The Intervention Probability Curve: Modeling the Practical Application of Threshold-Guided Decision-Making, Evaluated in Lung, Prostate, and Ovarian Cancers

Michael N. Kammer[1], Dianna J. Rowe[1], Stephen A. Deppen[1,2], Eric L. Grogan[1,2], Alexander M. Kaizer[3], Anna E. Barón[3], and Fabien Maldonado[1]

ABSTRACT

Background: Diagnostic prediction models are useful guides when considering lesions suspicious for cancer, as they provide a quantitative estimate of the probability that a lesion is malignant. However, the decision to intervene ultimately rests on patient and physician preferences. The appropriate intervention in many clinical situations is typically defined by clinically relevant, actionable subgroups based upon the probability of malignancy. However, the "all-or-nothing" approach of threshold-based decisions is in practice incorrect.

Methods: Here, we present a novel approach to understanding clinical decision-making, the intervention probability curve (IPC). The IPC models the likelihood that an intervention will be chosen as a continuous function of the probability of disease. We propose the cumulative distribution function as a suitable model. The IPC is explored using the National Lung Screening Trial and the Prostate Lung Colorectal and Ovarian Screening Trial datasets.

Results: Fitting the IPC results in a continuous curve as a function of pretest probability of cancer with high correlation ($R^2 > 0.97$ for each) with fitted parameters closely aligned with professional society guidelines.

Conclusions: The IPC allows analysis of intervention decisions in a continuous, rather than threshold-based, approach to further understand the role of biomarkers and risk models in clinical practice.

Impact: We propose that consideration of IPCs will yield significant insights into the practical relevance of threshold-based management strategies and could provide a novel method to estimate the actual clinical utility of novel biomarkers.

Introduction

Difficult decisions are required after the detection of a suspicious lesion. Diagnostic prediction models are useful in this decision-making process, as they provide a quantitative estimate of the probability that the lesion is malignant (1). The appropriate intervention in many clinical situations is guided by evidence-based recommendations provided by experts and relevant professional societies, which typically define clinically relevant threshold-based subgroups, such as low, intermediate, or high probability. For example, there are diagnostic prediction models for lung (2–4), ovarian (5, 6), thyroid (7), and prostate cancer (8), which typically incorporate patient medical history, clinical data, imaging and ultrasonographic characteristics of the lesion, and other biomarkers into a single probability score of likelihood of cancer. The probability thresholds are set to optimize the tradeoff between true positives, true negatives, false positives, and false negatives, and the risk–benefit ratios of interventions considered (9). However, in practice, the decision to intervene, whether through surgery, medications, further imaging, or behavioral change, is a function of patient and physician preferences made in the context of a shared decision-making process that incorporates factors beyond that probability estimate (10, 11). There are indeed many other important considerations that ultimately dictate which intervention is in the patient's best interest, such as the patient's age, comorbidities, quality of life, or financial situation (12, 13).

Thus, while conventional wisdom would expect clinical practice to be aligned with threshold-based recommendations (**Fig. 1**), practical data demonstrate that this is not the case. But while the "all-or-nothing" approach of threshold-based decisions is not strictly followed, a less linear correlation between these thresholds and actual patient decisions does seem to emerge. We present a novel approach to understanding clinical decision-making, the intervention probability curve (IPC). The IPC models the likelihood that an intervention will be chosen as a continuous function of the probability of disease. This function can be estimated using professional society guidelines or can be obtained by regression fitting to clinical data on intervention decisions.

Herein, we describe several general situations, and then fit the IPC to three large cohort studies: (i) the likelihood of invasive diagnostic procedure for indeterminate pulmonary nodules detected via low-dose CT scan in the National Lung Screening Trial (NLST; ref. 14), (ii) the likelihood of biopsy for ovarian abnormalities within the screening setting in the Prostate, Lung, Colorectal, and Ovarian (PLCO; ref. 15) study, and (iii) the likelihood of biopsy for prostate nodules detected within the screening setting in the PLCO study. We propose that consideration of IPC curves will yield significant insights into the practical relevance of threshold-based management strategies and could ultimately provide a clinically relevant method to estimate the clinical utility of novel candidate biomarkers.

[1]Vanderbilt University Medical Center, Nashville, Tennessee. [2]Tennessee Valley Healthcare Administration Nashville Campus, Nashville, Tennessee. [3]Colorado School of Public Health, University of Colorado Anschutz Medical Campus, Aurora, Colorado.

Corresponding Author: Fabien Maldonado, 2220 Pierce Avenue, Nashville, TN 37232. Phone: 615-936-8422; E-mail: fabien.maldonado@vumc.org

Cancer Epidemiol Biomarkers Prev 2022;31:1752–9

doi: 10.1158/1055-9965.EPI-22-0190

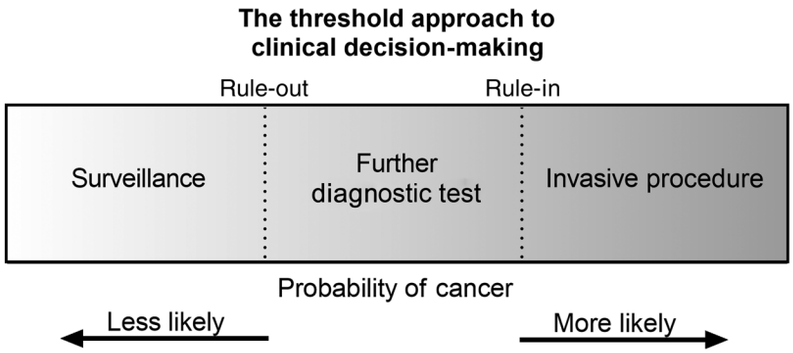

Figure 1.
Decision thresholds and the probability of cancer. The rule-out threshold is the probability of cancer below which careful observation with serial CT is warranted. The rule-in threshold is the probability of cancer above which invasive procedures, such as biopsy or surgery, are warranted. Diagnostic testing with PET, follow-up imaging, or biomarker test is warranted for probabilities of cancer between these two extremes. The location of these thresholds is based upon the rates of false positives and false negatives, the benefits of true positives and true negatives, the difficulty and cost of procedures, and patient preferences for treatment.

