Libertas Academica
FREEDOM TO RESEARCH

Open Access: Full open access to this and thousands of other papers at http://www.la-press.com.

Cancer Informatics

Supplementary Issue: Classification, Predictive Modelling, and Statistical Analysis of Cancer Data (A)

Combined Benefit of Prediction and Treatment: A Criterion for Evaluating Clinical Prediction Models

Dean Billheimer[1,2], Eugene W. Gerner[3], Christine E. McLaren[4] and Bonnie LaFleur[5]

[1]Agricultural and Biosystems Engineering, College of Agriculture and Life Sciences, [2]The BIO5 Institute, The University of Arizona, Tucson, AZ. [3]Cancer Prevention Pharmaceuticals, Tucson, AZ. [4]Department of Epidemiology and Genetic Epidemiology Research Institute, University of California Irvine, Irvine, CA. [5]Ventana Medical Systems, Tucson, AZ.

ABSTRACT: Clinical treatment decisions rely on prognostic evaluation of a patient's future health outcomes. Thus, predictive models under different treatment options are key factors for making good decisions. While many criteria exist for judging the statistical quality of a prediction model, few are available to measure its clinical utility. As a consequence, we may find that the addition of a clinical covariate or biomarker improves the statistical quality of the model, but has little effect on its clinical usefulness. We focus on the setting where a treatment decision may reduce a patient's risk of a poor outcome, but also comes at a cost; this may be monetary, inconvenience, or the potential side effects. This setting is exemplified by cancer chemoprevention, or the use of statins to reduce the risk of cardiovascular disease. We propose a novel approach to assessing a prediction model using a formal decision analytic framework. We combine the predictive model's ability to discriminate good from poor outcome with the net benefit afforded by treatment. In this framework, reduced risk is balanced against the cost of treatment. The relative cost–benefit of treatment provides a useful index to assist patient decisions. This index also identifies the relevant clinical risk regions where predictive improvement is needed. Our approach is illustrated using data from a colorectal adenoma chemoprevention trial.

KEYWORDS: predictive modeling, decision analysis, model evaluation, chemoprevention

SUPPLEMENT: Classification, Predictive Modelling, and Statistical Analysis of Cancer Data (A)

CITATION: Billheimer et al. Combined Benefit of Prediction and Treatment: A Criterion for Evaluating Clinical Prediction Models. *Cancer Informatics* 2014:13(S2) 93–103 doi: 10.4137/CIN.S13780.

RECEIVED: March 12, 2014. **RESUBMITTED:** June 29, 2014. **ACCEPTED FOR PUBLICATION:** July 1, 2014.

ACADEMIC EDITOR: JT Efird, Editor in Chief

TYPE: Methodology

FUNDING: Authors disclose no funding sources.

COMPETING INTERESTS: EWG reports that he is a co-founder, stock holder and paid consultant to Cancer Prevention Pharmaceuticals, and holds two patents (8,329,636 and 6,258,845) licensed to Cancer Prevention Pharmaceuticals. All other authors disclose no competing interests.

CORRESPONDENCE: bonnie.jean.lafleur@gmail.com

This paper was subject to independent, expert peer review by a minimum of two blind peer reviewers. All editorial decisions were made by the independent academic editor. All authors have provided signed confirmation of their compliance with ethical and legal obligations including (but not limited to) use of any copyrighted material, compliance with ICMJE authorship and competing interests disclosure guidelines and, where applicable, compliance with legal and ethical guidelines on human and animal research participants. Provenance: the authors were invited to submit this paper.

Introduction – Assessing Prediction Models to Support Clinical Decisions

What is the situation? A fundamental problem of medical decision making is that of prognosis.[1] The patient and clinician must decide which among the available treatments is likely to lead to the best outcome *for that particular patient*. When there is heterogeneity in individual patient's risk for poor outcome, reliance on the population 'mean' treatment effect may be of limited value. We seek a personalized prediction of patient health trajectories for the different treatments under consideration. In addition, many medical treatments come with both expected and unintended consequences (eg, monetary cost, inconvenience, side effects). Optimal treatment decisions must weigh a patient's likely benefits against the risk and severity of these consequences. In the following, we focus on situations in which the different treatment choices affect the probability of patient outcomes. Our goal is to evaluate the quality of prediction models or rules in the presence of uncertainty about outcomes.

Statistical model selection and prediction assessment are long-standing problems in the field of statistics (see eg,[2,3]). Much effort has been focused on the statistical properties of predictive models and their predictions. Common evaluation criteria include the Brier score and the area under

the receiver operating characteristic (ROC) curve. However, the clinical benefit of an improved predictive model remains difficult to assess. New measures are emerging which seek to quantify the clinical utility of predictions. These include reclassification measures (net reclassification improvement and integrated discrimination improvement[4]), as well as decision curves[5] and relative utility curves.[6,7] The decision curve analysis quantifies the clinical utility of a diagnostic prediction model by incorporating harms and benefits into an optimal decision threshold. The advantage of the Vickers and Elkin (VE)[5] approach is that a risk probability threshold can be used to "both categorize patients as positive or negative and to weight the false-positive and false-negative classifications".[8] Baker et al.[6] extend decision curves ideas to evaluate the relative expected maximum utility. This is the ratio of expected utility achieved by a risk prediction model to that obtained by perfect prediction. A key idea in both Ref. 5 and 6 is that the importance of harms and benefits may differ from patient to patient. Both approaches consider a range of thresholds appropriate to a particular diagnostic situation.

What is our solution? We propose a novel approach to evaluating prediction models using a decision analytic framework. Our work stems from the observation that a prediction model is clinically useful only if it changes a treatment decision and the prediction-supported treatment improves the patient's outcome compared to that which would have occurred with the original treatment choice. The clinical utility of prediction relies on the availability of better treatment options. Our approach combines a predictive model's ability to discriminate good from poor outcome with the benefits afforded by treatment. It also includes the (potential) negative consequences of treatment. We term this combination of predictive model and treatment efficacy the "combined benefit" (CB) of predictive treatment. We focus on a setting where the proposed treatment reduces a patient's risk (probability) of a poor primary outcome. The choice of whether to take a chemopreventive agent is our motivating example.

To preview our result, consider the probabilities of acquiring disease when doing nothing or taking a treatment, p_N and p_T, respectively. With each choice there is an associated "cost" (money, side effects, inconvenience), C_N and C_T. Also, we must consider the patient's utility (a valuation of a patient's preferences for different outcomes) for acquiring disease. Let U_0 be the patient's utility for no disease, and U_D that for acquiring disease. Then, the standard decision rule[9] is that treatment should be selected only if

$$p_N - p_T > \frac{C_T - C_N}{U_0 - U_D} \tag{1}$$

That is, the reduction in risk of disease is greater than the cost-to-benefit ratio of treatment.

