The Pennsylvania State University

The Graduate School

GENDER CONCORDANCE OF DOCTOR AND PATIENT DURING TOTAL BODY SKIN EXAMINATIONS

A Thesis in

Industrial Engineering

by

Michael Kassabian

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Master of Science

August 2022

The thesis of Michael Kassabian was reviewed and approved by the following:

Ling Rothrock
Professor, Industrial and Manufacturing Engineering
Thesis Advisor

Scarlett Miller
Affiliate Assistant Professor of Information Sciences and Technology
Assistant Professor, School of Engineering Design, Technology and Professional Programs and Industrial and Manufacturing Engineering

Steven Landry
Industrial and Manufacturing Engineering
Head of the Department

ABSTRACT

In a previous study out of the Penn State School of Medicine, 10 participants (5 faculty, 5 residents) were recorded performing Total Body Skin Exams (TBSE) on two healthy patients. These exams were analyzed for body segment order, duration of exam, and movements of both examiners and patients. From observations of the 20 resulting TBSEs, a standard procedure was developed that would have examiners start at the top of the patient and work their way down, examining the face, arms, torso, and legs before moving to the back and repeating the process. The idea was that the specific body segment order and short transition distances between segments would result in a more accurate, more efficient exam.

While often unconscious, gender bias in healthcare can have a large impact on a patient's quality of life. As such, it is important to determine what aspects of healthcare can be susceptible to these biases and the best way to prevent them. In this thesis, the effects of a training protocol on total body skin exams are analyzed to determine if the performance of a skin exam could be impacted by gender bias. Specifically, doctor-patient pairs were grouped based on gender agreement between the two, and exam time and accuracy were recorded. It was hypothesized that examiners untrained in this new protocol would perform better exams on those patients that had their same gender than those of another gender, but that no difference would be seen between the male and female patients of those examiners trained in the new exam protocol.

To test this new protocol, 31 participants were recruited to examine two model patients. Participants were then randomly assigned to either the control group (n=15) or the experimental group (n=16). Those in the experimental group were trained on the previously developed TBSE protocol, and those in the control group were not. All participants were instructed to complete a skin exam of the whole body. Examinations were recorded with special eye-ware (Tobii Glasses) that contain cameras and eye tracking technology. Together the footage of the exam and eye gaze

data show where on the patient's body participants were looking at any given time during the exam. Participants were analyzed on how many body segments were examined and for how long to get a total exam completeness proportion, and a total exam time.

The low sample size resulted in some low powered analyses. This reduces the strength of any conclusions drawn, but interesting relationships can still be gleaned from the data that could be used in further studies.

From the data collected, it was determined that female examiners in the control group performed significantly longer exams on female patients than on male patients but performed significantly more complete exams on male patients than the female patients. This could indicate a slight bias towards male patients. No other demographic of examiner showed a significant difference in exam time or completeness between male and female patients. This seemed to indicate that female examiners in the trained group were more consistent between patients than those in the control group.

Overall, these results show that participation in a training protocol was associated with reduced differences between male and female patients. That is to say that there was no significant difference shown between male and female patients of experimental group examiners where there was in the control group. More research needs to be done to improve the effects of a small sample size such as low power, but there are promising results so far.

If more subjects are included in future studies, the conclusions made here can be verified. Consistent TSBE results for the trained group of examiners is a good thing even if the original inconsistencies did not arise from gender bias. This method of implementing protocols for simple examinations could then be applied to other areas of healthcare while more long term, proven gender bias mitigation techniques are employed. It is unclear if any actual gender bias was improved here but providing consistent results across patients benefits everyone.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to acknowledge the help of my colleagues, and my advisor in the Human Performance Assessment and Modeling (HPAM) Lab, as well as my thesis readers for their advice and feedback on this paper. I would also like to acknowledge the work that researchers at Penn State's School of Medicine did in laying the groundwork for this thesis and recruiting the participants.

