Pattern Recognition Letters
journal homepage: www.elsevier.com

Receiver Operating Characteristic Curves with an Indeterminacy Zone Giovanni Parmigiani

Giovanni Parmigiani[a,**]

[a]*Dana Farber Cancer Institute, 450 Brookline Avenue, Boston 02115, U.S.A.*
Harvard T.H. Chan School of Public Health, 677 Huntington Avenue, Boston 02115, U.S.A.

ABSTRACT

This work extends Receiver Operating Characteristic (ROC) curve to the situation where some cases, falling in an intermediate "indeterminacy zone" of the predictor, are not classified. It addresses two challenges: definition of sensitivity and specificity bounds for this case; and summarization of the large number of possibilities arising from different choices of indeterminacy zones.

1. Introduction

Receiver Operating Characteristic (ROC) curves help with the visual assessment of the performance of classifiers. Fawcett (2006) reviews the field and points out that "ROC graphs are commonly used in medical decision making, and in recent years have been used increasingly in machine learning and data mining research".

I consider here the basic case of binary classification using a continuous score, such as a classification probability, or a quantitative biomarker. Traditionally, classification is simply implemented by a cutoff dichotomizing the score. In more recent applications, classification may includes an intermediate area of indeterminacy, which I will call *gray zone*.

For a famous example, Parker et al. (2009) present the PAM50 risk predictor of breast cancers, which provides a continuous risk score. In clinical applications, this score is most often split into three categories: low, intermediate and high. Women in the low and high categories are directed to specific clinical strategies. Women in the intermediate category are considered on a case by case basis by their clinicians. From an algorithmic standpoint, the intermediate group is not classified. Similarly, machine learning algorithms for classification of pathology and radiology images may allow for certain areas to be routed to further human examination. In these cases indeterminacy helps with practical implementation, by handling safe cases algorithmically and complex ones by human intervention.

Here I describe an algorithm for visualizing bounds on sensitivity/specificity pairs, for short *grayROC*, to assess the performance range of classifiers allowing for a region of indeterminacy, or gray zone. I try to address two challenges. The first is the definition of sensitivity and specificity bound when there is indeterminacy. The second is the visual summarization of the large number of possibilities arising from different choices of gray zones.

2. Algorithm

Consider a validation study of n labeled subjects, with scores x_i, $i = 1, \ldots, n$. Without loss, let the first n_0 subjects ($0 < n_0 < n$) have label 0 and the remaining n_1 have label 1. Also, low levels of the score are taken to predict class 0. The proportion of 1's in the target population is π, and may differ from the validation study proportion n_1/n, for example if the design of the validation study is a case-control.

A gray zone is defined by the interval (c_L, c_U). The extremes are the lower and upper cutoff. Cases with score below c_L are classified as 0's. Cases above c_U are classified as 1's. The rest remain unclassified.

Users of the grayROC need to specify a maximum tolerated percentage of unclassified cases, γ, based on the tradeoffs present in the practical application at hand. Let g_j be the number of class j points falling in the gray zone. A gray zone (c_L, c_U) satisfies the γ-constraint if the proportion of cases in the gray zone is less than γ, that is if $(g_0 + g_1)/n < \gamma$. A

**Corresponding author:
e-mail: gp@jimmy.harvard.edu (Giovanni Parmigiani)

gray zone (c_L, c_U) satisfies the target population γ-constraint if $((1-\pi)g_0 + \pi g_1)/n < \gamma$.

The grayROC algorithm is a model-free visualization. The basic building blocks are bounds on the cumulative frequencies associated with a given gray zone (c_L, c_U).