To assess a prediction model or treatment rule, we propose the CB criterion. This combines model-based predictions of acquiring disease (p_N and p_T) with costs and benefits associated with treatment.

$$\text{Combined benfit} = f_0 - f_T\left(\frac{C_T - C_N}{U_0 - U_D}\right) \tag{2}$$

where f_0 denotes the fraction of eligible patients who subsequently do not develop disease, and f_T the fraction who are treated. To evaluate a prediction model, the CB criterion may be considered a function of the cost–benefit ratio. The model influences the criterion through estimates of p_N and p_T, and their subsequent effect on treatment decisions and the fraction of patients treated. A model results in a larger CB if it correctly identifies patients who will benefit from treatment, and those who will not.

The cost–benefit ratio may be considered a patient-specific threshold for selecting treatment. Competing prediction models and/or treatment rules can be compared at each threshold value. It is possible for one model to provide greater benefit when the treatment cost is high, but a different model to be superior with low treatment cost. Further, patients have heterogeneous attitudes toward treatment cost and benefit. Thus, identifying a relevant range of treatment thresholds is key to evaluating competing prediction models. We note that individual patients do not benefit directly from the proposed CB framework. The benefit is indirect, and is achieved through the use of decision support models tuned to problem-specific costs and benefits.

Links to similar approaches. Our approach follows directly from an application of decision analysis, and is related to several results reported previously. Observe that the decision rule above is related to a widely used measure of clinical effectiveness, the "number needed to treat" (NNT).[10] This is the number of patients who must be treated to prevent one patient's disease. The form of this measure is

$$\text{NNT} = \frac{1}{p_N - p_T}$$

Clearly, NNT is the reciprocal of the standard decision rule (eqn. 1, above), but in our approach, it is scaled by the relative benefit and cost of treatment.

Also, our approach is similar to Vickers and Elkin[5] and to Baker et al.[6] for evaluating diagnostic prediction.

$$\frac{p}{(1-p)} = \frac{\text{Loss of overtreatment}}{\text{Benefit of treating diseased}}$$

where p is an individual's probability of disease. All three approaches rely on a formal decision analysis framework, and all consider a relevant region of risk, which is most useful for clinical decision making.

However, our approach differs in several important ways. First, we are concerned with problems in which the proposed treatment reduces the risk of disease. This leads to a criterion based on the difference in risk probabilities. Conversely, Ref. 5 and 6 consider the problem of diagnosis, and their criterion follows the odds ratio. Second, our CB measure relies on both the predictive model and the costs and benefits of the treatment. VE's "net benefit" criterion combines predictive accuracy with the costs of misclassification. Finally, CB makes use of utilities from both treated and untreated patients, whereas net benefit considers only patients with a positive diagnosis.[11] By comparison, Baker et al.[6] developed a relative utility curve, which compares the performance of a risk prediction model with that achieved by perfect prediction. They also propose a "test threshold": the minimum number of tests that would be traded for a true positive while maintaining non-negative expected utility.

Other perspectives. Our approach relies on a Bayesian perspective of decision making under uncertainty.[12,13] Specifically, it allows personalistic, subjective probabilities and utilities. Despite a scientific history since the 1930s,[14,15] there remain both practical difficulties and philosophical foundation controversy regarding this approach. Practically, evaluating and quantifying each patient's cost–benefit ratio (eg, in eqn. 1) is a key challenge. Both costs and benefits are composed of multiple objectives, and contribute to patient's highly personalistic utility valuations. In addition, the philosophical foundations of Bayesian decision theory have been criticized for their subjective nature, behavioristic decision making (rather than scientific inference), and reliance on semi-empirical, a priori reasoning.[16,17]

The next section introduces a motivating example in the area of colorectal adenoma chemoprevention. We make use of data from a clinical trial[18] evaluating a drug treatment to prevent adenoma recurrence. This trial exhibits key features that motivate our approach, and is an informative example for evaluating a predictive model. Note, however, we do not consider this as an analysis of the trial data. Because formal decision analysis is frequently omitted from informatics, biostatistics, and epidemiology training, Section 3 reviews the principles involved. Section 4 develops the CB measure, and Section 5 demonstrates its use with the adenoma chemoprevention trial data. Finally, we discuss ramifications of using formal decision analysis techniques to evaluate patient treatment decisions.

Example: Chemoprevention of Colorectal Adenoma

To motivate development, we consider a chemoprevention trial to prevent recurrence of colorectal adenomas.[18] This trial was hugely successful in recurrence prevention, and has multiple features which make it informative for methodologic examination. We use data from this clinical trial to motivate development of the methods, and to demonstrate use of the predictive model CB analysis.

Difluoromethylornithine (DFMO) and sulindac clinical trial overview. Three hundred seventy-five patients with a history of resected adenoma were randomly assigned to an oral chemopreventive, DFMO plus sulindac, or placebo following a stratified randomization scheme. Colonoscopies were performed at baseline and three years post-randomization. An independent data safety and monitoring board recommended early-stopping of the study for treatment efficacy. There were 267 evaluable patients: 129 in the placebo arm and 138 assigned to treatment with DFMO. Adenoma recurrence was 41% in the placebo group, and only 12% for patients treated with DFMO (risk ratio 0.30, 95% confidence interval 0.18–0.49, $P < 0.001$).

Trial safety: side effects with chemopreventive treatment. Any chemopreventive may increase the risk of side effects and adverse events. The DFMO treatment suggests small increases in risk of several side effects (shown in Table 1). None of the treatment groups comparisons reached statistical significance ($P < 0.05$). Nevertheless, any trend toward greater risk with DFMO is the same for the reported conditions. This suggests that we should consider the (predicted) benefit of DFMO treatment and weigh it against potential side effects in considering patient treatment decisions.

Decision problem components. Suppose we now consider treating a new patient with a resected adenoma. The patient has the choice of taking a chemoprevention therapy (DFMO + sulindac) to prevent recurrence. "Should this patient take DFMO + sulindac or not?"

The patient's decision may involve (at least) the following questions:

- What is the patient's risk of adenoma recurrence, say, in 3 years?
- If chemoprevention is chosen, what is the risk of recurrence?
- With chemoprevention, what are the risks and severity of side effects?
- Are there additional treatment risks without chemoprevention (such as risk associated with more colonoscopies)?

The usual statistics of trial reporting (OR = 0.3, $P < 0.001$) are informative about the average response to treatment, but do not tell us about individual patient's risk and benefit. If patients are heterogeneous for baseline risk,

Table 1. Reported frequency of side effects and adverse events (AE) from the DFMO plus sulindac trial, Meyskens et al. 2008.

EVENT	PLACEBO	DFMO + SUL	RISK RATIO
AE w/Hosp.	17%	22%	1.3
Cardiovascular	12%	15%	1.2
Gastrointestinal	8%	13%	1.7
15 dB Hearing Loss	10%	18%	1.9

treatment benefit, or risks of side effects, we need a more personalized approach.

Fundamentals of Decision Analysis

When faced with a decision in the context of uncertain risk and benefit, we rely on Bayesian decision analysis to provide a principled, coherent approach. We provide only a brief overview of the process. For textbook accounts of general Bayesian decision analysis, see eg, Ref. 13 and 19. For a text focusing on medical decisions see Ref. 20. Also, Ref. 9 provides a readable introduction to the implementation of evidence-based medicine as Bayesian decision-making.

A decision analysis explicitly recognizes multiple components of a decision problem. We outline the components and their parallel in the chemoprevention example.

