Impact of Different Study Populations on Reader Behavior and Performance Metrics: Initial Results

Brandon D. Gallas[1], Etta Pisano[2,3], Elodia Cole[2], Kyle Myers[1]

[1] FDA/CDRH/OSEL/DIDSR, Silver Spring, MD

[2] Beth Israel Deaconess Medical Center, Boston, MA

[3] Harvard Medical School, Harvard University, Boston, MA

Abstract

The FDA recently completed a study on design methodologies surrounding the **V**alidation of **I**maging **P**remarket **E**valuation and **R**egulation called VIPER. VIPER consisted of five large reader sub-studies to compare the impact of different study populations on reader behavior as seen by sensitivity, specificity, and AUC, the area under the ROC curve (receiver operating characteristic curve). The study investigated different prevalence levels and two kinds of sampling of non-cancer patients: a screening population and a challenge population. The VIPER study compared full-field digital mammography (FFDM) to screen-film mammography (SFM) for women with heterogeneously dense or extremely dense breasts. All cases and corresponding images were sampled from Digital Mammographic Imaging Screening Trial (DMIST) archives. There were 20 readers (American Board Certified radiologists) for each sub-study, and instead of every reader reading every case (fully-crossed study), readers and cases were split into groups to reduce reader workload and the total number of observations (split-plot study). For data collection, readers first decided whether or not they would recall a patient. Following that decision, they provided an ROC score for how close or far that patient was from the recall decision threshold. Performance results for FFDM show that as prevalence increases to 50%, there is a moderate increase in sensitivity and decrease in specificity, whereas AUC is mainly flat. Regarding precision, the statistical efficiency (ratio of variances) of sensitivity and specificity relative to AUC are 0.66 at best and decrease with prevalence. Analyses comparing modalities and the study populations (screening vs. challenge) are still ongoing.

Purpose

In this paper we present a study that investigates the effects of prevalence and enrichment on reader behavior as measured by AUC, sensitivity, and specificity in laboratory studies. We refer to this study as VIPER, **V**alidation of **I**maging **P**remarket **E**valuation and **R**egulation. The VIPER study was born from the desire to validate the use of small controlled lab studies instead of large prospective clinical trials to compare a new imaging modality to a reference imaging modality in studies to support FDA clearance or approval of medical imaging devices.

Medical Imaging 2017: Image Perception, Observer Performance, and Technology Assessment,
edited by Matthew A. Kupinski, Robert M. Nishikawa, Proc. of SPIE Vol. 10136, 101360A

One of the largest imaging studies conducted at the time of this study's conception (2010) was the Digital Mammographic Imaging Screening Trial [1, 2]. DMIST was designed to compare full-field digital mammography (FFDM) to screen-film mammography (SFM), pooling the effects and results of five different FFDM units and six different SFM units. The endpoints of this trial were sensitivity, specificity, and AUC, the area under the receiver operating characteristic curve (ROC). The trial was sized to detect an AUC difference of 0.06 between FFDM and SFM with 5% Type I error and 80% power. This requirement and the low prevalence of breast cancer in the screening population drove the study to be an extremely large study, enrolling 49,528 women and obtaining all relevant information for 42,760 of them employing 153 radiologists. Consequently, the trial was quite expensive at $26 million, a cost that was taken on by the National Cancer Institute to push the technology over clinical implementation hurdles. It was generally believed that such a study was beyond what should be borne by any one FFDM manufacturer.

The overall objective of this work is to determine whether a significant difference in performance between two imaging modalities measured in a large prospective clinical trial can also be achieved in a much smaller, controlled, cancer-enriched lab study. To this end, we designed VIPER to compare FFDM to SFM for women with heterogeneously dense and extremely dense breasts: Breast Imaging-Reporting and Data System (BIRADS) 4th edition breast density classifications of 3 and 4. We chose this subgroup because DMIST found FFDM was significantly better than SFM, and there were a lot of cancer cases in the DMIST archive that we could sample from. For this group AUC for FFDM was 0.15 better than the AUC for SFM, and the BIRADS-based sensitivity for FFDM was 0.14 better than that for SFM with specificity essentially unchanged.

In this work, we report on another objective of VIPER: to investigate the effects of changing the study population on radiologist behavior and performance endpoints AUC, sensitivity, and specificity (point estimates and variance estimates). Specifically, we investigated three levels of prevalence (enrichment) and two types of study populations ("screening" and "challenge"). For the VIPER "screening" studies, the study populations of non-cancer women reflect the clinical screening diagnosis scores (predominately BIRADS 1,2,3 and a few BIRADS 0). For the VIPER "challenge" studies, the study populations of non-cancer women only include women with BIRADS 0 screening diagnosis scores. The "challenge" study populations are expected to stress and challenge the modalities under study by enriching with women without cancers who were, in some sense, misdiagnosed.