First, the most favorable bound on these frequencies is calculated assuming perfect discrimination within the gray zone. Imagine an oracle would take care of the points in the gray zone on behalf of the classifier, by moving them to the extremes of the gray zone so that they can be classified correctly. Formally, define the starred scores as follows:

$$\begin{array}{llll} \text{if} & x_i \notin (c_L, c_U) & \text{then} & x_i^* = x_i \\ \text{if} & i \le n_0, x_i \in (c_L, c_U) & \text{then} & x_i^* = c_L \\ \text{if} & i > n_0, x_i \in (c_L, c_U) & \text{then} & x_i^* = c_U. \end{array}$$

Let I_A be the indicator of the set A, and define the cumulative frequencies:

$$F_0^*(c_L, c_U) = \sum_{i=1}^{n_0} I_{x_i^* < (c_L + c_U)/2} \tag{1}$$

$$F_1^*(c_L, c_U) = \sum_{i=n_0+1}^{n} I_{x_i^* < (c_L + c_U)/2}. \tag{2}$$

Conversely, the least favorable frequencies are constructed considering the worst case scenario for the points within the gray zone. Imagine now that a saboteur may be in charge of the points in the gray zone, by moving them to extremes of the gray zone, so that they are all classified incorrectly. This would result in the "daggered" scores, defined as:

$$\begin{array}{llll} \text{if} & x_i \notin (c_L, c_U) & \text{then} & x_i^\dagger = x_i \\ \text{if} & i \le n_0, x_i \in (c_L, c_U) & \text{then} & x_i^\dagger = c_U \\ \text{if} & i > n_0, x_i \in (c_L, c_U) & \text{then} & x_i^\dagger = c_L. \end{array}$$

Now define the cumulative frequencies:

$$F_0^\dagger(c_L, c_U) = \sum_{i=1}^{n_0} I_{x_i^\dagger < (c_L + c_U)/2} \tag{3}$$

$$F_1^\dagger(c_L, c_U) = \sum_{i=n_0+1}^{n} I_{x_i^\dagger < (c_L + c_U)/2}. \tag{4}$$

We can form a large number of starred and daggered pairs of cumulative frequencies satisfying the γ-constraint. The grayROC algorithm simplifies the visualization of these pairs by grouping them, and selecting a single higher and lower limit within each group, as follows.

Consider the r observed unique ranked values of the biomarker $x_{(1)}, \ldots x_{(r)}$. These points will constitute the set of possible values for the extremes (c_L, c_U) of the gray zone. Now define the midpoints between two consecutive values as $c_j = (x_{(r-1)} + x_{(r)})/2$ for $j = 2, \ldots, r$. For each c_j, consider the set of (c_L, c_U) pairs built by first adding the two neighboring observed points on either side, then the next two and so forth. This process continues as long as the gray zone satisfies the γ-constraint. If one of the extremes of the distribution is reached, the process continues on the other side. Among the resulting intervals, the grayROC chooses the "best" for visualization, defined as follows. For each (c_L, c_U), it eliminates the cases in the gray zone and then computes the area under the ROC curve (AUC, Bradley (1997)) using the classified cases only. The (c_L, c_U) pair maximizing the AUC so defined is $(c_L^*(c_j), c_U^*(c_j))$. The generating c_j is not necessarily the midpoint of this interval, but will be contained in it. If multiple gray zones are tied in this maximization, the algorithm minimizes gray zone width among optima. In this way, gray zones are not used in regions where discrimination is not helped by not classifying cases.

Then, the upper limits are defined by the set of points

$$\left(1 - F_1^*(c_L^*(c_j), c_U^*(c_j)), 1 - F_0^*(c_L^*(c_j), c_U^*(c_j))\right) \tag{5}$$

as c_j varies. Conversely, the lower limits are defined by the set of points

$$\left(1 - F_1^\dagger(c_L^*(c_j), c_U^*(c_j)), 1 - F_0^\dagger(c_L^*(c_j), c_U^*(c_j))\right). \tag{6}$$

for $j = 2, \ldots, r$. To implement, define the degenerate gray zones $(x_{(i)}, x_{(i)})$ and $(x_{(i)}, x_{(i+1)})$ as the empty set.

