January 2014

Measuring Performance Of Physicians In The Diagnosis Of Endometriosis Using An Expectation-Maximization Algorithm

Susan Jin
Yale University, jinsu2190@gmail.com

Measuring Performance of Physicians in the Diagnosis of Endometriosis using an Expectation-Maximization Algorithm

Susan Jin

Master of Public Health Thesis

Biostatistics

Yale School of Public Health

Spring 2014

Abstract

The quality of clinical studies rests on the reliability of the disease diagnosis, and it is important to assess various factors associated with the ability of a physician to provide an accurate diagnosis. Endometriosis is a gynecological disorder in women, which has typically been difficult to diagnose and assess accurately. We focus on the analysis of data collected in the Physician Reliability Study of the *Eunice Kennedy Shriver* National Institute of Child Health and Human Development (NICHD), National Institutes of Health (NIH), on the agreement between physicians in obstetrics and gynecology in the diagnosis of endometriosis. In the study, 12 gynecologists of three levels of professional experience reviewed the surgical intrauterine images of 156 patients and provided a diagnosis for each patient. The objective of our analysis is to investigate the performance of the physicians in diagnosing endometriosis and examine whether there are statistically significant differences in average diagnostic performance among the three groups of gynecologists in the study: international academic experts, regional expert surgeons, and residents. Given the diagnostic rating of each physician expert for every patient (including missing diagnoses), we propose an expectation-maximization (EM) algorithm to infer the true patient disease status, and measure the performance of each physician in diagnosing endometriosis. This is achieved by estimating the true disease status, and then calculating the sensitivity and specificity of each physician rater in diagnosing the disease. The results show that, although there is a marked difference in performance among the physicians, there is no significant difference among the three different groups of experts. This approach can be broadly used to estimate the sensitivity and specificity of a diagnostic test, when the true disease status is not known.

Acknowledgements

This research uses data of the *Eunice Kennedy Shriver* National Institute of Child Health and Human Development (NICHD), National Institutes of Health (NIH). The Physician Reliability Study, part of the larger Endometriosis: Natural History, Diagnosis and Outcomes (ENDO) Study, and subsequent research projects have been funded and conducted by the NIH NICHD. I would like to express my gratitude in being given the opportunity to conduct this research. I would also like to thank all of the members of the NIH NICHD Division of Intramural Population Health Research (DIPHR) for their support and guidance during my internship in the summer of 2013 and on my research projects.

My work on this study has been conducted under the guidance of Dr. Zhen Chen and Dr. Aiyi Liu of the Biostatistics and Bioinformatics Branch (BBB) of the NIH NICHD and Dr. Hongyu Zhao, Professor of Public Health (Biostatistics) at the Yale School of Public Health. I would like to thank my mentors for everything that they have taught me and all of the help and guidance that they have provided me over this time.

Table of Contents

List of Tables

List of Figures

Introduction

Endometriosis is a gynecological disorder in women in which the cells that normally line the uterus appear and grow on other areas of the body outside of the uterus. The cells grow on the ovaries, fallopian tubes, outer surface of the uterus, bladder, and other areas of the body[1]. The disease commonly causes various symptoms of pain and possibly infertility. Endometriosis is a common health problem for women, and occurs in over five million women in the United States[2]. The disease is often diagnosed and staged through surgical visualization, in which physicians review operative images to make judgments about the staging and treatment of the disease[3]. However, the accurate diagnosis and staging of endometriosis is subject to considerable error. It has been suggested that many physicians find the disease difficult to diagnose and treat, and there is often disagreement on the diagnosis and treatment of endometriosis[3, 4].

In this thesis, we consider data collected in the Physician Reliability Study (PRS), a trial conducted by the *Eunice Kennedy Shriver* National Institute of Child Health and Human Development (NICHD), National Institutes of Health with the goal of studying the degree of agreement among physicians in obstetrics and gynecology in the diagnosis of endometriosis. The PRS investigated the reliability of the diagnoses given by different groups of gynecologists who were provided the same amount of patient information at a given setting[3, 5].

