Learning Curves in

Emergency Ultrasonography

By

Kaitlyn Brady

A Project Report

Submitted to the Faculty

of

WORCESTER POLYTECHNIC INSTITUTE

In partial fulfillment of the requirement for the

Degree of Master of Science

in

Applied Statistics

December 2012

APPROVED:

_____

Dr. Joseph D. Petruccelli, Project Advisor

_____

Dr. Bogdan Vernescu, Department Head

Abstract

This project utilized generalized estimating equations and general linear modeling to model learning curves for sonographer performance in emergency ultrasonography. Performance was measured in two ways: image quality (interpretable vs. possible hindrance in interpretation) and agreement of findings between the sonographer and an expert reviewing sonographer. Records from 109 sonographers were split into two data sets—training (n=50) and testing (n=59)—to conduct exploratory analysis and fit the final models for analysis, respectively. We determined that the number of scans of a particular exam type required for a sonographer to obtain quality images on that exam type with a predicted probability of 0.9 is highly dependent upon the person conducting the review, the indication of the scan (educational or medical), and the outcome of the scan (whether there is a pathology positive finding). Constructing family-wise 95% confidence intervals for each exam type demonstrated a large amount of variation for the number of scans required both between exam types and within exam types. It was determined that a sonographer's experience with a particular exam type is not a significant predictor of future agreement on that exam type and thus no estimates were made based on the agreement learning curves. In addition, we concluded based on a type III analysis that when already considering exam type related experience, the consideration of experience on other exam types does not significantly impact the learning curve for quality. However, the learning curve for agreement is significantly impacted by the additional consideration of experience on other exam types.

Acknowledgements

I would like to express my deepest gratitude to my project and academic advisor, Dr. Joseph Petruccelli. He has provided me the wonderful opportunity to work on three real-world statistical application projects, helping me develop my analytical and written skills, culminating in this project. His guidance, insight, and patience proved invaluable.

I also wish to thank Dr. David Blehar and Dr. Romolo Gaspari at the University of Massachusetts School of Medicine for presenting the motivation for this project and providing the data. Their practical experience and contributions to this project provided the necessary medical relevance of this study.

Table of Contents

Table of Tables

Table of Figures

1 Introduction

Ultrasonography is an imaging technique that uses high-frequency sound waves to assist internal medical evaluations of the body (American College of Radiology [ACR], 2011). Emergency ultrasonography (also referred to as point of care, bedside, focused, clinical and physician performed ultrasonography) is a practice used to answer targeted questions with the flexibility to be conducted in any setting, and in situations limited by time and patient condition (American College of Emergency Physicians [ACEP], 2008). To date there are no set standards in place to classify when a sonographer may be considered competent and qualified to perform emergency ultrasonography (Hertzberg et al., 2000; Jang, Ruggeri, Dyne, & Kaji, 2010).Consequently, educational and training courses may vary between programs and departments.

Recommended guidelines have been published providing suggestions as to how much experience an individual should gain in order to be deemed qualified. For exam specific skills, these suggestions range from 25-50 (ACEP, 2008) scans of each exam type through the course of a training program to 500 (ACR, 2011) scans for specializing in one exam type without taking part in a training program. For general qualification (inclusion of all exam types, as opposed to specialized experience) through a hands-on training program, suggestions range from 150-250 (ACEP, 2008) scans to 500 (ACR, 2011) scans.

Guidelines are often not founded on (or supplemented with) empirical evidence, but instead are based on expert opinion and popular consensus (Gaspari, Dickman, & Blehar, 2009). In response to this, many studies have been conducted to assess common guidelines (Hertzberg et al., 2000; Gaspari et al., 2009; Jang et al., 2010). Results from these studies vary greatly, from supporting a requirement of just 25 scans for right upper quadrant (RUQ) scans (Gaspari et al., 2009), to suggesting that 50 RUQ scans is not enough to improve sonographer accuracy (Jang et al., 2010). Another study found 200 overall exams (inclusive of all exam types) to be insufficient to deem a sonographer competent (Hertzberg, 2010).

The discrepancies between findings suggest a need for further investigation into the learning curves of sonographers in emergency ultrasonography. Until now, guidelines and studies have focused on identifying a specific level of experience at which performance significantly increases above that of past experience levels, as opposed to identifying a point at which proficiency reaches a targeted level (Jang et al., 2010). This study focuses on the latter—estimating the number of scans required to achieve a specified proficiency level.

The American College of Radiology (2012) proposes a unique set of competency guidelines for each ultrasound exam type, suggesting that the learning curve for ultrasonography depends upon the exam being performed. Thus, this study examines each exam type individually, constructing the learning curves for each exam type based on experience on the specified exam type. In addition, we consider the implications of the added consideration of experience on exams beyond that of the specified exam type. We use two measurements of performance: image quality and agreement with the attending expert reviewer, and set the required proficiency level at 0.9 based on a recommendation from Jang et al. (2012).

1.1 Objectives

This project aims to model the learning curves of sonographers for ultrasonography in an emergency department for twelve exam types. In particular, we seek to identify which factors affect the shape of the learning curve and the implications these factors have on performance. We examine two measures of performance—quality and agreement—and utilize the modeled learning curves to estimate how many scans are required to obtain adequate performance with a specified predicted probability.

1.2 Research Questions

The questions to be answered are as follows:

1a. For each of the 12 exam types, how many scans of that exam type must a sonographer perform to obtain quality images with a predicted probability of 0.9?

1b. For each of the 12 exam types, how many scans of that exam type must a sonographer perform to obtain agreement with the attending reviewer with a predicted probability of 0.9?

2a. Is the learning curve for quality for an exam type affected by experience with other exam types?

2b. Is the learning curve for agreement for an exam type affected by experience with other exam types?

In the process of answering these questions, we also seek to identify possible confounding factors, including reviewer, educational indication, and pathology.

2 The Data

The data for this study was provided by the University of Massachusetts School of Medicine and consists of records for resident students at four University of Massachusetts Memorial Health Care emergency departments. Students in all four departments underwent the same training program, following the same protocol and using the same machines, while working interdepartmentally.

2.1 Terminology

The following terms and concepts will be referred to throughout this report.

- *Sonographer*: The resident (student) who is conducting the scan.
- *Reviewer*: The experienced, non-student, expert sonographer reviewing and rating the scan.
- *Initial Reading*: The sonographer's interpretation of his or her scan.
- *Final Reading*: The reviewer's interpretation of a sonographer's scan.
- *Quality Rating:* An ordinal variable for the quality score of a scan's images, ranging from 1-8, assigned by the reviewer. A rating of 1 indicates no interpretable images. A rating between 2 and 4 indicates possible limitation of interpretation due to poor image quality. A rating between 5 and 7 indicates interpretable images but with room for improvement. A rating of 8 indicates perfect technique without any errors.
- *Quality*: An indicator variable for whether a scan's quality rating was greater than 4. This variable equals 1 when the rating is at least 5 and represents an interpretable scan. It equals 0

when the rating is less than 5 and represents possible limitations in interpreting the scan. We refer to a sonographer obtaining quality on a particular scan when this variable equals 1.

- *Agreement:* An indicator variable for whether a sonographer's initial findings coincided with those of the attending reviewer's final reading.
- *Educational Indication*: An indicator variable for whether or not a scan was performed solely for educational (non-medical) purposes. It equals 1 for educational scans and 0 for medical scans.
- *Pathology:* An indicator variable for whether or not a scan resulted in positive findings by either the sonographer or reviewer.

2.2 Preparation

The data was collected over a 5 year period (03/2007-05/2012), recording the performance of 109 new sonographers beginning with their first scan. For a given sonographer, the order in which scans were performed was maintained down to the date. The database did not allow for discrimination of order for scans taking place on the same date. It was determined through preliminary exploration of multiple permutations of the same-date scans that order of scans performed on the same date did not change initial findings. Thus, the original order of appearance in the database was preserved.

One scan was deleted from the data set where the quality rating equaled 0 and the initial sonographer and final reviewer readings were missing, as it was believed to be incorrectly recorded data. There were several situations for which scans were excluded from analysis, but counted towards experience. There was one reviewer who conducted only 3 reviews over the 5-year period—these scans were excluded from analysis as they were thought to be 'test cases' and not formal reviews. Scans with no initial or final readings were excluded from agreement analyses (agreement = missing). Of the 26 exam types recorded, two of these were condensed into one exam type (lower extremity-left and lower-extremity-right were combined into lower extremity) and an additional 13 were excluded from analysis as they were procedural types or types not taught in the department and were thus considered unrepresentative. A list of all exam types is located in Appendix A. Following the data cleansing, a total of 55,731 scans consisting of 12 exam types were considered for quality analyses and 51,134 for agreement analyses.

The final step in preparing the data was to randomly split the data into a training set and a testing set to prevent data snooping. A computer generated random method was employed to split the data roughly in half. Each sonographer was assigned a random number from 0 to 1 and those with a number less than 0.5 were assigned to the training set, while the remainders were assigned to the testing set. A total of 50 sonographers were included in the training set and 59 in the testing set. Throughout the study, the training set was used for exploration, while the testing set was used to obtain final model fits and conduct analysis.

3 Methodology

3.1 Generalized Estimating Equations (GEE)

Since the data represent repeated measures, ordinary regression techniques which treat all observations as independent are not appropriate. It is reasonable to expect that performance on scans from an individual sonographer may be correlated and therefore we may consider the set of each sonographer's scans to be a separate cluster.

3.1.1 The GEE Model

The generalized estimating equations model is defined as follows:

$Let\ y_{ij}\ represent\ the\ j^{th}\ observation\ of\ the\ i^{th}\ sonographer, j = 1, \dots, n_i\ and\ i = 1, \dots, s.$

$Let\ x_{ijl}\ represent\ the\ j^{th}\ measurement\ of\ the\ i^{th}\ sonographer\ on\ variable\ l, l = 1, \dots, p.$

$Define\ \boldsymbol{Y}_i = \left[Y_{i1}, \dots Y_{in_i}\right]', \boldsymbol{x}_{ij} = \left[x_{ij1}, \dots, x_{ijp}\right]',\ \boldsymbol{\mu}_i = \left[\mu_{i1}, \dots, \mu_{in_i}\right]', \boldsymbol{V}_i = covariance\ matrix\ of\ \boldsymbol{Y}_i\ .$

$Then\ \widehat{\boldsymbol{\beta}}\ is\ the\ solution\ to\ the\ GEE\ given\ by$

$$\boldsymbol{S}(\boldsymbol{\beta}) = \sum_{\boldsymbol{i}} \left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}\right) \boldsymbol{V}_i^{-1}\left(\boldsymbol{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})\right) = \boldsymbol{0}$$

$where\ \mu_i(\boldsymbol{\beta}) = g^{-1}(\boldsymbol{x_i'\beta}).$

Since this study deals with binary response variables, the GEE approach is applied to a general linear model where $g(p) = ln\left(\frac{p}{1-p}\right)$ is the logit link function.

3.1.2 Purpose of GEE

Generalized estimating equations allow the model builder to specify an assumed covariance structure to account for correlation within clusters, while still providing interpretations over the average population. As opposed to the maximum likelihood method which assumes a multinomial distribution for the responses and becomes impractical for a large number of observations or variables, GEE's use a quasi-likelihood method in model fitting which does not assume a specific distribution. Thus, GEE's allow for a computationally simpler method for fitting regression models which utilize repeated measures data.

3.1.3 Assumptions of GEE

Generalized estimating equations require four assumptions:

1. The responses are from a distribution in the exponential family with a specified mean, $\mu = E(Y)$, and variance which is a function of the mean, $V(\mu)$.

2. The mean is a linear function of the predictors (this may be obtained through the use of a link function).

3. A covariance structure for the responses, $\mathbf{Y}_i$, must be specified, meaning that a working guess of the correlation structure is required. (Some possible options include independent, m-dependent, exchangeable, unstructured, and autoregressive (1)).

4. Any missing data are either missing completely at random (the probability that an observation is missing is independent of that observation's value) or that the data are missing at random (the probability that an observation is missing depends only on the observed values) and the estimating equations can be weighted by response probabilities.

One of the benefits of the GEE approach is that parameter estimates are robust to the specification of the working correlation structure. Thus, estimates are consistent even if the wrong structure is chosen. However, the empirically estimated standard errors are sensitive to the specification and tend to underestimate and have greater variability than the true values.

For this study, we reasonably assume assumptions one and two have been met. For assumption three, after data exploration, an independent working correlation structure is assumed. Lastly, it may not be reasonable to assume that missing data are missing completely at random (we do not weight the GEE and therefore do not consider the alternative of missing at random). It may or may not be the case that missing data as a result of no initial and/or final reading are missing completely at random. Additionally, it might be the case that the exam types deleted from analysis are more or less likely to show quality or agreement. For modeling the learning curves this is not an issue, as we limit our results to the exam types studied. However, for testing the significance of adding non exam experience to the model, we introduce the missing data. Thus, we proceed with caution in our results, under the consideration that missing data may not be reasonably assumed to be missing completely at random.

3.2 Inverse Interval Estimation

The classical use of a regression model is to predict a response value based on an explanatory variable. Sometimes however, as is the case in this project, it is of interest to determine which value of the explanatory variable is required to obtain a specified value of the response variable. In this case, one may solve the regression problem for the explanatory variable.

For a multiple logistic regression model where r is the quantitative explanatory variable of interest with β_r its corresponding coefficient, $\boldsymbol{x}$ is a p x 1 vector of the additional explanatory variables in the model with corresponding coefficients $\boldsymbol{\beta_x}$, and desired probability p, we may write the log odds resulting from the regression model as

$$log\left(\frac{p}{1-p}\right) = \beta_r r + \boldsymbol{x}^T\boldsymbol{\beta}$$

Solving yields

$$r = \frac{log\left(\frac{p}{1-p}\right) - \boldsymbol{x}^T\boldsymbol{\beta}}{\beta_r} = \frac{\mu_a}{\mu_b}\text{, estimated by } \hat{r} = \frac{a}{b}$$

where

$$\hat{\mu}_a = a = log\left(\frac{p}{1-p}\right) - \boldsymbol{x}^T\widehat{\boldsymbol{\beta}} \text{ and } \hat{\mu}_b = b = \hat{\beta}_r$$

Whenever possible, confidence limits should be provided along with point estimates. Thus, a Wald-type interval is provided for the above estimate, $\hat{r}$.

It follows from

$$\begin{pmatrix} a \\ b \end{pmatrix} \underset{\sim}{^{apx.}} N\left(\begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix}, \begin{pmatrix} V_{aa} & V_{ab} \\ V_{ab} & V_{bb} \end{pmatrix}\right),$$

that

$$a - yb \underset{\sim}{^{apx.}} N(0, \hat{\sigma}^2) \text{ with } \hat{\sigma}^2 = V_{aa} - 2\hat{y}V_{ab} + \hat{y}^2V_{bb} \text{ ,}$$

which provides the following level $1-\alpha$ confidence interval:

$$b^{-1}(a - \hat{\sigma}z, a + \hat{\sigma}z), if\ b > 0$$
$$b^{-1}(a + \hat{\sigma}z, a - \hat{\sigma}z), if\ b < 0$$

where $z = z_{1-\alpha/2}$ is the $1 - \frac{\alpha}{2}$ quantile of the standard normal distribution.

3.3 Area under the Receiver Operating Characteristic Curve

For a binary response model, a receiver operating characteristic (ROC) curve is a plot of the proportion of true positives (events predicted to be events) versus the proportion of false positives (nonevents predicted to be events), where the cut point for determining whether a predicted probability signifies an event is varied for each point on the ROC curve. The area under the ROC curve (AUC) is a measure of a model's predictive ability and ranges from 0 to 1. An AUC of 0.5 is equivalent to the predictive ability of a random model, values below 0.5 correspond to a predictive ability worse than that of a random model, and values above 0.5 correspond to better predictive ability than a random model. Thus, higher values above 0.5 suggest better predictive models. We estimate the AUC from a SAS macro that utilizes pair wise deletion Mann-Whitney statistics (SAS Institute, 2012).

3.4 Type III Analysis

A type III analysis tests the significance of a model term's effect above and beyond the effects of all other terms included in the model. The general method for this type of analysis is using sums of squares. However since we use generalized estimating equations with only quasi-likelihood functions in this study, there is a nuance to the type III method. For a GEE model, a score test is used to test the null hypothesis H_0: $\boldsymbol{L}'\boldsymbol{\beta} = \boldsymbol{0}$ against H_1: $\boldsymbol{L}'\boldsymbol{\beta} \neq \boldsymbol{0}$, where $\boldsymbol{L}'$ is a contrast for the type III test. Consider the regression parameters obtained from solving the restricted model $\boldsymbol{L}'\boldsymbol{\beta} = \boldsymbol{0}$, call them $\widehat{\boldsymbol{\beta}}$. Let $\boldsymbol{S}(\boldsymbol{\beta})$ represent the values of the GEE evaluated at $\widehat{\boldsymbol{\beta}}$. Then the generalized score statistic is defined as

$$T = \boldsymbol{S}(\widehat{\boldsymbol{\beta}})' \sum_m \boldsymbol{L}(\boldsymbol{L}' \sum_e \boldsymbol{L})^{-1} \boldsymbol{L} \sum_m \boldsymbol{S}(\widehat{\boldsymbol{\beta}})$$

where $\sum_m$ is the model-based estimated covariance matrix and $\sum_e$ is the empirically estimated covariance matrix. The p-value for the test statistic, T, is computed using a Chi-Square distribution with r degrees of freedom, where r is the degrees of freedom for the parameter of interest.

3.5 Bonferroni Adjustment for Multiplicity

Due to the extreme number of confidence intervals being computed in the study (240 for quality), the Bonferroni method was used to adjust for multiplicity. In general, to obtain a set of k confidence intervals with a family-wise $100(1-\alpha)\%$ confidence level, each confidence interval is computed at an individual $100\left(1-\frac{\alpha}{k}\right)\%$ level.

4 Analysis

The analysis was comprised of four main steps: exploratory analysis, model building, limit calculation, and hypothesis testing. The first step utilized the training data set, the second step utilized both the training and testing data sets, and the last two steps utilized the testing data set.