Fix y to be either 0 or 1. The sequences defined by $F_y^*(c_L^*(c_j), c_U^*(c_j))$ and $F_y^\dagger(c_L^*(c_j), c_U^*(c_j))$ as j varies in $2, \ldots, r$ do not necessarily define proper cumulative distributions, as they would in a standard ROC analysis. Rather the intent is to provide bounds to the sensitivity / specificity pairs available over a range of possible gray area strategies.

Starred and daggered curves are calculated using both classified and unclassified samples. The exclusion of the unclassified samples only affects the calculation of $(c_L^*(c_j), c_U^*(c_j))$.

In summary, the algorithm's steps to produce the data needed for plotting a grayROC graph are as follows:

Data: biomarker measurements and labels
Result: all cutoffs and cumulative frequencies pairs
compute set of candidate cutoff points;
compute midpoints of resulting partition;
for *each midpoint* **do**
  while *gray zone satisfies γ-constraint* **do**
    enlarge gray zone;
    evaluate AUC on classified cases only;
  end
  choose smallest gray zone limits with largest AUC ;
  compute starred & daggered cumulative frequencies ;
end

Algorithm 1: The grayROC procedure for computation of upper and lower limits in expressions (5) and (6).

I explored an alternative implementation where the lower and upper limit of the gray area are used in turn to index the AUC optimization, instead of the midpoints. Upper and lower limits can produce markedly different results. Bounds are less stable than the midpoints when sample sizes are small. Nonetheless, this strategy provides a different view of the overlap in the tails, and may turn out to be useful in some applications.

3. Illustration

To illustrate the application and interpretation of the grayROC, I consider a gene expression biomarker for the prediction

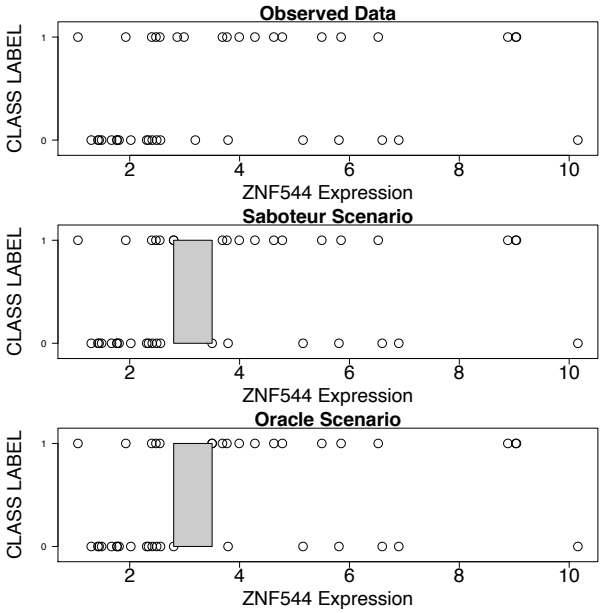

Fig. 1. Dotplots of biomarker levels by class as observed (top) and in the hypothetical scenarios used in the construction of the grayROC plot. The gray zone is $(2.8, 3.5)$**. In the "oracle" scenario, class 1 points in the gray zone are moved to the upper limit** 3.5 **while the class 0 points in the gray zone are moved to** 2.8**. The reverse is true in the "saboteur" scenario.**