Materials and Methods

Predictive model calculations

Lung cancer screening

The NLST dataset was obtained from the NCI. The NLST is a multicenter, randomized controlled trial comparing low-dose helical CT with chest radiography for the screening of current and former heavy smokers for lung cancer (14). All low-dose CT acquisitions were interpreted at the screening center by radiologists approved to read for the NLST. Images were reviewed for the presence of lung nodules, masses, or other abnormalities suspicious for lung cancer, and the detection of any of these was described as a positive screening result. Diagnostic evaluation was performed outside the context of the NLST according to standard of care, although information on diagnostic evaluation performed in response to a positive screening result was collected. Diagnostic procedures considered "invasive" included: transthoracic CT-guided, bronchoscopic or surgical lung biopsy. Data from nodules detected within the CT arm of the trial were used to calculate the Mayo Clinic Model probability of cancer, which incorporates the patient's age, ever-smoker vs. never smoker status, and prior cancer, in addition to lung nodule largest diameter, presence of spiculation, and lobe location into a single probability score from 0 to 1. For patients with a cancer diagnosis, the data from the screening visit immediately preceding the cancer diagnosis was used. For patients with no cancer diagnosis, the first screening visit with a reported CT abnormality was used.

Prostate cancer screening

The PLCO cohort study was used to assess the management of abnormalities detected at screening for prostate and ovarian cancers (15). The PLCO cohort comprises over 150,000 participants aged 55 to 74 years old at time of consent who were randomized into a screening arm and standard-of-care arm.

Within the screening arm of the PLCO prostate dataset, men ($n =$ 38,340) were screened yearly for up to 5 years. Each screen included serum PSA level quantification and a digital rectal exam (DRE). The screening results were used to calculate the Prostate Biopsy Collaborative Group (PBCG) probability, a prediction model of prostate malignancy, modified from the known Prostate Cancer Prevention Trial (PCPT) probability calculator (8). The PBCG probability model incorporates age, serum PSA concentration, family history of prostate cancer, and results from a DRE into a single probability score from 0 to 1. The PBCG probability was calculated for each screening visit, resulting in multiple results per participant. Each screening visit was treated as an independent probability assessment. Probability calculations were performed in R version 4.0.3 (16).

Ovarian cancer screening

Within the screening arm of the PLCO ovarian dataset, women ($n =$ 39,105) were screened yearly for up to 5 years with each screen including a transvaginal ultrasound (TVU) and cancer antigen 125 (CA-125) quantification. The results of these screenings were used to calculate the Risk of Malignancy Index (RMI), as described previously (5, 6). Briefly, the RMI incorporates menopausal status, serum CA-125 concentration, and categorical descriptors of the mass on ultrasound including solid areas, multiloculated cysts, bilateral lesions, ascites, and intra-abdominal masses in a single prediction model with a value that can range from 0 to infinity, although a typical score for a benign mass is RMI < 25. The RMI was calculated for each screening visit with an abnormality, resulting in multiple results per participant. For patients with an intervention procedure, the RMI from the screening visit that was associated with decision to perform the procedure was used. For patients with no invasive procedure, the RMI from the first visit with a nonzero RMI was used.

Fitting the IPC

The IPC describes the likelihood that an intervention was in practice selected based on the pretest probability of a given malignancy as defined by its respective clinical prediction model. A curve was fit to the data using a logistic regression, with binary outcome variables (1 = intervention, 0 = no intervention). A custom IPC curve based on the cumulative distribution function (derived below) was fit to the data using nonlinear least squares regression, using the pretest probability of cancer as the independent variable and the intervention outcome (1 = intervention, 0 = no intervention) as the independent variable. To fit the IPC to binned clinical data, several approaches were used. In the first approach, patients were grouped into equal-population bins based upon the probability of cancer, and then the proportion of patients in each group undergoing intervention was calculated. For each group, the average probability of cancer was used as the dependent variable and was plotted against the proportion of patients undergoing intervention. The cumulative distribution function (CDF)-based IPC (derivation presented below) was fit to this data using nonlinear least squares regression. In the second approach, patients were grouped into equal-width bins, using 20 bins ranging from 0 to 1 for lung and prostate cancers, and with 33 bins ranging from 0 to 500 for ovarian cancer. The number of patients with interventions within each bin was divided by the total number of patients in each bin. The CDF-based IPC was fit to this data also using nonlinear least squares regression. In the third approach, the binning process was repeated 100 times using bootstrap sampling (sampling with replacement) and used to fit the CDF-based IPC. In the fourth approach, to smooth the curves, on each of the 100 repetitions, 1%

Gaussian noise was added to every patient's probability, effectively dithering the signal. To achieve this, a random number was drawn from a Gaussian distribution with mean 0 and standard deviation (SD) 1, multiplied by 0.01, then added to the probability of cancer. The average proportion in each of these 100 repetitions for each bin was used to fit the IPC function. Histogram binning, repeated sampling, and dithering were performed in MATLAB R2020b (MathWorks, Natick, MA), and the IPC was fit using GraphPad Prism (GraphPad Software, San Diego, CA). All R^2 presented are approximate, as calculated according to Kvalseth's method (17). Research datasets contained deidentified data and this project was deemed nonhuman research by the institutional review board (IRB) and the need for consent was waived.

Data use, access, and availability

All data was obtained from the NCI and is publicly available. NLST data use was approved by ECOG-ACRIN (NCI Protocol number A6654T4). PLCO data use was approved by the Cancer Data Access System (project PLCO-778). All studies were conducted in accordance with the Declaration of Helsinki. Studies at VUMC required no identified information or biospecimens, and therefore was exempt from IRB approval. All analysis can be performed using standard statistical analysis software.