4.1 Data Exploration

Prior to performing any analyses, it is important to understand the data and identify possible underlying trends or limitations. In our exploration, we used graphics and descriptive statistics to investigate individual sonographer performance, aggregate sonographer performance, and possible interactions between performance and non-experience related variables.

4.1.1 Descriptive Statistics

Within the training set, 80.85% of all scans performed resulted in quality, while 94.39% resulted in agreement. Of the twelve exam types, most of the scans conducted were FAST exams (25.14%), with the fewest being Cardiac (Code) and Ophthalmic (Ocular) (0.37% and 0.43%, respectively). A total of 32.05% of scans were for purely educational purposes, and a total of 15.59% all scans resulted in positive findings by either the learning sonographer or expert reviewer. Of the five reviewers, reviewer A conducted the vast plurality of the reviews (44.4%).

A more detailed breakdown of the distribution of all variables may be found in Appendix B.

4.1.2 Individual Sonographers

Using the training data set, exploration of individual sonographers' performance for both quality and agreement shows variation from sonographer to sonographer, although there is a common trend for performance on agreement to be higher than performance on quality. Figure 1 shows graphs for a sample of three sonographers, depicting the proportion of scans resulting in quality and the proportion of scans resulting in agreement versus exam experience. The values at a particular level of exam experience (summing over all exam types) were calculated as follows:

$$p\_Q = \frac{\sum_j q_j}{n_q}, \qquad p\_A = \frac{\sum_j a_j}{n_a}, \qquad j = 1, \ldots, 12$$

where

$$a_j = \begin{cases} 1 \; if \; agreement = 1 \; for \; exam \; type \; j \\ 0 \; if \; agreement = 0 \; for \; exam \; type \; j \end{cases}, \qquad q_j = \begin{cases} 1 \; if \; quality = 1 \; for \; exam \; type \; j \\ 0 \; if \; quality = 0 \; for \; exam \; type \; j \end{cases}$$

$$n_a = \sum_j v_{aj}, \qquad n_q = \sum_j v_{qj}$$

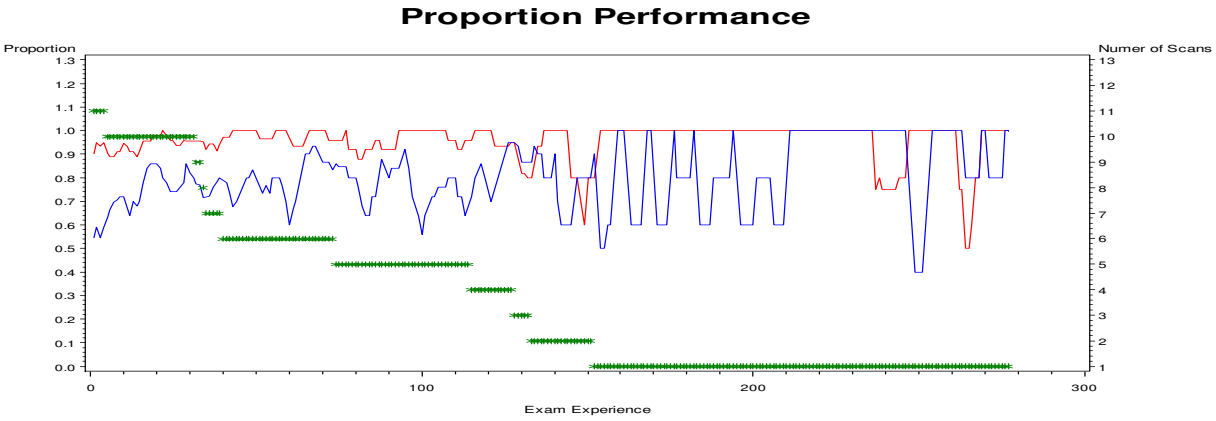
$$v_{aj} = \begin{cases} 1 \; if \; agreement = 1 \; or \; 0 \; for \; exam \; type \; j \\ 0 \; if \; agreement = missing \; for \; exam \; type \; j \end{cases}, \qquad v_{qj} = \begin{cases} 1 \; if \; quality = 1 \; or \; 0 \; for \; exam \; type \; j \\ 0 \; if \; quality = missing \; for \; exam \; type \; j \end{cases}$$

That is, for each performance measure, the number of scans conducted at a given experience level resulting in quality (agreement) were summed and the count then divided by the total number of scans (scans without missing values for agreement, in the case of agreement) conducted at that level.

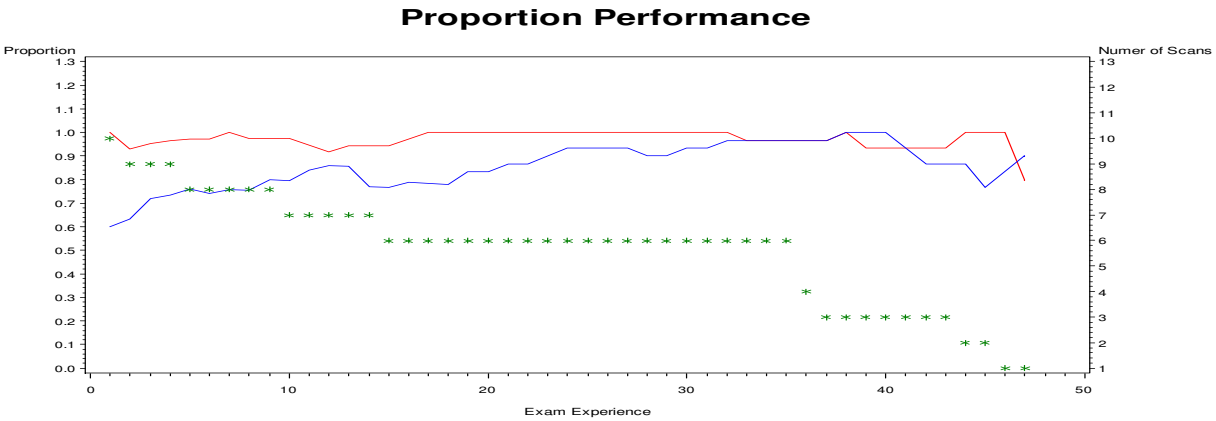
Note that it is common for n_q and n_a to decrease as exam experience increases, as the number of scans performed at a given experience level will naturally decrease. For example, a sonographer may perform eight different exam types ten times each, but only four different exam types fifty times each (i.e. eight scans at experience level ten, but only four scans at experience level fifty).

Next, a five-term backwards moving average was calculated from the above and was plotted to reduce variability, particularly variability due to a decreasing number of observations with increased exam experience.

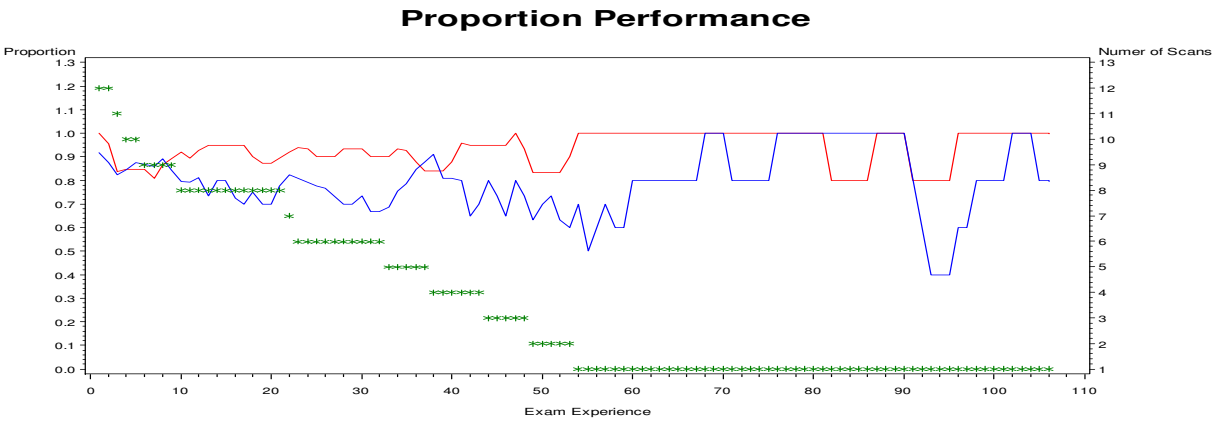
In all three graphs, it can be seen that the proportion of agreement tends to be higher than the proportion of quality. There is some variation in all three graphs, suggesting that a sonographer's performance may not be consistent or improving over time, but may fluctuate. For graphs (a) and (b), a slight upward trend in proportion of quality scans can be seen up to an exam experience of about 30 and 13, respectively. However, for graph (c) it appears that there is a decreasing trend in quality over the first 20 or so scans. The curves for agreement are more stable than those for quality, showing little to no upward or downward trends. It may be the case that when fitting models for quality and agreement, exam experience is not a significant term.



(a)



(b)



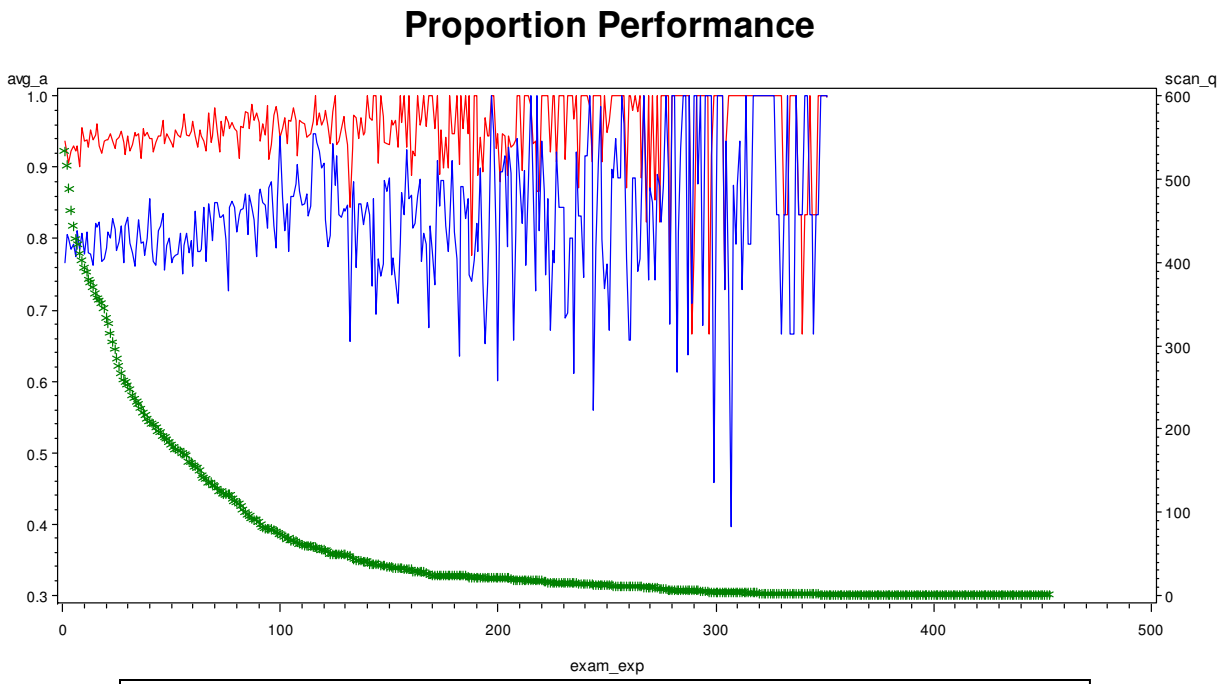
(c)

Key	Blue: Proportion Quality	Left axis: Proportion
	Red: Proportion Agreement	Right axis: Number of scans
	Green: Number of scans	x-axis: exam experience level

Figure 1: Individual sonographer proportion performance plot

4.1.3 Aggregate Sonographer Performance

Since there is such variation between individual sonographers, it is also of interest to explore performance across all sonographers. Figure 2 depicts average sonographer performance across exam experience. The values were calculated as those for figure 1 in section 4.1.2, but summing over all sonographers rather than just one. Once again, a five-term backwards moving average is plotted to reduce variation and smooth the plot.



Key		
	Blue: Proportion Quality	Left axis: Proportion
	Red: Proportion Agreement	Right axis: Number of scans
	Green: Number of scans	x-axis: exam experience level

Figure 2: Aggregate proportion performance plot

Compared to the individual curves, these aggregate curves show slightly less variation (except for where number of observations is very low) and a slight upward trend in quality performance through roughly experience level 120. The first 100 or so scans also seem to show a minor increase in proportion of scans resulting in agreement. Perhaps there is indeed a slight positive relationship between performance and exam experience.

We seek to examine the learning curves of various exam types. Thus, we now consider the breakdown of aggregate sonographer performance by exam type in figures 3 to 14.

