trials to individual patients: the need for risk stratification. *Jama*. 2007; 298(10):1209–12. [PubMed: 17848656]
3. Kent DM, et al. Assessing and reporting heterogeneity in treatment effects in clinical trials: a proposal. *Trials*. 2010; 11:85. [PubMed: 20704705]
4. Kravitz RL, Duan N, Braslow J. Evidence-based medicine, heterogeneity of treatment effects, and the trouble with averages. *Milbank Q*. 2004; 82(4):661–87. [PubMed: 15595946]
5. Espinoza MA, et al. The value of heterogeneity for cost-effectiveness subgroup analysis: conceptual framework and application. *Med Decis Making*. 2014; 34(8):951–964. [PubMed: 24944196]
6. Basu A, Meltzer D. Value of information on preference heterogeneity and individualized care. *Med Decis Making*. 2007; 27(2):112–127. [PubMed: 17409362]
7. van Gestel A, et al. The role of the expected value of individualized care in cost-effectiveness analyses and decision making. *Value Health*. 2012; 15(1):13–21. [PubMed: 22264967]
8. Claxton K. The irrelevance of inference: a decision-making approach to the stochastic evaluation of health care technologies. *J Health Econ*. 1999; 18(3):341–64. [PubMed: 10537899]
9. Van Calster B, Vickers AJ. Calibration of risk prediction models: impact on decision-analytic performance. *Med Decis Making*. 2015; 35(2):162–169. [PubMed: 25155798]
10. The GUSTO investigators. An international randomized trial comparing four thrombolytic strategies for acute myocardial infarction. *N Engl J Med*. 1993; 329(10):673–682. [PubMed: 8204123]
11. Mark DB, et al. Cost effectiveness of thrombolytic therapy with tissue plasminogen activator as compared with streptokinase for acute myocardial infarction. *N Engl J Med*. 1995; 332(21):1418–24. [PubMed: 7723799]

12. Kent DM, et al. An independently derived and validated predictive model for selecting patients with myocardial infarction who are likely to benefit from tissue plasminogen activator compared with streptokinase. *Am J Med.* 2002; 113(2):104–111. [PubMed: 12133748]
13. Kent DM, et al. Tissue plasminogen activator was cost-effective compared to streptokinase in only selected patients with acute myocardial infarction. *J Clin Epidemiol.* 2004; 57(8):843–52. [PubMed: 15485737]
14. Kent DM, et al. Risk and treatment effect heterogeneity: re-analysis of individual participant data from 32 large clinical trials. *Int J Epidemiol.* 2016
15. Gerstein HC, et al. Effects of intensive glucose lowering in type 2 diabetes. *N Engl J Med.* 2008; 358(24):2545–59. [PubMed: 18539917]
16. Wyse DG, et al. A comparison of rate control and rhythm control in patients with atrial fibrillation. *N Engl J Med.* 2002; 347(23):1825–33. [PubMed: 12466506]
17. Major outcomes in high-risk hypertensive patients randomized to angiotensin-converting enzyme inhibitor or calcium channel blocker vs diuretic: The Antihypertensive and Lipid-Lowering Treatment to Prevent Heart Attack Trial (ALLHAT). *Jama.* 2002; 288(23):2981–97. [PubMed: 12479763]
18. Major outcomes in moderately hypercholesterolemic, hypertensive patients randomized to pravastatin vs usual care: The Antihypertensive and Lipid-Lowering Treatment to Prevent Heart Attack Trial (ALLHAT-LLT). *Jama.* 2002; 288(23):2998–3007. [PubMed: 12479764]
19. A randomized, controlled trial of aspirin in persons recovered from myocardial infarction. *Jama.* 1980; 243(7):661–9. [PubMed: 6985998]
20. The Bypass Angioplasty Revascularization Investigation (BARI) Investigators. Comparison of coronary bypass surgery with angioplasty in patients with multivessel disease. *N Engl J Med.* 1996; 335(4):217–25. [PubMed: 8657237]
21. A trial of the beta-blocker bucindolol in patients with advanced chronic heart failure. *N Engl J Med.* 2001; 344(22):1659–67. [PubMed: 11386264]
22. A randomized trial of propranolol in patients with acute myocardial infarction. I. Mortality results. *Jama.* 1982; 247(12):1707–14. [PubMed: 7038157]
23. Echt DS, et al. Mortality and morbidity in patients receiving encainide, flecainide, or placebo. The Cardiac Arrhythmia Suppression Trial. *N Engl J Med.* 1991; 324(12):781–8. [PubMed: 1900101]
24. The Lipid Research Clinics Coronary Primary Prevention Trial results. I. Reduction in incidence of coronary heart disease. *Jama.* 1984; 251(3):351–64. [PubMed: 6361299]
25. The Diabetes Control and Complications Trial Research Group. The effect of intensive treatment of diabetes on the development and progression of long-term complications in insulin-dependent diabetes mellitus. *N Engl J Med.* 1993; 329(14):977–86. [PubMed: 8366922]
26. The effect of digoxin on mortality and morbidity in patients with heart failure. *N Engl J Med.* 1997; 336(8):525–33. [PubMed: 9036306]
27. Knowler WC, et al. Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. *N Engl J Med.* 2002; 346(6):393–403. [PubMed: 11832527]
28. Berkman LF, et al. Effects of treating depression and low perceived social support on clinical events after myocardial infarction: the Enhancing Recovery in Coronary Heart Disease Patients (ENRICH) Randomized Trial. *Jama.* 2003; 289(23):3106–16. [PubMed: 12813116]
29. Bostom AG, et al. Homocysteine-lowering and cardiovascular disease outcomes in kidney transplant recipients: primary results from the Folic Acid for Vascular Outcome Reduction in Transplantation trial. *Circulation.* 2011; 123(16):1763–70. [PubMed: 21482964]
30. Di Bisceglie AM, et al. Prolonged therapy of advanced chronic hepatitis C with low-dose peginterferon. *N Engl J Med.* 2008; 359(23):2429–41. [PubMed: 19052125]
31. Hypertension Detection and Follow-up Program Cooperative Group. Five-year findings of the hypertension detection and follow-up program. I. Reduction in mortality of persons with high blood pressure, including mild hypertension. *Jama.* 1979; 242(23):2562–71. [PubMed: 490882]
32. Eknoyan G, et al. Effect of dialysis dose and membrane flux in maintenance hemodialysis. *N Engl J Med.* 2002; 347(25):2010–9. [PubMed: 12490682]
33. Multiple Risk Factor Intervention Trial Research Group. Multiple risk factor intervention trial. Risk factor changes and mortality results. *Jama.* 1982; 248(12):1465–77. [PubMed: 7050440]