1. The decision maker (DM): patient (and her physician).
2. The set of actions available to DM: take DFMO + sulindac or not.
3. The possible outcomes or consequences that may be uncertain: adenoma recurrence, adverse events, hearing loss, carcinoma.
4. Information or evidence that may be relevant: DFMO and sulindac chemoprevention trial
5. Utility, an assessment of the DM's preferences for the different outcomes: weighs disease recurrence against possible side effects of medication. This also considers less well defined factors such as the requirement of taking daily medication, or increased risk from more colonoscopies. Patients' utilities vary substantially by individual.

The DM's goal is to choose among the possible actions to achieve the best outcome. "Best" is defined by the probability weighted outcome preferences; this is maximum expected utility.

More formally, consider the set of actions $\mathcal{A} = \{a_1, a_2, \cdots, a_k\}$ available to the DM, and that $z \in Z$ are the uncertain outcomes. The choice of a_i induces a probability distribution on Z that may depend on (nuisance) parameters $\theta \in \Theta$. We denote this by

$$p_a(z \mid \theta)$$

The information available about θ is denoted by x, and may be represented by $p(\theta \mid x)$. Finally, the DM's preferences for the different outcomes are described by a *utility function*, $u(z, a)$, which values the different outcomes z for action a (from Ref. 20 p. 55).

The expected utility for each potential action, a_i, may be computed, conditional on information x.

$$U(a) = \int_z \int_\Theta u(z,a)\, p_a(z \mid \theta)\, p(\theta \mid x) \mathrm{d}\theta \mathrm{d}z$$

The best action is the a_i that maximizes expected utility.

Note that we may rearrange the equation, and integrate

$$U(a) = \int_z u(z,a) \int_\Theta p_a(z \mid \theta)\, p(\theta \mid x) d\theta dz \quad (3)$$

$$= \int_z u(z,a) p_a(z \mid x) \mathrm{d}z \quad (4)$$

where $p_a(z \mid x)$ is the (posterior) predictive distribution of outcome z, given information x, when action a is taken. Now the meaning of the equation is clear. We choose the action that maximizes the weighted average of the outcome utilities. The weights correspond to the predictive distribution of outcomes when action a is taken (for each patient).

Combining Risk Prediction and Treatment Benefit

We develop the prediction–treatment CB criterion. To ease interpretation, we describe development in terms of the adenoma chemoprevention example. For this development, we assume that a model is available to predict the probability of adenoma recurrence. This model accounts for differences in baseline risk associated with patient-specific covariates, and for differences in risk associated with chemopreventive treatment. In the next section, we describe one modeling approach to predict heterogeneous probability of recurrence.

For each person, we estimate the reduction in probability of adenoma recurrence associated with DFMO treatment. Let $p_{\mathrm{N}i}$ denote the probability of recurrence for patient i with no treatment, and $p_{\mathrm{T}i}$ the probability with DFMO treatment. If the risk reduction with treatment $(p_{\mathrm{N}i} - p_{\mathrm{T}i})$ is large enough, then treatment is indicated.

We also posit a benefit of avoiding disease recurrence: $U_{0i} - U_{\mathrm{D}i,}$ where U_{0i} denotes the patient's utility of no recurrence, and $U_{\mathrm{D}i}$ their utility of disease recurrence. Similarly, each patient incurs a loss associated with treatment (side effects, inconvenience, cost): $C_{\mathrm{T}i} - C_{\mathrm{N}i}$.[1] Consider this the "cost" of treatment minus the "cost" of no treatment. Clearly costs are more than just monetary.

A standard decision analysis result (Ashby and Smith, 2000) says to treat only if

$$p_{\mathrm{N}i} - p_{\mathrm{T}i} > \frac{C_{\mathrm{T}i} - C_{\mathrm{N}i}}{U_{0i} - U_{\mathrm{D}i}} = \text{cost : benefit ratio}$$

We define the *indifference threshold* (δ_i) to be the probability difference where left and right hand sides of the inequality above are equal.

$$p_{\mathrm{N}i} - p_{\mathrm{T}i} > \frac{C_{\mathrm{T}i} - C_{\mathrm{T}i}}{U_{0i} - U_{\mathrm{D}i}} = \delta_i$$

[1]Side effects should also be considered under uncertain outcomes, and their risk modeled. For simplicity we consider them fixed for each patient.

Thus, we treat only if predicted risk reduction is greater than δ_i. In the next subsection we compare the patient-specific index δ_i against a threshold (δ) to classify patient's treatment decisions.

CB. Now consider a fixed risk reduction threshold S. We may create a table describing treatment choice and outcome for the population of patients eligible for treatment. Note that patients with large risk reduction, $p_{Ni} - p_{Ti} > \delta$, are treated, while those with small risk reduction, $p_{Ni} - p_{Ti} < \delta$, are not. Table 2 illustrates the treatment decision process.

The table entries a, b, c, and d denote the fractions of people treated/not treated, and the fractions with adenoma recurrence/no recurrence. Note that $a + b + c + d = 1$. If all patients are treated, then $a = p_T$, probability of recurrence among treated. If none are treated, then $c = p_N$, probability of recurrence among untreated.

Now, for a fixed δ the expected benefit (expected utility) of the combined treatment and prediction model is

$$\text{Expected Benefit} = (a+c)U_D + (b+d)U_0 - (a+b)C_T - (c+d)C_N$$

Consider this the average benefit per person.

To derive CB, we perform some algebra adding and subtracting $(b + d)U_D$ and $(a + b)C_N$. After rearranging and collecting terms we obtain the following expression:

$$\text{Expected benefit} = (b+d)\left(U_0 - U_D\right) - (a+b)\left(C_T - C_N\right) + (a+b+c+d)\left(U_D - C_N\right)$$

Note that the last term is constant for all values of δ, and can be ignored for decision making. Finally, we divide by $U_o - U_D$ (assume $U_o - U_D > 0$, adenoma recurrence is not the preferred outcome). This results in the CB criterion.

$$\text{Combined benefit}(\delta) = (b+d) - (a+b)\left(\frac{C_T - C_N}{U_0 - U_D}\right) \quad (5)$$

$$= (b+d) - (a+b)\delta \quad (6)$$

For any risk reduction, $\delta = p_N - p_T$, the CB criterion [CB(δ)] is the fraction of people who do not recur, less the fraction who are treated, weighted by the relative cost of treatment. This is the average benefit per person after adjusting for the cost of treatment. Note that if everyone is treated (ALL), then $CB(\delta) = 1 - p_T - \delta$. As a function of δ, this is a line with slope –1. If no one is treated (NONE), then $a = b = 0$, and $CB(\delta) = 1 - p_N$, the fraction of nontreated patients who do not recur.

Table 2. Treatment decision and outcomes for a specific value of δ.