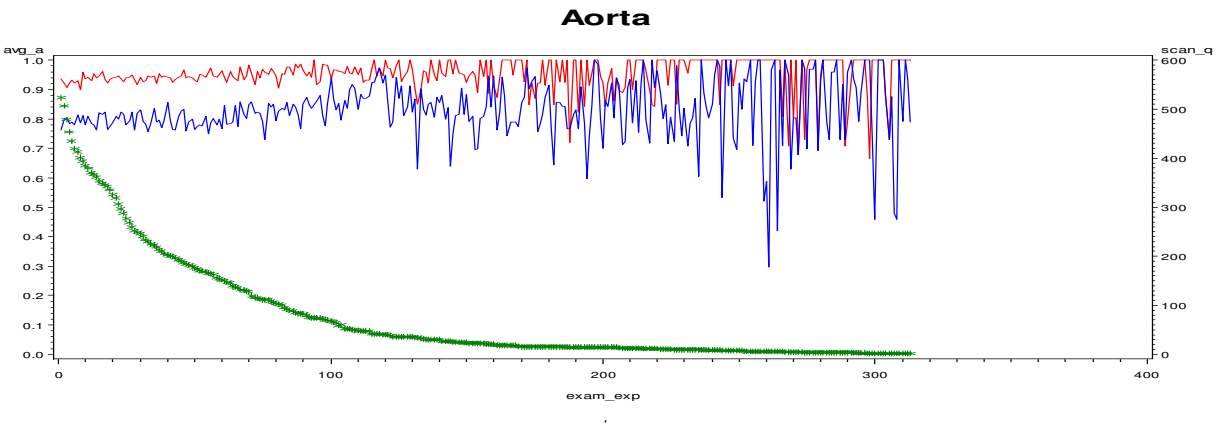


Figure 3: *Aorta proportion performance plot*

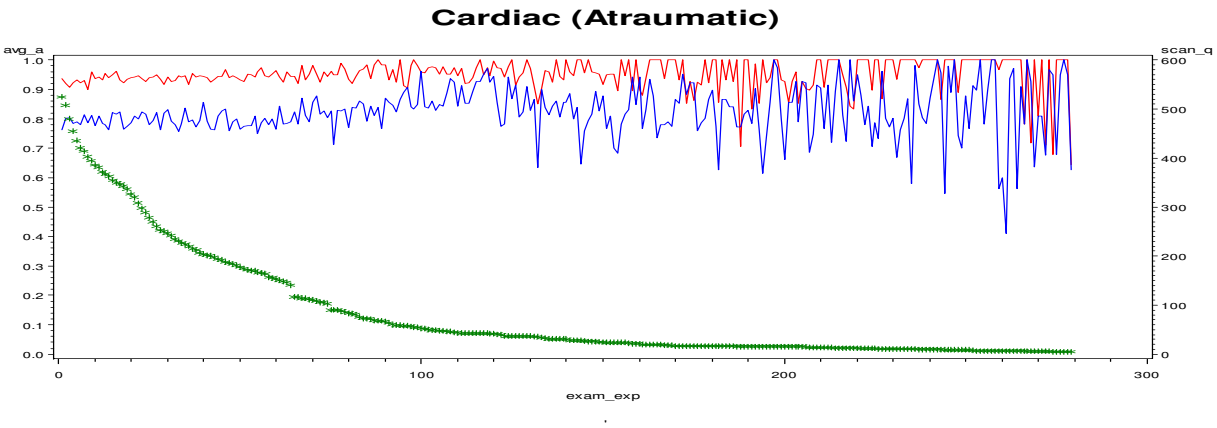


Figure 4: Cardiac (Atraumatic) proportion performance plot

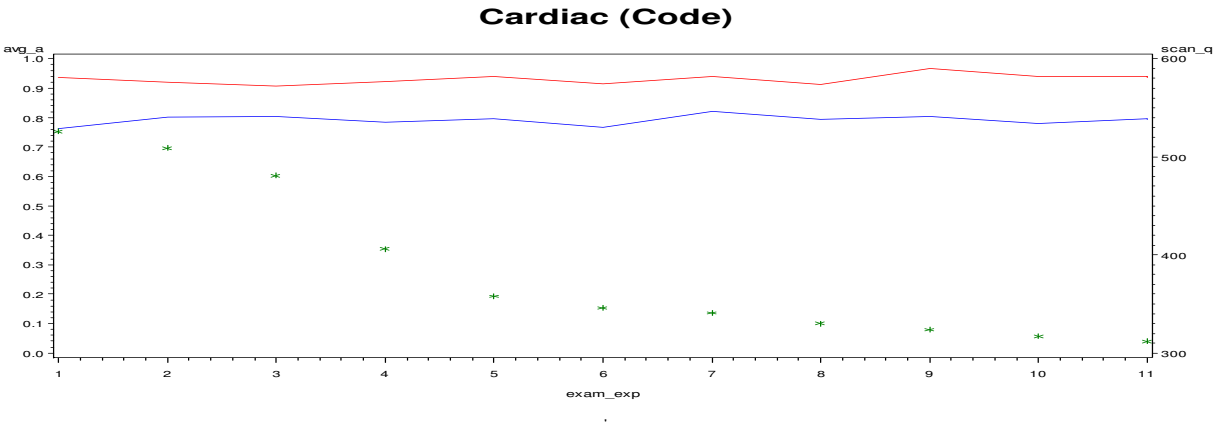


Figure 5: Cardiac (Code) proportion performance plot

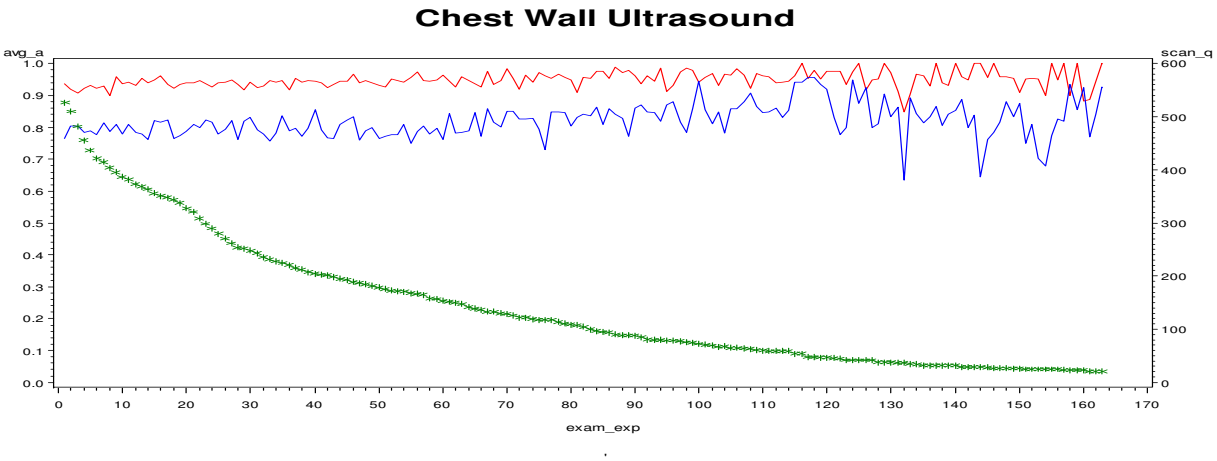


Figure 6: Chest Wall proportion performance plot

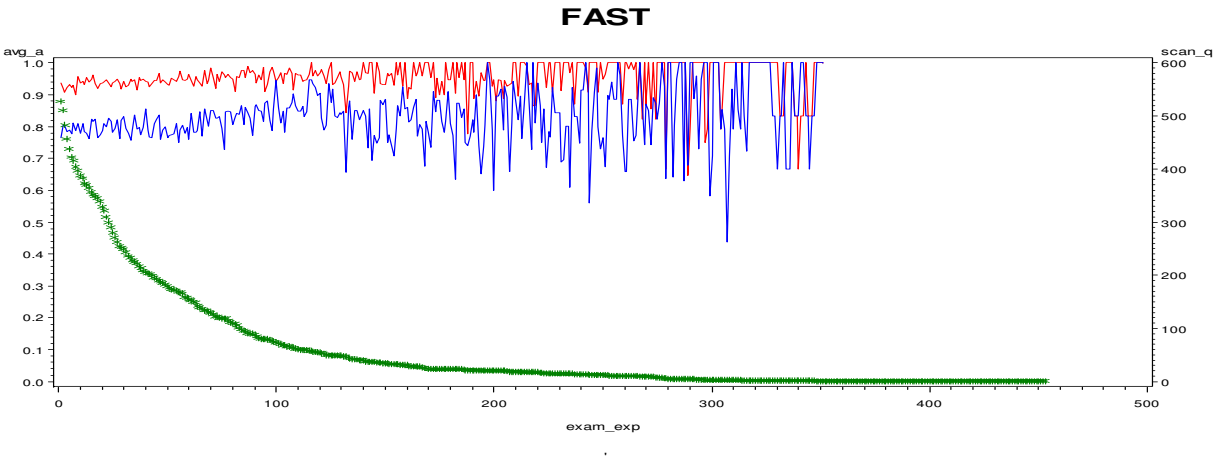


Figure 7: *FAST proportion performance plot*

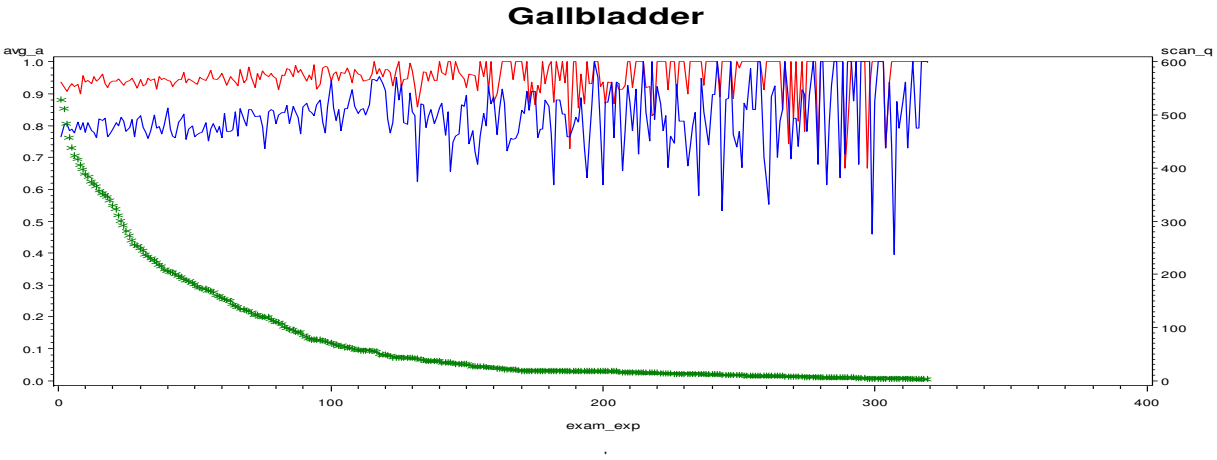


Figure 8: Gallbladder proportion performance plot

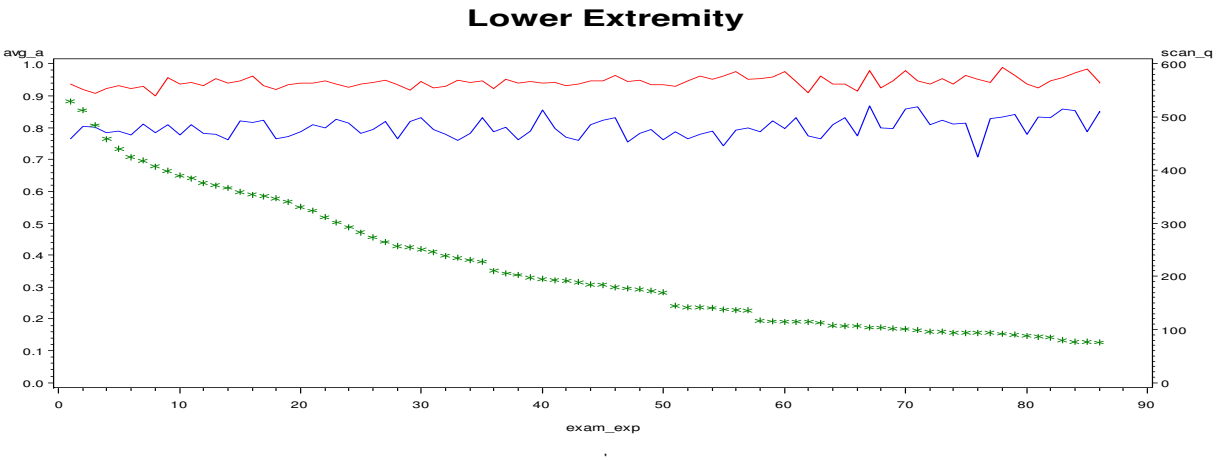


Figure 9: Lower extremity proportion performance plot

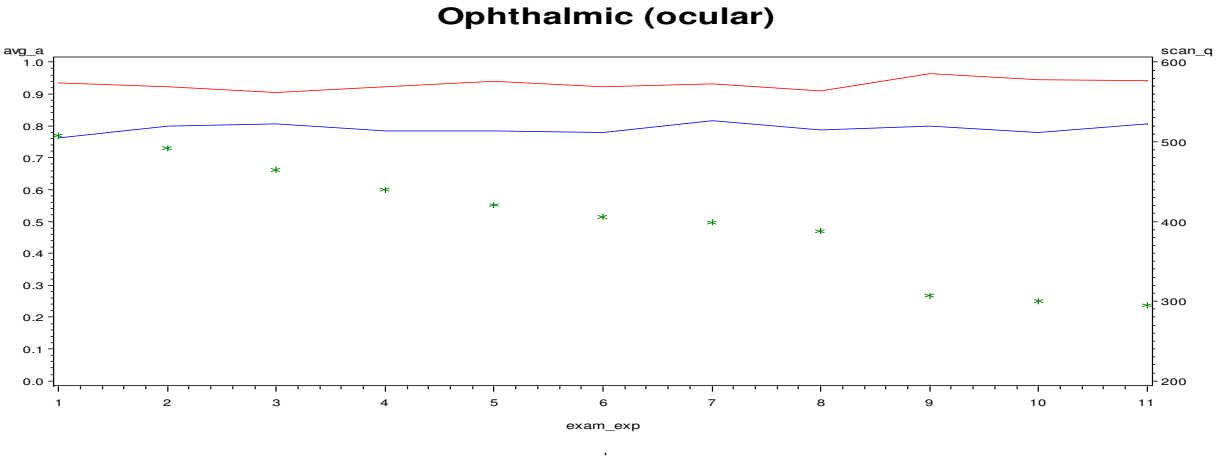


Figure 10: Ophthalmic (ocular) proportion performance plot

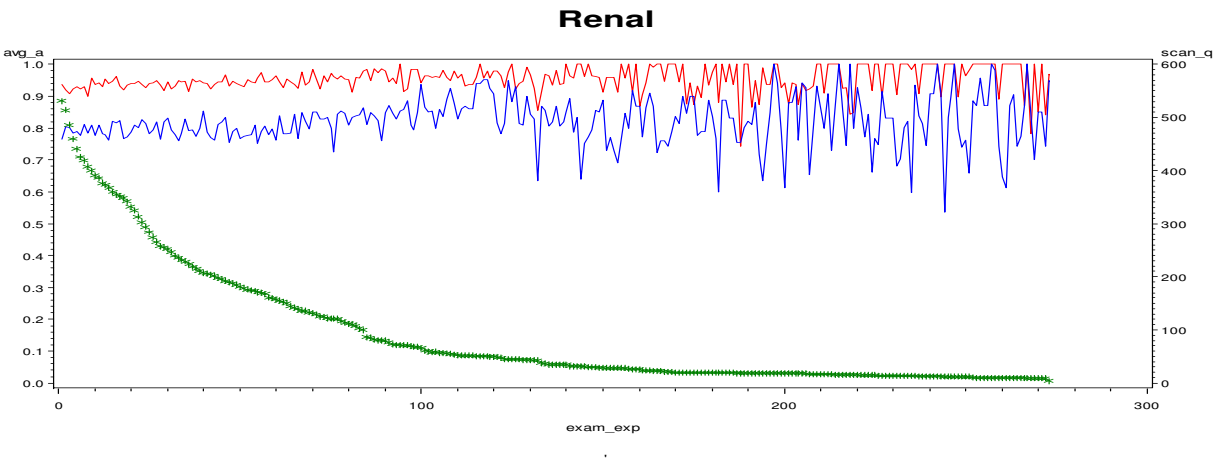


Figure 11: Renal proportion performance plot

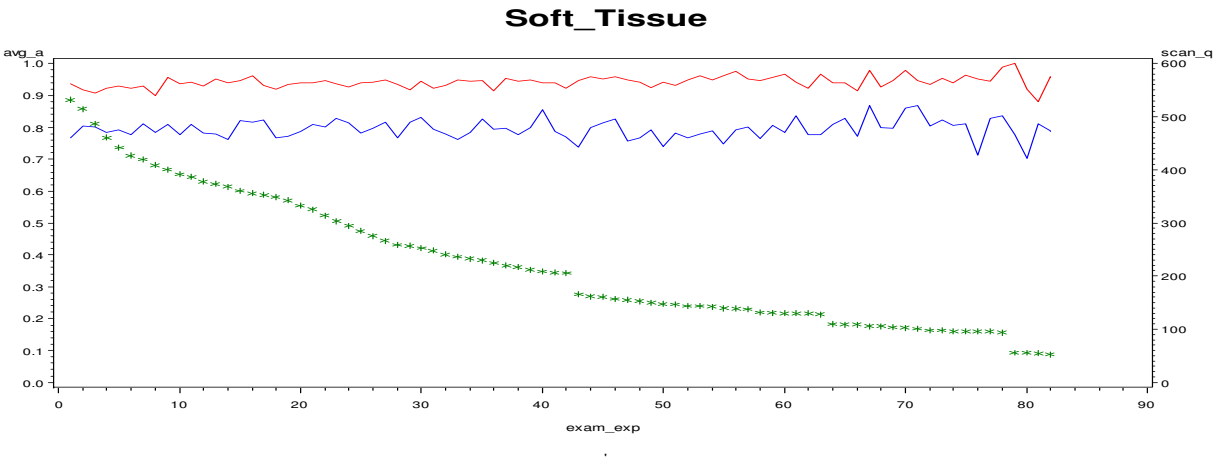


Figure 12: Soft tissue proportion performance plot

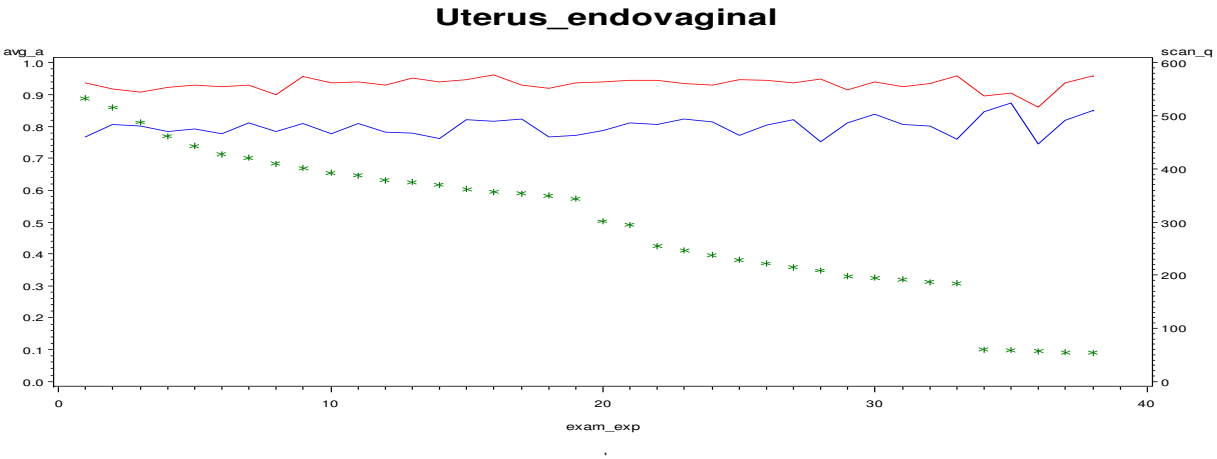


Figure 13: Uterus (endovaginal) proportion performance plot

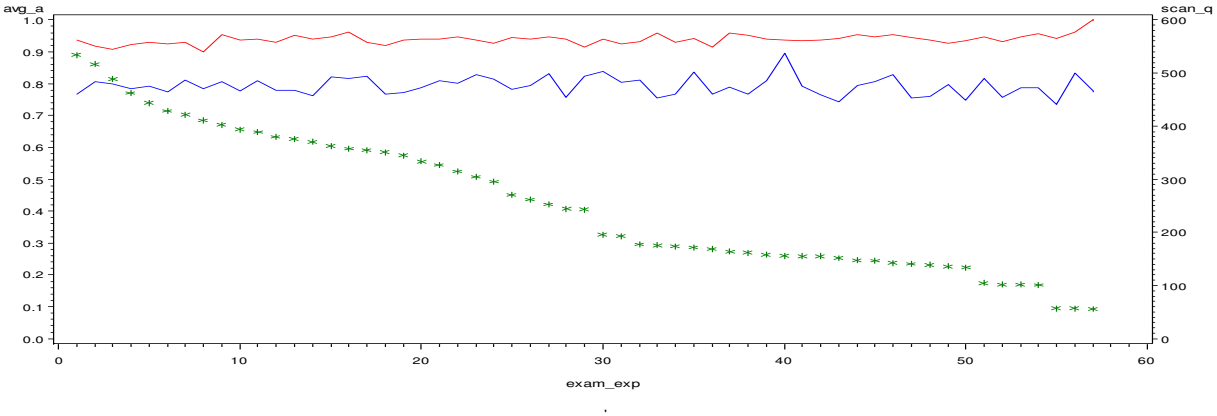


Figure 14: Uterus (trans-abdominal) proportion performance plot

Key		
	Blue: Proportion Quality	Left axis: Proportion
	Red: Proportion Agreement	Right axis: Number of scans
	Green: Number of scans	x-axis: exam experience level

The general trend in figures 3-14 correspond with figure 2 for all exam types considered. For all exam types, proportion agreement remains higher than proportion quality, with minor to no increase over experience. Of interest is how these plots depict the amount of experience gained by sonographers on any given exam type and how much this varies between types. For instance, both cardiac (code) and ophthalmic (ocular) scans are limited to a maximum experience level of 11 scans within the training data set, whereas some other exams have experience levels of up to 300 or even 400 scans. Thus, some of the learning curves to be modeled are likely to be more precise than others due to larger samples.

4.1.4 Potential Confounders

Initial investigation of individual and average sonographer performance suggests potentially little to no association between exam experience and performance on quality and agreement, although it was hypothesized that performance increases with exam experience. This may be due to extraneous variation resulting from a number of confounding variables. Since the original database contains information on the additional variables of reviewer, educational indication, and pathology, the relationship between these variables and performance was also explored.

4.1.4.a Reviewer

Each scan was reviewed by one of five reviewers. It is not unreasonable to propose that different reviewers have different reviewing habits. To explore this possibility, the distribution of quality and agreement outcomes are computed for each reviewer, as shown in table 1.