The Physician Reliability Study (PRS) is comprised of a random sample (n = 156) of women from the larger NICHD Endometriosis: Natural History, Diagnosis and Outcomes (ENDO) study[3]. For the PRS, 12 physicians in obstetrics and gynecology (OB/GYNs) were recruited to determine whether or not each of the 156 patients had endometriosis based on observation of intrauterine images of these women. The recruited physicians included 3 groups

of 4 each with different levels of professional experience. Of the recruited physicians, 4 were considered international academic experts (IE) in the field, 4 were regional expert surgeons (RE), and 4 were residents (RD). Each physician diagnosed the presence or absence of endometriosis in each woman. For each of the 156 patients, each of the 12 physicians gave a rating of either 1, if he/she thought the patient had the disease, 0, if he/she thought the patient did not have the disease, or 9, if he/she thought that a diagnosis could not be determined from the given information.

In this study, we aim to evaluate the performance of and the agreement between the three groups of physicians with different levels of professional experience in diagnosing endometriosis. This can be accomplished by estimating the sensitivity and specificity of diagnosis of each physician expert. However, a major challenge in statistical inference of sensitivity and specificity for our problem is that the true disease status of each patient is unknown to us. That is, we do not have the ground truth for each patient, and the diagnostic decision for any patient can be different based on different doctors. Therefore, the true disease status of each patient needs to be inferred from the observed data, along with the sensitivity and specificity estimates. In addition, in our analysis, we need to appropriately deal with the "missing" values of 9, the rating that the diagnosis could not be determined, because this rating reflects a decision made by the physician that can differ between physicians.

In order to utilize the undetermined ratings and account for the lack of a true disease status, we propose an expectation-maximization (EM) algorithm to obtain estimates of the probabilities that each rater will give a certain rating conditional on the estimated disease status of the patient[6, 7]. Then, based on the sensitivity and specificity estimates for each physician, we are able to assess their accuracy in diagnosing endometriosis. Our general methodology can

effectively evaluate the performance of raters when the true disease status is not known, and has broader applications to other scientific problems with similar set-ups.

Materials and Methods

Data Set

The data set consists of 12 physician experts who each gave a diagnostic rating to 156 patients. For each patient, each physician gave a rating of either 1, indicating that the patient has the disease, 0, indicating that the patient does not have the disease, or 9, indicating that the disease status could not be determined based on the given information. Thus, the data set is a 156 row by 12 column matrix comprised of 1s, 0s, and 9s.

Expectation-maximization Algorithm

We assume that the ratings given by each physician expert reflects on their degree of confidence in their diagnosis, where diagnoses of 1 or 0 suggest that they have more confidence in their decision and the undetermined ratings of 9 reflect uncertainty that lies between 1 and 0. We believe that the undetermined ratings are still informative, since different physicians made different determinations on whether or not a diagnosis could be made based on the same intrauterine images. In order to account for the undetermined ratings and unknown patient disease status, we will use an expectation-maximization (EM) algorithm[7] to obtain estimates of the probabilities that each rater will give any of the three ratings given the disease status of a patient. The following sections are the explanation of the EM algorithm.

Section 1 – Notation

We have the original data set, a 156 x 12 matrix X comprised of values of 0, 1, or 9. The entry X_{ij} of X denotes the diagnosis for patient *i* given by expert *j*.

$$X_{ij} = \begin{cases} 1 & \text{expert } j \text{ says patient } i \text{ has the disease} \\ 0 & \text{expert } j \text{ says patient } i \text{ does not have the disease} \\ 9 & \text{expert } j \text{ cannot determine whether patient } i \text{ has the disease} \end{cases}$$

where $i = 1, \ldots, 156$ indexes patients and $j = 1, \ldots, 12$ indexes physicians.

Furthermore, let T_i denote the *true* disease status of patient i, where $T_i = 1$ means that the patient has the disease and $T_i = 0$ means that the patient does not have the disease.

To examine how well each physician performed in relation to the others, the following probabilities are of key interest.

$$P_{jl1} = P(X_{ij} = l \,|T_i = 1)$$

$$P_{jl0} = P(X_{ij} = l \,|T_i = 0)$$

where $l = 1, 0, \text{or } 9$ rating.

In this notation, P_{j11} is the probability of physician *j* giving a diagnosis of the presence of endometriosis when the patient truly has the disease (true positive) and P_{j00} is the probability of the physician giving a diagnosis of the absence of endometriosis when the patient truly does not have the disease (true negative). These are the probabilities of expert *j* giving the correct diagnosis.

