OBSERVER PERFORMANCE METHODS FOR DIAGNOSTIC IMAGING

Foundations, Modeling and Applications with R-Examples

Dev P. Chakraborty, PhD

Chapter 01: Preliminaries

This chapter provides background material, starting with diagnostic interpretations occurring everyday in hospitals, which define diagnostic tasks. The process of imaging device development by manufacturers is described, stressing the role of physical measurements in optimizing the design. Device approval for marketing requires measurements falling under the rubric of observer performance studies, the subject of this book. Once the device is deployed, hospital-based medical physicists use phantom quality control measurements to maintain image quality. Lacking the complexity of clinical images, phantom measurements may not detect clinical image quality degradation. Model observers, that reduce the imaging process to mathematical models and formulae, are intended to bridge the gap. Unlike physical, phantom and model observer measurements, observer performance methods measure the effect of the entire imaging chain, including the critical role of the radiologist. Four observer-performance paradigms are described that account for localization information to different extents. Physical and observer performance methods are put in the context of a hierarchy of efficacy levels, where the measurements become increasingly difficult, but more clinically meaningful, as one moves to higher levels. An overview of the book, which is split into four parts, is presented and suggestions are made on how to best use it. An online appendix introduces the reader to R/RStudio, which are used as pedagogical tools throughout the book, and RJafroc, an R-package encapsulating all analytical methods described in this book. It was coded mainly by Xuetong Zhai, MS. It extends the capabilities of the author's earlier Windows software (JAFROC at http://www.devchakraborty.com). All online material now resides at https://expertcadanalytics.com.

Part A

Chapter 02: The binary data collection paradigm

This chapter introduces the binary data collection paradigm, where the ground truth has two values, non-diseased and diseased, while the radiologist's response is likewise limited to two values: Yes – the patient is diseased or No – the patient is non-diseased. It is also termed the "Yes/No" paradigm. It leads to a 2 x 2 decision vs. truth state table, and the definitions of true positive, false positive, and their complements, false negative and true negative, respectively. An analogy is used to explain the terms sensitivity and specificity. Corresponding fractions true positive fraction (TPF) and false positive fraction (FPF) are defined. Sensitivity and specificity are estimated by TPF and (1 – FPF), respectively. The concepts of disease-prevalence, both in population and laboratory studies are explained. These are used to develop expressions for positive and negative predictive values (PPV and NPV), which are more relevant to clinicians than sensitivity and specificity. The expressions are illustrated with R code using values typical for screening mammography. It is shown that overall accuracy is a poor measure of performance as it is strongly influenced by prevalence. The reasons why PPV and NPV are irrelevant to laboratory studies are noted. The binary paradigm is the starting point for understanding the more common ratings paradigm described below.

Chapter 03: Modeling the binary paradigm

Described in this chapter is 2-parameter statistical model for the binary task. It introduces the fundamental concepts of *decision variable* and *decision threshold* (the latter, denoted zeta, is one of the parameters of the model) that pervade this book. Varying experimental conditions, like disease prevalence, cost and benefits of decisions, can induce the observer to alter the decision threshold. The other parameter is the separation mu of two unit variance normal probability density functions (*pdfs*). The normal distribution, and associated *pdf* and sampling from it, is illustrated with R-examples. Expressions for sensitivity and specificity are derived. The receiver-operating characteristic (ROC) plot of TPF vs. FPF is introduced. The area AUC under the ROC curve is a measure of performance that is independent of decision threshold. It also avoids ambiguity associated with comparing pairs of sensitivity and specificity values. The meaning of mu as the perceptual signal-to-noise ratio (pSNR) is compared to a SNR parameter in computerized analysis of mammography phantom images (CAMPI). The concept of random sampling is introduced by examining the dependence of variability of the operating point on the numbers of cases. Expressions are derived for 95% confidence intervals on an operating

point. An online appendix contains the 2nd part of the R tutorial, and more complex examples of code to be used in this book.