Reviewer	% Quality	% Agreement	% Missing Agreement
A	81.12%	85.65%	9.85%
B	78.09%	81.13%	12.11%
C	78.78%	88.05%	6.10%
D	91.13%	85.22%	9.40%
E	75.77%	91.43%	4.34%

Table 1: Distribution by reviewer

There is a difference among the reviewers in the percentage of quality scans (a range of 15%) and agreement of readings (a range of 10%). In addition, there is a difference between reviewers in whether or not a scan received an initial or final reading and was consequently included in analysis (three times as many scans are excluded from analysis when the reviewer is ‘B’ than when the reviewer is ‘E’). It is possible that reviewer is a confounding factor in a sonographer’s performance and thus should be considered for inclusion in the final models.

4.1.4.b Educational Indication

The reason for performing each scan is recorded, one of these potential reasons being ‘educational – unbillable’. This type of educational scan is one which is done purely for educational purposes, meaning the patient did not present with any symptoms. It is of interest to determine whether there is a difference between whether a scan results in quality or agreement based on if it was performed for purely educational purposes or for medical purposes with a patient presenting symptoms. The distribution of quality and agreement by indication (educational or medical) is displayed in table 2.

Indication	% Quality	% Agreement	% Missing Agreement
Educational – unbillable	81.14%	90.41%	6.56%
Medical	80.73%	83.90%	10.00%

Table 2: Distribution by educational indication

There does not seem to be a difference between the percentage of educational and medical scans resulting in quality. However, 6% more of the educational scans resulted in agreement than those conducted for medical purposes. Educational indication may be a confounding factor (particularly for agreement) and should be considered for inclusion in the final models.

4.1.4.c Pathology

As part of the initial and final readings of a scan, it is recorded whether a positive finding for an ailment/medical condition/etc. was found. This is known as pathology, thus a pathology (+) scan is one in which either the sonographer or reviewer yield a positive finding; a pathology (-) scan will be referred to as one in which neither the sonographer nor the reviewer yielded a positive finding. It is of interest to check whether the distribution of quality and agreement varies by pathology (see table 3).

Initial/Final Reading	% Quality	% Agreement	% Missing Agreement
Pathology (+)	84.88%	67.30%	6.16%
Pathology (-)	80.12%	89.44%	9.41%

Table 3: Distribution by pathology

While there appears to be only a small difference in the percentage of quality or missing agreement scans for pathology (+) and pathology (-) scans (4% and 3%, respectively), there is a larger difference between the two on agreement. About 22% more of the pathology (-) scans resulted in agreement than the pathology (+) scans. Pathology should be considered for inclusion in the models, and in particular for the agreement model.

4.2 Model Selection

The variables listed in table 4 were considered for inclusion in both models.

Numeric	Categorical	Interaction
exam experience	exam type	exam type*exam experience
	reviewer	reviewer*exam experience
	educational indication	educational indication*exam experience
	pathology	pathology*exam experience

Table 4: Model selection

Model selection was performed on the training data set using backwards elimination and a Wald-type test utilizing empirically estimated standard errors with a threshold value of 0.05. A qualitative variable with more than two levels was considered for removal only if all of its levels were non-significant. In addition to this, the quasi-likelihood information criterion (QIC) and the QIC_U (an estimate of the QIC) for each model fit were compared along with confidence intervals of the area under the receiver operating curve (AUC). Since the objective was to model the learning curves of sonographers over exam

experience, exam experience was included in the final model by default. Tables 5 and 6 summarize each step of the elimination process for the quality and agreement models, respectively.

Parameter identified for removal (p-value)	QIC	QIC_U	95% lower AUC	95% upper AUC
educational indication*exam experience (0.0886)	24527.9799	24468.6518	0.6633	0.6794
N/A	24530.6127	24468.9161	0.6633	0.6794

Table 5: Quality model selection

Parameter identified for removal (p-value)	QIC	QIC_U	95% lower AUC	95% upper AUC
educational indication (0.3578)	7149.5977	7140.4798	0.8666	0.8888
educational indication*exam experience (0.5347)	7148.3859	7139.3654	0.8674	0.8894
exam type*exam experience (>0.10)	7147.2155	7137.6514	0.8671	0.8892
reviewer*exam experience (>0.15)	7138.1787	7126.6194	0.8659	0.8882
N/A	7136.3698	7123.7680	0.8652	0.8876

Table 6: Agreement model selection

Based on the results shown in tables 5 and 6, the final forms of both models were chosen to include the variables listed in table 7.

Quality Variables	Agreement Variables
exam experience	exam experience
exam type	exam type
reviewer	reviewer
educational indication	pathology
pathology	
exam type*exam experience	
reviewer*exam experience	
pathology*exam experience	

Table 7: Selected model forms

4.3 Model Validation

Both the quality and agreement models were validated by fitting the model forms selected in table 7 to the testing data set. The QIC, QIC_U, and estimated AUC were checked, residuals plotted, and parameter estimates tested and compared to those obtained on the training data set.

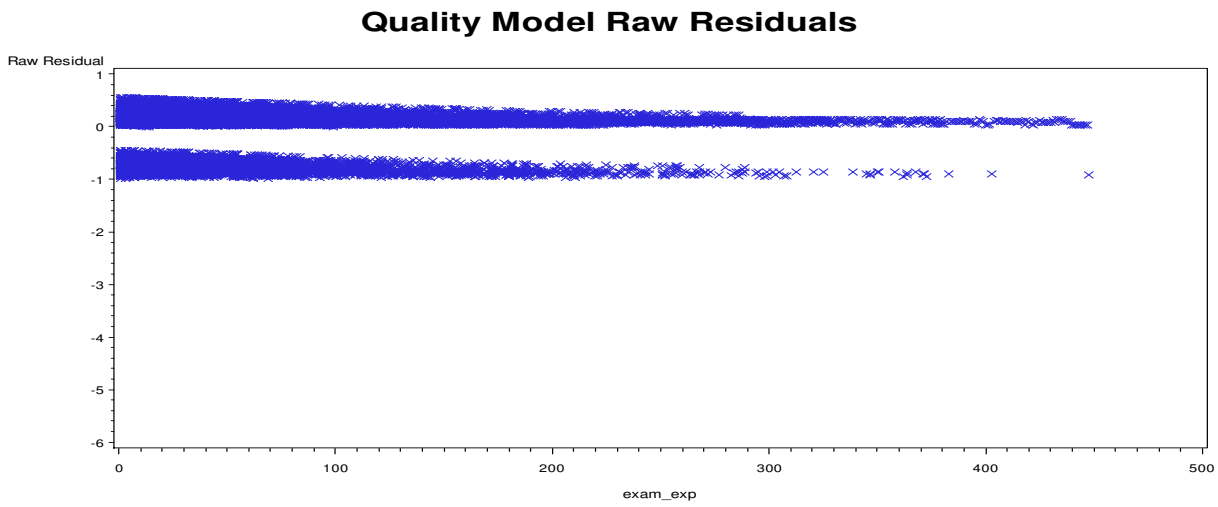
4.3.1 Quality

The first three quantitative checks of model fit returned favorable results, particularly in relation to being consistent with those obtained from the training set: QIC = 26461.2506, QIC_U = 26377.3152, and a 95% confidence interval for AUC: (0.6766, 0.6920).

Plots of the raw and Pearson residuals (see figures 15 and 16) suggest that the model tends to over predict scans without quality, and have more variation and larger residuals for lower levels of experience.

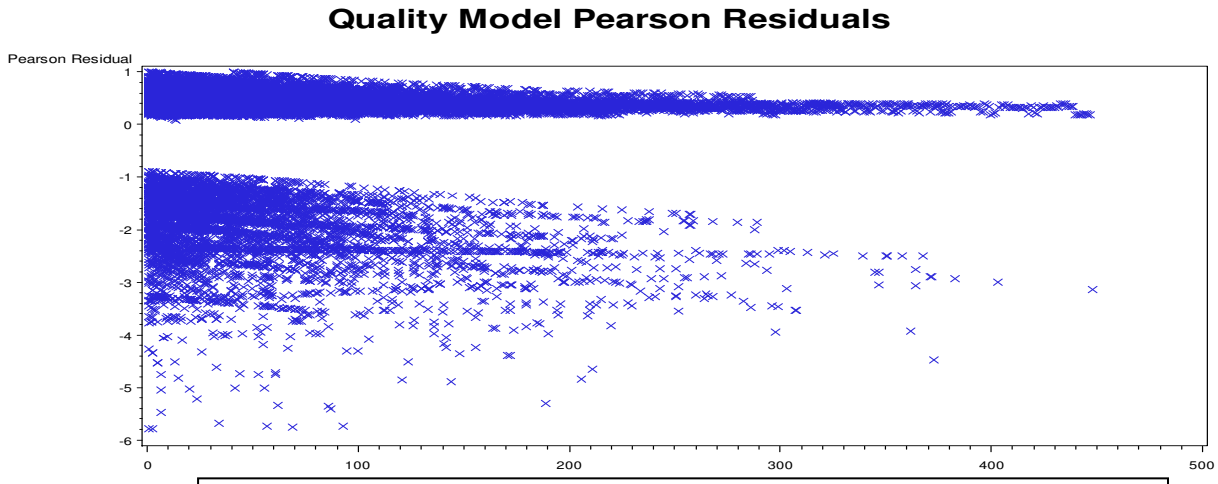
The raw (r_i) and Pearson (r_{pi}) residuals plotted were calculated using the following equations:

$$r_i = y_i - \mu_i \text{ and } r_{pi} = (y_i - \mu_i)\sqrt{\frac{1}{V(\mu_i)}}$$



Key	Vertical axis: raw residual Horizontal axis: exam experience

Figure 15: Raw residuals for quality model



Key	Vertical axis: Pearson residual Horizontal axis: exam experience

Figure 16: Pearson residuals for quality model

Using a Wald test for the individual parameters, only the interaction term between pathology and exam experience was no longer significant (p-value = 0.3901), although it was still positive. (Once again, a

qualitative variable with more than two levels was considered significant if at least one of its levels was significant.) Also different from the fit on the training set is that the coefficient for exam experience changed from significantly negative on the training set ($\hat{\beta} = -0.0150, p = 0.0136$) to significantly positive on the testing set ($\hat{\beta} = 0.0139, p = 0.0093$). To compensate for this, the number of interaction terms for exam type and exam experience that were negative increased from 1 to 9. Thus, while the information criteria and AUC of the final model suggest a fair fit, the actual parameter estimates are cause for concern.

This discovery raised the question of whether the model is so highly influenced by individual sonographers that the model fit is strongly dependent upon the data split or if the data split obtained here happened to be an unrepresentative one. To check this, we randomly split the full data set ten more times using the same method as previously to obtain twenty new data sets. The final model form was fit to all twenty sets and the parameter estimates checked against the three discrepancies which were observed for the original training and testing sets. Table 8 summarizes the outcomes for these new data sets.

Model Term	**# (-) and Significant**	**# (+) and Significant**	**# (-) and insignificant**	**# (+) and insignificant**
exam experience	0	14	1	5
exam experience*pathology	0	8	1	11
exam experience*exam type Distribution of # (-), regardless of significance	**1** 0	**3** 1	**6** 1	**7** 1	**8** 7	**9** 9	**10** 1

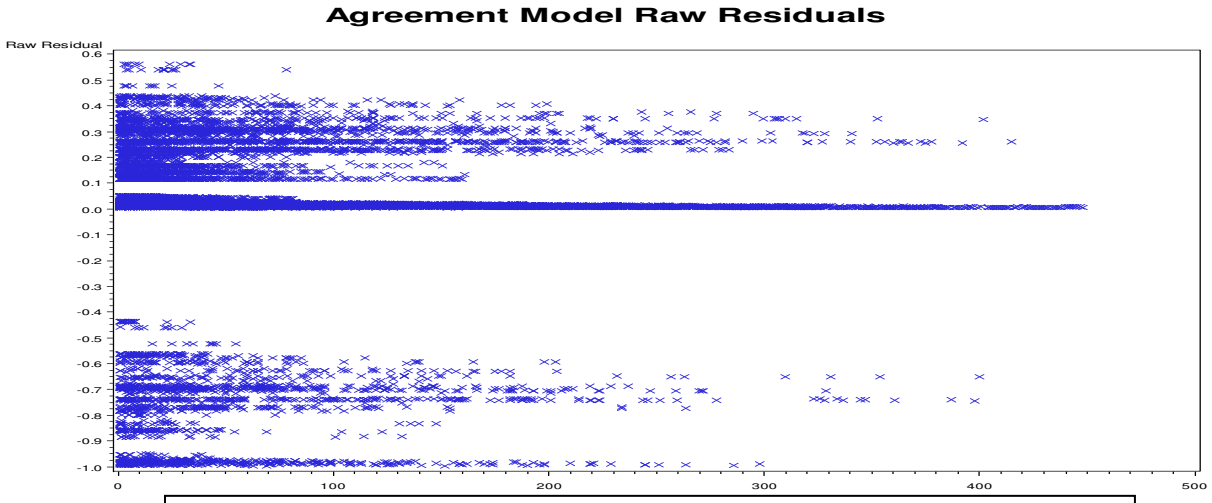
Table 8: Quality: counts of outcomes for 20 new data sets

We see that out of the twenty new data sets none replicated the significantly negative coefficient for exam experience as was seen in the training set, whereas the testing set result of a significantly positive coefficient was replicated 14 times. The training set outcome of a significantly positive coefficient for pathology and exam experience was replicated 8 times and the testing set outcome of an insignificant coefficient was replicated 11 times. Lastly, noting that negative coefficients for the interaction term of exam type and exam experience tended to correspond with a positive coefficient for exam experience and visa versa, the most frequent number of negative coefficients for the interaction term was 9 (replicated 9 times), supporting the finding from the testing set. This suggests that while the interaction between pathology and exam experience may be highly dependent upon the data split, the observed training set results for exam experience and exam experience interacted with exam type may have been a remote occurrence. Thus, the replications of data splitting and model fitting suggest that the testing data set is representative of the entire data set and strengthens the validity of the final model.

4.3.2 Agreement

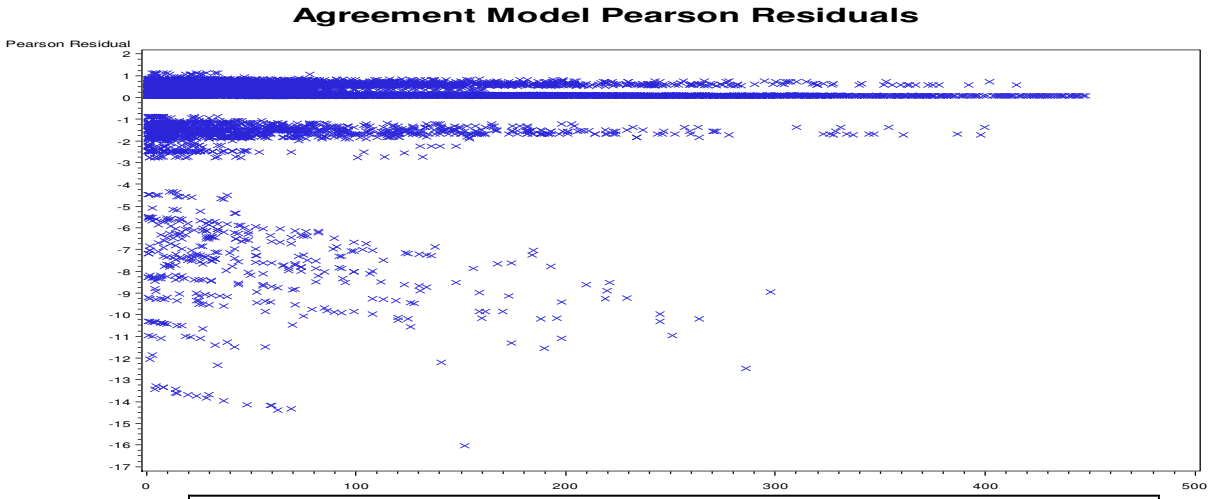
The first three quantitative checks of model fit returned favorable results: QIC = 8803.2173, QIC_U = 8794.1512, and a 95% confidence interval for AUC: (0.8652, 0.8876).

A plot of the raw residuals (see figure 17) suggests that the model may be less accurate in predicting agreement for earlier exam experience levels. A plot of the Pearson residuals (see figure 18) shows a large cluster of scans with high residuals (relative to the others, with $r_p < -4$), corresponding to scans resulting in disagreement between the sonographer and reviewer. Upon investigation of this cluster, it was revealed that all of these scans were negative for pathology. Thus, this model fit may be questionable for scans not showing pathology.



Key Vertical axis: raw residual
Horizontal axis: exam experience

Figure 17: Raw residuals for agreement model



Key Vertical axis: Pearson residual
Horizontal axis: exam experience

Figure 18: Pearson residuals for agreement model

Using a Wald test for the individual parameters, two parameters showed a discrepancy between the training data set fit and the testing data set fit. The term for exam type 'soft tissue' was no longer significant (p-value = 0.16308) and the interaction term between pathology and exam experience was no longer significant (p-value = 0.3901). For both model fits, exam experience was insignificant (while negative for the training set and positive for the testing set). As in the case of the quality model, to explore the discrepancies, the agreement model was fit to the twenty random data splits, as summarized in table 9.

Model Term	# (-) and Significant	# (+) and Significant	# (-) and insignificant	# (+) and insignificant
exam experience	1	0	10	9
soft tissue	14	0	6	0
exam experience*pathology	0	10	0	10

Table 9: Agreement: counts of outcomes for 20 new data sets

We see that out of the twenty new data sets, ten replicated the exam experience result of the training set, while nine replicated that of the testing set, where all nineteen reached the conclusion that the exam experience term was insignificant. Only one data set's model fit returned a significantly negative result for exam experience. For the soft tissue exam parameter, fourteen data sets replicated the result of the training set, while six replicated that of the testing set. Lastly, for the interaction term of exam experience and pathology, half of the data sets replicated the training result and half replicated the testing result. This suggests that while our model may be valid in its conclusion of an insignificant relationship between exam experience and agreement, as determined by a Wald test for the model parameter, it may be susceptible to data splits for other parameter estimates, particularly those for the soft tissue exam and the interaction between exam experience and pathology.

4.4 Model Interpretation

Following the validation of the two models, the parameter estimates were examined to understand the effects of the various factors.