Chapter 1

INTRODUCTION

In an equitable healthcare system, every patient should get the care they need regardless of race, ethnicity, sex, gender, sexual orientation, socioeconomic status, or religion. This, however, is not always the case. Whether by direct prejudice or unconscious bias, not all those seeking medical services will receive the same quality of care. If this difference in quality of care is related to an individual's identity, it may be attributed to a bias on the part of the service provider. Discriminatory biases can be conscious differences in treatment based on an assessment of an individual’s worth (explicit bias) or automatic reactions based on positive and negative associations with an individual’s identity (implicit bias) [1]. One example of a discriminatory bias that can be explicit or implicit is gender bias.

Gender bias is the different and unjust treatment of one gender over another, based on gender identity or presentation [2]. Often in the context of healthcare, gender bias can also be used to refer to unfair treatment based on differences of biological sex [3]. This different treatment in healthcare can manifest in different ways. Gender blindness sees research performed on male participants with the assumption that male and female bodies will respond the same. Women may be excluded from drug studies for the sake of reproductive health, with results from male trials extrapolated to all users. This male only test sample leads to medications that are undertested on female users [3]. There can also be a strong bias for males in healthcare. Women’s diagnoses may take longer and be less accurate than men’s because they are sometimes not taken seriously by medical professionals [4]. Women’s diseases may also be seen as less important. In a Norwegian survey of over 800 medical professionals and students, subjects ranked coronary heart disease as more important than arthritis and fibromyalgia. The study suggested that the lower ranks of the latter two could have something to do with the fact that sufferers of these diseases are more likely to be women [5]. Male bias does not mean, however, that men will always receive better healthcare than women. Gender role ideology can also negatively impact rendered healthcare for men and women [3].

Men may not receive the full benefit of preventative care because they are not made aware of possible preventative screenings or preventative health habits [4]. Men’s reproductive and mental health care may also be lacking because women are more strongly associated with reproduction [3]. This can be seen in the lack of knowledge about how prostate cancer in men can affect mental health and gender identity [6].

These examples point to gender biases in healthcare negatively affecting both men and women. It has appeared in diagnoses, procedures, and examinations [4,7,8,9]. So, it is important to understand where shortcomings from these biases may appear in healthcare and how they can be identified.

In a previous study out of the Penn State School of Medicine, 10 participants (five students, five faculty) were observed performing total body skin exams (TBSEs) on a male and female patient. Researchers were interested in exam duration, exam completeness, patient movement, examiner movement, and time spent on transitions between segments. From observing the participants, a protocol detailing a specific order in which body segments should be examined was developed to improve accuracy and efficiency. The reduced transition distances between body segments (shown in figure 1-1) are meant to reduce exam duration and the specific order is meant to help examiners remember to check every segment thus improving exam completeness [10].

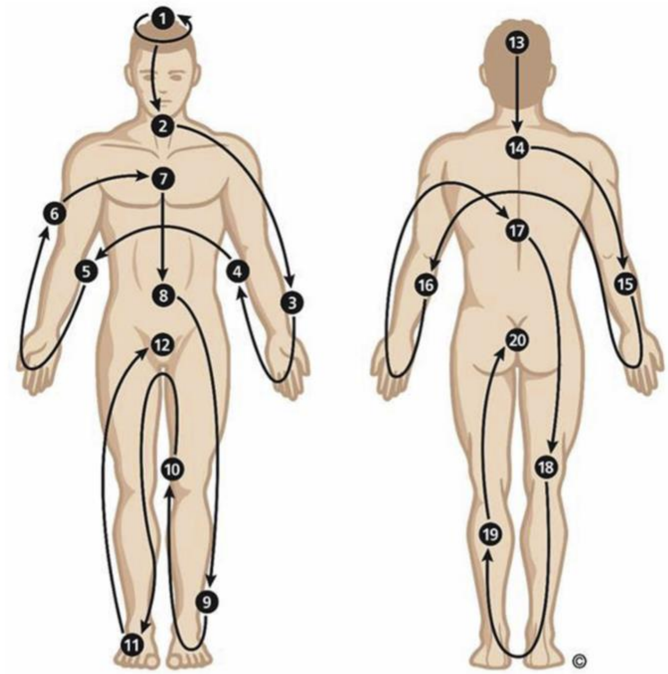

Figure 1-1: Exam order established by researchers at the Penn State School of Medicine. Examination begins at 1 with the scalp and ears and progresses through the numbered body segments to 20 [10].