Results

Derivation of the IPC

When the decision to intervene is guided by the probability of disease, the proportion of patients that undergo the intervention at a given likelihood of disease should follow a cumulative distribution function, where low probability patients undergo interventions at a lower rate than patients at higher probability. **Figure 2** shows several functions we evaluated as potential IPCs. In **Fig. 2A**, increasing probability of disease correlates with an increased probability that an intervention will occur. **Figure 2B**, conversely, reflects a threshold-based decision-making process in which transitions from low-to-intermediate and intermediate-to-high-probability result in a step function, an 'all-or-nothing' type of process. This is the type of IPC one would expect on the basis of professional society guidelines which strongly recommend intervention above a certain probability threshold (a *rule-in* threshold, T_I) and discourage use of the intervention below a certain probability threshold (a *rule-out* threshold, T_O). In the intermediate probability group (between these thresholds) the function between T_O and T_I may be flat if an intervention is recommended independently of the probability of disease (**Fig. 2B**, solid line). However, if within that intermediate probability subset the rate of intervention increases with probability, the IPC between T_O and T_I may increase monotonically (**Fig. 2C**, dotted line).

A model that includes uncertainty around these decision thresholds is likely closer to clinical practice. This can be modeled as a sigmoidal distribution: at low probability (below T_O) the intervention is unlikely to occur, and at high probability (above T_I) the intervention is likely, and the likelihood of intervention increases with probability between these thresholds. In this scenario, the curve exhibits a smooth transition over the thresholds (**Fig. 2D**). Here a useful mathematical approximation is the cumulative normal distribution function (CDF, **Eq. A**), which is the integral of the normal distribution:

$$IP_x = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{\frac{-(x-\mu)^2}{2\sigma^2}} dx \qquad \text{(A)}$$

where σ is the SD and μ is the mean of the normal distribution. To obtain the appropriate values in relation to the decision thresholds, we can use $\sigma = (T_I - T_O)/2$ and $\mu = (T_I + T_O)/2$.

It is also important to account for situations in which an intervention may proceed despite low probability of disease, such as in the case of a young, healthy person with a suspicious lesion. For such a patient, the probability of adverse complications from an invasive procedure

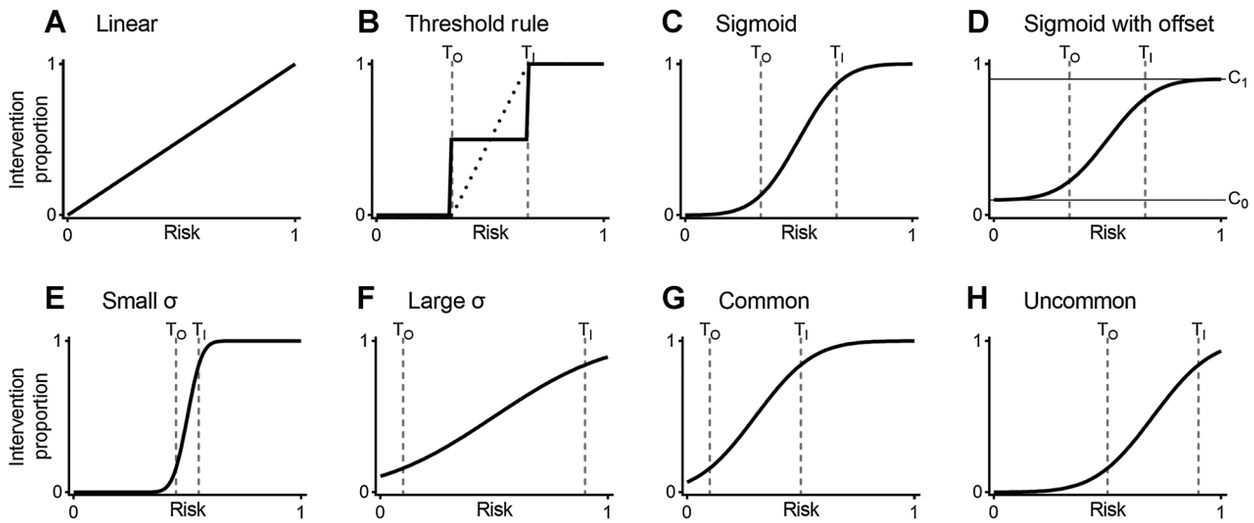

Figure 2.

General examples of IPCs. The likelihood that a patient will undergo an intervention generally scales with increasing probability of disease (Risk). This likelihood can increase linearly with probability (**A**), or the intervention will definitely occur above a *rule-in* threshold and definitely not occur below a *rule-out* threshold (**B**), where patients between the thresholds undergo the intervention at a constant rate (solid line) or at an increasing rate with probability (dotted line). Threshold-based decision-making, which accounts for other factors besides likelihood of disease can be modeled as a sigmoid curve (**C**). When a portion of the population undergoes the intervention or does not undergo it regardless of probability, these offsets can be accounted for (**D**). On the basis of this curve, four general situations can be described: (**E**) a small sigma, where the thresholds are close together; (**F**) a large sigma, where the thresholds are far apart; (**G**) a common intervention, where all but the lowest probability patients undergo the intervention; and (**H**) an uncommon intervention, where only the highest probability patients undergo the intervention.

may be minimal, but the benefit from a curative procedure could extend life by decades. Conversely, in some cases the intervention may not be warranted even in very high-probability cases as in the case of an elderly patient with multiple comorbidities. For such a patient, an invasive procedure would carry greater risks and the risk–benefit ratio would be unfavorable. To account for these "offsets" along the IPC, the constant C_0 represents the proportion of patients that would undergo intervention regardless of probability, and C_1 represents the proportion that would never undergo the intervention regardless of probability. The IPC can be modified to include these offsets as in Equation B:

$$IP_x = \frac{(1 - C_0 - C_1)}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{\frac{-(x-\mu)^2}{2\sigma^2}} dx + C_0 \quad \text{(B)}$$

The IPC parameters (σ, μ, C_0, and C_1) can be estimated from guidelines or fit to historical data of intervention decisions using nonlinear regression against the distribution of patients who underwent an intervention.