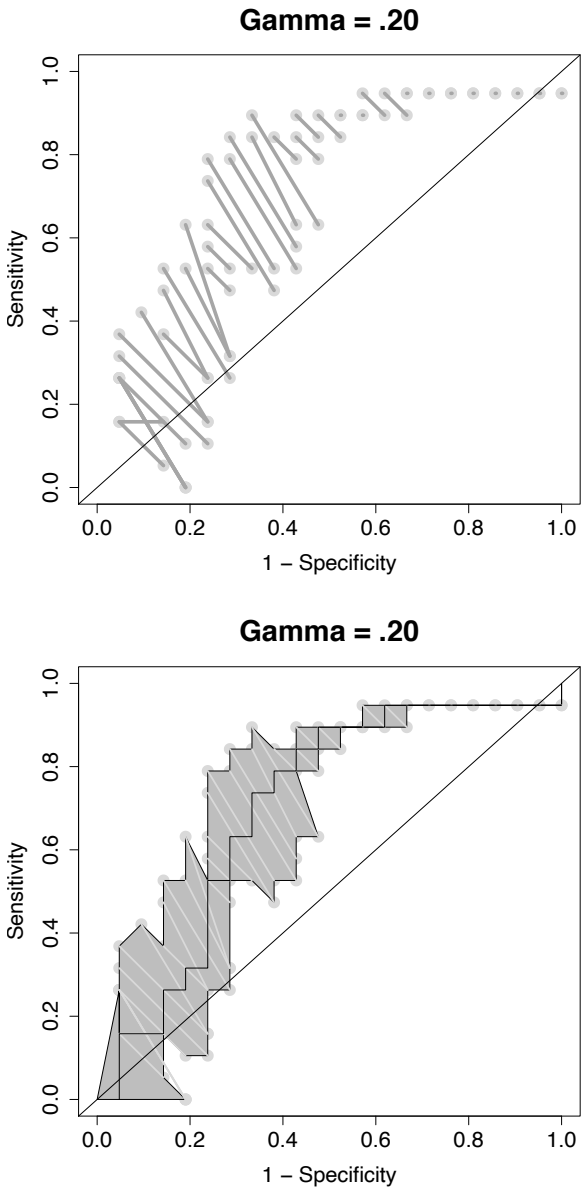

Fig. 2. grayROC displays at maximum tolerated percentage of unclassified cases, γ**, of** $.2$**. The top panel shows segments connecting starred and daggered points corresponding to the same** c_j**. The segments collapse to a point when the optimal gray area for the corresponding** c_j **is empty. The bottom panel shows, in addition, the area between the two curves defined by connecting the starred and daggered points. The thinner line corresponds to the standard ROC curve.**

of suboptimal (class 0) versus optimal (class 1) surgical debulking in ovarian cancer patients. Data are available from the `CuratedOvarianData` Bioconductor package by Ganzfried et al. (2013). Clinical and biological background can be found in Riester et al. (2014). The specific biomarker presented here reflects the transcriptional level of the gene ZNF544, as measured using an Agilent microarray by Yoshihara et al. (2012).

Figure 1 shows the observed biomarker levels by class. Higher level of expression are generally associated with optimal debulking (class 1). Figure 1 also illustrates the type of hypothetical scenarios that enter as building block in the construction of the grayROC, to visually represent the definitions of x^* and $x^\dagger$.

Each of hypothetical scenarios in Figure 1 enter the optimization used to find the $c_U^*(c_j)$'s. These in turn are used to form the starred and daggered sensitvity and specificity bounds. Figure 2 shows segments connecting starred and daggered points corresponding to the two bounds associated with the same c_j. These can be used to explore potential gray area strategies. Say one is interested in a classifier with approximately 80% specificity and 70% sensitivity. ZNF544 does not reach this performance. The upper points inform us that if one were allowed to pass 20% of suitably chosen observations to the oracle, than ZNF544 could reach close to the desired sensitivity/specificity trade-off. It also informs us that if the same observations were passed to the saboteur, the sensitivity and specificity would drop close to the diagonal line of no discrimination.

Figure 2 also shows, in the bottom panel, the region defined by the starred points as the upper limit, and by the daggered points as the lower limit. Points within the region are not easily interpretable in terms of the optimization of the previous section. The shading is purely a visual aid.

Figure 3 shows grayROC visualizations corresponding to four additional choices of γ.

Figure 2 also illustrates that the region defined by the upper and lower limits in the grayROC algorithm is not necessarily convex.