In the summer of 2021, a follow up study was performed by the Penn State School of Medicine in collaboration with the Human Performance Assessment and Modeling (HPAM) Lab to test the speed and accuracy of their new examination order against a control. In this experiment, patient skin health was not recorded, nor were examiners expected to assess patient health. For this reason, the term accuracy will not be used as an assessment of examiners' ability to find skin abnormalities, but instead as a description of the number of body segments examined out of the total. The terms exam completion and exam accuracy will be used interchangeably in this thesis.

The school of medicine created the experimental design, recruited participants, and established the training technique to be used for the experimental group of participants. Members of the HPAM lab carried out the experimental design, being present for all tests to answer questions, collect data, and ensure the experiment was performed properly. 31 Participants (29 students, 2 medical professionals) were brought in to examine one male and one female model patient each. It was the researchers' goal to determine if there was an improvement in examiner performance in those participants that were taught this specific examination order.

It is the further goal of this thesis to determine what if any differences there may be in the collected speed and accuracy data that can be attributed to the gender interaction between doctor and patient. Additionally, if gender bias is established to be a factor in the performance of a TBSE, a secondary goal is to determine if the implementation of an exam protocol could reduce differences of speed and accuracy seen. A look at previous studies may shed some light onto the way that gender bias manifests in other areas of healthcare as well as ways to mitigate it.

Literature Review

Gender, and gender interactions may impact healthcare at varying levels. Individually, patients and examiners may view healthcare differently depending on their own gender [7,11,12,13]. However, gender interaction of the healthcare provider and the patient together may also impact the quality of care provided [7,8,9].

Patients' own view of health and healthcare may be colored by their gender [11]. In a sun health survey of Australian college students, a relationship was found in female students that correlated amount of sun exposure behaviors with amount of sun health and skin cancer knowledge. This relationship did not exist in male students. There also seemed to be less sun protection behaviors exhibited by male students [11]. This would seem to suggest that female students in the study were more aware of and concerned about preventative measures than the male students.

A study in parts of Eastern North America and Puerto Rico conducted over the phone, surveyed participants about behaviors and beliefs surrounding cancer screenings. Men in the study were found to have participated in cancer screenings less often than women in the study. They were also significantly less open to the idea of cancer screenings in the moment. It was only once specific scenarios were introduced detailing who would be performing the screenings and what they would entail that those men reacted significantly more positively [12]. The study suggested that the reason for these relationships may be that women are encouraged to participate is cancer screening much more often than men are, citing regular health checkups, ad campaigns, and other organizations like the NFL raising awareness.

Some findings also indicate a difference between how male and female physicians operate. In a longitudinal study of 509 patients over the course of a year, it was reported that female doctors tend to focus more on preventative measures and psychosocial counseling while male doctors focused more on getting medical history statements and performing physical exams. They also found that patients of female doctors were significantly more satisfied than those of male doctors [13].

Another study looking into satisfaction surveyed 10,205 patients about their experiences with their doctors' indicating types of services rendered, whether they chose their doctor themselves, and their satisfaction level on 5 metrics. On satisfaction, it was found that female patients who chose female doctors were the least satisfied whereas male patients of a female doctor were the most satisfied [7]. This would seem to indicate a difference in patient expectations depending on patient-doctor gender interaction. While both the gender of the patient and gender of the doctor may impact care, gender concordance or discordance may be a more complete way to view issues of gender bias. This study also found that female doctors focus more on preventative health [7]. This may have to do with the fact that the majority of patients that chose their own doctors were female and of those female patients, a majority chose female doctors. Since there are more expectations for women to get preventative health screenings this may lead to increased preventative health measures performed by female doctors.