Chapter 04: The ratings paradigm

In this chapter the more commonly used *ratings* method of acquiring ROC data is described. It yields greater definition to the underlying ROC curve than just one operating point obtained in the binary task, and is more efficient. In this method, the observer assigns a rating to each case, where the rating is an ordered label summarizing the confidence of the observer that the case is diseased. Described is a typical ROC counts table, actually an observed clinical dataset, and how operating points (i.e., pairs of FPF and TPF values) are calculated from it. A labeling convention for the operating points is introduced, as is notation for the observed integers in the counts table. The rules for calculating operating points are expressed as formulae and implemented in R. The ratings method is contrasted to the binary method, in terms of efficiency and practicality. A theme occurring repeatedly in this book, that the *ratings are not numerical values but rather they are ordered labels* is illustrated with an example. A method of collecting ROC data on a 6-point scale is described that has the advantage of yielding an unambiguous single operating point. The forced choice paradigm is described. Two current controversies are addressed: one on the utility of discrete (e.g., 1 to 6) vs. quasi-continuous (e.g., 0 to 100) ratings and the other on the applicability of a clinical screening mammography-reporting scale for ROC analyses. The author recommends the 1 – 6 discrete scale and favors using cancer yield to reorder clinical ratings such as BI-RADS.

Chapter 05: Empirical AUC

The empirical AUC is defined as the area under the empirical ROC plot, which consists of straight-line segments connecting adjacent operating points, including the trivial ones at (0,0) and (1,1). It has the advantage of being independent of any modeling assumptions and the availability of analytical methods for determining its sampling behavior. Its main disadvantage is that it yields no insight into factors limiting the observer's performance. Empirical AUC based analysis is often termed non-parametric. Notation is introduced for labeling individual cases that is used in subsequent chapters. Formulae are presented for computing empirical operating points from ratings data. The formulae are illustrated using the dataset introduced in Chapter 03. An important theorem, often termed Bamber's theorem, relating the empirical area under the ROC to a formal statistic, known as the Wilcoxon, is derived. The importance of the theorem derives from its applications to non-parametric analysis of ROC data. An online appendix describes details of calculating the Wilcoxon statistic using an R-

coded example. Since the empirical AUC always yields a number, the researcher could be unaware about unusual behavior of the empirical ROC curve, so it is always a good idea to plot the data and look for evidence of large extrapolations. An example would be data points clustered at low FPF values, which imply a large AUC contribution, unsupported by intermediate operating points, from the line connecting the uppermost non-trivial operating point to (1,1).

Chapter 06: The binormal model

The equal variance binormal model was described in Chapter 02. In Chapter 04 it was shown, for a single clinical dataset, that the unequal-variance binormal model fitted the data better. This is a universal finding in many ROC studies dating to the 1960s, not just limited to medical imaging. The main aim of this chapter is to demystify statistical curve fitting. It starts with a description of the binormal model, and how it accommodates data binning. The invariance of the binormal model to arbitrary monotone transformations of the ratings is demonstrated with an **R** example – this gives the model unusual robustness with respect to deviations from strict normality. Expressions for sensitivity and specificity are derived. Two notations used to characterize the binormal model are explained. Expressions for the *pdfs* of the binormal model are derived. A simple linear fitting method is illustrated – this used to be the only recourse a researcher had before Dorfman and Alf's seminal 1969 publication. The maximum likelihood method for estimating parameters of the binormal model is detailed and an R-implementation is compared to a website implementation of ROCFIT, a commonly used software for fitting the binormal model to ROC data. Validation of the fitting method is described, i.e., how can one be confident that the fitting method, which makes normality and other assumptions, is valid for a dataset arising from an unknown distribution. An Appendix has a detailed derivation, originally published in a terse paper on the partial area under the ROC curve, a special case of which yields the total area under the binormal model fitted curve. An online appendix describes methods for debugging R code and displaying plots. Another appendix describes calculation of variance of binormal fitted AUC.