Methods

The digital (FFDM) and screen-film mammograms (SFM) included in this study are from the American College of Radiology Imaging Network's Digital Mammographic Imaging Screening Trial (DMIST) image archives. Eligibility for inclusion in this study required that the BIRADS breast density classification documented in DMIST be 3 or 4 by SFM. The total number of cases for each category that we were able to obtain both screen-film and digital mammograms were as follows: 129 cancer cases (determined within 455 days of initial imaging), 158 cases scored

BIRADS 0 at screening on the SFM, 158 cases scored BIRADS 0 at screening on the FFDM, 144 cases scored BIRADS 1 or 2 at screening on SFM, and 133 cases scored BIRADS 1 or 2 on FFDM. There were no BIRADS 3 cases in our dataset as BIRADS 3 score classifications at screening during DMIST were rare.

Given the available cases, we designed five VIPER sub-studies. The target per-reader prevalences were as follows:

```
##                prevalence    N1perReader    N0perReader
## Screening11          0.11             20            153
## Screening29          0.29             32             77
## Screening50          0.50             30             30
## Challenge29          0.29             32             78
## Challenge50          0.50             30             30
```

The lowest prevalence VIPER sub-study was 11%. For comparison, the prevalence in DMIST was 0.8% for 455 days of follow-up [1]; DMIST prevalence was nearly a factor of 15 smaller than the prevalence of VIPER.

Notice that three of the five sub-studies are denoted as "screening" studies and the other two are "challenge" studies. The screening studies generally reflect the clinical screening population for the women without cancer in that there are significantly more BIRADS 1-3 than BIRADS 0 cases in the dataset. The women without cancer in the challenge studies were all BIRADS 0. Per-reader populations in the VIPER sub-studies for the women without cancer were as follows:

```
##                BIRADS123perReader    BIRADS0perReader    ratio
## Screening11                   138                  16    8.625
## Screening29                    69                   8    8.625
## Screening50                    26                   4    6.500
## Challenge29                     0                  78    0.000
## Challenge50                     0                  30    0.000
```

The ratio of BIRADS 1-3 cases for every BIRADS 0 case in DMIST was found to be 10.7 BIRADS 1-3 for every BIRADS 0. This ratio is the same whether determined by SFM or FFDM screening results. The BIRADS sub-populations listed above for VIPER were based on DMIST SFM and FFDM screening results in equal proportions.

The VIPER sub-studies used split-plot designs [3]. A split-plot design allows a reduction in the workload of individual readers and a reduction in the total number of reader evaluations. This reduction comes without sacrificing a lot of statistical precision compared to a fully-crossed study design (every reader reads every case in all modalities). Each VIPER sub-study had 20 readers split into four groups of five. The cases in each sub-study were also split into 4 groups, though the number of BIRADS 1-3 cases was limited and there was overlap of these cases across groups to achieve the low prevalence screening sub-study. The total case counts in the VIPER sub-studies were as follows:

```
##                N1    N0
## Screening11    80   339
```

```
## Screening29   127   308
## Screening50   119   120
## Challenge29   127   312
## Challenge50   119   120
```

All readers were American Board of Radiology certified, MQSA qualified, and have clinically interpreted at least 50 film mammograms and 50 digital mammograms as part of their residency or practice. Readers were allowed to participate in more than one sub-study as long as they were assigned to a group with no overlap in cases to read. Ultimately, there were 44 readers across all the studies.

For each sub-study, the cases with and without cancer were each split evenly into two groups. There were two sessions for each reader. Each was preceded by training, which included a description of the study population. In the first session, half the cases were read by FFDM and half were read by SFM. In the second session, the cases were read in the opposite modality from the first session. The minimum washout was 27 days, the mean was 68.19 days, and the median was 49.5 days.

For data collection, readers first decided whether or not they would recall a patient. Following that decision, they provided an ROC score (101-point scale) for how close or far that patient was from the recall decision threshold. Consequently, the recall operating point is embedded in and consistent with the ROC scores and runs from 0 to 201. The "No" recall decision corresponds to scores 0 to 100. The "Yes" recall decision corresponds to scores 101 to 201.

All the analyses were produced with iMRMC-v2p8, which can be found at https://github.com/DIDSR/iMRMC/releases . MRMC stands for multi-reader, multi-case; the radiologists are the readers and the patients are the cases. The iMRMC software performs MRMC analyses of the data. This should be understood to mean that the results (AUC, sensitivity, and specificity) are reader- and case-averaged quantities, and that variances and confidence intervals of the results account for the variability from the readers and the cases. All AUC estimates from the iMRMC software are the non-parametric versions; all variances utilize U-statistics and account for reader and case variability [3, 4, 5].

