Identification of distributions for risks based on the first moment and c-statistic: with application in decision modeling, sample size calculations, and value-of-information analysis

## Abstract

Decision analytic modeling, sample size calculation, and value-of-information analysis for risk prediction models often require specification of the distribution of predicted risks in the target population. However, this is seldom reported in conventional model development and validation studies. On the other hand, indirect information on this distribution is often available in terms of expected value and c-statistic. Methods that can reconstruct the distribution of risks from such information can have broad applicability. Here, we show that for a wide class of probability distributions for risks, knowing the expected value and c-statistic is enough to identify the distribution. The derivation motivates numerical algorithms for mapping a given pair of expected value and c-statistic to the parameters of specified two-parameter distributions for probabilities. In a simulation study, we evaluated the numerical accuracy of such algorithms for common families of distributions for risks (Beta, Logit-normal, and Probit-normal). In a case study we show how this approach can be used to enable Monte Carlo sampling from the posterior distribution of parameters that determine the distribution of predicted risks. The accompanying **mcmapper** R package encapsulates these developments.

## Highlights

* Most risk modeling studies report on outcome prevalence and discriminatory performance of the model (area under the receiver operating characteristic / c-statistic).
* Decision analytic modeling, sample size calculation, and value-of-information analysis for risk prediction models often require specification of the distribution of predicted risks.
* We show that for a wide class of probability distributions for risks, knowing the outcome prevalence and the model’s c-statistic is enough to identify the distribution of risks.
* The accompanying open-access software provides numerical algorithms for such mapping for common distributions for risk.

## Background

In many instances, one might desire to know the distribution of individualized risks in a population of interest. Consider, for example, a decision analysis aimed at finding the optimal cut-off on predicted risks of lung cancer, to identify high-risk individuals for referral to computed tomography. A key parameter for such an analysis is the proportion of individuals classified as high-risk at a given cut-off value, a quantity that depends on the distribution of predicted risks. In addition to decision-analytic modeling, sample size calculations and recently proposed value-of-information (VoI) analysis methods for external validation of risk prediction models require specifying a distribution for predicted risks (Riley et al. 2021; Pavlou et al. 2021; Sadatsafavi et al. 2023).

In decision-analytic modeling, identifying a distribution based on its moments (e.g., mean and variance) or confidence intervals is a common practice. However, in risk modeling studies, the variance of, or confidence bounds around, predicted risks is seldom reported. Riley et. al. note that sometimes the report provides some clue, for example, through a histogram of predicted risks Riley et al. (2024); but parameterizing the distribution of risks based on quantitative performance metrics can be a more objective approach. Fortunately, the majority of risk prediction development or validation studies report the c-statistic of the model summarizing its discriminatory power(Collins et al. 2024). Generally speaking, the c-statistic provides indirect information about the spread of predicted risks (models with low c-statistic have concentrated mass, while those with high c-statistic are generally more variable). Given that common distributions for modeling probabilities (e.g., Beta, or Logit-normal) are indexed by two parameters, it is intuitive that within a given family of such distributions, knowing the expected value and c-statistic should uniquely identify the distribution. But to the best of our knowledge, no rigorous proof has hitherto been offered.

This work builds on such intuition, and proves the uniqueness of such identification. We provide open-source software that implements numerical mapping algorithms based on such identification, examine its face validity through brief simulation studies, and its utility via a case study.

## Notation and definitions

Let be the strictly monotonic cumulative distribution function (CDF) from the family of distributions of interest with support on . Let be a random draw from this distribution and the corresponding response variable. Let be the first moment of , and its c-statistic. is the probability that a random draw from s among ‘cases’ (those with ) is larger than a random draw from s among ‘controls’ (those with ). Formally, with and being two pairs of predicted risks and observed responses, with .

We define as *quantile-identifiable* if knowing any pair of quantiles fully identifies the distribution. Common two-parameter distributions for probabilities, such as Beta (), Logit-normal (, where ) and Probit-normal ( where is the standard Normal CDF) satisfy the strict monotonicity and quantile-identifiability. The quantile-identifiability of the Beta distribution was established by Shih Shih (2015). For the Logit-normal and Probit-normal distributions, it is immediately deduced from the monotonic link to the Normal distribution and the quantile-identifiability of the latter.