On the other hand, P_{j01} is the probability of the physician giving a diagnosis of no endometriosis when the patient truly has the disease (false negative) and P_{j10} is the probability of the physician giving a diagnosis of endometriosis when the patient truly does not have the disease (false positive). These represent the probabilities of the expert *j* giving the wrong diagnosis.

Finally, P_{j91} and P_{j90} are the probabilities of the physician being unable to determine the diagnosis when the patient truly has the disease and truly does not have the disease, respectively. These represent the probabilities of the expert *j* deciding that the diagnosis cannot be determined based on the provided information.

Since the true disease status for each patient is unknown, the EM algorithm will be implemented to estimate these six probabilities for each physician expert. First, the T matrix for the patient disease status is constructed.

Denote a 156 row x 2 column matrix T for the disease status of the patients, where the two column values in each row, t_{i0} and t_{i1}, will represent the disease status of the patient *i*. The value t_{i0} will be the probability of the patient not having the disease and t_{i1} will be the probability of the patient having the disease. When $t_{i0} = 0$ and $t_{i1} = 1$, the patient *i* has the disease with probability 1. When $t_{i0} = 1$ and $t_{i1} = 0$, the patient *i* does not have the disease with probability 1. The sum of the two probabilities, t_{i0} and t_{i1}, will always be 1 ($t_{i0} + t_{i1} = 1$).

Through successive iterations of the EM algorithm, the disease status of the patient will be calculated as the probability of having the disease (where $0 < t_{i1} < 1$) and the probability of not having the disease (where $0 < t_{i0} < 1$).

Section 2 – Initialization of T Matrix and Starting Probabilities

First, the patient status matrix T is initialized by counting the number of 1s and 0s in each row X_i (the ratings of all 12 experts for each patient) of X. The matrix T is initialized such that a 0 in the first column T_1 means that the patient does not have the disease and a 1 in the second column T_2 means that the patient has the disease. If more physicians determined that patient *i* has the disease than does not have the disease, that is, the number of 1s in the row X_i is greater than the number of 0s, then patient *i* is initially assigned as having endometriosis ($t_{i0} = 0$ and $t_{i1} = 1$). On the other hand, if the number of 0s in the row is greater than the number of 1s, the patient is initially assigned as not having endometriosis ($t_{i0} = 1$ and $t_{i1} = 0$). In the case that the number of 0s and 1s in the row is the same, the patient is initially assigned 0.5 in both columns of the T matrix, and neither has nor does not have endometriosis ($t_{i0} = 0.5$ and $t_{i1} = 0.5$).

Using the initial matrix T, the initial estimates of the probabilities P_{jl1} and P_{jl0} are obtained through the following equations:

$$P_{jl1} = \frac{\sum_{i,X_{ij}=l} t_{i1}}{\sum_{i=1}^{156} t_{i1}}$$

$$P_{jl0} = \frac{\sum_{i,X_{ij}=l} t_{i0}}{\sum_{i=1}^{156} t_{i0}}$$

Section 3 – Updating T Matrix with Probabilities P_{jl1} and P_{jl0}

Then, the initial estimates of the probabilities P_{jl1} and P_{jl0} are used to update the patient status matrix T. For patient *i*, we start with:

$$t_{i0}^{(0)} = \mathrm{P}(T_{i0} = 1) = \frac{1}{2}$$

$$t_{i1}^{(0)} = \mathrm{P}(T_{i1} = 1) = \frac{1}{2}$$

For expert 1, let $X_{i1} = l$, then we update t_{i0} and t_{i1} by:

$$t_{i0}^{(1)} = \mathrm{P}(T_{i0} = 1 \mid X_{i1} = l) = \frac{\mathrm{P}(X_{i1} = l \mid T_{i0} = 1)\mathrm{P}(T_{i0} = 1)}{\mathrm{P}(X_{i1} = l)} = \mathrm{P}_{1l0} t_{i0} / \mathrm{P}(X_{i1} = l)$$

$$t_{i1}^{(1)} = \mathrm{P}(T_{i1} = 1 \mid X_{i1} = l) = \frac{\mathrm{P}(X_{i1} = l \mid T_{i1} = 1)\mathrm{P}(T_{i1} = 1)}{\mathrm{P}(X_{i1} = l)} = \mathrm{P}_{1l1} t_{i1} / \mathrm{P}\left(X_{ij} = l\right)$$

The common denominator $\mathrm{P}(X_{i1} = l)$ is omitted, because the final t_{i0} and t_{i1} will be scaled to $t_{i0} + t_{i1} = 1$. Then the following updates are done.