34. McConnell JD, et al. The long-term effect of doxazosin, finasteride, and combination therapy on the clinical progression of benign prostatic hyperplasia. *N Engl J Med.* 2003; 349(25):2387–98. [PubMed: 14681504]
35. Hochman JS, et al. Coronary intervention for persistent occlusion after myocardial infarction. *N Engl J Med.* 2006; 355(23):2395–407. [PubMed: 17105759]
36. Braunwald E, et al. Angiotensin-converting-enzyme inhibition in stable coronary artery disease. *N Engl J Med.* 2004; 351(20):2058–68. [PubMed: 15531767]
37. Bulger EM, et al. Out-of-hospital hypertonic resuscitation following severe traumatic brain injury: a randomized controlled trial. *Jama.* 2010; 304(13):1455–64. [PubMed: 20924011]
38. Bulger EM, et al. Out-of-hospital hypertonic resuscitation after traumatic hypovolemic shock: a randomized, placebo controlled trial. *Ann Surg.* 2011; 253(3):431–41. [PubMed: 21178763]
39. SHEP Cooperative Research Group. Prevention of stroke by antihypertensive drug treatment in older persons with isolated systolic hypertension. Final results of the Systolic Hypertension in the Elderly Program (SHEP). *Jama.* 1991; 265(24):3255–64. [PubMed: 2046107]
40. The SOLVD Investigators. Effect of enalapril on mortality and the development of heart failure in asymptomatic patients with reduced left ventricular ejection fractions. *N Engl J Med.* 1992; 327(10):685–91. [PubMed: 1463530]
41. The SOLVD Investigators. Effect of enalapril on survival in patients with reduced left ventricular ejection fractions and congestive heart failure. *N Engl J Med.* 1991; 325(5):293–302. [PubMed: 2057034]
42. The TIMI Study Group. Comparison of invasive and conservative strategies after treatment with intravenous tissue plasminogen activator in acute myocardial infarction. Results of the thrombolysis in myocardial infarction (TIMI) phase II trial. *N Engl J Med.* 1989; 320(10):618–27. [PubMed: 2563896]
43. Ioannidis JP, Garber AM. Individualized cost-effectiveness analysis. *PLoS Med.* 2011; 8(7):e1001058. [PubMed: 21765810]
44. Meltzer D. Addressing uncertainty in medical cost-effectiveness analysis implications of expected utility maximization for methods to perform sensitivity analysis and the use of cost-effectiveness analysis to set priorities for medical research. *J Health Econ.* 2001; 20(1):109–129. [PubMed: 11148867]

Figure 1.
Individual Outcome Risk Distribution by C-statistic

Figure 2.

Net benefit of treating all members of the population compared to no treatment as a function of population outcome prevalence and the willingness-to-pay (WTP) threshold

Figure 3.
ICER distribution changes with c-statistic

Figure 4.
EVIC as a function of c-statistic at several levels of mortality risk and WTP thresholds

Figure 5.
Predictive model miscalibration compared to perfectly accurate population ICER: EVIC values at different levels of c-statistic, mean mortality risk, and miscalibration

Table 1

Illustrative example assumptions

Symbol	Meaning	Values
$QALY(R_{Death})$	Expected QALY gain with tPA vs. streptokinase in MI as a function of baseline risk of death	
R_{Death}	Baseline risk of death	0.05 – 0.2
RR_{tPA}	Relative risk of death with tPA vs. streptokinase in MI	85%
$Q_{NoDeath}$	Incremental QALYs gained if death averted	12 QALYs
Q_{tPA}	Incremental QALY loss from treatment, independent of outcome	0
$Cost(R_{Death})$	Expected cost with tPA vs. streptokinase in MI as a function of baseline risk of death	
$Cost_{NoDeath}$	Savings if death averted	\$0
$Cost_{tPA}$	Cost of tPA vs. streptokinase in MI, independent of outcome	\$2,000