TREATMENT	DEVELOP DISEASE	NO DISEASE
Treated; $p_{Ni} - p_{Ti} > \delta$	a	b
Untreated; $p_{Ni} - p_{Ti} < \delta$	c	d

Use of CB. The relative cost of treatment, δ, is a useful index to aid treatment decisions. At the indifference threshold, δ may be interpreted as both the relative cost of treatment and (predicted) risk reduction necessary to justify treatment. For treatments with a small relative cost (eg, taking a multivitamin), only a small reduction in risk is needed to accept treatment. Conversely, when the relative cost is high (eg, prophylactic colonectomy), then the risk reduction must be large to justify treatment. *Each medical decision has a relevant range of δ values.* We may think of this range spanning the patients' tolerance to risk of poor outcome and to treatment cost.

CB can be used to compare different prediction models or rules, as well as the treat ALL and treat NONE decision rules. Prediction models enter CB(δ) through the computed values p_{Ni} and p_{Ti}. For a fixed risk reduction, different prediction models will perform better or worse at actually classifying patient outcome. We may compute CB(δ) for each prediction model (or rule) across δ values, focusing on the range relevant to the clinical decision. Models with larger CB provide greater benefit. We note three key features of CB.

1. We care only about a specific range of δ values for each decision. Better prediction outside that range is not clinically relevant.
2. CB may be improved by better identification of patients likely to be helped by treatment.
3. CB is also improved by identifying patients unlikely to benefit from treatment.

Predicting a Patient's Risk of Recurrence

We outline our procedure for predicting risk of adenoma recurrence; the details are given in Appendix 1. The goal of the CB criterion is to evaluate the clinical relevance of a prediction model and treatment decisions based on the predictive distributions. The model developed for our adenoma example is intended to illustrate the procedure. It is not intended as an exhaustive analysis of adenoma recurrence.

The primary outcome is adenoma recurrence after three years of follow-up. We model the probability of recurrence using logistic regression with Bayesian model averaging (BMA[21]). BMA accounts for uncertainty in the selection of the prediction model, as well as in the model coefficients. This approach has been shown to improve model predictive performance, and appears less prone to overfitting than alternative procedures.

We fit separate models for placebo- and DFMO-treated patients. In each model, potential predictors include patient demographics (age, sex, body mass index [BMI], aspirin use), as well as characteristics of their baseline adenoma. These characteristics include:

- Location: proximal or distal colon
- Large adenoma (>1 cm)
- Number of adenomas
- Villous (yes/no)

Table 3. Distribution of BMA logistic regression coefficients for placebo patients. Results average over 30 best models retained by BMA.

	PROB $\beta \neq 0$	E[β]	SD[β]
Intercept	1.00	–1.41	0.60
Number of adenomas	0.66	0.31	0.26
Location (proximal)	0.58	0.63	0.62
Aspirin use (yes)	0.19	0.18	0.40
Sex (male)	0.04	0.03	0.17
Sex * Aspirin	0.39	0.40	0.56

Potential molecular (PGE2, putrescine, spermidine) and genotypic (*Ode* and *Fmo3*) biomarkers were also considered. None of these, however, was found to be predictive of recurrence. They are not considered further.

BMA results overview. For patients receiving placebo, the model average fitting summary is shown in Table 3. The second column, Pr($\beta \neq 0$), sums the posterior probabilities across models that include a given predictor. Unlike *P*-values, larger probabilities indicate a greater role in prediction. The number of adenomas and adenoma location at baseline are important predictors of recurrence. These predictors exhibit substantial probability of inclusion in prediction at 0.66 and 0.58, respectively. In addition, aspirin use among male patients adds to predictive ability [Pr($\beta \neq 0$) = 0.39]. Note, however, that aspirin use was very different among males and females, and it is unclear whether this represents an independent effect of aspirin use. See Appendix 1 for details and further interpretation.

Figure 1 shows the posterior predictive probability of recurrence for patients assigned to the placebo group. We observe substantial heterogeneity of recurrence risk ranging from 25% to about 75%. The error bars indicate uncertainty associated with modeling. These regions indicate 66% (black) and 95% (gray) posterior predictive probability. For DFMO-treated patients, none of the predictors has substantial probability of model inclusion. With DFMO treatment, our best prediction is that all patients have about 12% risk of recurrence. This inability to detect important predictors of recurrence is likely because of the small number of recurrences among treated patients (17 of 138). These posterior predictive probabilities will be used in the calculation of the CB criterion. For each patient, they represent our best estimates of p_N and p_T, respectively.

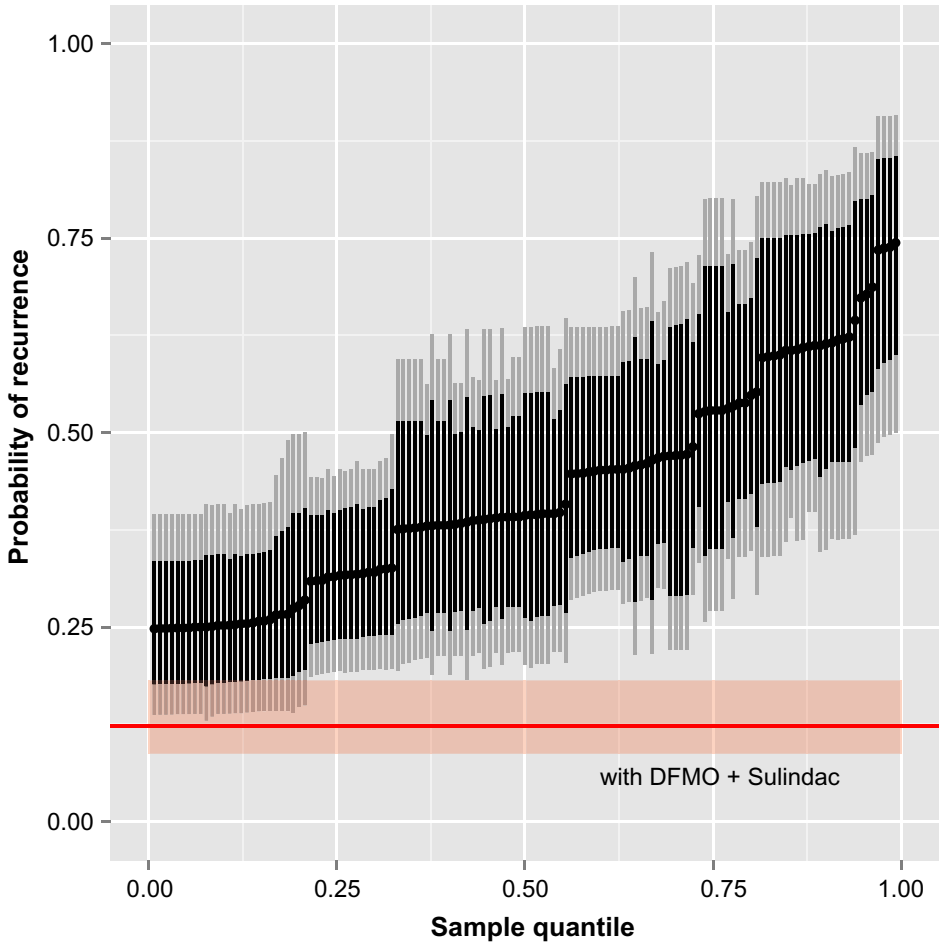

Figure 1. Predicted probability of recurrence for patients with placebo treatment. Center point is the Bayesian model average prediction. Error bars show 66% (black) and 95% (gray) model uncertainty intervals. Orange line denotes the predicted recurrence with DFMO plus sulindac treatment (with 95% credible region).