Several observations should be noted in this analysis. The first concerns the size of σ. When σ is small, the thresholds are close together, and the IPC looks more like a binary decision cutoff (**Fig. 2E**). When it is large, the thresholds are far apart, and the IPC looks more like a linear function (**Fig. 2F**). The second set of patterns is defined by μ. When μ is small, the IPC corresponds to a situation in which all but the lowest probability patients get the intervention (a "common intervention scenario," **Fig. 2G**). A large μ corresponds to a situation in which only the highest probability patients get the intervention (an "uncommon intervention scenario," **Fig. 2H**).

Fitting the IPC using indeterminate pulmonary nodules

The NLST dataset contains follow-up data for 26,722 patients assigned to a low-dose CT screening arm, of which ∼6,000 had screen-detected pulmonary nodules (14). Patients with a positive screen, defined as a pulmonary nodule ≥4 mm in largest diameter, were followed by standard of care, typically consisting of either follow-up CT or an invasive biopsy. From this dataset, we calculated the probability of malignancy as defined by the Mayo Clinic model and extracted the decision to perform an invasive tissue biopsy (2). In total, 5,792 patients in the CT arm of the NLST had a pulmonary nodule with the requisite information to calculate the Mayo Clinic model probability. Of these, 757 had a diagnostic procedure related to the pulmonary nodule, which involved a biopsy. The Mayo Clinic model was derived in a population of 629 patients, with an average age of 59.8, where 71% of the population was a current or former smoker. Of these 629, 146 had cancer, a prevalence of 23%. In comparison, the NLST cohort used here to evaluate the IPC had an average age of 61, a lower prevalence of cancer of 17%, and 100% current or former smoker rate, as smoking history was an inclusion criterion within the trial.

Using this dataset, a histogram of all patients (**Fig. 3A**) and a histogram of patients undergoing the intervention (**Fig. 3B**) were constructed. It is noteworthy that most patients were low probability, and very few patients had a probability of cancer above 30% probability of cancer. Then, the proportion of patients in each group was calculated by dividing the number of intervention patients by the total number of patients (**Fig. 3C**). The proportion of patients undergoing intervention increases as probability of cancer increases. To fit an IPC to this raw distribution, two methods were used. First, a logistic

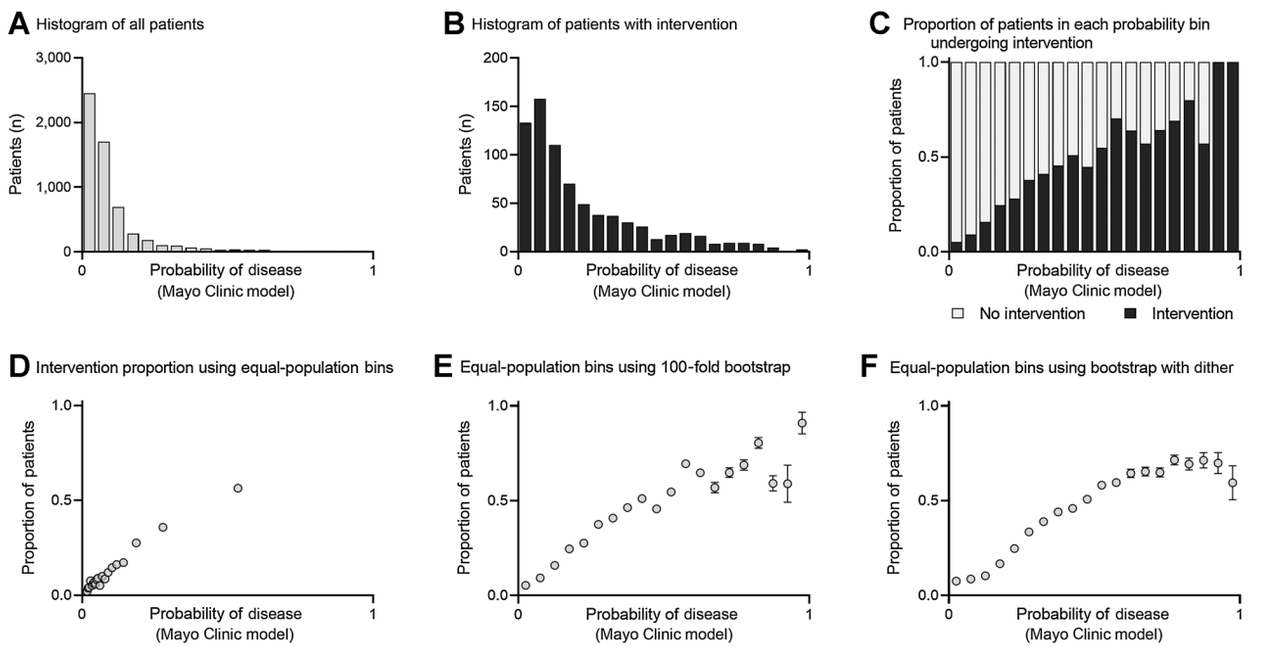

Figure 3.

Fitting the IPC to clinical data. **A,** A histogram of all patients with pulmonary nodules (6–30 mm in largest diameter) in the NLST dataset. **B,** A histogram of patients from **A** who underwent a diagnostic intervention. **C,** The proportion of patients in each bin who underwent a diagnostic intervention. The proportion of patients undergoing diagnostic intervention can be further analyzed by fitting a predetermined function (the IPF) derived from an understanding of the underlying probability models (**D**). The IPF can also be fit to binned proportions of patients, using groups defined by equal numbers of patients in each bin (**E**) or equal-width bins using bootstrap sampling with a dither effect applied on each bootstrap fold (**F**).

regression was fit to the data, using a binary output of no intervention versus intervention. This resulted in a statistically significant correlation (likelihood ratio test $P < 0.0001$), although this approach did not enable analysis of the correlation with thresholds or constant offsets (C_0 and C_1). Then, the data was used to fit **Equation B** by nonlinear least squares regression (**Fig. 3D**), resulting in a fitted rule out threshold of 0.077 [95% confidence interval (CI), 0.019–0.135], a rule in threshold of 0.400 (95% CI, 0.345–0.454), a C_0 of 0 (unable to fit a 95% CI), and a C_1 of 0.622 (95% CI, 0.565–0.679). This approach produced a true $R^2 = 0.1432$, although the low R^2 is expected due to the binary outcome data fit to a continuous independent variable (**Fig. 3E**).