Results

As we present our results, please recall that the VIPER study populations are restricted to women with dense breasts (BIRADS density 3, Heterogeneously Dense, and 4, Extremely Dense) and cancer is determined within 455 days of initial imaging. Also, sensitivity and specificity are based on the binary reader “recall” and “do not recall” decisions, whereas the ROC curves combined the numeric scores collected after the “recall” and “do not recall” decisions. We show reader-specific results and then average these to get reader-averaged results. Please note that in the figures and tables we sometime use the abbreviation for the true positive fraction and the true negative fraction in place of sensitivity (TPF) and specificity (TNF).

In Figure 1 we show the performance results for FFDM across the different sub-study populations for the AUC, sensitivity, or specificity. The error bars are 95% confidence

intervals based on the t-distribution where the variance estimate is given by U-statistics [4] and the degrees of freedom are approximated [3].

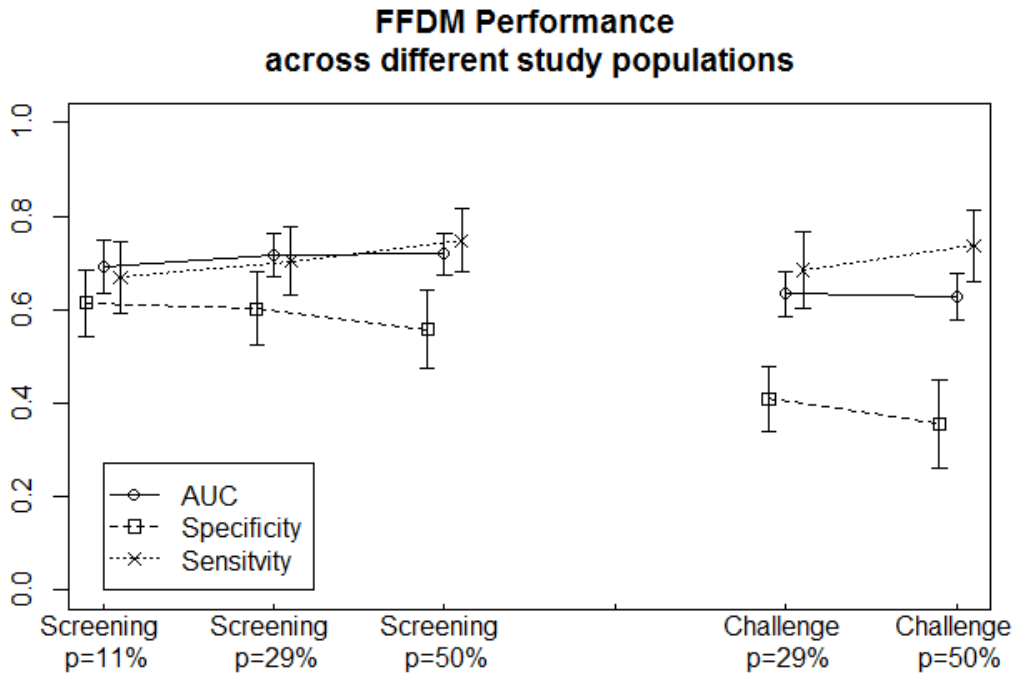

Figure 1: AUC, sensitivity, and specificity for FFDM across different study populations. Error bars are 95% MRMC confidence intervals. They account for reader and case variability.

In the screening sub-studies, as we increase prevalence, we see a mild increase in sensitivity and a mild decrease in specificity, whereas AUC remains relatively flat. In the challenge sub-studies, the impact of prevalence on sensitivity and specificity is more pronounced. Additionally, the sensitivity measured in the challenge study was relatively unchanged from that in the screening study, controlling for prevalence, whereas AUC and specificity are reduced. These results are expected of an observer that is maximizing a cost-benefit analysis: increasing prevalence causes readers to call more cases positive, moving a reader's operating point up and to the right [6], with little impact on AUC.

For comparison, the DMIST FFDM performance results were 0.78 for AUC, 0.57 for sensitivity, and 0.91 for specificity. These performance results are based on pooling the reader data, not averaging reader-specific performance results. AUC was based on the seven-point malignancy scores. Sensitivity and specificity are based on the BIRADS scores [7] dichotomized as negative (score of 1, 2, or 3) and positive (score of 0, 4, 5). We choose to compare the BIRADS-based sensitivity and specificity since BIRADS ratings are the clinical management scores, making them similar to the “recall” and “do not recall” clinical action decisions collected in the VIPER study.