Results – CB Curves to Assess Prediction

We use the BMA results of the previous section to demonstrate the CB curve method. We use point estimates for disease probabilities and patient fractions a, b, c, and, d based on the observed clinical trial data.[18] While there is some danger of over-optimistic assessment, recall that BMA is robust to overfitting. As with all prediction model assessment, use of an independent test set would provide a more reliable approach.

Figure 2 shows the CB curve [CB(δ)] for the BMA prediction model of adenoma recurrence (blue line). Small values of δ correspond to low cost treatments (those with mild side effects), while large values are associated with high treatment cost. The dashed (black) line corresponds to the CB of treating ALL patients, while the horizontal dotted line denotes the benefit of treating NONE. At $\delta = 0$ (no cost of treatment), the benefit of treating ALL vs. NONE is denoted by the vertical distance between lines ($0.88 - 0.59 = 0.29$). This is the difference in non-recurrence probabilities in treated and placebo arms.

At treatment cost $\delta = 0.29$ (the observed reduction in recurrence), the treat ALL and treat NONE lines cross. Thus, if the cost of treatment is equivalent to 0.29 adenoma recurrences, there is no net benefit to treating all patients (compared with treating none).

The figure shows that for treatment thresholds between 0.13 and 0.50, the BMA prediction provides substantial benefit compared with the treat ALL and treat NONE strategies. This benefit is provided by not treating selected patients with small treatment-related reductions in risk of recurrence.

Note that CB for the BMA prediction and treat ALL strategies coincide for treatment thresholds $\delta < 0.13$. This occurs because the prediction model cannot reliably identify patients with recurrence probabilities less than 0.25 ($p_T = 0.12$ and $\delta = 0.13$; threshold $\approx 0.25 = 0.12 + 0.13$).

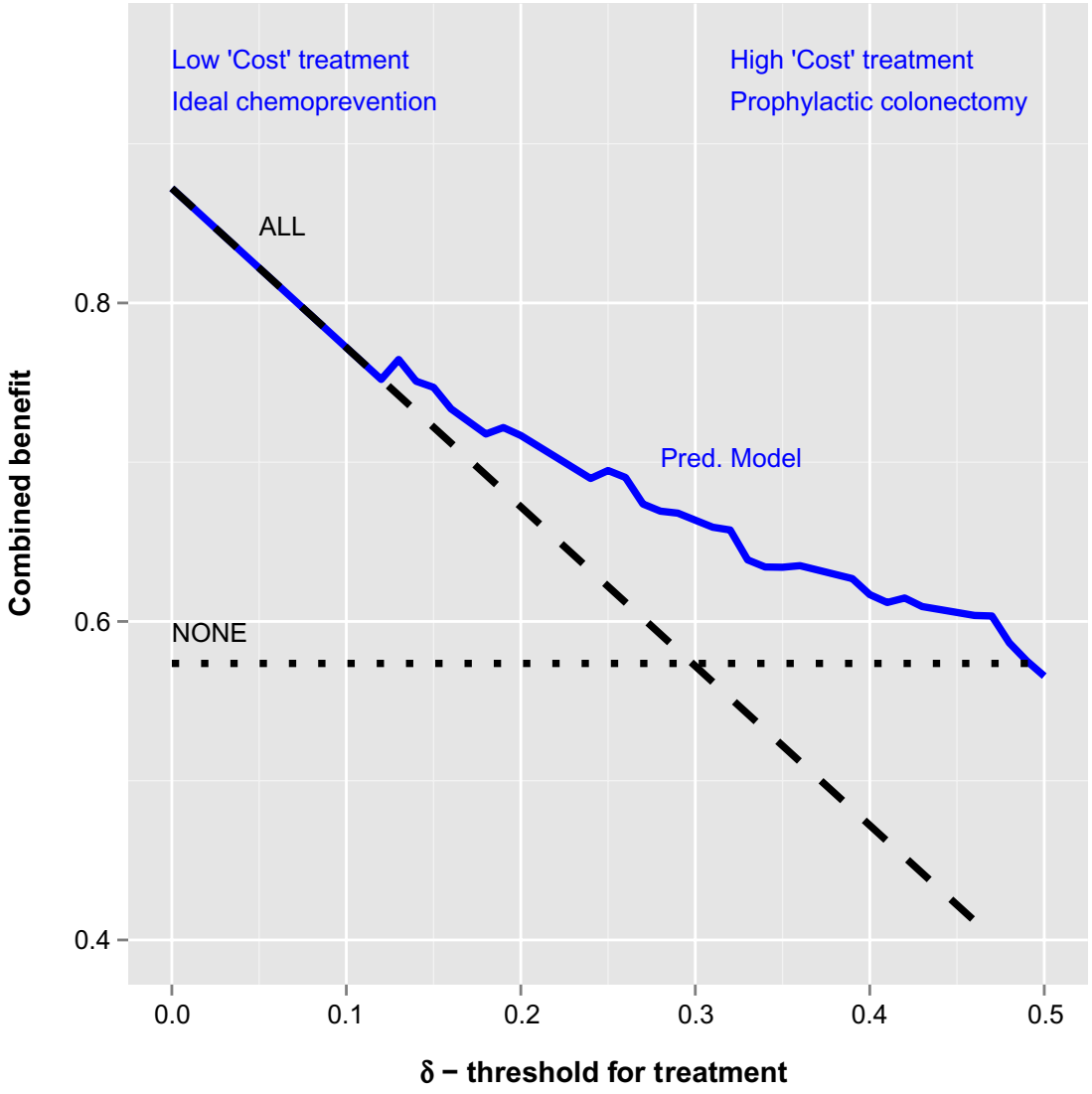

Figure 2. The CB of prediction and treatment (Y axis) for different treatment thresholds δ (X axis). The CB of the BMA prediction model of adenoma recurrence is denoted by the blue line. The dashed (black) line corresponds to the CB of treating ALL patients, while the horizontal dotted line denotes the benefit of treating NONE.