A different curve fitting strategy grouped patients by probability of cancer using several approaches. The first approach tested was a grouping of patients into bins of equal population, then calculating the proportion of patients undergoing intervention in each group (**Fig. 3E**). When fit to **Equation B** using nonlinear least-squares regression, the equal-population bin strategy yielded an R^2 of 0.983. The fitted coefficients had very wide 95% CI, because the highest risk data point had an x-axis value of 0.54, so the impact of intervention decisions for high-risk patients was not considered in the curve fit. In the second approach, an equal-width bin approach was used, where patients were binned in increments of 0.05 (**Fig. 3F**). Patients were sampled using bootstrap sampling, repeated 100 times. On each sampling fold, the probability of malignancy was dithered by adding random noise. When fit to **Equation B** using least squares regression, the bootstrap sampling with dither approach yielded an $R^2 = 0.983$ (**Fig. 4A**). This approach allowed robust calculation of the parameters and CIs: a rule out threshold of 0.0926 (95% CI, 0.055–0.171), a rule in threshold of 0.517 (95% CI, 0.471–0.571), a C_0 of 0 (95% CI, 0–0), and a C_1 of 0.680 (95% CI, 0.652–0.711). These thresholds match approximately with the American College of Chest Physicians (ACCP) guidelines of 0.65 for rule in and 0.05 for rule out (18). The fitted value of C_1 of 0.68 implies that even at highest probability of malignancy, 32% of patients did not undergo biopsy.

Prostate nodules

The prostate screening abnormalities dataset from the PLCO study was used to calculate the PBCG model (8). Of the total patients in the screening arm, 30,456 patients contained variables to calculate the PBCG probability score. A transrectal biopsy was used as the invasive procedure. Of the 30,456 participants, 4,872 had a transrectal biopsy related to an abnormal screening visit. When grouped using equal-width bins and bootstrap sampling with dither, the fitted thresholds were $T_0 = 0.495$ (95% CI, 0.475–0.532) for rule out, $T_1 = 0.916$ (95% CI, 0.883–0.936) for rule out, C_0 of 0 (95% CI, 0–0.020), C_1 of 1 (95% CI, 0.963–1), with an $R^2 = 0.994$ (**Fig. 4B**). The population used to derive the PBCG model included 5,992 patients, with an average age of 64.7, median PSA level of 6 ng/mL, a 57% rate of normal DRE, and a 32% prevalence of cancer. In comparison, the population used to evaluate the IPC in this context (the PLCO) had a median age of 62.7 and a prevalence of cancer of 11%. The National Comprehensive Cancer Network (NCCN) advises consideration of initial biopsy in men with higher than 2.6 ng/mL PSA levels and that a 6–12 core biopsy panel of the nodule of interest be performed. Biopsy-cores are then used to calculate a Gleason Score (range of 2–10), which is used to evaluate prognosis and treatment (8). The fitted rule in $T_0 = 0.495$ and $T_1 = 0.916$ for rule out suggests that patients with 50% or higher probability score most commonly underwent a transrectal biopsy.

Adnexal masses and ovarian abnormalities

The PLCO ovarian dataset has follow-up data on over 150,000 screening visits, including blood CA-125 levels and an ultrasound report. The RMI was calculated according to the method described previously (6). Of the 39,104 women in the screening arm, 37,283 contained variables to calculate the RMI. Transvaginal biopsy linked to an abnormal TVU was used as the invasive procedure. Of the 37,283 participants, 3,008 had a biopsy related to an abnormal screening visit. When grouped using equal-width bins, bootstrap sampling, and dither, the fitted thresholds are $T_0 = 169$ (95% CI, 153–183) for rule out and $T_1 = 443$ (95% CI, 422–470) for rule in (**Fig. 4C**), a C_0 of 0.028 (95% CI, 0.009–0.043), and a C_1 of 0.62 (95% CI, 0.590–0.660), with an R^2 of 0.970. The population that was used to train the RMI included 143 patients, 42 of which had a malignancy (a prevalence of cancer of 29%). The population had an average age of 52, and 58% of the population was postmenopausal. When compared with the population used to evaluate the IPC (the PLCO dataset), there were 303 malignancies (a prevalence of cancer of 0.8%), an average age of 62, and nearly every participant (98.9%) was postmenopausal. The Royal College of Obstetrician and Gynecologists (RCOG) guidelines use the RMI to triage women as low (RMI < 25), moderate (25–250), or high (above >250; ref. 5). The fitted rule in threshold at $T_0 = 169$ lands within the RCOG moderate probability range. At highest probability, only 62% of patients underwent the invasive diagnostic procedure.

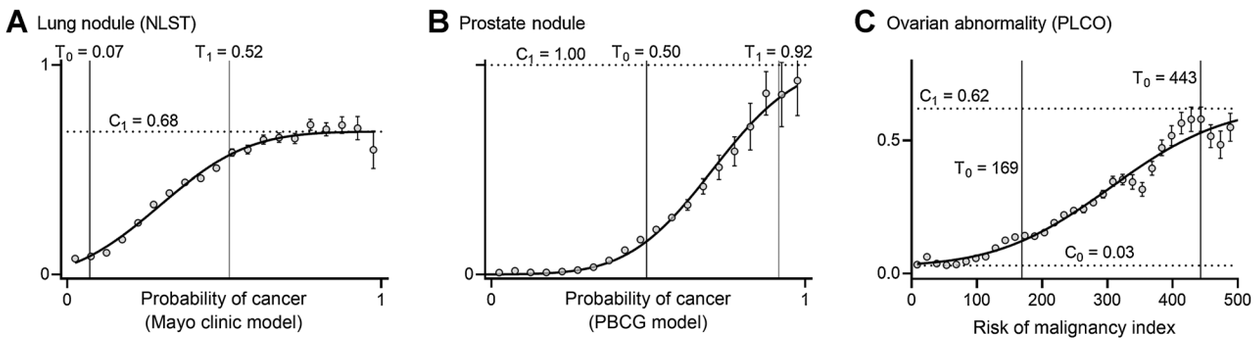

Figure 4.
The IPC fit to clinical data. **A,** Pulmonary nodules, data from the NLST. **B,** Prostate nodules, data from the PLCO study. **C,** Pelvic masses and adnexal masses.