Update with expert 1:

$$t_{i0}^{(1)} = \mathrm{P}_{1l0} t_{i0}^{(0)}$$

$$t_{i1}^{(1)} = \mathrm{P}_{1l1} t_{i1}^{(0)}$$

Update with expert 2:

$$t_{i0}^{(2)} = \mathrm{P}_{2l0} t_{i0}^{(1)}$$

$$t_{i1}^{(2)} = \mathrm{P}_{2l1} t_{i1}^{(1)}$$

For the update with each expert j, we have the formula:

$$t_{i0}^{(j)} = \mathrm{P}_{jl0} t_{i0}^{(j-1)}$$

$$t_{i1}^{(j)} = \mathrm{P}_{jl1} t_{i1}^{(j-1)}$$

The last update with expert 12 is:

$$t_{i0}^{(12)} = \mathrm{P}_{12l0} t_{i0}^{(11)}$$

$$t_{i1}^{(12)} = \mathrm{P}_{12l1} t_{i1}^{(11)}$$

We scale the final t_{i0} and t_{i1} so that the probabilities add up to 1, and the new values for T_i are:

$$t_{i0}^{(\text{new})} = t_{i0}^{(12)} / (t_{i0}^{(12)} + t_{i1}^{(12)})$$

$$t_{i1}^{(\text{new})} = t_{i1}^{(12)} / (t_{i0}^{(12)} + t_{i1}^{(12)})$$

This process is repeated for each patient *i*, such that we have a new, updated matrix of patient status $T^{(new)}$ which is still a 156 row by 2 column matrix.

Section 4 – Updating Probabilities P_{jl1} and P_{jl0} with T Matrix

With the new patient status matrix $T^{(new)}$, the probabilities P_{jl1} and P_{jl0} are updated.

$$P_{jl1}^{(\text{new})} = \frac{\sum_{i, X_{ij}=l} t_{i1}^{(\text{new})}}{\sum_{i=1}^{156} t_{i1}^{(\text{new})}}$$

$$P_{jl0}^{(\text{new})} = \frac{\sum_{i, X_{ij}=l} t_{i0}^{(\text{new})}}{\sum_{i=1}^{156} t_{i0}^{(\text{new})}}$$

Now the new probabilities $P_{jl1}^{(\text{new})}$ and $P_{jl0}^{(\text{new})}$ can be used again in the procedure described in section 3 to obtain a new T matrix. Then, the new T matrix will be used again in the

procedure described in section 4 to obtain new probabilities. Additional iterations are performed and this cycle will continue until $P_{jl1}^{(\text{new})}$ and $P_{jl0}^{(\text{new})}$ converge to the final probabilities.

The final estimates $P_{jl1}^{(\text{final})}$ and $P_{jl0}^{(\text{final})}$ will be our final estimates for the six probabilities for each physician expert.

Results

For the data that we analyzed, the probabilities stabilized after around 51 iterations. The final probability estimates are presented in Tables 1 and 2, as follow. Table 1 presents the probabilities of each physician giving every diagnosis, given that the patient has the disease. Table 2 presents the probabilities of each physician giving every diagnosis, given that the patient does not have the disease.

Table 1. Underlying status is the patient has the disease.

Physician	P(physician says no disease | patient does have disease)	P(physician says yes disease | patient does have disease)	P(physician says cannot determine | patient does have disease)
IE 1	0.06	0.94	0.00
IE 2	0.02	0.81	0.17
IE 3	0.02	0.87	0.11
IE 4	0.05	0.77	0.18
RE 1	0.00	0.78	0.22
RE 2	0.14	0.86	0.00
RE 3	0.14	0.72	0.14
RE 4	0.00	0.97	0.03
RD 1	0.03	0.52	0.45
RD 2	0.02	0.91	0.08
RD 3	0.07	0.81	0.12
RD 4	0.14	0.82	0.04

Table 2. Underlying status is the patient does not have the disease.