There are many differences between the VIPER and DMIST studies that could explain the performance differences. We will explore these in another paper. Here, it is worth noting that the prevalence is much larger in VIPER than DMIST, and it appears to have the same impact on sensitivity and specificity across these studies as within VIPER. Increasing prevalence leads to an increase in sensitivity (DMIST 0.57 to VIPER 0.69) and a decrease in specificity (DMIST 0.91 to VIPER 0.61).

In Figure 2, we show the operating points for all 20 readers using FFDM in the three screening sub-studies. The figure also shows the reader-averaged ROC curves [8] and reader-averaged operating points. It is amazing how the wide range of reader performance and operating points can be consolidated in the coherent and expected story: increasing prevalence causes readers to call more cases positive, moving a reader's operating point up and to the right [6] with little impact on AUC.

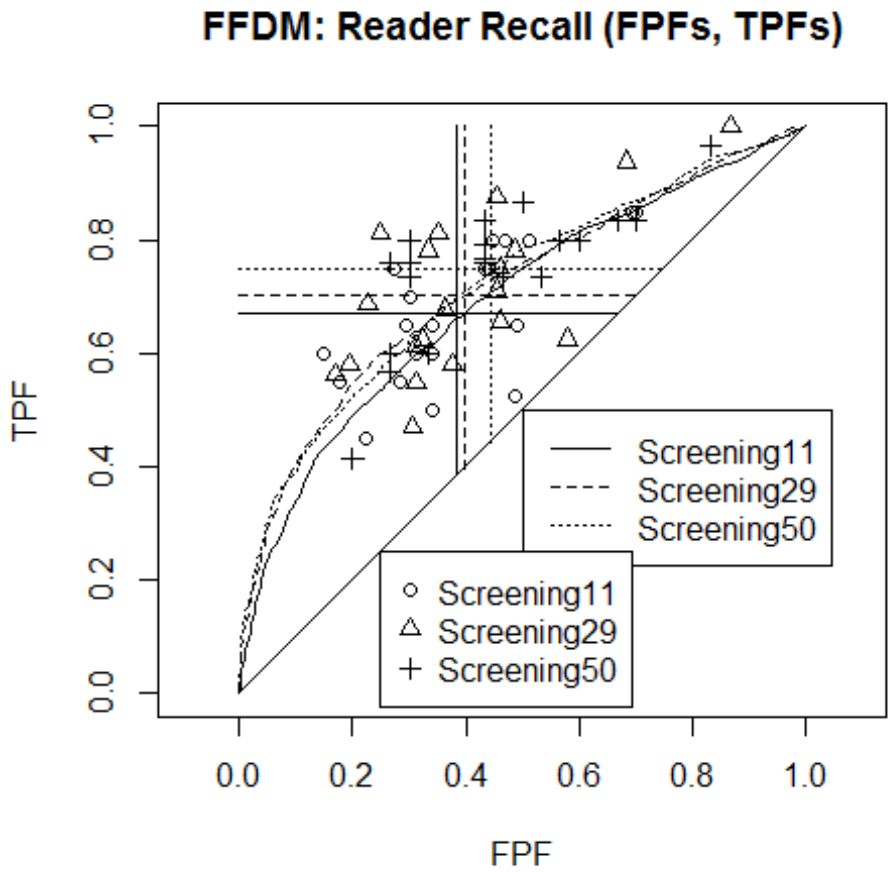

Figure 2: This figure shows reader-specific operating points (symbols), reader-averaged operating points (crossing vertical and horizontal lines), and reader-averaged ROC curves. The diagonal line is the chance-guessing line.

In the table below we show the statistical efficiency of sensitivity and specificity relative to AUC for the designed experiments. Statistical efficiency is the ratio of the variance of one divided by (relative to) the other. We see that the efficiencies of both sensitivity and specificity are less than one, are less than the efficiency of AUC. Furthermore, the efficiencies relative to AUC decrease as prevalence approaches 50%. This means that the estimation of sensitivity and specificity is less precise than AUC and gets worse as prevalence approaches 50%. The precision of these different endpoints can be compared because they live on the same scale, are close, and are not near the extremes 1.0 or 0.0.

```
##              N0perReader effSpecFFDM N1perReader effSensFFDM
## Screening11          153       0.656          20       0.553
## Screening29           77       0.335          32       0.380
## Screening50           30       0.281          30       0.433
## Challenge29           78       0.470          32       0.344
## Challenge50           30       0.275          30       0.416
```

Discussion

Regarding the impact of prevalence, an early study evaluated its effect on AUC, sensitivity, and specificity [9]. That study also evaluated the impact on the confidence ratings. However, the results are not strong as the experiment was very small and MRMC analyses were not performed (statistical analyses did not account for both reader and case variability). There is a one study of the impact of prevalence on AUC in medical imaging [10] and one on the impact on the confidence ratings [11]. These studies are larger and do MRMC analyses. There is also a study of the impact of prevalence on sensitivity and specificity in medical imaging [12]. This study was large, but did not do MRMC analyses. Also, we are not aware of any studies comparing the impact of different study populations (screening vs. challenge). VIPER is a large-scale investigation of these issues.