Discussion

The decision to intervene on a suspicious lesion via invasive diagnostic procedures incorporates the probability of disease, but also depends on many other factors. While quantitative predictive models incorporate easily quantifiable clinical variables and biomarkers, the other relevant decision factors are more difficult to quantify, as for example, the reluctance to undergo an invasive procedure or a patient's fear of dying of cancer (9, 11, 19). On the provider side, physicians may be more influenced by financial incentives or fear of litigation. Most clinical prediction models incorporate clinical variables and rarely include patient and physician preferences. However, a complete picture of the clinical landscape surrounding patient management, evaluation of healthcare expenditures and benefits, and the formulation of management guidelines should take all relevant factors into account. Thus, assessing these clinical decisions using the IPC is likely to yield significant insights into these factors, some of which may represent actionable areas for improvement. Analysis of the IPC would inform professional societies or government agencies on actual clinical practice, which would allow an assessment of the congruence of current guidelines and clinical practice. Substantial incongruence of threshold-based recommendations and actual decisions in practice could reflect either barriers to implementation of evidence-based recommendations still viewed as normative, in which case future research efforts should be focused on addressing these barriers, or alternatively may reveal a need to revise these recommendations to accord with practical realities.

We took several approaches to fitting the IPC to the clinical data. First, fitting to the raw data proved challenging, as the distribution of patients was highly skewed towards low risk in the NLST dataset (**Fig. 3A**) and in the PLCO dataset as well. Therefore, any fitting to the raw data provided poor resolution of the behavior of the curve at higher probabilities of malignancy. To improve upon this, several grouping strategies were employed, and ultimately the equal-width bins using bootstrap sampling with dither provided the most robust, reproducible approach with high R^2 values and defined CIs (**Fig. 3F**).

Analysis of the IPC would be helpful in raising important questions and guiding future research: for example, in the NLST, 32% of patients at highest probability of disease did not undergo the recommended invasive diagnostic procedure. This could be due to a variety of reasons, including: (i) the accuracy of longitudinal imaging; (ii) the risk of harm from invasive procedure; (iii) financial reasons; (iv) patient preference due to age or quality of life; (v) the disease is too advanced to be amenable to cure; or (vi) patients underwent nonsurgical treatments, such as stereotactic body radiotherapy without histologic confirmation. The clinical reality is likely a combination of several factors, but the cause could lead to specific areas of research focus: if (i) optimizing longitudinal follow-up strategies. If (ii), improving the safety and efficacy of the invasive procedures. If (iii), development of low-cost diagnostic measures. If (iv), it is the treatments that should be the focus. If (v), earlier detection and screening programs would provide maximum benefit. These explanations, of course, are not mutually exclusive of one another, but illustrate how IPCs may reveal implementation issues that could be addressed through targeted research efforts.

Perhaps as importantly, the IPC may help improve diagnostic predictive models in ways likely to be most impactful. For example, analysis of intervention decision data may reveal that physicians are likely to overtreat a specific condition, and therefore a highly sensitive biomarker to rule-out disease may be useful, while a highly specific biomarker may add little to clinical practice.

The choice of function to model the IPC based on clinical data should be further investigated. Here, the normal CDF is investigated because the normal distribution appears repeatedly in population-based studies of disease probabilities. For predictive models derived using logistic regression, the normal CDF is an appropriate choice because the logistic function can be derived assuming the underlying populations are normally distributed. Therefore, it represents a natural starting point for a mathematical description of clinical management in these populations. Both the Mayo Clinic model (lung) and PBCG (prostate) models are logistic regression models that result in probability scores from 0 to 1, where the probability score is the log-odds of disease transformed into a probability. The RMI (ovarian cancer), however, represents a numeric score that is not directly related to the log-odds of disease. We postulate that any continuous biomarker or probability model, regardless of range, can be fit with an IPC curve, but there may be situations where the normal CDF (proposed here) does not appropriately model the data. In addition, this study only considers monotonically increasing functions, but there may be clinical situations that are best modeled as non-monotonic functions, especially as the analysis becomes more granular, and this would require a function that was not derived from a CDF. For example, in this analysis of NLST data, all invasive procedures were considered as the intervention. However, in the intermediate probability group, diagnostic fluorodeoxyglucose (FDG) PET scans are occasionally performed prior to an invasive procedure (3). The IPC for diagnostic PET scan in this scenario may increase at low probability as probability increases, but then decrease as probability approaches 1. The reason for this would be that once the probability of cancer is high enough, clinicians may prefer to forgo the diagnostic PET and move straight to biopsy (note that PET scans are also performed for staging purposes, so that the IPC integrating staging PET scan may look quite different). In addition, there may be a "hook effect," where the IPC decreases at very high probability, due to assessment that the disease may have progressed far enough that the patient opts for palliative care, or because the lesion is so obviously malignant that the patient elects to move directly to surgical resection or stereotactic body radiotherapy.