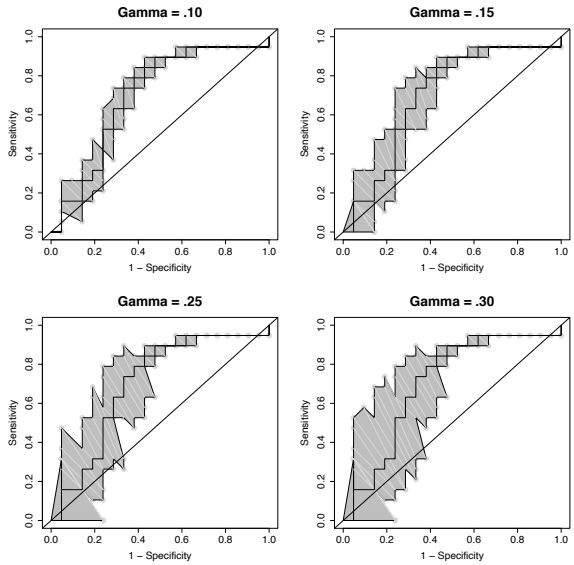

Fig. 3. grayROC displays for ZNF544 at maximum tolerated percentage of unclassified cases, γ, of .1 **(top left)** .15 **(top right)** .25 **(bottom left) and** .30 **(bottom right.) The thinner line corresponds to the standard ROC curve.**

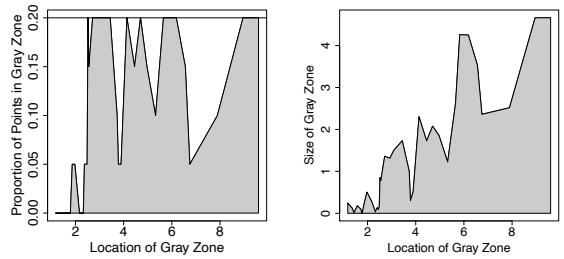

Fig. 4. Proportion of points falling in the gray zone (left) and width of the gray zone in the biomarker scale (right) as a function of c_j at $\gamma = .2$.

If $\gamma = 0$ the grayROC region collapses to the standard ROC line, also drawn in Figures 2 and 3.

In regions where the two class-specific distributions have little overlap, say left of 2, there can be little or no advantage in allowing for a gray zone. Conversely, where the density of biomarker points in the two classes is similar, a gray zone has the potential to improve the practical implementation of the biomarker. Figure 4 depicts this trade-off by elucidating where in the biomarker range the gray area is useful. Only in a narrow range of values does the grayROC algorithm needs to make full use of the 20% of data points allowed for the gray zone (top panel).

Lastly, Figure 5 shows grayROCs for four additional genes, chosen in part to illustrate less common features. Regions can be disjoint, when stretches of non-empty gray areas are followed by stretches of empty gray areas. Often this is associated with lack of monotonicity in the likelihood ratio of the two conditional biomarker distributions.

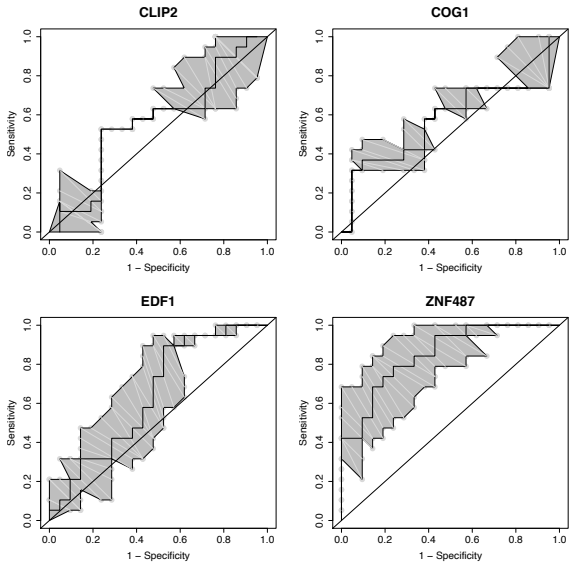

Fig. 5. grayROC displays at maximum tolerated percentage of unclassified cases, γ, of .2 **for the four genes indicated at the top of each panel. The thinner line corresponds to the standard ROC curve.**

ZNF487 exemplifies a biomarker with relatively good discrimination. The upper bounds indicates that correct reclassification of as few as 20% of cases could lead to high discrimination. This reclassification could be achieved by biomarkers that prove effective in the gray zone for ZNF487. The lower bound indicates that, if unclassified observations are handled poorly, the performance suffers, but discrimination remains above chance by a clear margin even with a gray area of 20%.