Chapter 07: Sources of variability affecting AUC

Irrespective of how it is estimated, AUC is a realization of a random variable, and as such is subject to variability. Identified are three sources of variability affecting AUC estimates: case-sampling, between-reader and within-reader. Unless strict replication of readings is used, the within-reader component cannot be separated from the others. Modern methods can analyze datasets without the need for such replication. AUC depends on a case-set index $\{c\}$, $c = 1,2,...,C$, where each case-set is a collection of randomly sampled specified numbers of

non-diseased and diseased cases. Described are two resampling methods, the jackknife and the bootstrap, of estimating variability of almost any statistic. Estimation of jackknife and bootstrap case-sampling variability, using R-code, is compared to an analytic non-parametric method that is applicable to the empirical AUC. The concept of a calibrated data simulator is introduced and illustrated with a simple example. A source of variability not generally considered, namely threshold variability, is introduced, and a cautionary note is struck with respect to indiscriminate usage of the empirical AUC.

Part B

Chapter 08: Hypothesis testing

This chapter describes, at a simple level, a widely used tool in statistics. Hypothesis testing is the process of dichotomizing the possible outcomes of a statistical study and using probabilistic arguments to choose one option over the other. The specific application considered in this chapter is how to decide whether an estimate of AUC is consistent with a pre-specified value. The concepts of null and alternative hypotheses, and Type I and Type II errors are introduced and illustrated with R examples. The role of pre-specified values of alpha and beta as controls on Type I and Type II errors, respectively, is explained, as are the ubiquitous terms "p-value" and "statistical power". The plot of empirical power (ordinate) vs. empirical alpha, termed by Metz the "ROC within and ROC" is explained. The reasons behind the conventional choices of alpha = 5% and beta = 20% are discussed. Very different values are adopted in other fields such as particle physics and cosmology. While comparing a single reader's performance to a specified value is not a clinically interesting problem, and the next two chapters describe methods for analyzing interpretations by a group of readers of a common set of cases in typically two modalities, the basic concepts remain the same.

Chapter 09: Dorfman-Berbaum-Metz-Hillis (DBMH) analysis

Interest is generally in comparing performance of a group of readers interpreting a common set of cases in two or more treatments. Such data is termed *multiple reader multiple case* (MRMC). This chapter describes the Dorfman-Berbaum-Metz (DBM) method. The other method, due to Obuchowski and Rockette, is the subject of the following chapter. Both have been substantially improved by Hillis. The DBMH approach, implemented in RJafroc, uses jackknife-derived pseudovalues as stand-ins for individual case-level figures of merit and standard analysis of variance (ANOVA) methods. Starting with mean squares, directly calculated from the pseudovalues,

the procedure is to compute a ratio that is distributed under the null hypothesis (NH) of no treatment effect, as an F-statistic with unit expected value. This permits testing the significance of the observed value of the F-statistic, calculation of the p-value and confidence intervals for the figure of merit difference. Expressions for non-centrality parameters, used in Chapter 11 for sample size estimation, are derived. Demonstrations with R software are used to illustrate the analysis including how to validate the analysis by showing that it has the expected rejection rate under the NH. The meaning of pseudovalues is presented as evidence of why the basic assumption of the analysis is valid for the empirical AUC.

Chapter 10: Obuchowski-Rockette-Hillis (ORH) analysis

This chapter describes the Obuchowski-Rockette method introduced in 1995, subsequently shown to be substantially equivalent to the DBMH method. Unlike the DBMH method, in the OR method the figure of merit is modeled directly. The two models have the same number of parameters, except that the covariance of the error term contains four terms, which account for different types of variances present in MRMC data. Since it holds the key to understanding the method, the error covariance is explained in detail by using the case-set {c} index. Resampling methods are used to estimate the error covariance terms. Described are random-reader random-case, fixed-reader random-case and random-reader fixed-case analyses. Applications of the method using the capabilities of the RJafroc package are given. The corresponding code is brief as most of the complexity is cloaked in the package. An online appendix contains a longer but more transparent implementation that follows the formulae. Another online appendix describes single-modality multiple-reader analysis, which is used in Chapter 22 to compare CAD to a group of radiologists interpreting the same cases.