4.4.1 Quality

Table 10 displays the each parameter's estimate, standard error, lower and upper 95% Wald confidence interval bounds, test statistic, and p-value for a Wald test of significance.

Parameter	Estimate	Stderr	LowerCL	UpperCL	Z	ProbZ
Intercept	0.636325	0.160127	0.322481	0.950169	3.973866	7.07E-05
Exam experience	0.013871	0.005332	0.00342	0.024322	2.601387	0.009285
Reviewer A	0.588747	0.080366	0.431231	0.746262	7.325779	2.38E-13
Reviewer B	0.40919	0.09123	0.230383	0.587997	4.485281	7.28E-06
Reviewer C	0.224477	0.088395	0.051227	0.397727	2.539484	0.011102
Reviewer D	1.310132	0.115187	1.084369	1.535895	11.37392	0
Reviewer E	0	0	0	0		
Aorta	0.279593	0.176241	-0.06583	0.625018	1.586425	0.112643
Cardiac (Atraumatic)	-0.21296	0.174007	-0.55401	0.128087	-1.22387	0.221003
Cardiac (Code)	0.72082	0.427454	-0.11698	1.558615	1.686307	0.091737
Chest Wall Ultrasound	1.791897	0.210746	1.378842	2.204952	8.502624	0
FAST	1.132802	0.183617	0.772918	1.492685	6.16936	6.86E-10
Gallbladder	0.439911	0.177549	0.091922	0.787899	2.477693	0.013224
Lower Extremity	0.435591	0.227104	-0.00952	0.880707	1.918024	0.055108
Ophthalmic (ocular)	-0.0534	0.46789	-0.97045	0.863649	-0.11413	0.909136
Renal	1.065037	0.184647	0.703136	1.426938	5.767968	8.02E-09
Soft Tissue	1.165881	0.207792	0.758616	1.573147	5.610799	2.01E-08
Uterus (endovaginal)	0.351101	0.24484	-0.12878	0.83098	1.434001	0.151572
Uterus (trans-abdominal)	0	0	0	0		
pathology (-)	-0.42815	0.084776	-0.5943	-0.26199	-5.05031	4.41E-07
pathology (+)	0	0	0	0		
educational (1)	-0.22805	0.057993	-0.34172	-0.11439	-3.93235	8.41E-05
educational (0)	0	0	0	0		
Reviewer A*exam exp	-0.00175	0.000647	-0.00301	-0.00048	-2.6996	0.006942
Reviewer B*exam exp	0.001239	0.001375	-0.00146	0.003934	0.901224	0.367469
Reviewer C*exam exp	0.001754	0.001024	-0.00025	0.003761	1.712295	0.086842
Reviewer D*exam exp	-0.00324	0.00102	-0.00524	-0.00124	-3.17466	0.0015
Reviewer E*exam exp	0	0	0	0		
Aorta*exam exp	-0.01153	0.005469	-0.02225	-0.00081	-2.1078	0.035048
Cardiac (Atraumatic)*exam exp	-0.00891	0.005574	-0.01983	0.00202	-1.59764	0.110123
Cardiac (Code)*exam exp	0.187515	0.103695	-0.01572	0.390754	1.808338	0.070554
Chest Wall Ultrasound*exam exp	-0.01079	0.005786	-0.02213	0.00055	-1.86484	0.062204
FAST*exam exp	-0.01255	0.005365	-0.02307	-0.00204	-2.3396	0.019305
Gallbladder*exam exp	-0.00851	0.005585	-0.01945	0.00244	-1.52307	0.127741
Lower Extremity*exam exp	-0.00275	0.008525	-0.01945	0.013963	-0.32204	0.747421
Ophthalmic (ocular)*exam exp	-0.00191	0.134516	-0.26555	0.261737	-0.01419	0.988678
Renal*exam exp	-0.01063	0.005541	-0.02149	0.000226	-1.91909	0.054972
Soft Tissue*exam exp	0.006819	0.008959	-0.01074	0.024379	0.761176	0.446552
Uterus (endovaginal)*exam exp	-0.00251	0.012585	-0.02717	0.022162	-0.19906	0.842217
Uterus (trans-abdominal)*exam exp	0	0	0	0		
pathology (-)*exam exp	0.000775	0.000902	-0.00099	0.002544	0.85941	0.390114
pathology (+)*exam exp	0	0	0	0		

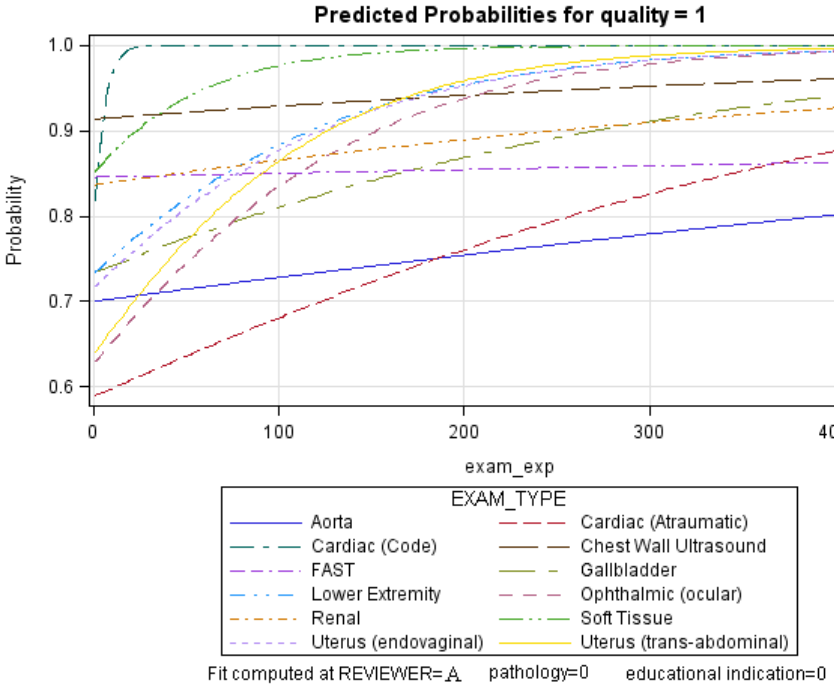
Table 10: Parameter estimates for the quality model

We now provide interpretations for a subset of these parameter estimates to gather some insight into the behavior of the model. Each estimate's interpretation is under the condition that all other factors in the model are held fixed. For interaction terms, interpretations are made with regards to a change in experience level, under the condition that the qualitative factor in the interaction term is fixed at the specified level for that term.

- For each additional pathology (+) gallbladder scan reviewed by Reviewer E, the odds of quality increase multiplicatively by $e^{0.013871-0.00851} = 1.005$. That is, they increase by 0.5%.
- The odds of obtaining quality on a medical, pathology (+) FAST exam (for a fixed level of exam experience) reviewed by reviewer E are $e^{1.132802} = 3.104$ times those for a medical, pathology (+) uterus (trans-abdominal) exam (at the same fixed level of exam experience) also reviewed by reviewer E. The odds are 310.4% higher.
- The odds of obtaining quality on an educational, pathology (+) uterus (trans-abdominal) exam (at a fixed level of exam experience) reviewed by reviewer E are $e^{-0.22805} = 0.7961$ times those for a medical, pathology (+) uterus (trans-abdominal) exam (at the same fixed level of exam experience) also reviewed by reviewer E. The odds are 20.39% lower.

The other parameter estimates may be interpreted in a similar manner. It is important to note that for levels of the qualitative effects with large (>0.05) p-values, interpretations should be made stating that there is not enough evidence to suggest a difference between the odds of quality for that effect level and the reference level of the effect (the reference level is identified as that which corresponds to all estimates being either 0 or missing).

Figure 19 provides a partial graphical representation of the model. Of the 240 learning curves, 12 are plotted, each curve corresponding to a different exam type, with exam experience along the horizontal axis and predicted probability of obtaining quality on the vertical axis. All other covariates have been fixed at the levels for which they are most represented: reviewer = Reviewer A, educational indication = 0 (medical), and pathology = 0 (pathology (-)). All of the quality learning curves are in Appendix D.



Key	Vertical axis: predicted probability Horizontal axis: exam experience

4.4.2 Agreement

Table 11 displays the each parameter's estimate, standard error, lower and upper 95% Wald confidence interval bounds, test statistic, and p-value for a Wald test of significance.

Parameter	Estimate	Stderr	LowerCL	UpperCL	Z	ProbZ
Intercept	2.002624	0.145034	1.718363	2.286885	13.80799	0
Exam experience	6.46E-05	0.000673	-0.00125	0.001383	0.096045	0.923485
Reviewer A	0.024335	0.102467	-0.1765	0.225166	0.237492	0.812275
Reviewer B	-0.49709	0.12182	-0.73585	-0.25833	-4.08054	4.49E-05
Reviewer C	-0.41228	0.099084	-0.60648	-0.21808	-4.16096	3.17E-05
Reviewer D	-0.15606	0.104068	-0.36003	0.047908	-1.49962	0.133714
Reviewer E	0	0	0	0		
Aorta	-1.10562	0.163006	-1.42511	-0.78613	-6.78271	1.18E-11
Cardiac (Atraumatic)	-1.22451	0.164167	-1.54627	-0.90275	-7.45893	8.73E-14
Cardiac (Code)	-0.21215	0.353651	-0.90529	0.480996	-0.59988	0.548588
Chest Wall Ultrasound	-0.74798	0.224415	-1.18783	-0.30813	-3.33302	0.000859
FAST	-0.99271	0.165281	-1.31665	-0.66876	-6.00617	1.9E-09
Gallbladder	-0.81996	0.14959	-1.11315	-0.52677	-5.48136	4.22E-08
Lower Extremity	-1.75704	0.195975	-2.14114	-1.37293	-8.96562	0
Ophthalmic (ocular)	-1.243	0.461968	-2.14844	-0.33755	-2.69065	0.007131
Renal	-1.20293	0.142162	-1.48156	-0.9243	-8.46168	0
Soft Tissue	-0.22038	0.157999	-0.53005	0.089297	-1.39479	0.16308
Uterus (endovaginal)	-0.96377	0.168772	-1.29456	-0.63299	-5.71051	1.13E-08
Uterus (trans-abdominal)	0	0	0	0		
pathology (-)	3.157853	0.121973	2.918791	3.396915	25.88983	0
pathology (+)	0	0	0	0		
pathology (-)*exam exp	0.002319	0.001369	-0.00036	0.005002	1.693793	0.090305
pathology (+)*exam exp	0	0	0	0		

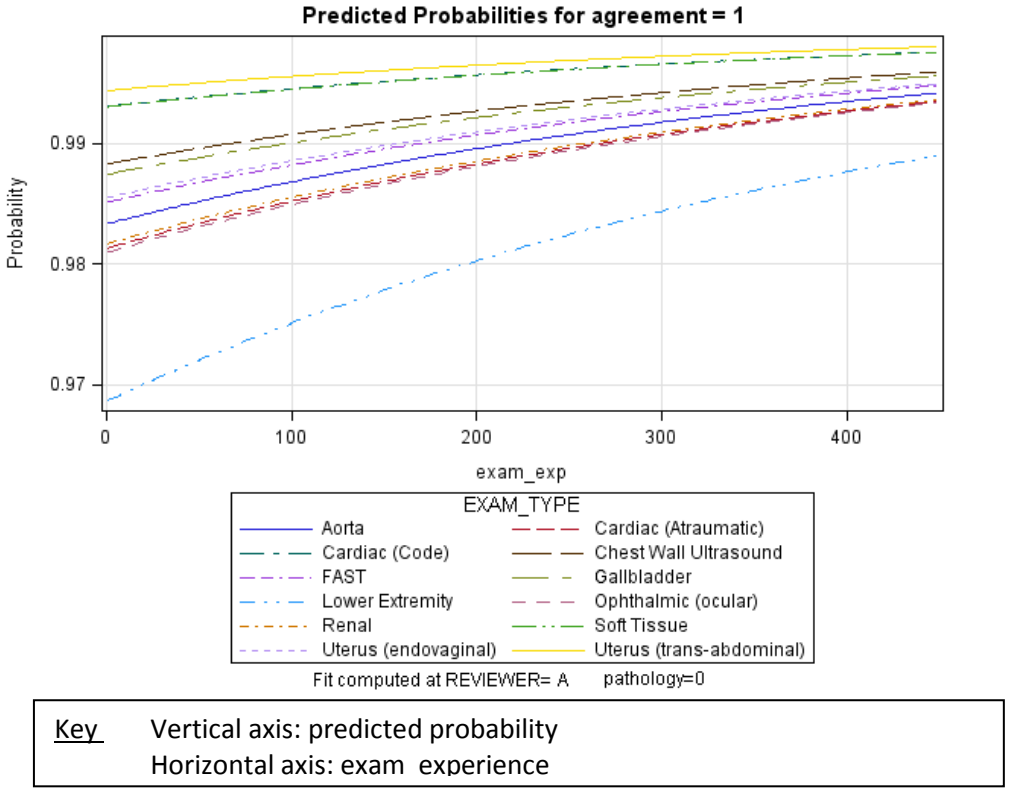
Table 11: Parameter estimates for the agreement model

We now provide interpretations for a subset of these parameter estimates. Each estimate's interpretation is under the condition that all other factors in the model are held fixed. For interaction terms, interpretations are made with regards to a change in experience level, under the condition that the qualitative factor in the interaction term is fixed at the specified level for that term.

- There is insufficient evidence to suggest that for each additional scan of a particular exam type, the odds of agreement change.
- The odds of obtaining agreement on a pathology (+) chest wall scan reviewed by reviewer E are $e^{-0.74798} = 0.4733$ times those for a pathology (+) uterus (trans-abdominal) scan reviewed by reviewer E. That is, they are 52.67% lower.
- The odds of obtaining agreement on a pathology (+) uterus (trans-abdominal) exam reviewed by reviewer C, for a fixed level of experience, are $e^{-0.41228} = 0.6621$ times those for a pathology (+) uterus (trans-abdominal) exam reviewed by reviewer E. The odds are 33.79% lower.

The other parameter estimates may be interpreted in a similar manner.

Figure 20 provides a partial graphical representation of the model. Of the 120 learning curves, 12 are plotted, each curve corresponding to a different exam type, with exam experience along the horizontal axis and predicted probability of obtaining agreement on the vertical axis. The remaining covariates have been fixed at the levels for which they are most represented: reviewer = Reviewer A and pathology = 0 (pathology (-)).



Key	Vertical axis: predicted probability Horizontal axis: exam experience

Figure 20: Sample agreement learning curves

4.5 Research Question 1

Recall that the goal of research question 1 was to identify how many scans of a particular exam type are required for a sonographer to obtain quality (agreement) with a predicted probability of 0.9. Since it was determined that exam experience is not a significant predictor of agreement, there is not a certain number of scans required to obtain agreement. Thus, we only considered the number of scans to obtain quality with a predicted probability of 0.9.

Due to the nature of the data and the significant confounding factors, there is no single estimate for the number of scans required. Instead, there are a total of 240 (12 exams * 5 reviewers * 2 levels of indication * 2 levels of pathology) estimates, each pertaining to a different combination of the factor levels. This resulted in a total of 20 estimates and confidence intervals for each exam type.

As we considered each exam type individually, it was reasonable to treat each exam type's set of 20 curves as a single unit, distinct from the other exam types' curves. Thus we grouped estimates corresponding to a particular exam type. Each exam type's set of 20 confidence intervals were constructed at a family-wise 95% confidence level. That is, using the Bonferroni adjustment, each of the

20 intervals was individually constructed at a 99.75% confidence level. The resulting point estimates and confidence intervals for each exam type are located in Appendix C.

Table 12 provides a broad summary of each exam type's twenty estimates for the quality model. This table is meant purely for summary purposes and not for drawing statistical conclusions. To obtain each row, first the 20 confidence intervals for the specified exam type were sorted by ascending order of their respective point estimates. Next, the lower bound of the first ordered confidence interval and the upper bound of the last ordered confidence interval were noted, respectively, in the 'Lowest' and 'Highest' table columns. This gives an overall summary of the range of the 20 confidence intervals for that exam type. Next, the 25^{th} and 75^{th} percentiles, according to point estimate, were identified and their lower and upper confidence bounds noted, respectively, in the 25^{th} and 75^{th} percentile columns. This provides a summary of the range of the middle 50% of the intervals, according to point estimate.

It is observed for quality that there is much variation, both among confidence intervals for a single exam type and across exam types. Some exam types appear to have less variation than others. For example, the ranges for soft tissue, cardiac (code), and uterus (trans-abdominal) scans are much smaller than those for renal, chest wall, and aorta scans.

Wider confidence intervals are less useful, as they provide both less precision as well as impractical numbers (e.g. it is not reasonable to require a sonographer to perform $8.12*10^9$ scans of a particular exam type in order to be considered proficient). These wider intervals may be due to flatter learning curves as a consequence of the inverse interval process (obtaining a confidence interval around a point on the y-axis for a nearly horizontal curve results in selecting points on the x-axis which are farther apart).

Exam	Lowest	Highest	25th percentile	75th percentile
Aorta	0	9,853.00	0	314
Cardiac (Atraumatic)	0	1279	123	1279
Cardiac (Code)	0	18	0	10
Chest Wall	0	359,089	0	121
FAST	0	7,687	0	1146
Gallbladder	0	403	54	311
Lower Extremity	0	267	5	183
Ophthalmic (Ocular)	0	5,108	0	3,479
Renal	0	8.13*10^9	0	246
Soft Tissue	0	83	0	42
Uterus (endovaginal)	0	350	37	247
Uterus (trans-abdominal)	0	231	40	172

Table 12: Summary of quality inverse intervals

4.6 Research Question 2

The objective of research question 2 was to determine whether non exam experience, that is, experience with exam types other than the one being conducted, contributes to the learning curve of a

particular exam type, given that the learning curve already takes into account exam experience for that exam type and any significant confounders as identified in the model selection process. To answer this question a type III analysis was performed to test for the significance of adding the non exam experience variable to the models selected in section 4.2.

The model for quality was not significantly improved by the addition of the non exam experience term ($\chi^2 = 2.76, p = 0.0968$), whereas the model for agreement was significantly improved ($\chi^2 = 4.76, p = 0.0291$). Thus, non exam experience impacts the learning curve for agreement, but not for quality.