4. Discussion

I am not aware of a good visualization approach to examine classification algorithms that allow for an area of indeterminacy. I hope the grayROC will prove of practical help.

A grayROC visualization depends on the specification of the proportion γ of cases falling in the indeterminacy zone. The grayROC is, by design, sensitive to γ. Also, the influence of γ will differ in each dataset. In general, a plausible choice of γ may reflect the trade-offs inherent to the practical implementation of the algorithm. A grayROC can help users quantify and communicate the consequences of adopting a specific γ.

A full decision analytic approach (Raiffa and Schleifer (1961)) for selecting upper and lower thresholds (and thus γ) is feasible if one is able to quantify the utility associated with classifications, as well as the utilities following assignment to the gray zone. While the grayROC is not a method for optimally selecting γ, it can assist if the decision can only be approached informally. For example, if indeterminate cases need to examined by a costly human reader for accurate classification, different grayROC plots at varying γ can be used to informally evaluate the trade-off between added accuracy and added cost.

The grayROC is helpful when all cases have a known binary label but some are not classified. This differs from multi-class

ROC analysis (e.g. Hand and Till (2001)), where the number of labels is greater than two. It also differs from semisupervised analyses (Chapelle et al. (2006)), where some cases are not labeled. Lastly, it differs from systems where binary labels and/or classifications are replaced by fuzzy set memberships.

Evangelista et al. (2005) consider ensemble methods for classification. They visualize properties of the ensembles using a single ROC curve based on aggregating multiple classifiers through fuzzy logic operators including T-conorms and T-norms (Jang et al. (1997)). They term this approach "fuzzy ROC". In related work Castanho et al. (2007) generalize traditional ROC analysis to evaluate a single fuzzy-rule-based system, not necessarily arising through ensemble learning.

There are many valid alternatives to ROC curves for investigating and visualizing the properties of a threshold-based classifier. These include Total Operating Characteristic (Pontius and Si (2014)), Decision Curve Analysis (Vickers and Elkin (2006)) and Detection Error Tradeoff, which plots the false rejection rate versus the false acceptance rate (Martin et al. (1997)). I hope that the ideas illustrated here may be helpful in generalizing these methods to classifiers with indeterminacy zones.

The grayROC is not a visualization of uncertainty about the ROC curve in the standard statistical sense. Both the upper and lower bound are themselves point estimates, and their variability could be address by simple resampling approaches. Yet visualizing both the set and uncertainty about the set boundaries could be challenging. Also, γ is expressed in terms of the (potentially rescaled) proportion of cases in the validation study, without consideration for uncertainty.

The oracle and saboteur scenarios are extreme. Variants of this algorithm could be constructed by further specifying bounds on the proportion of cases that could be correctly classify by a human if left in the gray area. Then instead of moving all the gray area points to extremes, these known proportions could be used to move only some of the points and achieve less extreme bounds. These classification proportion could potentially depend on the biomarker region.

From a statistical perspective, indeterminacy can also help characterize regions of the score with poor discriminatory ability. Thus, compared to fully deterministic approaches, allowing for indeterminacy may lead to a different evaluation of classifiers and different approaches to biomarker discovery.

Acknowledgments

Work supported by NIH-NCI grant 4P30CA006516-51 and NSF grant DMS-1810829. The companion repository at `https://github.com/gp1d/grayROC.git` includes: 1) an R package implementing Algorithm 1 and grayROC plots; 2) R markdown files to reproduce all the analyses in Section 3.

References

Bradley, A., 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern recognition 30 (7), 1145–1159.

Castanho, M. J. P., Barros, L. C., Yamakami, A., Vendite, L. L., 2007. Fuzzy Receiver Operating Characteristic Curve: An Option to Evaluate Diagnostic Tests. IEEE Transactions on Information Technology in Biomedicine 11 (3), 244–250.