Chapter 11: Sample size estimation for ROC studies

The topic addressed in this chapter is "how many readers and cases", usually abbreviated to "sample-size", should one employ to conduct an ROC study. As noted in Chapter 8, a typical design sets alpha equal to 5% and beta equal to 20%; the latter corresponds to 80% power. Sample size estimation involves: (a) performing a pilot study in order to determine the magnitudes of the variance components and (b) making an educated guess regarding the true performance difference between the two modalities, termed the anticipated effect size. Since the non-centrality parameter was defined in DBMH Chapter 9, in this chapter sample size estimation is illustrated using the DBMH method. Hillis has derived the conversions between DBMH and ORH parameters. An online appendix describes the corresponding implementation using the ORH method. Two examples, widely used in the methodology literature, are given using functions in RJafroc. A cautionary note, by Kraemer and

colleagues, regarding using the observed effect-size as the anticipated effect size, is summarized. Prediction accuracy of sample size estimation, investigated in a recent study, is summarized. The approach in this chapter is to utilize the confidence interval for the observed effect size as a guide in choosing the anticipated effect size. It is shown that the common practice of specifying effect-size as the difference of two AUCs can be misleading. Suggested is an alternative that uses the relative change in the separation parameter corresponding to a given AUC as a measure of effect-size.

Part C

Chapter 12: The free-response receiver operating characteristic (FROC) paradigm

This chapter introduces the FROC paradigm. Unlike the ROC paradigm, this paradigm allows for localization information, classified as correct or incorrect, to be used in the analysis. The FROC paradigm is shown to be a search task. The structure of FROC data, namely a random number of zero or more mark-rating pairs per image, is described. A proximity criterion is used to classify each mark as lesion localization or non-lesion localization. The use of ROC terminology such as true positive and false positive in the FROC context is discouraged. The FROC paradigm is placed in its historical context. The FROC paradigm is placed in its historical context and a key publication by Bunch et al is described. The FROC curve is introduced and a data simulator, implemented in R, is used to show its dependence on perceptual signal to noise ratio (pSNR). A "solar" analogy is used to explain, at an intuitive level, the dependence of the FROC curve on pSNR. Online appendices explain the code used to generate the plots, explain physical SNR measurements in the mammography quality control context and the "Bunch transforms" relating ROC and FROC curves.

Chapter 13: Empirical operating characteristics derivable from FROC data

This chapter describes various empirical operating characteristics that can be plotted from FROC date. The chapter is divided into two parts: formalisms and examples. A distinction is made between latent (suspicious regions, regardless of whether they are marked) and actual marks. The formalism describes the construction of operating points from the data. Calculation of corresponding AUCs is addressed in the next chapter. Formalism for the AFROC, the inferred ROC, the weighted AFROC (wAFROC), etc., is presented. R-code examples of "raw" and binned FROC, AFROC plots are shown. A misconception about location level true-negatives – actually an unmeasurable event – is explained. The rationale is presented for not using non-lesion localizations

on diseased cases in computing a measure of performance. Another misconception, that neglect of all but the highest rated NL on non-diseased cases sacrifices much statistical power, is addressed. It is recommended that FROC-curve based measures of performance not be used to assess performance in localization tasks. Evidence for this recommendation, based on simulation studies, is presented here and in Chapter 17. The AUC under the weighted AFROC is the recommended way of summarizing performance in localization tasks.