5 Discussion

This study provides insight into the discrepancies found between other research studies, discussions, and guidelines (ACEP, 2008; ACR, 2011; Gaspari et al., 2009; Hertzberg et al., 2000; Jang et al., 2010), which identify a wide range of experience levels for a sonographer to be deemed qualified. Our results show that sonographer performance is dependent not just upon experience, but also on a number of other factors, including exam type, reviewer, educational indication, and pathology.

Past research suggests that performance learning curves are unique to different sonographers (Hertzberg et al., 2000). The exploratory data analysis of this study supports this claim, as we saw varying levels of performance and trends for the individual proportion performance plots. However, our final analysis and model validation suggests that the learning curve for quality is only partially dependent upon sonographer, as shown by the twenty-two (including the original) repeated data splits and model fits. Only the interaction term between pathology and exam experience appeared susceptible to the data split for both quality and agreement, while the learning curve for soft tissue exams on agreement was also susceptible to the data split. For both quality and agreement, the influence of exam experience on the learning curve remained consistent for fifteen and twenty of the data splits, respectively.

The type III analysis of this study showed that non exam experience significantly impacts the learning curve for agreement, but not for quality (under the consideration of exam experience and identified confounding factors). That is, experience on other exam types may affect a sonographer's agreement performance for a particular exam type, but not his or her quality performance. Based on research by Gaspari et al. (2009), this may be interpreted as suggesting that technical (image quality) skills are exam specific, whereas interpretive (agreement) skills may carry over from one exam type to another.

Our results suggest that a sonographer's quality performance may be defined in terms of his or her exam experience, in combination with exam type, reviewer, educational indication, and pathology. We also suggest that when exam experience is taken into account, it is not necessary to also account for additional experience beyond that of the specified exam type. On the other hand, it is more appropriate to consider experience beyond that of the exam type of interest, along with exam type, reviewer, and pathology, in identifying a sonographer's agreement performance, rather than considering his or her specific exam experience. Thus, in addition to exam type and confounding factors such as reviewer,

educational indication, and pathology, the learning curve also depends upon the measurement of performance (quality or agreement).

6 Limitations

There are limitations to this study which must be considered in tandem with our results. The sample consists only of sonographers practicing at four University of Massachusetts Memorial Health Care emergency departments. Thus, the conclusions made here may not generalize beyond this population. In particular, it has been suggested that learning curves may be department dependent (Jang et al., 2010), as training programs and protocols differ between departments. Thus, caution should be heeded when extending these results to other departments. More so, since learning curves may be dependent upon the sonographer (Hertzberg et al., 2000), caution should also be used when extending these results beyond the sonographers upon whose data the models were built. Finally, it is not clear whether missing data may be assumed to be missing completely at random to meet the randomness assumption of the GEE modeling approach.

7 Future Work

Further research would benefit by extending the ideas of this study. First, due to a limited amount of time and a pre-selected subset of research questions, we focused on learning curves based on exam experience. It is of interest to also consider learning curves based on overall experience, particularly as suggested by our type III analysis for the agreement model which suggested that significant information can be gained by including non exam experience in the model. Second, we considered a predicted probability of 0.9. Research may be conducted to explore other possible predicted probabilities, considering more lenient or stringent requirements.

Also a consideration for future work is the controlling of potential confounders. We identified that a sonographer's performance on quality and agreement is dependent upon who is reviewing his or her scan. To obtain better control, and potentially fewer learning curves, future studies should use a single reviewer or reviewers with consistent reviewing techniques (possibly obtained through training or calibration) whenever possible.

8 Conclusions

We conclude that there are a number of important factors that contribute to a sonographer's performance, as measured by quality and agreement. Specifically, for quality these factors include exam experience, exam type, reviewer, educational indication, and pathology. For agreement these factors include non exam experience, exam type, reviewer, and pathology. The number of scans required for a sonographer to obtain quality with a predicted probability of 0.9 is dependent upon the combination of the contributing factors and varies widely. Since there is no statistically significant evidence of a relationship between exam experience and agreement, no estimate is made for the number of scans required for a sonographer to obtain agreement with a predicted probability of 0.9. There is insufficient

evidence to conclude that experience beyond that of the exam type of interest contributes significantly to the learning curve for quality. However, although exam experience may not be associated with agreement, it was found that exam experience beyond that of the exam type of interest does contribute significantly to the learning curve for agreement.

Appendix A: List of Exam Types

Table A.1 lists first all of the exam types for which data was collected and second all of the exam types included for analysis.

Exam Types Recorded	Exam Types Included
Aorta	Aorta
Bladder	Cardiac (Atraumatic)
Cardiac (Atraumatic)	Cardiac (Code)
Cardiac (Code)	Chest Wall Ultrasound
Chest Wall Ultrasound	FAST
FAST	Gallbladder
Gallbladder	Lower Extremity
Guided Vascular Common Femoral	Ophthalmic (ocular)
Guided Vascular Internal Jugular	Renal
Guided Vascular Peripheral Vein	Soft Tissue
Lower Extremity	Uterus (endovaginal)
Musculoskeletal Bony	Uterus (trans-abdominal)
Musculoskeletal Joint	
Musculoskeletal Muscle/Tendon	
Musculoskeletal Ultrasound	
Ophthalmic (ocular)	
Renal	
Soft Tissue	
Testicular Ultrasound	
Ultrasound Guided Central Line	
Ultrasound Guided Paracentesis	
Ultrasound Guided Thoracentesis	
Upper Extremity - Left	
Upper Extremity - Right	
Uterus (endovaginal)	
Uterus (trans-abdominal)	

Table A. 1: List of exam types

Appendix B: Data Distribution

Tables B.1 to B.5 describe the data distribution for quality and agreement for the entire training data set and then the training set broken down by exam type, reviewer, educational indication, and pathology. Note that percent missing corresponds to the percentage of scans missing an agreement value.

% Quality	*% Agreement*	*% Missing*
80.86	*85.99*	*8.9*

Table B. 1: Aggregate training set distribution

Exam Type	% of Total	% Quality	% Agreement	% Missing
Aorta	12.07	73.56	86.37	11.23
Cardiac (Atraumatic)	9.78	65.01	84.8	9.58
Cardiac (Code)	0.37	88.66	88.66	3.09
Chest Wall Ultrasound	11.5	91.39	91.52	6.58
FAST	25.14	85.52	86.97	8.16
Gallbladder	14.91	78.92	85.71	8.66
Lower Extremity	4.77	76.03	84.62	6.39
Ophthalmic (ocular)	0.43	67.26	79.65	11.5
Renal	11.85	85.64	86.91	6.1
Soft Tissue	4.05	90.34	79.29	11.05
Uterus (endovaginal)	1.9	79.41	73.47	19.21
Uterus (trans-abdominal)	3.24	70.81	76.98	18.95

Table B. 2: Exam type distribution

Reviewer	% of Total	% Quality	% Agreement	% Missing
A	44.4	81.12	85.65	9.85
B	16.09	78.09	81.13	12.11
C	12.47	78.78	88.05	6.1
D	12.81	91.13	85.22	9.4
E	14.23	75.77	91.43	4.34

Table B. 3: Reviewer distribution

Indication	% of Total	% Quality	% Agreement	% Missing
medical	67.95	80.73	83.9	10.01
educational	32.05	81.14	90.41	6.56

Table B. 4: Educational indication distribution

Pathology	% of Total	% Quality	% Agreement	% Missing
(-)	84.41	80.12	89.44	9.41
(+)	15.59	84.88	67.3	6.16

Table B. 5: Pathology distribution

Appendix C: Inverse Interval Estimates

Tables C.1 to C.12 provide the quality inverse interval estimates for each of the twelve exam types, with twenty estimates each, corresponding to each of the twenty learning curves for an exam type. The letters 'A', 'B', 'C', 'D', and 'E' identify the reviewer for whom the curve corresponds to. Inclusion of 'ed' in the curve name corresponds to educational scans (lack of 'ed' corresponds to medical). Inclusion of 'p' corresponds to pathology (+) scans (lack of 'p' corresponds to pathology (-) scans).

curve	scan	lower	upper
A_ed_p	1160	0	7350
A_p	1542	0	9853
A_ed	817	0	2136
A	983	0	2582
B_ed_p	244	0	496
B_p	308	0	636
B_ed	299	78	520
B	351	87	616
C_ed_p	258	68	449
C_p	314	76	552
C_ed	305	158	453
C	352	178	526
D_ed_p	33	0	314
D_p	0	0	127
D_ed	0	0	0
D	0	0	0
E_ed_p	547	0	1253
E_p	645	0	1488
E_ed	549	167	930
E	622	184	1059

Table C. 1: Aorta interval estimates

curve	scan	lower	upper
A_ed_p	369	0	754
A_p	439	0	905
A_ed	404	122	686
A	461	137	785
B_ed_p	220	79	362
B_p	257	89	426
B_ed	257	123	391
B	290	137	443
C_ed_p	231	116	347
C_p	265	131	399
C_ed	264	162	366
C	295	180	410
D_ed_p	269	0	830
D_p	401	0	1279
D_ed	357	0	818
D	448	0	1039
E_ed_p	358	116	599
E_p	404	128	679
E_ed	384	197	571
E	424	215	632

Table C. 2: Cardiac (Atraumatic) inverse interval estimates

curve	scan	lower	upper
A_ed_p	2	0	6
A_p	3	0	7
A_ed	4	0	8
A	5	0	11
B_ed_p	3	0	7
B_p	4	0	8
B_ed	5	0	10
B	6	0	13
C_ed_p	4	0	8
C_p	5	0	10
C_ed	6	0	13
C	7	0	15
D_ed_p	0	0	4
D_p	0	0	6
D_ed	-0	0	7
D	1	0	6
E_ed_p	5	0	10
E_p	6	0	13
E_ed	7	0	15
E	8	0	18

Table C. 3: Cardiac (Code) inverse interval estimates

curve	scan	lower	upper
A_ed_p	0	0	0
A_p	0	0	0
A_ed	0	0	0
A	0	0	74
B_ed_p	0	0	0
B_p	0	0	0
B_ed	0	0	29
B	4	0	70
C_ed_p	0	0	0
C_p	0	0	34
C_ed	0	0	58
C	36	0	80
D_ed_p	9819	0	359089
D_p	8366	0	305664
D_ed	0	0	0
D	0	0	0
E_ed_p	0	0	50
E_p	0	0	121
E_ed	52	0	111
E	111	17	204

Table C. 4: Chest wall interval estimates

curve	scan	lower	upper
A_ed_p	378	0	2442
A_p	0	0	433
A_ed	765	0	4007
A	1417	0	7687
B_ed_p	8	0	122
B_p	97	0	211
B_ed	135	30	239
B	203	38	367
C_ed_p	67	0	142
C_p	141	35	247
C_ed	165	91	238
C	224	123	325
D_ed_p	460	0	1157
D_p	342	0	836
D_ed	398	0	1146
D	198	0	523
E_ed_p	325	0	895
E_p	498	0	1420
E_ed	409	115	703
E	518	132	904

Table C. 5: FAST inverse interval estimates

curve	scan	lower	Upper
A_ed_p	148	45	250
A_p	211	59	362
A_ed	219	115	324
A	271	138	403
B_ed_p	108	54	162
B_p	143	71	215
B_ed	155	92	218
B	186	108	264
C_ed_p	126	79	174
C_p	158	99	218
C_ed	168	120	216
C	197	140	255
D_ed_p	0	0	60
D_p	19	0	142
D_ed	83	9	157
D	161	27	296
E_ed_p	209	108	311
E_p	252	126	378
E_ed	253	164	342
E	290	186	394

Table C. 6: Gallbladder inverse interval estimates

curve	scan	lower	upper
A_ed_p	58	7	108
A_p	82	8	156
A_ed	95	15	176
A	118	15	221
B_ed_p	58	18	98
B_p	77	23	131
B_ed	88	29	146
B	105	33	177
C_ed_p	70	25	116
C_p	88	29	147
C_ed	98	35	160
C	115	39	190
D_ed_p	0	0	27
D_p	6	0	50
D_ed	29	0	60
D	55	5	105
E_ed_p	102	20	183
E_p	122	22	222
E_ed	131	31	231
E	150	34	267

Table C. 7: Lower extremity inverse interval estimates

curve	scan	lower	upper
A_ed_p	101	0	3526
A_p	123	0	4316
A_ed	133	0	4354
A	153	0	5023
B_ed_p	92	0	2499
B_p	109	0	2976
B_ed	117	0	3029
B	134	0	3479
C_ed_p	102	0	2679
C_p	118	0	3111
C_ed	126	0	3155
C	142	0	3565
D_ed_p	35	0	1351
D_p	62	0	2479
D_ed	78	0	2898
D	102	0	3822
E_ed_p	135	0	4074
E_p	155	0	4689
E_ed	161	0	4587
E	179	0	5108

Table C. 8: Ophthalmic (ocular) inverse interval estimates

curve	scan	lower	upper
A_ed_p	0	0	142
A_p	91	0	278
A_ed	148	0	335
A	249	0	595
B_ed_p	20	0	82
B_p	71	10	132
B_ed	99	34	163
B	142	45	239
C_ed_p	55	7	103
C_p	101	35	166
C_ed	122	60	183
C	161	76	246
D_ed_p	1176475	0	8.13E+09
D_p	846982	0	5.85E+09
D_ed	0	0	0
D	0	0	197
E_ed_p	154	0	310
E_p	224	0	467
E_ed	231	52	410
E	288	56	519

Table C. 9: Renal inverse interval estimates

curve	scan	lower	upper
A_ed_p	0	0	15
A_p	2	0	25
A_ed	12	0	30
A	24	6	42
B_ed_p	0	0	22
B_p	10	0	27
B_ed	19	3	35
B	29	10	48
C_ed_p	8	0	26
C_p	18	3	33
C_ed	26	9	43
C	36	14	58
D_ed_p	0	0	0
D_p	0	0	0
D_ed	0	0	1
D	0	0	13
E_ed_p	20	3	36
E_p	31	10	52
E_ed	39	13	65
E	49	16	83

Table C. 10: Soft tissue inverse interval estimates

curve	scan	lower	upper
A_ed_p	65	0	159
A_p	89	0	223
A_ed	101	0	245
A	123	0	302
B_ed_p	64	0	136
B_p	82	0	176
B_ed	92	0	194
B	109	0	232
C_ed_p	76	0	158
C_p	93	0	197
C_ed	102	0	211
C	119	0	248
D_ed_p	0	0	45
D_p	16	0	59
D_ed	37	0	93
D	63	0	161
E_ed_p	107	0	247
E_p	127	0	296
E_ed	135	0	305
E	154	0	350

Table C. 11: Uterus (endovaginal) inverse interval estimates

curve	scan	lower	upper
A_ed_p	81	34	127
A_p	99	42	157
A_ed	109	48	170
A	127	54	199
B_ed_p	77	40	113
B_p	92	48	136
B_ed	100	53	147
B	114	60	169
C_ed_p	86	47	125
C_p	101	55	147
C_ed	108	60	156
C	122	67	177
D_ed_p	24	0	51
D_p	46	14	77
D_ed	60	23	97
D	80	30	130
E_ed_p	113	54	172
E_p	129	61	198
E_ed	136	67	206
E	152	73	231

Table C. 12: Uterus (trans-abdominal) inverse interval estimates

Appendix D: Quality Learning Curves

Figures D.1 to D.5 contain the 240 learning curves for quality performance.