The split-plot study design is still new and under investigated. Here we found it to be an efficient way to collect data, allowing us to conduct 5 sub-studies instead of just 1.

Conclusions

This work confirms expectations that AUC is invariant to prevalence whereas prevalence can impact sensitivity and specificity. This work also shows the impact of sampling from a "challenge" population instead of a screening population.

Bibliography

[1] Etta D. Pisano, Constantine Gatsonis, Edward Hendrick, Martin Yaffe, Janet K. Baum, Suddhasatta Acharyya, Emily F. Conant, Laurie L. Fajardo, Lawrence Bassett, Carl D'Orsi,

Roberta Jong, and Murray Rebner. Diagnostic performance of digital versus film mammography for breast-cancer screening. *N Engl J Med*, 353(17):1773–1783, 2005.

[2] Etta D. Pisano, Constantine A. Gatsonis, Martin J. Yaffe, R. Edward Hendrick, Anna N. A. Tosteson, Dennis G. Fryback, Lawrence W. Bassett, Janet K. Baum, Emily F. Conant, Roberta A. Jong, Murray Rebner, and Carl J. D'Orsi. American college of radiology imaging network digital mammographic imaging screening trial: Objectives and methodology. *Radiology*, 236(2):404–412, 2005.

[3] Nancy Obuchowski, Brandon D. Gallas, and Stephen L. Hillis. Multi-reader ROC studies with split-plot designs: A comparison of statistical methods. *Acad Radiol*, 19(12):1508–1517, 2012. Invited paper for Special Metz Memorial Issue I.

[4] Brandon D. Gallas and David G. Brown. Reader studies for validation of CAD systems. *Neural Networks Special Conference Issue*, 21(2-3):387–397, 2008. Invited manuscript for special conference issue.

[5] Brandon D. Gallas, Andriy Bandos, Frank Samuelson, and Robert F. Wagner. A framework for random-effects ROC analysis: Biases with the bootstrap and other variance estimators. *Commun Stat A-Theory*, 38(15):2586–2603, 2009.

[6] Charles E. Metz. Basic principles of ROC analysis. *Semin Nucl Med*, 8(4):283–298, 1978.

[7] Etta D. Pisano, Constantine Gatsonis, Edward Hendrick, Martin Yaffe, Janet K. Baum, Suddhasatta Acharyya, Emily F. Conant, Laurie L. Fajardo, Lawrence Bassett, Carl D'Orsi, Roberta Jong, and Murray Rebner. Diagnostic performance of digital versus film mammography for breast-cancer screening. Supplementary Material Online, 2005.

[8] Weijie Chen and Frank W. Samuelson. The average receiver operating characteristic curve in multi-reader multi-case imaging studies. *Br J Radiol*, 87:20140016, Jun 2014.

[9] T. K. P. Egglin and A. R. Feinstein. Context bias: A problem in diagnostic radiology. *JAMA*, 276:1752–1755, 1996.

[10] David Gur, Howard E. Rockette, Derek R. Armfield, Arye Blachar, Jennifer K. Bogan, Giuseppe Brancatelli, Cynthia A. Britton, Manuel L. Brown, Peter L. Davis, James V. Ferris, Carl R. Fuhrman, Sara K. Golla, Sanj Katyal, Joan M. Lacomis, Barry M. McCook, F. Leland Thaete, and Thomas E. Warfel. Prevalence effect in a laboratory environment. *Radiology*, 228(1):10–14, 2003.

[11] David Gur, Andriy I Bandos, Carl R Fuhrman, Amy H Klym, Jill L King, and Howard E Rockette. The prevalence effect in a laboratory environment: Changing the confidence ratings. *Acad Radiol*, 14(1):49–53, Jan 2007.

[12] Karla K Evans, Rosemary H Tambouret, Andrew Evered, David C Wilbur, and Jeremy M Wolfe. Prevalence of abnormalities influences cytologists' error rates in screening for cervical cancer. *Arch Pathol Lab Med*, 135:1557–1560, December 2011.