Chapter 14: Computation and meanings of empirical FROC FOM-statistics and AUC measures

Computational formulae are provided for AUCs under various operating characteristics defined in the previous chapter. The important ones, namely AUCs under the AFROC and the weighted AFROC are detailed. Formulae for other AUCs (e.g., FROC and inferred ROC) are in the online appendix to this chapter. Small dataset examples, which permit hand-calculations, are used to illustrate the AFROC and wAFROC AUC, allowing the reader to appreciate that the AFROC gives undue importance to cases with more lesions, while the wAFROC corrects this deficiency. Two theorems are derived: the first shows the equivalence between the empirical AUC under the wAFROC and a quasi-Wilcoxon statistic. This theorem is the free-response equivalent of Bamber's theorem relating the empirical ROC AUC to a Wilcoxon statistic. The second theorem derives an expression for the area under the straight-line extension from the empirical end-point to (1,1). The contribution of this area increases as the abscissa of the end-point of the wAFROC decreases, i.e., as more non-diseased cases are not marked and as ordinate of the end-point increases, i.e., as more lesions, especially those with greater weights, are marked. Other online appendices have details of the proofs and numerical-integration based demonstrations of the AUC vs. quasi Wilcoxon statistic equivalences.

Chapter 15: Visual Search Paradigms

Visual search is defined as grouping and assigning labels to different regions in the image, where the labels correspond to entities that exist in the real world. Two components of expertise in such tasks are identified. Search expertise is the ability to find lesions while avoiding finding non-lesions. Lesion-classification expertise is the ability to correctly classify found suspicious regions. Two experimental methods of studying visual search are described. By far the more common one involves measuring reaction time and accuracy in an image possibly containing a defined target and defined distractors. In the medical imaging context targets and distractors cannot be defined, as these are perceptual constructs existing in the radiologist's mind. The second paradigm involves measuring, using eye-tracking apparatus, where the radiologist is looking. The Kundel-Nodine schema for modeling visual search in the medical imaging context is described. It consists of two

stages, a brief global impression stage which uses peripheral vision to identify localized perturbations, the latent marks of the FROC paradigm. The second stage involves examining each found suspicious region using the fovea, and making a decision of whether to mark it. The 2-stage process bears a close resemblance to how CAD algorithms are designed. The Kundel-Nodine schema is the starting point of the Radiological Search Model (RSM) described in the following chapter.

Chapter 16: The Radiological Search Model (RSM)

The Radiological Search Model (RSM) is a statistical parameterization of the Nodine-Kundel model. It consists of a search stage, characterized by two parameters: lambda and nu. The lambda parameter is the mean of a Poisson distribution describing the distribution of the numbers of non-lesion localizations per case, with smaller values indicative of avoiding generating non-lesion localizations. The second parameter is the success probability of a binomial distribution, describing the probability that a true lesion will be found by the search mechanism, with values approaching unity indicative of the ability to find true lesions. The third parameter, mu, is the separation of two unit variance normal distributions, describing the distribution of sampled decision variables from non-lesion and true lesion sites, with larger values indicative of correct classifications of found suspicious regions. The total number of RSM parameters, namely three, is one more than conventional ROC models. This is because describing search requires two parameters and describing lesion-classification performance requires one parameter. A re-parameterization is described that converts the mu-dependent lambda and nu parameters to intrinsic parameters, which are independent of the mu parameter. Online appendices demonstrate Poisson and binomial sampling.

Chapter 17: Predictions of the Radiological Search Model

This chapter describes predictions of the RSM and how they compare with evidence. All RSM-predicted operating characteristics share a constrained end-point property. All other ROC models predict that the end-point, namely the uppermost non-trivial point on the ROC, reached at infinitely low reporting threshold, is (1,1), while the RSM predicts it does not reach (1,1). The reason is that the RSM allows the existence of cases with no latent sites and hence no z-samples. Expressions for the ROC operating point as a function of the three RSM parameters, the lesion distribution vector, and the reporting threshold, are derived. The accessible portion of the RSM-predicted ROC curve is proper. To fully account for performance, the area under a straight-line extension from the empirical end-point to (1,1) needs to be included. Similar comments apply to the AFROC and wAFROC. Unlike them, the FROC curve cannot be meaningfully extended beyond the observed end-point.