## Lemma

Consider , a family of probability distributions with support on [0,1], with the following characteristics:

* the CDF is strictly monotonic and has no jumps;
* the distribution is quantile-identifiable.

Then is uniquely identifiable from {, }.

## Proof

Details are provided in the Appendix. In brief, we establish the following system of equations for :

where is the set (typically a pair) of parameters indexing . The proof involves showing that this system has at most one solution.

## Implementation and simulation studies

We have implemented a set of numerical algorithms for finding the parameters of specified families of distributions of the above-mentioned class given a known mean and c-statistic in the accompanying [**mcmapper**](https://cran.r-project.org/package=mcmapper) R package Sadatsafavi (n.d.). Details of our implementation are provided in Supplementary Material - Section 1. Supplementary Material - Section 2 provides the results of brief simulation studies that document the satisfactory performance of these algorithms across a wide range of expected value and c-statistic ( and ).

## Case study

Consider a probabilistic micro-simulation model for the evaluation of a triaging algorithm that identifies individuals at high risk of death after an acute myocardial infarction (AMI) upon their arrival at an emergency department. Individuals are assigned a mortality risk from the distribution of this risk in the population. Parameters that index this distribution are inferred from finite-sample studies. Quantifying uncertainty around their values (i.e., second-order uncertainty) is required for a probabilistic decision analysis. The report on the development of the underlying risk prediction model often contains an estimate of the expected value of risk and the c-statistic of the model for predicting . Assuming that the model is moderately calibrated in the development population (i.e., ), the identifiability result in this work can directly be applied to recover the parameters of the distribution.

To illustrate this, we create a risk prediction model for 30-day mortality after an AMI based on data from 500 observations from the GUSTO-I clinical trial. This setup replicates a previous case study (where more details are provided) (Sadatsafavi et al. 2025). The model equation is

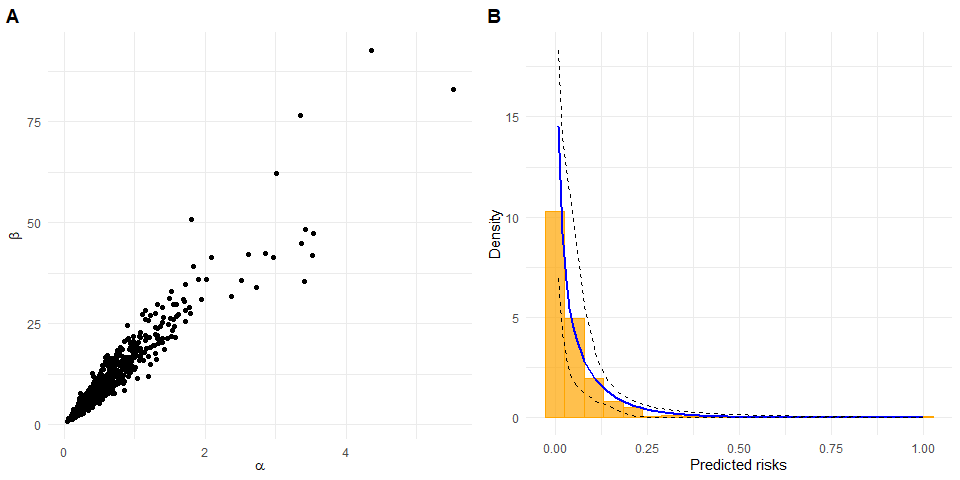
.

Out of 500, 28 experienced the outcome (). The c-statistic of the model in this sample was 0.842, with a 95% credible interval, quantified using the Bayesian bootstrap method(Gu, Ghosal, and Roy 2008) of 0.775-0.898.