Physician	P(physician says no disease | patient does not have disease)	P(physician says has disease | patient does not have disease)	P(physician says cannot determine | patient does not have disease)
IE 1	0.86	0.12	0.02
IE 2	0.55	0.08	0.37
IE 3	0.59	0.02	0.39
IE 4	0.56	0.03	0.41
RE 1	0.46	0.15	0.39
RE 2	0.83	0.17	0.00
RE 3	0.70	0.03	0.27
RE 4	0.44	0.26	0.30
RD 1	0.21	0.09	0.71
RD 2	0.42	0.18	0.40
RD 3	0.55	0.11	0.34
RD 4	0.59	0.21	0.19

For almost every physician, the probabilities of giving the correct diagnoses, which are true positives in Table 1 Column 2 and true negatives in Table 2 Column 1, are higher than the probabilities in the other two columns, which represent either giving the incorrect diagnoses or being unable to diagnose. There is an observed difference in performance among individual physicians, such as the first resident, who has a large probability (0.71) of being unable to make a diagnosis.

To further compare the performance of each physician and physician group, we plot the true positive probabilities against the false positive probabilities and the true negative probabilities against the false negative probabilities. In the plots, purple dots represent international experts, blue dots represent regional experts, and green dots represent residents. The physicians who perform better are those whose points fall closer to the upper left corner, indicating that the rater has a larger true probability and smaller false probability.

Figure 1. Plot of the true positive probabilities against the false positive probabilities.

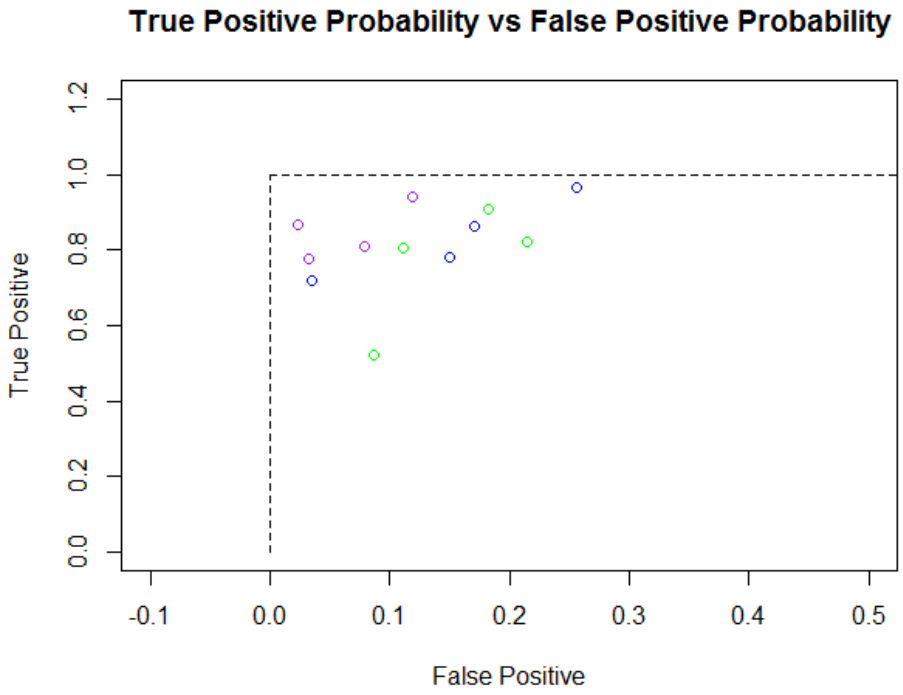
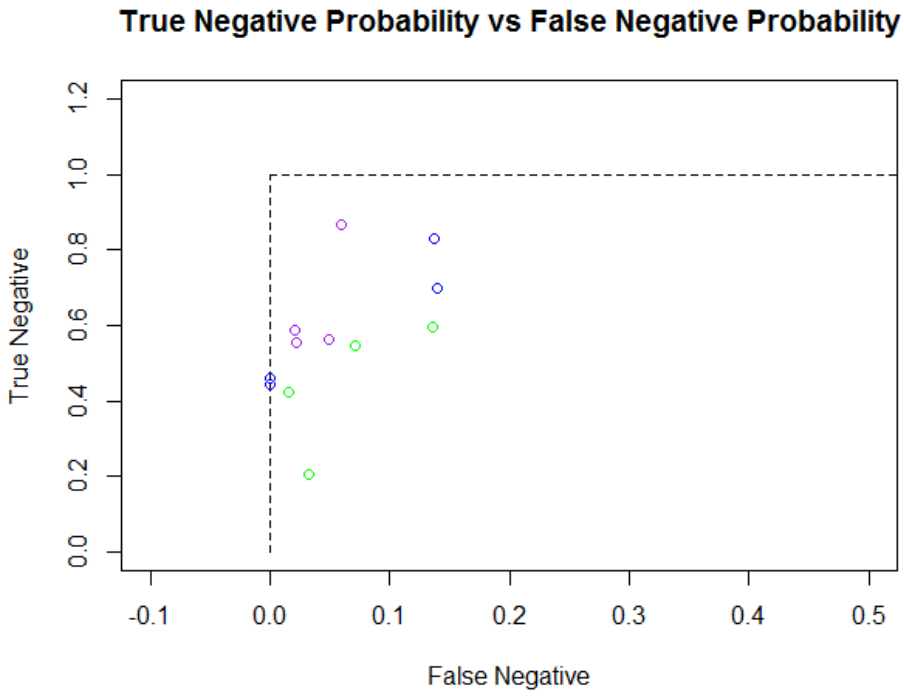