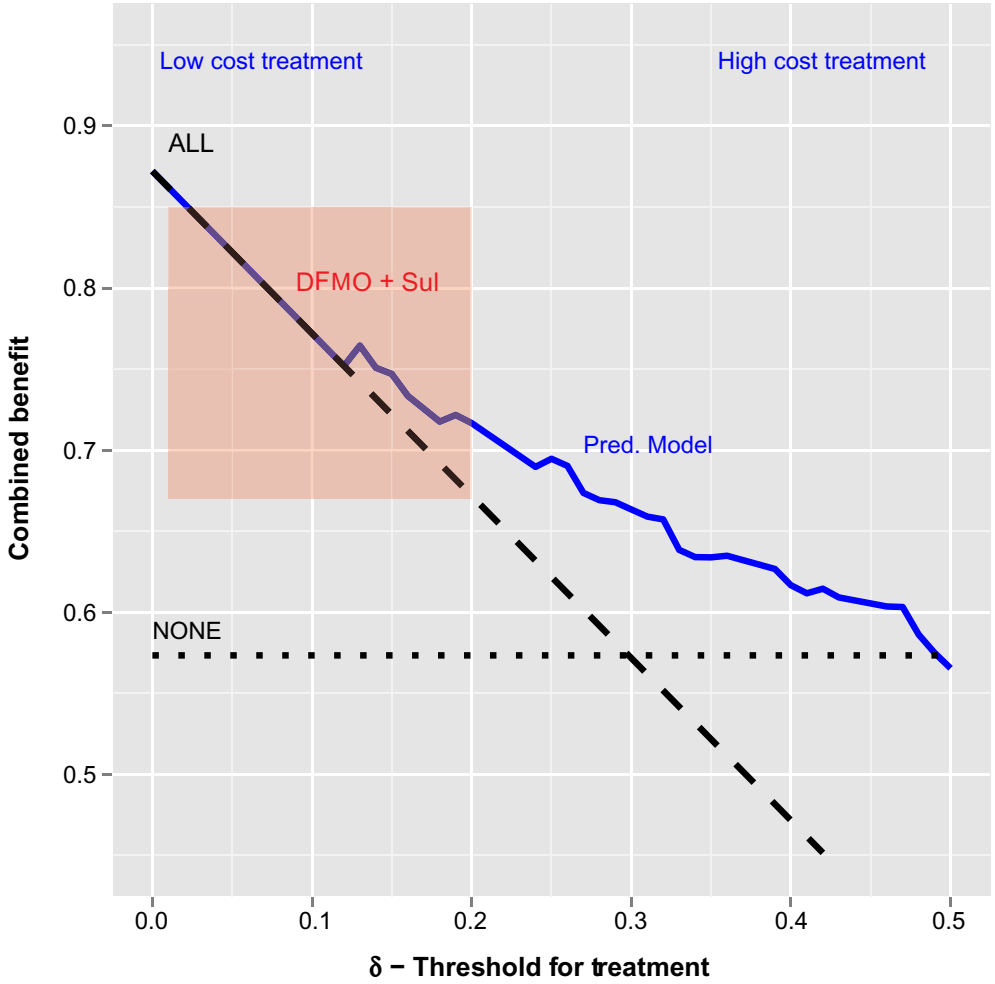

Figure 3. The relevant threshold region for DFMO plus sulindac treatment is indicated by the orange shaded region. Patients with recurrence risk reduction between 0.02 and 0.20 receive limited benefit with DFMO, and might prefer to avoid chemopreventive treatment. The BMA prediction model is relatively poor at identifying such patients.

What is the relevant range of thresholds (δ) associated with the DFMO plus sulindac treatment? Figure 3 shows the same CB curves with an approximate range of relevant treatment thresholds. The DFMO plus sulindac treatment may contribute to potentially serious side effects, but these are only weakly indicated by the trial data. Thus, we posit that small-to-moderate reductions in recurrence risk (0.02–0.20) are sufficient to indicate treatment. Note that the BMA prediction model provides only limited benefit at the upper end of this range. Among patients who are most averse to taking a chemopreventive, we may identify a few with low enough baseline risk to justify avoidance of treatment. This indicates that if we wish to improve prediction in this situation, we should focus on patients with low recurrence probabilities.

We next illustrate how CB can be used to compare different prediction models or rules. Rather than using the full prediction model, suppose we instead choose a risk cut-point and treat all patients exceeding that point. Figure 4 shows the CB curve when that risk probability is 0.40 (approximate frequency of recurrence in the placebo arm). With this simplified rule, we obtain much of the benefit afforded by the full BMA model, and vastly exceed the CB obtained by the treat ALL rule. This benefit is obtained by excusing low-risk patients from treatment. Note that this simple rule fixes each patient's decision threshold at $\delta = 0.28$.[2]

Finally, Figure 5 compares the performance of the full BMA prediction model with a restricted model that omits adenoma location (restricted model). This demonstrates how predictions based on different covariates (eg, biomarkers) can be compared. The inclusion of adenoma location provides a modest improvement in predictive performance. But, this improvement is realized primarily among smaller threshold values ($\delta < 0.33$). These smaller thresholds are more relevant for this chemoprevention treatment decision.

Discussion

Summary. We have developed a criterion that combines a patient's predicted outcomes under different treatment options with consideration of loss associated with the treatment.

[2]This threshold is greater than our proposed risk region for this treatment decision. We include this to demonstrate the use of CB.