If the original development sample was available, one could use sample-based methods, such as Bayesian bootstrapping, to quantify uncertainty around the distribution of predicted risks(Pasta, Taylor, and Henning 1999). Assuming we do not have access to the sample, we use the summary statistics above as the source of information. We parameterize the distribution of prevalence as , and that of the c-statistic as . The former is based on the standard Beta-Binomial Bayesian modeling, and latter is based on finding the two parameters of the Beta distribution whose mean and 97.5th percentile match, respectively, the point estimate and the upper bound of the credible interval reported above.

We use the Beta distribution to characterize predicted risks, which is indexed by two parameters and . Our goal is to quantify our uncertainty around the joint distribution of and . This can be achieved by repeatedly sampling from the assigned distributions to prevalence () and c-statistic (), applying our mapping algorithm, and recording the resulting and values. We note that this sampling algorithm implements independent draws from and , despite the fact that they are estimated from the same data. However, as shown in Supplementary Material - Section 3, the posterior distributions of and are independent if such distributions are constructed via Bayesian bootstrapping in the original sample.

**Figure 1** - panel A shows the scatter plot of 10,000 pairs of sampled and values. Panel B shows the histogram of the individualized risks from the development sample along with the density of the Beta distribution based on point estimates of prevalence and c-statistic. In addition, the 2.5th and 97.5th percentiles of the density values at each point are overlaid on the histogram.



**A**: Scatter plot of (, ) pairs. **B**: Histogram of predicted risks (bars), the density of the Beta distribution based on point estimates of prevalence and c-statistic (solid line), and central 95% boundaries of Beta densities from the Monte Carlo sample (dashed lines - note that dashed lines are not valid densities in themselves; rather, at each point they represent the 95% boundary of the density values from the Monte Carlo simulation)

## Remarks

We showed that strictly continuous distributions for risks that are quantile-identifiable can be identified by knowing their expected value and c-statistic. The identifiability motivates numerical mapping algorithms that are implemented into the publicly available **mcmapper** R package. In a case study we showed how summary statistics from a model development study can be used to enable uncertainty quantification around the parameters of the distribution of risks. In pursuing these developments we were motivated by the need for probabilistic modeling of the distribution of predicted risks in decision analysis, as well as several recent developments around sample size calculations and VoI analysis for risk prediction models(Riley et al. 2021; Sadatsafavi et al. 2023).

While the mapping algorithms performed well in our brief simulation studies, they might be numerically unstable at extreme ranges, especially for the c-statistic. This is because the CDF can be very flat (within the floating point precision of the computing device) in extended parts of its range in such cases, making numerical integration inaccurate. However, extreme values of the c-statistic are not plausible in real-world situations. A c-statistic that is very close to 1, for example, might indicate a complete separation of cases and controls. Reciprocally, a c-statistic that is very close to 0.5 signals no variability in risks. In such situations, the appropriateness of modeling the risk as a continuous distribution might be doubtful.

In assigning a distribution type to the distribution of risks (e.g. Beta in our case study), the onus is on the investigator to decide on the appropriateness of the family of distributions. Sometimes there is auxiliary information in a report, such as the histogram of predicted risks, that the investigator can draw on, and perform sensitivity analyses as required.

Overall, this approach provides objectivity to specifying the distribution of predicted risks, with applicability in health technology assessments, sample size calculations, and VoI analyses. Given the ubiquity of reporting prevalence and c-statistic in model development and validation studies, we expect this approach to play an important role in the future implementation and uptake of this approach.

# References

Collins, Gary S, Karel G M Moons, Paula Dhiman, Richard D Riley, Andrew L Beam, Ben Van Calster, Marzyeh Ghassemi, et al. 2024. “TRIPOD+AI Statement: Updated Guidance for Reporting Clinical Prediction Models That Use Regression or Machine Learning Methods.” *BMJ*, April, e078378. <https://doi.org/10.1136/bmj-2023-078378>.

Gu, Jiezhun, Subhashis Ghosal, and Anindya Roy. 2008. “Bayesian Bootstrap Estimation of ROC Curve.” *Statistics in Medicine* 27 (26): 5407–20. <https://doi.org/10.1002/sim.3366>.