An important application of IPC would be the analysis of the actual, "real-life" clinical utility of novel diagnostic biomarkers. The potential for clinical utility of a novel biomarker is sometimes estimated by using metrics of reclassification, which estimate the proportion of patients with intermediate probability of disease appropriately reclassified as low or high probability (20–23). For example, if the pretest probability of an indeterminate, but actually malignant nodule is "intermediate", and a biomarker is applied which updates that probability to "high", the reclassification is appropriate. A major limitation of this method, however, is that movement *within* these groups may be as, if not more clinically relevant than movement *across* thresholds. An example of this is demonstrated in **Fig. 5A**, which plots the Mayo Clinic Model probability for a patient with an indeterminate pulmonary nodule included in a recent study that evaluated a novel combined biomarker (blood- and radiomics-based; ref. 24) overlaid upon the IPC fit to the NLST dataset. This nodule was in reality benign, but with a calculated pretest probability of 55%. After the biomarker was applied, the patient's probability of cancer was updated to 17%. Despite being drastically different, these probabilities are both considered intermediate as defined by the ACCP. Thus, while intuitively this movement within the intermediate group would be expected to appropriately

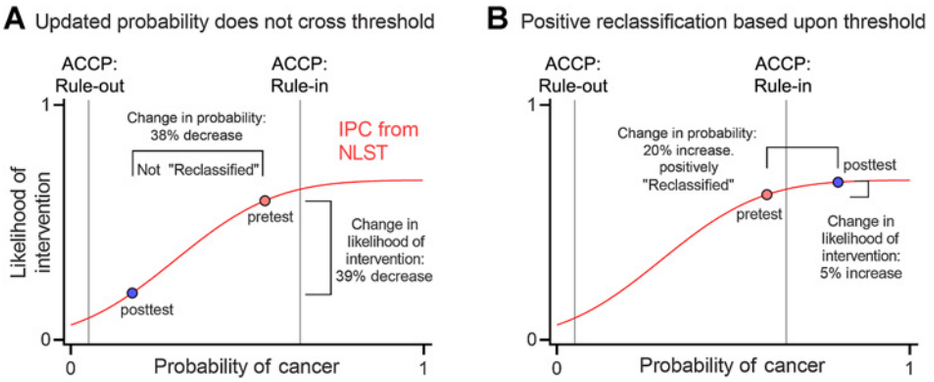

Figure 5.

Application of the IPC to the evaluation of the potential clinical utility of novel biomarkers within the context of pulmonary nodules. **A,** A patient with a Mayo model probability (red circle) of 55% was given the biomarker test, resulting in an updated probability of 17% (blue circle). On the basis of ACCP guidelines for the management of IPNs, the pretest probability was in the intermediate group, and the biomarker did not change this classification. However, on the basis of analysis of the IPC as fit to the NLST dataset (red curve), the likelihood of intervention would decrease from 59% to 20%, a reduction of 39%. **B,** A patient with pretest probability of 60% (Mayo, red circle) updated to 80% (biomarker, blue circle), where the change in probability crosses the "Rule-in" threshold, and therefore this patient is counted as a "positive reclassification" despite the IPC demonstrating that a likely estimate of change in the patient's management is a 5% increase in likelihood of intervention.

influence management decisions, reclassification metrics would not capture this benefit to the patient. Analysis of the IPC, however, reveals that application of this biomarker in a group of patients with pretest probability of 55% may decrease the likelihood of intervention from 59% to 20%—an absolute decrease of 39%—which may be clinically significant. Conversely, as illustrated in **Fig. 5B**, an appropriate reclassification by the biomarker from intermediate (60%) to high probability (80%) may be clinically irrelevant, as the likelihood of intervention would only increase by 5%. As such, the IPC could provide a more clinically relevant method of estimating the potential clinical utility of novel biomarkers.

One limitation of this analysis is that the use of these selected probability models was not mandated for the management decisions in these cohorts. The Mayo Clinic model and RMI were published and used clinically when the NLST and PLCO studies were ongoing, respectively, so physicians may have used them in their analysis, but it is not guaranteed. However, the PBCG model was not published until after the PLCO study was concluded. Despite this, in all three cases the models incorporate factors that are established predictors of malignancy, so that while the exact models may not have informed management of these patients, we argue that they serve as a reasonable foundation for our assessment here. In addition, there are several diagnostic probability models for each of these clinical cases, and our choice of the Mayo Clinic model for lung, RMI for ovarian, and PBCG for prostate cancer was based upon which models had been published and validated and for which available datasets including all necessary variables were available.

Another limitation of this study is the difference in population for the derivation of these models compared with the population in which the intervention data was available. This is a natural result of the design of such studies – biomarker development studies need definitive outcome data that has been adjudicated, typically by the invasive procedure, whereas the patients enrolled in screening studies do not undergo intervention unless the probability of malignancy is sufficiently high. Therefore, derivation cohorts typically have higher prevalence of cancer, and higher levels of known risk factors for cancer (such as age or smoking history).

Finally, another limitation is that all centers involved in the NLST and PLCO studies were tertiary care centers, and patient management may not reflect common practice in the community. Further analysis may reveal that the IPC differs between community clinics and tertiary care clinics, so future studies should focus on geographic variations (both national and international).

In conclusion, we have introduced the novel concept of IPC in decision-making for clinical management of patients with lesions suspicious for malignancy. We have demonstrated that the decisions to perform invasive procedures follow probability curves, that the cumulative distribution functions can be fit using clinical data resulting in high correlation, and that the variables in these curves correspond to professional society guidelines in three cancer types. Future studies of IPCs will help understand how biomarkers may be most impactful, identify barriers to optimal patient management, and clarify patient and physician preferences in comparison with professional society recommendations.

Authors' Disclosures

M.N. Kammer reports grants from NIH and personal fees from Meru Biotechnologies and Biodesix outside the submitted work; in addition, M.N. Kammer has a patent for Robust interferometer and methods of using same issued and licensed to Meru Biotechnologies and a patent for Free-solution response function interferometry issued and licensed to Meru Biotechnologies. S.A. Deppen reports grants from NCI during the conduct of the study as well as grants from NCI outside the submitted work. E.L. Grogan reports grants from NIH during the conduct of the study. A.E. Barón reports grants from NCI during the conduct of the study. No disclosures were reported by the other authors.

Authors' Contributions

M.N. Kammer: Conceptualization, data curation, formal analysis, investigation, visualization, methodology, writing–original draft. **D.J. Rowe:** Data curation, investigation. **S.A. Deppen:** Validation, methodology, writing–review and editing. **E.L. Grogan:** Validation, writing–review and editing. **A.M. Kaizer:** Investigation, methodology, writing–review and editing. **A.E. Barón:** Investigation, methodology, writing–review and editing. **F. Maldonado:** Conceptualization, supervision, methodology, writing–original draft.