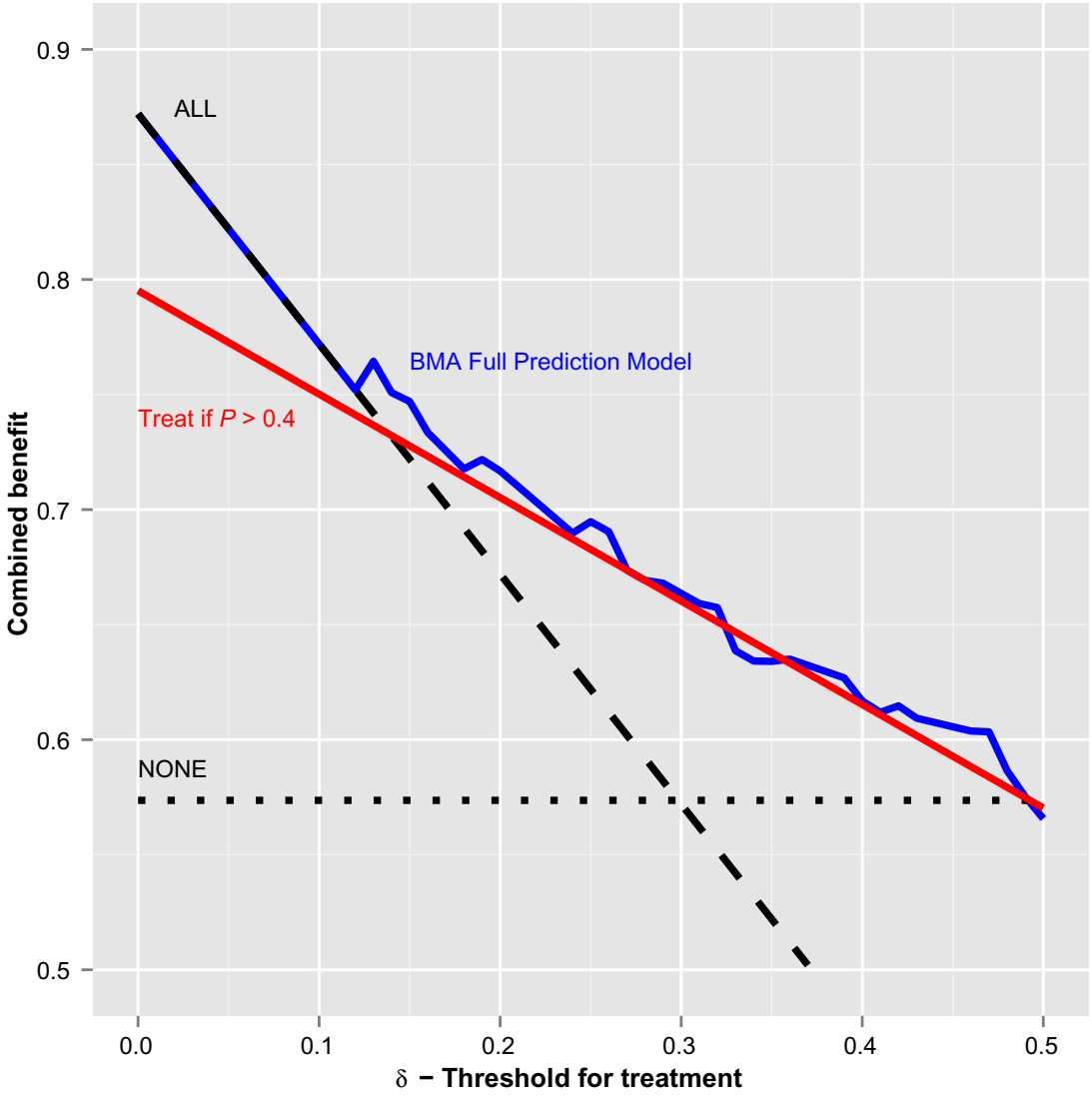

Figure 4. CB curve for a fixed decision probability of 0.40. This simpler rule achieves much of the benefit of the full BMA prediction. The equivalent threshold is $\delta = 0.28$.

The CB curve helps us focus on the relevant risk groups by considering only the range of risk reduction that is consistent with the relative cost of treatment. The CB curves can be used to compare different prediction models, the contribution of potential biomarkers to an existing model, and different treatment decision rules.

In our motivating example for chemoprevention of colorectal adenoma, we observe that there is substantial interpatient heterogeneity of recurrence risk among untreated patients. However, over the risk region of interest we are unable to identify patients who would benefit by avoiding treatment. This example demonstrates that clinically beneficial improvements in prediction (eg, new biomarkers) should identify patients with very low risk of recurrence – those who would benefit by avoiding treatment. While not addressed in our example, it would also be useful to identify patients with high risk of experiencing side effects associated with treatment.

For the medical community to fully embrace personalized medicine, we need improved approaches for assessing treatment decisions. These include improvements in

- predicting what will happen to individual patients,
- evaluating predictive models,
- incorporating treatment benefits and consequences, and
- understanding patient utilities for outcomes.

The decision analytic approach outlined above demonstrates how these components interact, and that evaluation of individual components in the absence of the others is incomplete. We argue that prediction–decision statistical approaches are more relevant for clinical decision support than *P*-value based inference for treatment efficacy.

Why not use clinical trials for benefit assessment? Clinical trials provide a wealth of information about patients

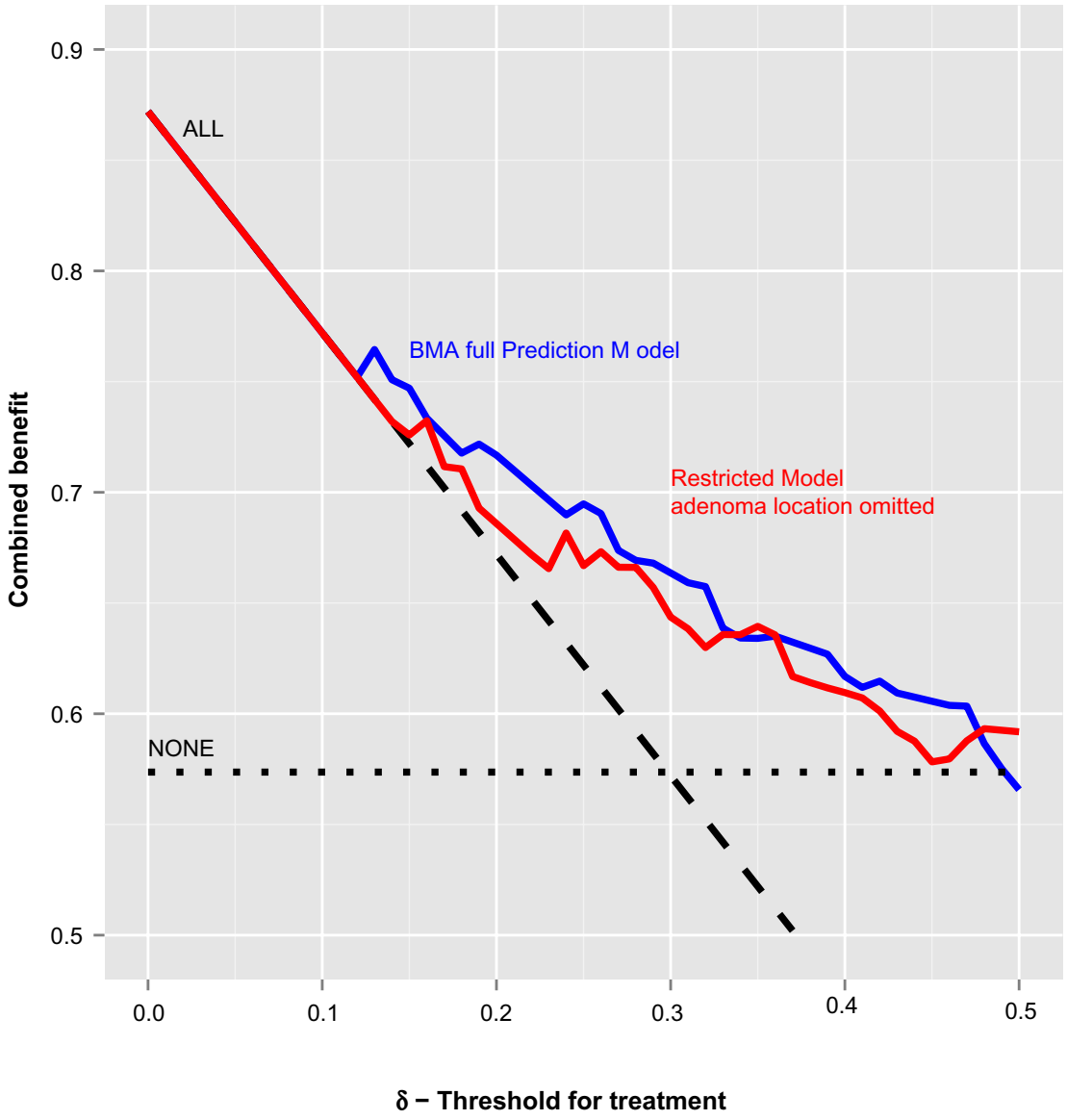

Figure 5. CB curves for BMA prediction with all covariates (blue) and for model averaged predictions with adenoma location omitted (restricted model, red). Note that the full model outperforms the restricted model up to $\delta = 0.33$. The two models exhibit similar performance at higher thresholds.

with disease or those susceptible to it. In addition, trials include a formal monitoring mechanism to assess outcomes, and to evaluate side effects. As we demonstrate, this information is useful for estimating patient outcome predictive distributions, and is necessary to evaluate clinical benefit (not just treatment efficacy). A slight expansion of current clinical trial protocols would include information about patient utilities. This additional information would allow a more complete picture of the benefits of treatment.