Chance level performance for the AFROC and the observer who does not mark any image is discussed. The latter, yielding AFROC-AUC = 0.5, is more informative - perfect performance on non-diseased cases - compared to the corresponding ROC (ROC-AUC = 0.5). Expressions are derived for search and lesion-classification performance, defined in Chapter 15, as functions of RSM parameters. These can be estimated using RSM-based curve fitting described in the next chapter. Evidence for the validity of the RSM is presented as is further evidence that the FROC is a poor descriptor of performance.

Chapter 18: Fitting RSM to FROC/ROC data and key findings

Because of degeneracy issues, where two RSM parameters appear in an inseparable combination, FROC curve based fitting is not possible with human observer data. However, it can be used for designer level CAD data, as in the initial-detection candidate-analysis (IDCA) method. Fitting the ROC curve, which breaks the degeneracy, has proven successful. The RSM-ROC fitting algorithm was applied, along with PROPROC and CBM, to 236 individual datasets, drawn from 14 MRMC datasets described in Online Chapter 24. A sampling of the three fits are presented in this chapter, and the rest can be viewed on the author's website. For each dataset, AUCs predicted by PROPROC or CBM, when plotted against RSM AUCs, showed a near unit-slope linearity through the origin, with $R^2 > 0.999$. Compared to the RSM, the PROPROC predictions were, on the average, 2.6% larger, while the CBM estimates were 1% larger. In some cases PROPROC, and to a lesser extent CBM, grossly overestimated AUC by performing a large extrapolation to (1,1) unsupported by the intermediate operating points. The near equality of all three estimates is explained in terms of uniqueness of ideal observer performance that each algorithm attempts to emulate. An inverse correlation was found between search and lesion-classification performances, suggesting that observers tend to compensate for deficiency in search by improved performance in lesion-classification, and vice-versa. Average search performance was 23%, average lesion-classification performance was 88% and average AUC was 79%. These values demonstrate that search expertise is the bottleneck limiting overall performance in diagnostic tasks.

Chapter 19: Analyzing FROC data and sample size estimation

This chapter describes analysis of multiple-reader multiple-case (MRMC) FROC datasets. Apart from the choice of figure of merit, analyzing FROC data is similar to analyzing ROC data. The DBMH and ORH methods are applicable to any scalar figure of merit. No assumptions are made regarding independence of ratings on the same case – a sometimes-misunderstood point. Analysis of a sample FROC dataset is demonstrated, including visualization of the relevant operating characteristic using RJafroc implemented

functions. Suggestions are made on how to report the results of a study (the suggestions apply equally to ROC studies). Single fixed factor analysis is described, followed by a newly developed crossed-treatment analysis, applicable when one has two treatment factors and their levels are crossed. Sample-size estimation for FROC studies is also not fundamentally different from that described for ROC studies. Using the RSM fitting method described in the previous chapter, NH values for the three RSM-parameters are derived. These allow relating a chosen ROC effect size to the equivalent AFROC effect size. The latter is usually larger, hence the increased power.

Part D

Chapter 20: Other, non RSM, methods of fitting proper ROC curves

The binormal model, widely used since the 1960s, to fit ROC data in various, mostly non-medical-imaging contexts, predicts an inappropriate chance line crossing and a "hook" that is a patently false prediction. While excuses have been made as to why this discrepancy can be ignored, researchers have been working on methods of fitting proper ROC curves, defined as those that do not show such inappropriate behavior. Related issues are data degeneracy and the ability to fit datasets with no interior ROC operating points. Expert radiologists frequently produce such datasets, and indeed the RSM predicts this behavior for experts. The RSM-based proper ROC curve-fitting method was described in Chapter 19. This chapter describes three classical methods for fitting proper ROC curves. The likelihood ratio is defined and is shown that an observer using a likelihood ratio based decision variable is ideal in the Neyman-Pearson sense. The proper ROC, implemented in PROPROC software, uses a complex transformation from the binormal model z-sample scale to a likelihood ratio scale. PROPROC cannot fit degenerate datasets. The contaminated binormal model (CBM) is simpler, has a resemblance to the RSM, and can fit practically any dataset, including degenerate ones. Finally, for historical completeness, the bigamma model is reviewed. Its basic assumption of a gamma distributed decision variable is inconsistent with the central limit theorem of statistics. Online appendices detail how to view ROC curves and their slopes, and how to plot PROPROC, CBM and bigamma model predicted ROC curves.