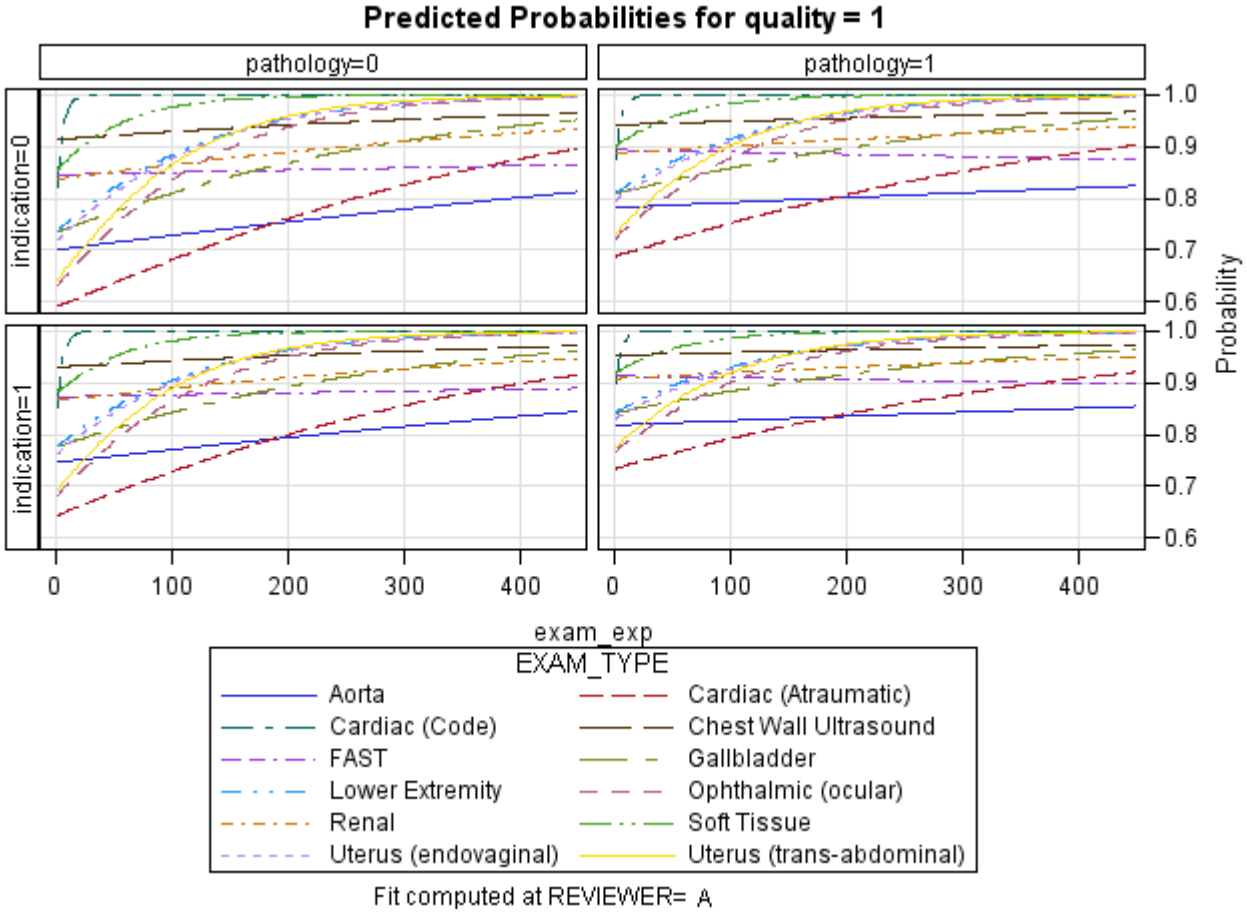


Figure D. 1: Reviewer A quality learning curves

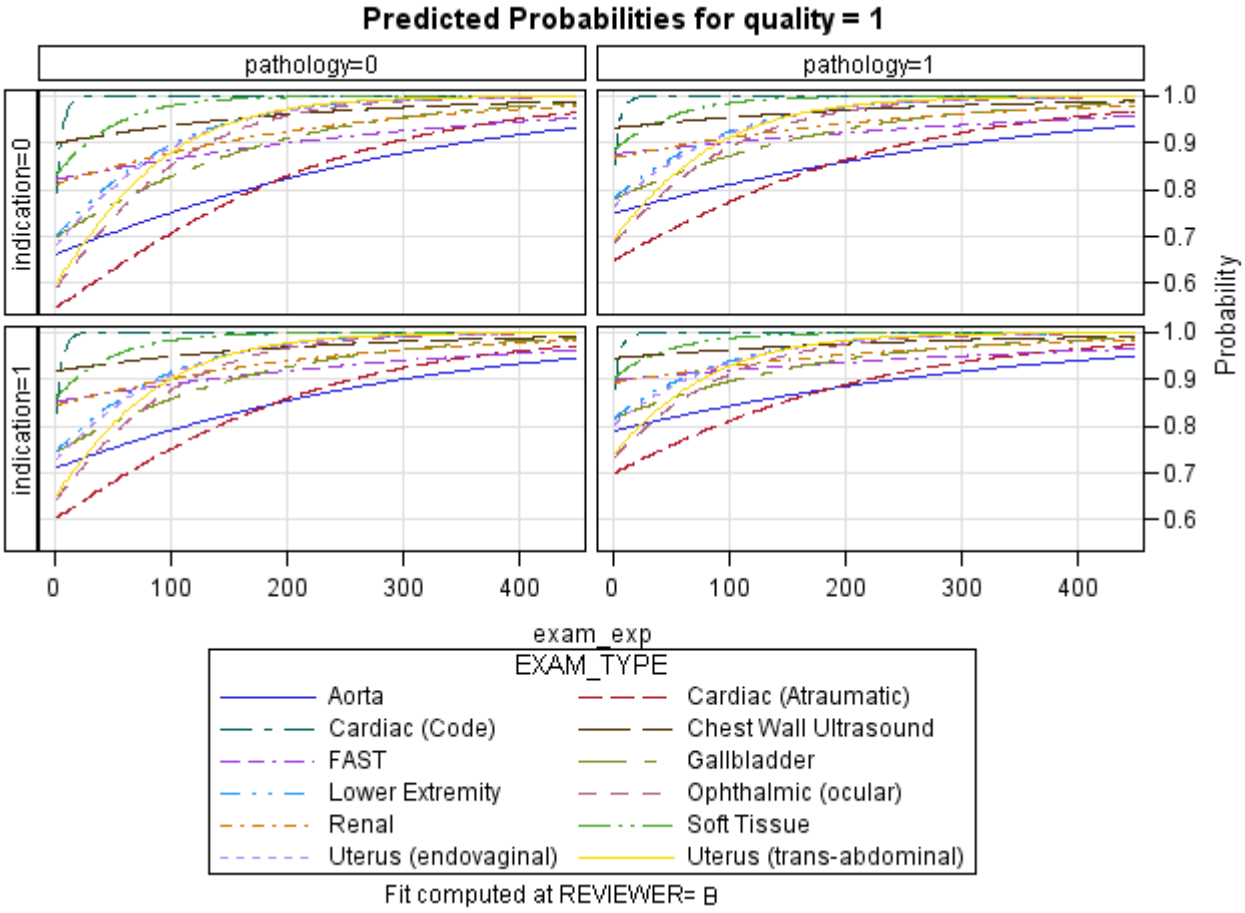


Figure D. 2: Reviewer B quality learning curves

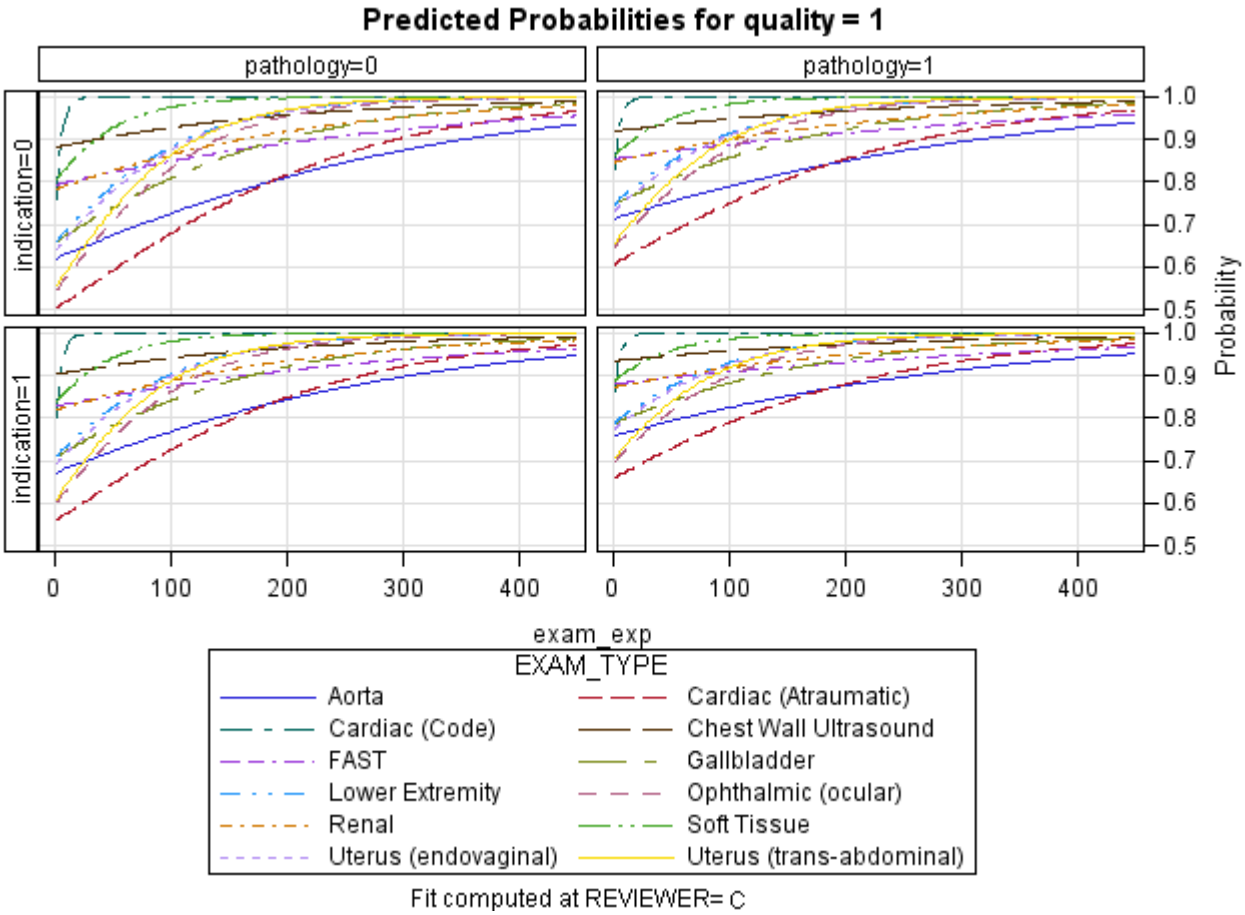


Figure D. 3: Reviewer C quality learning curves

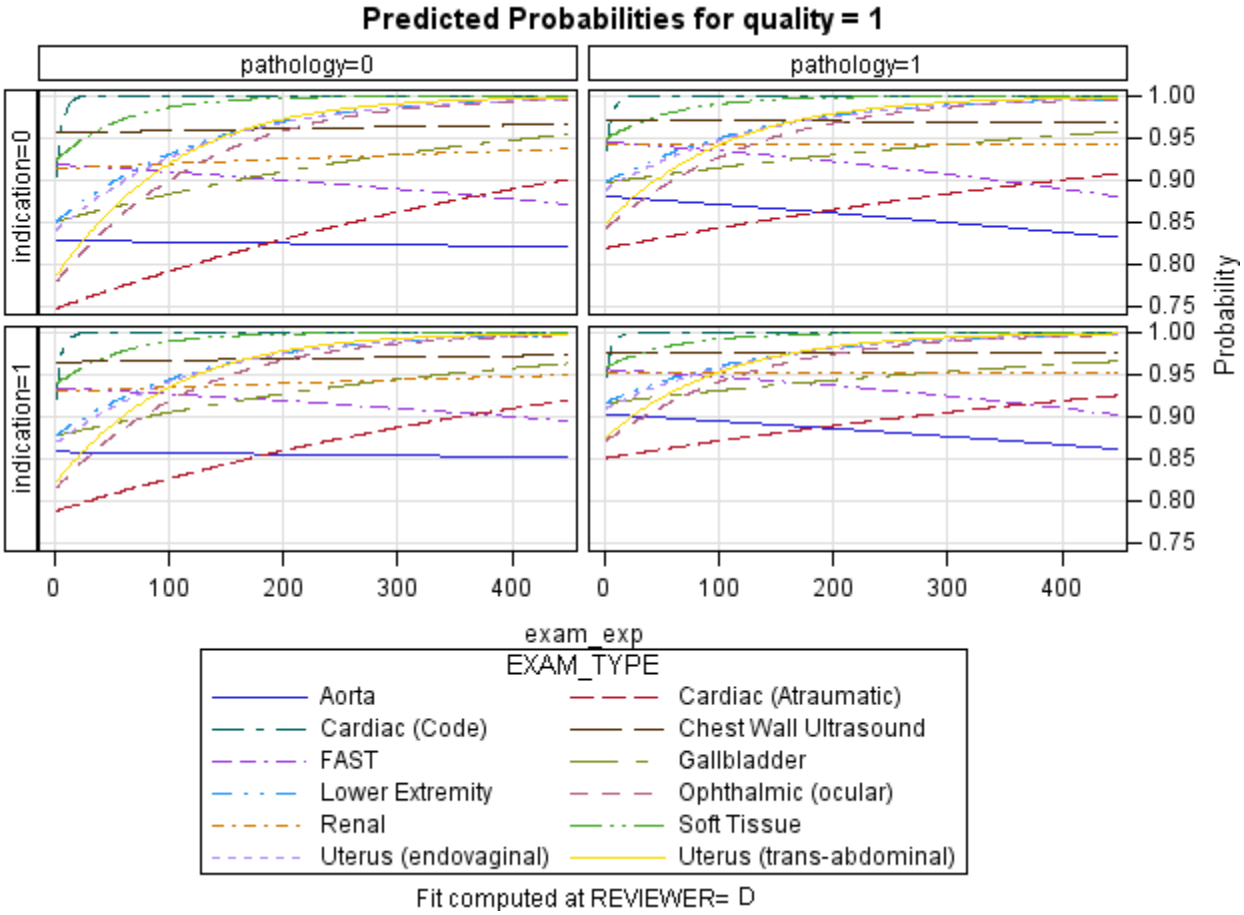


Figure D. 4: Reviewer D quality learning curves

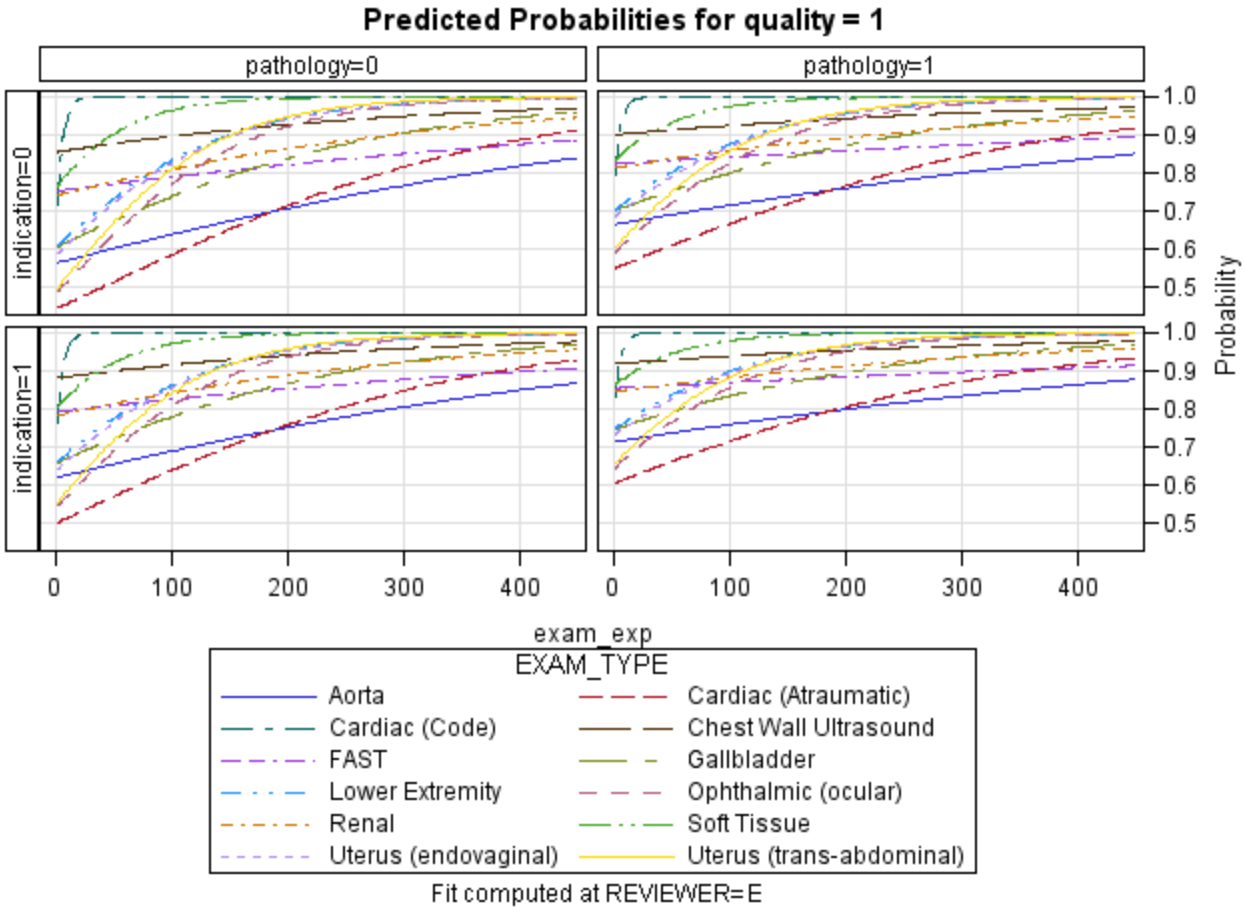


Figure D. 5: Reviewer E quality learning curves

Appendix E: Agreement Learning Curves

Figures E.1 to E.5 contain the 120 learning curves for agreement performance.