Figure 2. Plot of the true negative probabilities against the false negative probabilities.



Discussion

The results of the probability estimates show that there is an observable difference in performance among individual physicians, but, overall, there appears to not be a highly significant difference among the three different groups of physicians. Table 1 Column 2 represents the physician sensitivity and Table 2 Column 1 represents the physician specificity. In general, these probabilities are higher than the probabilities of the other columns, and the sensitivity and specificity of most physicians appears to be reasonably high. However, some physicians seem to have performed more poorly than others, such as resident 1, who has low sensitivity and low specificity, and has a high probability of being unable to make the diagnosis. In another example, regional expert 4 has high sensitivity, but lower specificity, and a higher probability of giving false positives.

Figure 1 and Figure 2 compare the performance of the three groups of physicians. In Figure 1, it appears that the international experts performed somewhat better than the other two groups, and the regional experts performed better than the residents. In Figure 2, the residents appear to have performed worse than the other two groups, but it is difficult to distinguish the performance of the international experts and regional experts. Since the undetermined diagnostic probabilities have not been incorporated into these plots, these interpretations should be taken with caution. There may be some difference between the three groups of physicians, but it does not appear to be particularly significant.

In conclusion, an expectation-maximization algorithm was used to estimate the sensitivity and specificity of twelve physicians that diagnosed endometriosis, without information of the true disease status of the patient. The EM algorithm seems to provide

reasonable estimates of the sensitivity and specificity of each physician. This approach has the potential to be applied to other problems with similar situations, where diagnostic ratings are available, but the true underlying status of the patients is not available. It can be broadly used to evaluate the performance of a diagnostic test, when the true disease status is not known.

References

1. Endometriosis: MedlinePlus. NIH National Library of Medicine. <http://www.nlm.nih.gov/medlineplus/endometriosis.html>.
2. Endometriosis fact sheet. Department of Health and Human Services, Office on Women's Health. <http://womenshealth.gov/publications/our-publications/fact-sheet/endometriosis.cfm>.
3. Schliep KC, Stanford JB, Chen Z, Zhang B, Dorais JK, Boiman Johnstone E, Hammoud AO, Varner MW, Louis GM, Peterson CM. Interrater and intrarater reliability in the diagnosis and staging of endometriosis. Obstet Gynecol. 2012; 120(1): 104-12.
4. Prentice A. Regular review: Endometriosis. BMJ. 2001; 323(7304): 93-5.
5. Xie Y, Chen Z, Albert PS. A crossed random effects modeling approach for estimating diagnostic accuracy from ordinal ratings without a gold standard. Stat Med. 2013; 32(20): 3472-85.
6. Zhang B, Chen Z, Albert PS. Latent class models for joint analysis of disease prevalence and high-dimensional semicontinuous biomarker data. Biostatistics. 2012; 13(1): 74-88.
7. Dempster AP, Laird NM, Rubin DB. Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society. 1977; 39(1): 1-38.