Our societal trend toward personalized medicine indicates that we need more information about "who to treat," and less focus on "which treatment to use." Such a shift in perspective would change the focus of clinical trials from drug superiority to one of patient benefit. This seems much more relevant for health care than the usual *P*-value based inference for efficacy.

Author Contributions

DB and BL conceived the concepts and wrote the first draft of the manuscript. DB, CEM, and BL analyzed the data. DB, EWG, CEM and BL contributed to the writing, made critical revisions and approved the final version. All authors reviewed and approved of the final manuscript.

REFERENCES

1. Hemingway H, Croft P, Perel P, et al. Prognosis research strategy (progress) 1: A framework for researching clinical outcomes. *BMJ*. 2013;346:e5595.
2. Geisser S. *Predictive Inference: An Introduction*. Chapman and Hall; 1993. New York.
3. Harrell FE. *Regression Modeling Strategies*. Springer; 2001. New York.
4. Pencina MJ, Agostino RBD, Agostino RBD, Vasan RS. Evaluating the added predictive ability of a new marker: From area under the roc curve to reclassification and beyond. *Stat Med*. 2008;27(2):157172.
5. Vickers AJ, Elkin EB. Decision curve analysis: A novel method for evaluating prediction models. *Med Decis Making*. 2006;26:565–74.
6. Baker SG, Cook NR, Vickers A, Kramer BS. Using relative utility curves to evaluate risk prediction. *J R Stat Soc Ser A Stat Soc*. 2009;172(4):729–48.

7. Baker SG, Van Calster B, Steyerberg EW. Evaluating a new marker for risk prediction using the test tradeoff: An update. *Int J Biostat*. 2012;8(1). 1–37.
8. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models. *Epidemiology*. 2010;21(1):128–38.
9. Ashby D, Smith AFM. Evidence-based medicine as bayesian decision-making. *Stat Med*. 2000;19:3291–305.
10. Laupacis A, Sackett D, Roberts R. An assessment of clinically useful measures of the consequences of treatment. *N Engl J Med*. 1988;318(26):1728–33.
11. Rousson V, Zumbrun T. Decision curve analysis revisited: overall net benefit, relationships to roc curve analysis, and application to case-control studies. *BMC Med Inform Decis Mak*. 2011;11(45):l-9.
12. Savage LJ. *The Foundations of Statistics*. Dover; 1972. New York.
13. French S, Inusia DR. Statistical Decision Theory, Kendall's Library of Statistics, 9. Arnold Publishers; 2000. New York.
14. Ramsey FP. *The Foundations of Mathematics and Other Essays*. London: Routledge & Kegan Paul; 1931.
15. De Finetti B. La prevision, ses lois logiques, ses sources subjectives. *Ann Inst Henri Poincare*. 1937;7:1–68.
16. Albert M. Bayesian rationality and decision making: A critical review. *Anal Kritik*. 2003;25:101–17.
17. Birnbaum A. The neyman-pearson theory as decision theory, and as inference theory; with a criticism of the lindley-savage argument for bayesian theory. *Synthese*. 1977;36:19–49.
18. Meyskens FL, McLaren CE, Pelot D, et al. Difluoromethylornithine plus sulindac for the prevention of sporadic colorectal adenomas: A randomized placebo-controlled, double-blind trial. *Cancer Prev Res*. 2008;1(1):32–8.
19. Bernardo JM, Smith AFM. *Bayesian Theory*. Wiley; 1994. New York.
20. Parmigiani G. Modeling in Medical Decision Making: A Bayesian Approach. John Wiley and Sons; 2002. Chichester.
21. Hoeting JA, Madigan D, Raftery AE, Volinsky CT. Bayesian model averaging: A tutorial. *Stat Sci*. 1999;14(4):382–417.
22. Barbieri MM, Berger JO. Optimal predictive model selection. *Ann Stat*. 2004;32(3):870–97.
23. Raftery A, Hoeting J, Volinsky C, Painter I, Yeun KY. BMA: Bayesian model averaging. 2013. R Package Version 3.16.1. http://CRAN.R-project.org/package=BMA.
24. R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing; 2012.

Appendix 1

The primary outcome is adenoma recurrence after three years of follow-up. We model the probability of recurrence using logistic regression with Bayesian model averaging (BMA[21]). BMA accounts for uncertainty in the selection of the prediction model, as well as in the model coefficients. This approach has been shown to improve model predictive performance, both theoretically and empirically.[22] It also appears to be less prone to overfitting than alternative procedures.

We fit separate models for placebo- and DFMO-treated patients. In each model, potential predictors include patient demographics (age, sex, BMI, aspirin use), as well as characteristics of their baseline adenoma. These characteristics include:

- Location: proximal or distal colon
- Large (>1 cm)
- Number of adenomas
- Villous (yes/no)

Potential molecular (PGE2, putrescine, spermidine) and genotypic (*Ode* and *Fmo3*) biomarkers were also considered. None of these, however, was found to be predictive of recurrence. They are not considered further.

All patient demographics and adenoma characteristics were included as potential predictors. Patient age was included using a restricted cubic spline to allow nonlinear association with recurrence probability. In addition, age-by-sex and aspirin use-by-sex interactions were also considered as potential predictors. BMA was implemented using the BMA package[23] in R[24]. The prior probability of inclusion of each predictor was 0.3. Sensitivity analysis indicated that similar results were obtained over a range of prior values (0.10–0.50). Models with posterior probability greater than 0.005 were retained for inclusion in prediction. Posterior predictive distributions were estimated using Monte Carlo sampling and each patient's observed covariate values. Monte Carlo sampling was conducted by first selecting a retained model according to posterior model probabilities, and then drawing parameter values from that model's posterior (parameter) distribution.

BMA results. Among patients assigned to placebo treatment, we find substantial uncertainty about the "best" logistic regression model. BMA has great advantage in this situation. The single best regression model accounts for only 12% of posterior model probability. The top five models account for 52% of posterior probability. The top 30 models were retained to estimate recurrence probability. Table A1 summarizes coefficients, averaging over these top 30 models.

The table shows the results for selected predictors. The posterior probability of inclusion (Prob $\beta \neq 0$) is the sum of posterior model probabilities, which include a given predictor. Thus, we see that the number of adenomas and the location of the adenoma at baseline have greater than 50% posterior probability. The posterior expected value and standard deviation of the coefficients reflect a mixture of distributions in which the coefficient is included ($\beta \neq 0$) or excluded ($\beta = 0$). We also see that sex and aspirin use, separately, have modest effects on prediction of recurrence. However, their interaction effect has substantial posterior probability (nearly 40%). We observed much greater aspirin use among men than among women. The estimated interaction may be related to this unequal distribution.

Table A1. Distribution of logistic regression coefficients for placebo patients. Results average over 30 best models retained by BMA.

	PROB $\beta \neq 0$	E[β]	SD[β]
Intercept	1.00	-1.41	0.60
Number of adenomas	0.66	0.31	0.26
Location (proximal)	0.58	0.63	0.62
Aspirin use (yes)	0.19	0.18	0.40
Sex (male)	0.04	0.03	0.17
Sex * Aspirin	0.39	0.40	0.56