Chapelle, O., Schlkopf, B., Zien, A., 2006. Semi-supervised learning. MIT Press, Cambridge.

Evangelista, P. E., Embrechts, M. J., Bonissone, P., Szymanski, B. K., July 2005. Fuzzy roc curves for unsupervised nonparametric ensemble techniques. In: Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005. Vol. 5. pp. 3040–3045 vol. 5.

Fawcett, T., 2006. An introduction to ROC analysis. Pattern Recognition Letters 27 (8), 861 – 874, ROC Analysis in Pattern Recognition.
URL `http://www.sciencedirect.com/science/article/pii/S016786550500303X`

Ganzfried, B. F., Riester, M., Haibe-Kains, B., Risch, T., Tyekucheva, S., Jazic, I., Wang, X. V., Ahmadifar, M., Birrer, M. J., Parmigiani, G., Huttenhower, C., Waldron, L., 2013. curatedOvarianData: clinically annotated data for the ovarian cancer transcriptome. Database (Oxford) 2013, bat013, pMCID: PMC3625954.
URL `http://dx.doi.org/10.1093/database/bat013`

Hand, D. J., Till, R. J., Nov 2001. A simple generalisation of the area under the ROC curve for multiple class classification problems. Machine Learning 45 (2), 171–186.
URL `https://doi.org/10.1023/A:1010920819831`

Jang, J.-S. R., Sun, C. T., Mizutani, E., 1997. Neuro-Fuzzy and soft computing: A computational approach to learning and machine intelligence. Prentice-Hall.

Martin, A., Doddington, G., Kamm, T., Ordowski, M., 1997. The DET curve in assessment of detection task performance. Tech. rep., DTIC.

Parker, J. S., Mullins, M., Cheang, M. C. U., Leung, S., Voduc, D., Vickery, T., Davies, S., Fauron, C., He, X., Hu, Z., Quackenbush, J. F., Stijleman, I. J., Palazzo, J., Marron, J. S., Nobel, A. B., Mardis, E., Nielsen, T. O., Ellis, M. J., Perou, C. M., Bernard, P. S., Mar. 2009. Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes. Journal of Clinical Oncology 27 (8), 1160–1167.

Pontius, R. G. J., Si, K., 2014. The total operating characteristic to measure diagnostic ability for multiple thresholds. International Journal of Geographical Information Science 28 (3), 570–583.
URL `https://doi.org/10.1080/13658816.2013.862623`

Raiffa, H., Schleifer, R., 1961. Applied Statistical Decision Theory. MIT Press, Cambridge.

Riester, M., Wei, W., Waldron, L., Culhane, A. C., Trippa, L., Oliva, E., Kim, S.-H., Michor, F., Huttenhower, C., Parmigiani, G., Birrer, M. J., Apr 2014. Risk prediction for late-stage ovarian cancer by meta-analysis of 1525 patient samples. J Natl Cancer Inst.
URL `http://dx.doi.org/10.1093/jnci/dju048`

Vickers, A. J., Elkin, E. B., 2006. Decision curve analysis: a novel method for evaluating prediction models. Medical Decision Making 26 (6), 565–574.

Yoshihara, K., Tsunoda, T., Shigemizu, D., Fujiwara, H., Hatae, M., Fujiwara, H., Masuzaki, H., Katabuchi, H., Kawakami, Y., Okamoto, A., Nogawa, T., Matsumura, N., Udagawa, Y., Saito, T., Itamochi, H., Takano, M., Miyagi, E., Sudo, T., Ushijima, K., Iwase, H., Seki, H., Terao, Y., Enomoto, T., Mikami, M., Akazawa, K., Tsuda, H., Moriya, T., Tajima, A., Inoue, I., Tanaka, K., 2012. High-risk ovarian cancer based on 126-gene expression signature is uniquely characterized by downregulation of antigen presentation pathway. Clinical Cancer Research 18 (5), 1374–1385.
URL `http://clincancerres.aacrjournals.org/content/18/5/1374`