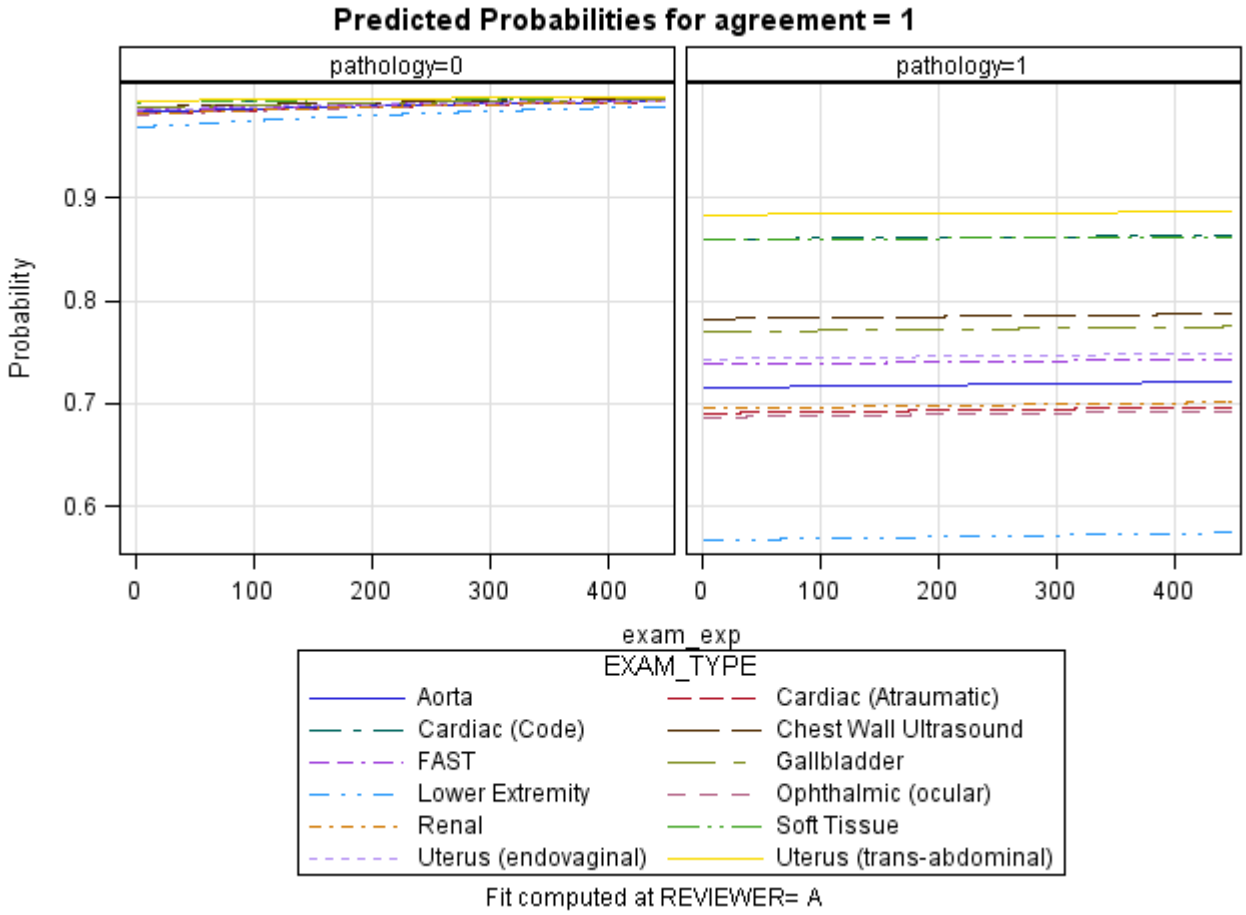


Figure E. 1: Reviewer A agreement learning curves

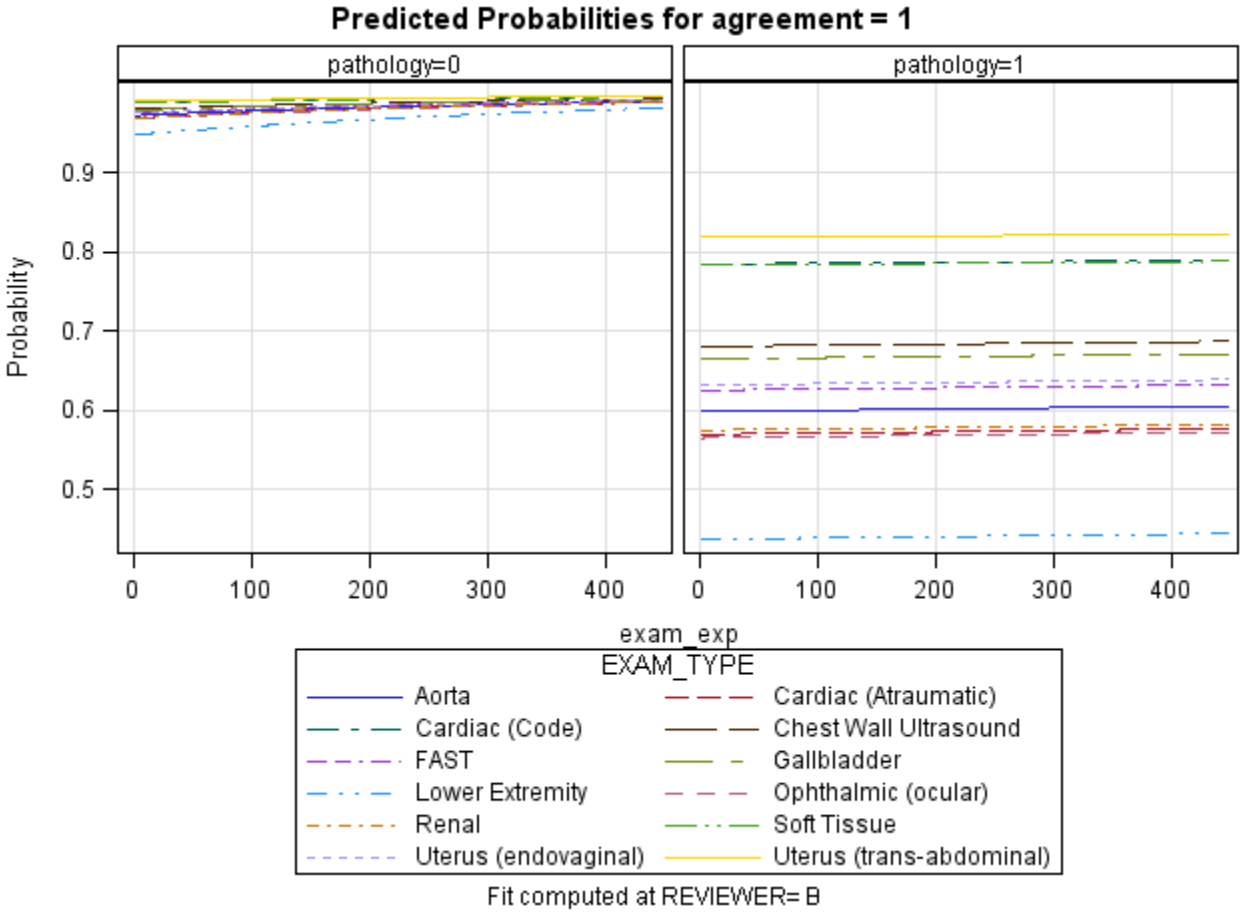


Figure E. 2: Reviewer B agreement learning curves

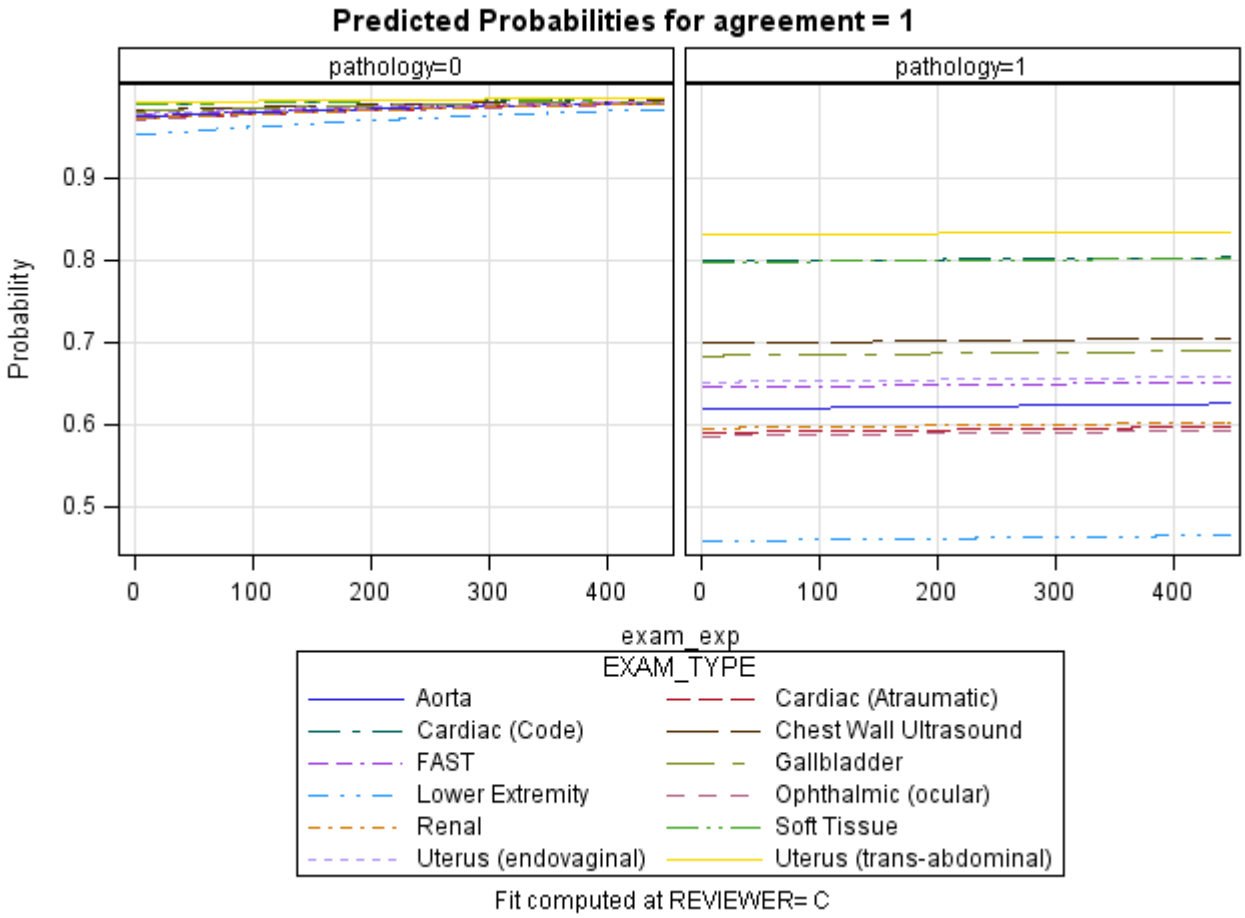


Figure E. 3: Reviewer C agreement learning curves

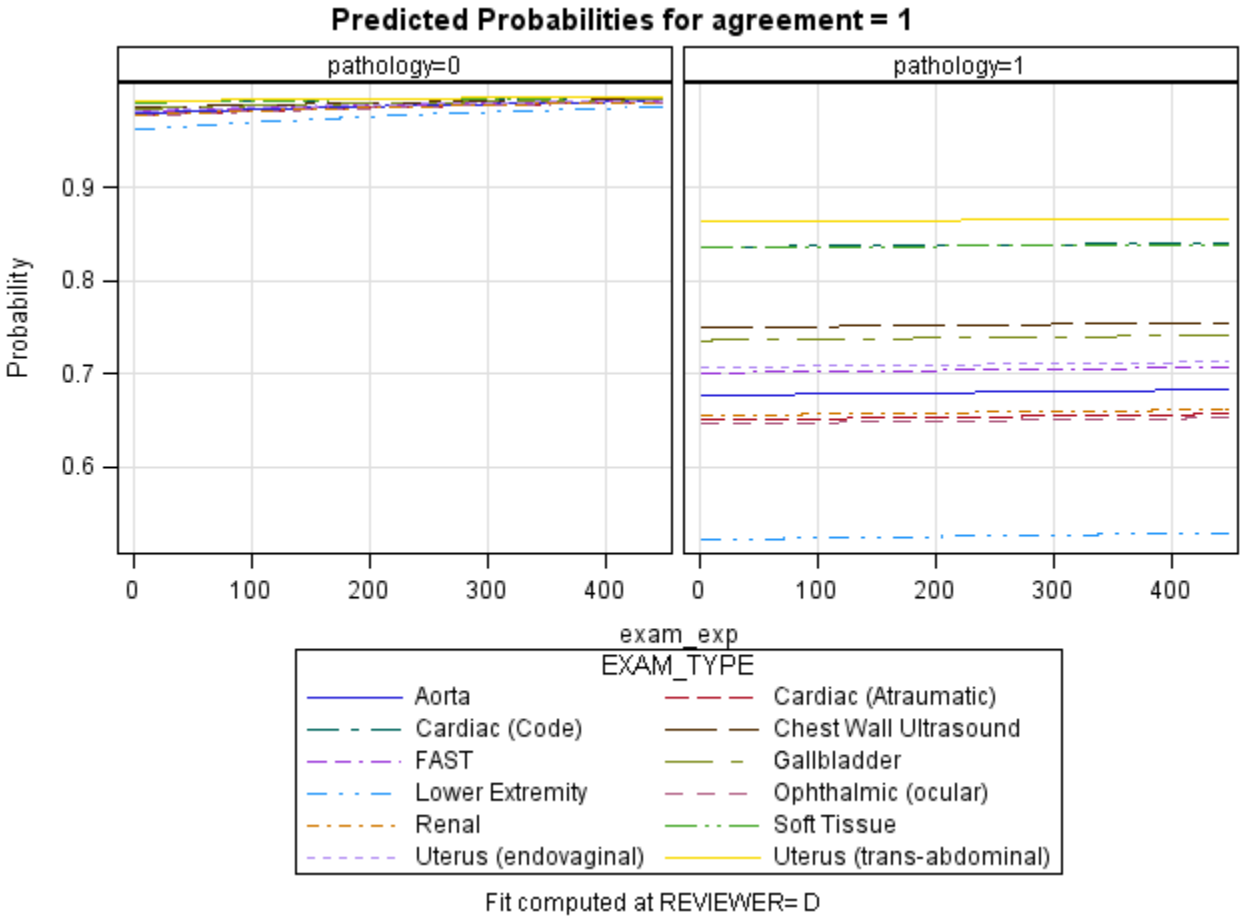


Figure E. 4: Reviewer D agreement learning curves

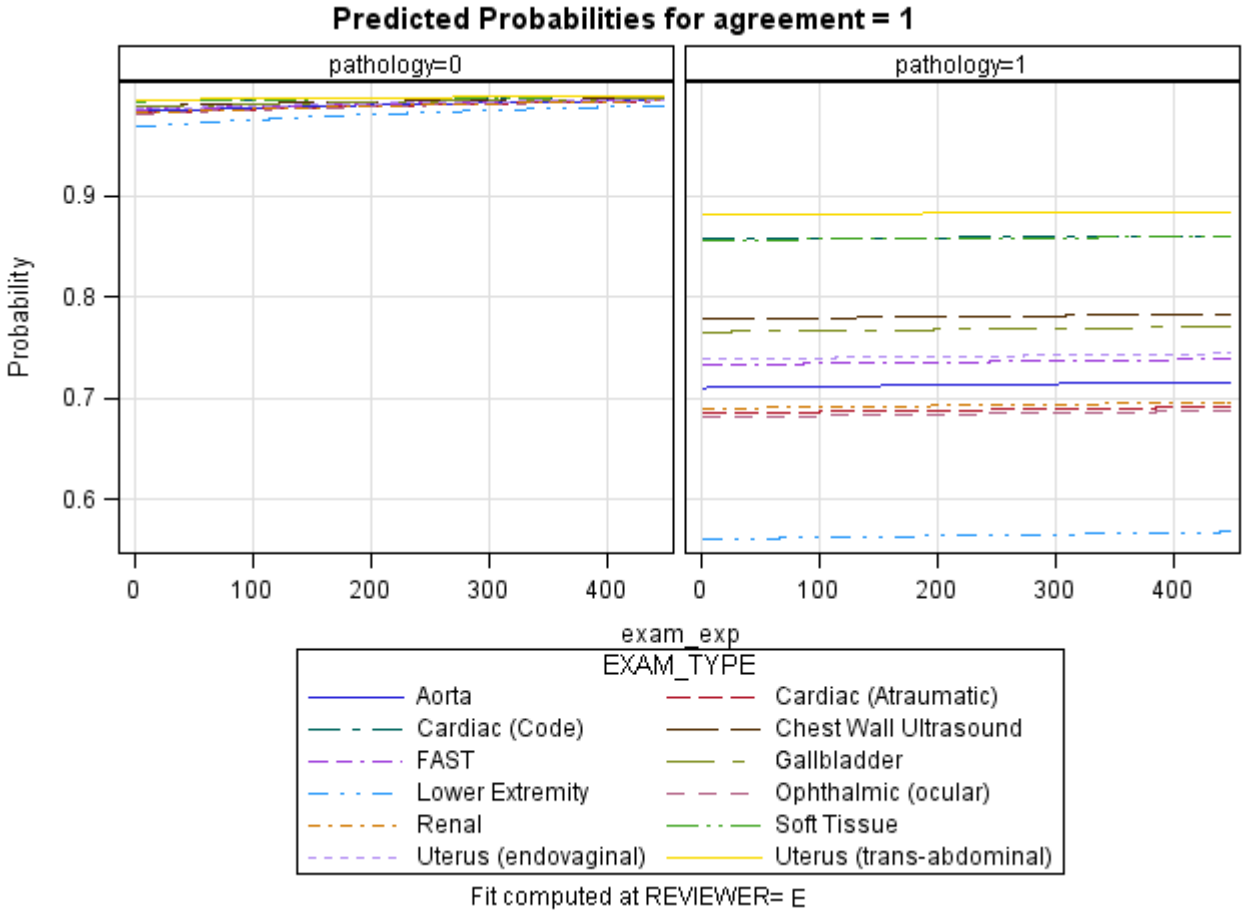


Figure E. 5: Reviewer E agreement learning curves

References

Agresti, A. (2007). An *introduction to categorical data analysis.* Hoboken, NJ: John Wiley & Sons, Inc.

American College of Emergency Physicians (2008). Policy statement: Emergency ultrasound guidelines. Dallas, TX: American College of Emergency Physicians.

American College of Radiology (2011). ACR-SPR-SRU practice guideline for performing and interpreting diagnostic ultrasound examinations. Retrieved from http://www.acr.org/~/media/13B896B9F4844E3082E7D7ED66AFC148.pdf

American College of Radiology (2012). Ultrasound guidelines. Retrieved from http://www.acr.org/Quality-Safety/Standards-Guidelines/Practice-Guidelines-by Modality/Ultrasound

Gaspari, J. R., Dickman, E., Blehar, D. (2009) Learning curve of bedside ultrasound of the gallbladder. *The Journal of Emergency Medicine, 37(1)*, 51-56.

Hertzberg B. S., Kliewer M. A., Bowie, J.D., Carroll, B. A., DeLong, D. H., Gray, L., & Nelson, R. C. (2000). Physician training requirements in sonography: How many cases are needed for competence? *American Journal of Roentgenology, 174 (5)*, 1221-1227.

Jang, T. B., Ruggeri, W., Dyne, P., & Kaji, A. H. (2010). The learning curve of resident physicians using emergency ultrasonography for cholelithiasis and cholecystitis. *The Society for Academic Emergency Medicine, 17 (11)*, 1247-1252.

SAS Institute Inc. (2009). *SAS/STAT® 9.2 User's guide, Second Edition.* (2nd ed.) Cary, NC: SAS Institute Inc.

SAS Institute Inc. (2012). ROC: Nonparametric comparison of areas under correlated ROC curves [SAS macro]. Available from http://support.sas.com/kb/25/017.html#pur

Stevens, J. R. (2012). *STAT 5810/6810 - Handout #4.1-b: Inverse intervals in logistic regression* [PDF document]. Retrieved from Lecture Notes Online Web site http://www.stat.usu.edu/jrstevens/stat6810/