Acknowledgments

This work was supported by the NCI at the NIH (U01CA186145 to P.P. Massion, U01CA152662 to E.L. Grogan) and the Martineau Family Foundation (to M.N. Kammer). The funders had no access to the study data or results prior to manuscript submission.

To Pierre Massion, our inspiration, mentor, teacher, counselor, colleague, friend, and guiding light, may we honor his legacy in our continued work to improve outcomes in lung cancer through our study of lung cancer development, biomarkers, and early detection. Every step forward will be taken down a path you have paved. May we always be inspired by patients, driven by science, and empowered by each other.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received February 20, 2022; revised May 11, 2022; accepted June 16, 2022; published first June 22, 2022.

References

1. Deppen SA, Grogan EL. Using clinical risk models for lung nodule classification. Semin Thorac Cardiovasc Surg 2015;27:30–5.
2. Swensen SJ, Silverstein MD, Ilstrup DM, Schleck CD, Edell ES. The probability of malignancy in solitary pulmonary nodules: application to small radiologically indeterminate nodules. Arch Intern Med 1997;157:849–55.
3. Herder GJ, van Tinteren H, Golding RP, Kostense PJ, Comans EF, Smit EF, et al. Clinical prediction model to characterize pulmonary nodules: validation and added value of 18F-fluorodeoxyglucose positron emission tomography. Chest 2005;128:2490–6.
4. Maldonado F, Varghese C, Rajagopalan S, Duan F, Balar AB, Lakhani DA, et al. Validation of the BRODERS classifier (Benign versus aggRessive nODule Evaluation using Radiomic Stratification), a novel HRCT-based radiomic classifier for indeterminate pulmonary nodules. Eur Respir J 2021;57:2002485.
5. van den Akker PA, Zusterzeel PL, Aalders AL, Snijders MP, Samlal RA, Vollebergh JH, et al. External validation of the adapted Risk of Malignancy Index incorporating tumor size in the preoperative evaluation of adnexal masses. Eur J Obstet Gynecol Reprod Biol 2011;159:422–5.
6. Jacobs I, Oram D, Fairbanks J, Turner J, Frost C, Grudzinskas JG. A Risk of Malignancy Index incorporating CA 125, ultrasound, and menopausal status for the accurate preoperative diagnosis of ovarian cancer. Br J Obstet Gynaecol 1990; 97:922–9.
7. Smith-Bindman R, Lebda P, Feldstein VA, Sellami D, Goldstein RB, Brasic N, et al. Risk of thyroid cancer based on thyroid ultrasound imaging characteristics: results of a population-based study. JAMA Intern Med 2013;173:1788–96.
8. Ankerst DP, Straubinger J, Selig K, Guerrios L, De Hoedt A, Hernandez J, et al. A contemporary prostate biopsy risk calculator based on multiple heterogeneous cohorts. Eur Urol 2018;74:197–203.
9. Ost DE, Gould MK. Decision-making in patients with pulmonary nodules. Am J Respir Crit Care Med 2012;185:363–72.
10. Wood DE, Kazerooni EA, Baum SL, Eapen GA, Ettinger DS, Hou L, et al. Lung Cancer Screening, Version 3.2018, NCCN Clinical Practice Guidelines in Oncology. J Natl Compr Canc Netw 2018;16:412–41.
11. Wiener RS, Slatore CG, Gillespie C, Clark JA. Pulmonologists' reported use of guidelines and shared decision-making in evaluation of pulmonary nodules: a qualitative study. Chest 2015;148:1415–21.
12. Maiga AW, Deppen SA, Massion PP, Callaway-Lane C, Pinkerman R, Dittus RS, et al. Communication about the probability of cancer in indeterminate pulmonary nodules. JAMA Surg 2018;153:353–7.
13. Tanner NT, Porter A, Gould MK, Li XJ, Vachani A, Silvestri GA. Physician assessment of pretest probability of malignancy and adherence with guidelines for pulmonary nodule evaluation. Chest 2017;152: 263–70.
14. Aberle DR, Adams AM, Berg CD, Black WC, Clapp JD, Fagerstrom RM, et al. Reduced lung-cancer mortality with low-dose computed tomographic screening. New Engl J Med 2011;365:395–409.
15. Gohagan JK, Prorok PC, Hayes RB, Kramer B-S. The prostate, lung, colorectal and ovarian (PLCO) cancer screening trial of the National Cancer Institute: History, organization, and status. Control Clin Trials 2000;21: 251S–72S.
16. R Core Team. R: A Language and Environment for Statistical Computing. Version 4.0.3 [software]. 2020 Available from: R Foundation for Statistical Computing <https://www.R-project.org/>.
17. Kvalseth TO. Cautionary note about R2. Am Stat 1985;39:279–85.
18. Gould MK, Donington J, Lynch WR, Mazzone PJ, Midthun DE, Naidich DP, et al. Evaluation of individuals with pulmonary nodules: when is it lung cancer? Diagnosis and management of lung cancer, 3rd ed: American College of Chest Physicians evidence-based clinical practice guidelines. Chest 2013; 143:e93S–e120S.
19. Phadke S, Vander Weg M, Itani N, Grogan N, Ginader T, Mott S, et al. Breast cancer patient preferences for test result communication. Breast J 2019;25: 1326–7.
20. Paynter N, Cook N. A bias-corrected net reclassification improvement for clinical subgroups. Med Decis Making 2013;33:154–62.
21. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology 2010;21:128–38.
22. Cook NR, Ridker PM. Advances in measuring the effect of individual predictors of cardiovascular risk: the role of reclassification measures. Ann Intern Med 2009;150:795–802.
23. Pencina MJ, D'Agostino RB Sr, D'Agostino RB Jr, 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: 157–72.
24. Kammer MN, Lakhani DA, Balar AB, Antic SL, Kussrow AK, Webster RL, et al. Integrated biomarkers for the management of indeterminate pulmonary nodules. Am J Respir Crit Care Med 2021;204:1306–16.