Chapter 21: The bivariate binormal model and its software implementation (CORROC2)

Metz and Kronman introduced the bivariate binormal model around 1980 to analyze paired ROC datasets. Its software implementation (CORROC2) has been used in over one hundred publications, but it is not well documented in the archival literature. While not necessary to analyze MRMC datasets, it is the only known method, until a recent advance, to measure correlations at the latent decision variable level between paired interpretations. The correlations are needed to design a calibrated simulator – an example is given in Chapter 23. A calibrated simulator is essential to proper validation of any proposed analysis method. The bivariate sampling model and visualizing the bivariate probability density function are illustrated with R examples. Parameter estimation of the bivariate binormal model is described. Practical details of CORROC2 software are provided as well as how this Windows software can be run in an OS X environment. The details of the code are in online appendices. A recent advance, which replaces the binormal model with the contaminated binormal model, is described. It too yields the desired correlations but is robust with respect to degenerate datasets that cannot be fitted by CORROC2.

Chapter 22: Comparing performance of standalone CAD to a group of radiologists interpreting the same cases

Since in the US a low "second-reader" bar for CAD to be considered a success has been adopted, standalone CAD performance is rarely measured. The author analyzed a dataset from one of the few published studies in mammography where standalone performance of CAD was compared to a group of expert radiologists interpreting the same cases. The published analysis is extended to account for case-variability. By considering the difference in performance between each radiologist and CAD, the problem reduces to the single modality multiple reader analysis described in Chapter 10, except that the NH is that the average difference performance is zero. The method is applicable to any scalar figure of merit that can be calculated from the data. The use of partial area measurements in CAD research is strongly discouraged. It is shown that these yield ambiguous results and moreover ignore some of the data. R code is presented for fixed–case and random-case analysis, the former duplicating the analysis in the Hupse-Karssemeijer 2013 publication. Allowing case to be random increases, as expected, the widths of confidence intervals and p-values. The AUC under the LROC yields a p-value of 0.0349, significant at alpha = 0.05 but the corresponding AUC under the ROC did not yield a significant difference ($p = 0.321$). The difference is attributed to the increased statistical power of LROC relative to ROC, which is afforded by using location information.

Chapter 23: Design and calibration of a single-modality multiple-reader decision variable simulator and using it to validate the proposed CAD analysis method

Co-author: Xuetong Zhai, MS

This chapter leverages a recent advance, namely the bivariate contaminated binormal model (BCBM), which allows estimation of decision variable correlations between two readers interpreting the same cases. Unlike the bivariate binormal model for which software (CORROC2) was developed more than three decades ago, the software implementation of BCBM, named CORCBM (for correlated CBM), can fit degenerate datasets. The BCBM-based simulator generates samples, one per reader, for each case, from a multivariate normal distribution with specified mean vector and covariance matrix. The calibration of these parameters to statistically resemble the CAD vs. 9-radiologist Hupse-Karssemeijer dataset is described and implemented in R. The simulator was used to generate 2000 samples under the null hypothesis condition. For each simulation covariance and variance of the Obuchowski–Rockette single-modality method was estimated and a NH-rejection was recorded if the p-value was below 0.05. The observed fraction of rejections was close to 0.05, thereby validating the analysis of the previous chapter. The averages of the covariance and the variance were within the corresponding bootstrap confidence intervals, calculated from the original dataset. This shows that the simulator is statistically identical to the original dataset. A realistic simulator is key to proper validation of any proposed method of analyzing ROC data.