Pasta, David J., Jennifer L. Taylor, and James M. Henning. 1999. “Probabilistic Sensitivity Analysis Incorporating the Bootstrap: An Example Comparing Treatments for the Eradication of Helicobacter Pylori.” *Medical Decision Making* 19 (3): 353–63. <https://doi.org/10.1177/0272989x9901900314>.

Pavlou, Menelaos, Chen Qu, Rumana Z. Omar, Shaun R. Seaman, Ewout W. Steyerberg, Ian R. White, and Gareth Ambler. 2021. “Estimation of Required Sample Size for External Validation of Risk Models for Binary Outcomes.” *Stat Methods Med Res* 30 (10): 2187–2206. <https://doi.org/10.1177/09622802211007522>.

Riley, Richard D., Thomas P. A. Debray, Gary S. Collins, Lucinda Archer, Joie Ensor, Maarten van Smeden, and Kym I. E. Snell. 2021. “Minimum Sample Size for External Validation of a Clinical Prediction Model with a Binary Outcome.” *Stat Med* 40 (19): 4230–51. <https://doi.org/10.1002/sim.9025>.

Riley, Richard D., Kym I. E. Snell, Lucinda Archer, Joie Ensor, Thomas P. A. Debray, Ben van Calster, Maarten van Smeden, and Gary S. Collins. 2024. “Evaluation of Clinical Prediction Models (Part 3): Calculating the Sample Size Required for an External Validation Study.” *BMJ* 384 (January): e074821. <https://doi.org/10.1136/bmj-2023-074821>.

Sadatsafavi, Mohsen. n.d. “The Mcmapper R Package (Development Version).” Accessed November 8, 2024. <https://cran.r-project.org/package=mcmapper>.

Sadatsafavi, Mohsen, Tae Yoon Lee, Laure Wynants, Andrew J. Vickers, and Paul Gustafson. 2023. “Value-of-Information Analysis for External Validation of Risk Prediction Models.” *Medical Decision Making: An International Journal of the Society for Medical Decision Making* 43 (5): 564–75. <https://doi.org/10.1177/0272989X231178317>.

Sadatsafavi, Mohsen, Andrew J. Vickers, Tae Yoon Lee, Paul Gustafson, and Laure Wynants. 2025. “Expected Value of Sample Information Calculations for Risk Prediction Model Validation.” *Medical Decision Making* 45 (3): 232–44. <https://doi.org/10.1177/0272989x251314010>.

Shih, Nenghui. 2015. “The Model Identification of Beta Distribution Based on Quantiles.” *Journal of Statistical Computation and Simulation* 85 (10): 2022–32. <https://doi.org/10.1080/00949655.2014.914513>.

# Appendix: Proof of the main lemma

First, we use the standard result that for as a non-negative random variable, . Thus, knowing is equivalent to knowing the area under the CDF. Next, applying Bayes’ rule to the distribution of among cases () and controls () reveals that the former has a PDF of and the latter , where is the PDF of . Thus we have

where and . Integration by parts for the first term and a change of variable for the second term result in

i.e., among the subset of s with the same , is monotonically related to . As such, the goal is achieved by showing that {, } uniquely identifies . We prove this by showing that two different CDFs and from the same family and with the same cannot have the same .

Given that both CDFs are anchored at (0,0) and (1,1), are strictly monotonic, and have the same area under the CDF but are not equal at all points, they must cross. However, they can only cross once, given the quantile-identifiability requirement (if they cross two or more times, any pairs of quantiles defined by the crossing points would fail to identify them uniquely).

Let be the unique crossing point of the two CDFs, and let be the CDF value at this point. We break into two parts around :

Without loss of generality, assume we label the s such that when . In this region, due to the CDFs strictly increasing, . As such, replacing by the larger positive quantity will increase this term. As well, in the region, , and . As such, replacing by the smaller positive quantity will also increase this term. Therefore we have

and the term on the right-hand side is zero because of the equality of the area under the CDFs. Therefore, , establishing the desired result.