A 1999 survey of medical students in their final year of med school found that students were more aware of and sympathetic to the concerns of patients of their same gender. They also reported feeling less comfortable performing more intimate examinations on those patients of the opposite gender [8]. Even in case studies of actual procedures, gender concordance/discordance between patient and healthcare provider saw significant effects on care. A study of 1,320,108 patients and 2,937 surgeons found that sex discordance between patient and surgeon was associated with a significant increase in adverse outcomes defined as readmission, complications, and death. The group that was most susceptible to these adverse outcomes was female patients of male surgeons [9].

Gender biases have found their way into health care at the levels of health behaviors promoted to patients, consultations, trust of patients in their doctors, and even in performance of certain procedures [11,7,9]. These biases may be unconscious, but they need to be unlearned if patients are to receive equitable care.

The first step to bias mitigation is bias awareness. After that, individuals need to address and begin to combat stereotypes by interacting with people from other backgrounds and intentionally including people with different background from their own in social and professional environments [14]. This can be a long process so quickly reducing some of the negative effects of gender bias would be very beneficial. One way to quickly combat unconscious biases in healthcare practices could be to put more specific procedures into place.

In a study of practicing nurses and their behavior related to protocols, more experienced nurses were found to rely more on their intuition than on established protocols. Where protocols were found to be the most helpful were with inexperienced nurses, in uncommon situations, to boost confidence, and to support decision making [15]. Having already established rules and standard operating procedures allows individuals to make decisions more confidently where a new or uncomfortable situation may cause doubt.

In the case of the follow up TBSE study to test the new examination order. The majority of the participants were students with only those who had been on rotation or who had learned during extracurricular activities having any experience with skin exams. For those students with little or no experience, they may feel less comfortable performing total body skin exams especially on patients of another gender as seen in the 1999 survey of senior med students [8]. Since gender bias can lead to longer and less accurate diagnoses [4], it is possible that unfamiliarity with skin exams on patients of the opposite gender may lead to longer and less accurate exams. These less experienced students participating in the follow up study may however be the perfect people to benefit from established protocols to improve speed and accuracy.

Research Objectives and Significance

The objective of this thesis is to determine what relationship exists between the gender concordance of doctor and patient, and the quality of care rendered during Total Body Skin Exams. The doctor patient groups can be split into four groups or dyads; male examiners and male patients, female examiners and female patients, male examiners and female patients, and female examiners and male patients. The first two dyads are the gender concordant dyads, and the last two are the gender discordant dyads. Based on the literature review, an especially intimate exam could lead to feelings of discomfort for gender discordant dyads. This could lead to slower, less accurate skin exams.

The first objective is to examine the relationship between speed and accuracy. A positive relationship would suggest that longer exams are associated with more accuracy and shorter exams are associated with less accuracy. The literature however suggests that gender bias may lead to a negative relationship with longer exams being associated with less accuracy.

The next objective is to determine if there are speed and accuracy differences between gender concordant and discordant dyads. While no specific test for gender bias was given to participants, a difference in the way an examiner treats their male and female patients could point to gender bias indirectly. If a gender difference does seem to point to a difference in speed and accuracy, the dyads can be broken into smaller groups of the treatment and control examiners.

The third objective would then be to determine if differences found in the second research question could be attributed to either the control or trained groups. If the trained group is shown to have fewer differences in speed and accuracy than their control group counterparts, then the new examination protocol could be useful for establishing a more equitable level of care.

Expected Contributions

The results of this thesis contribute to the ongoing exploration of gender bias in healthcare. This thesis is specifically focused on a total body skin exam and provides data on the impact gender of examiner and patient may have on accuracy and duration of skin exams. This thesis also aims to examine the impact that training may have on exam accuracy and duration in the context of different examiner and patient gender combinations. Results of this analysis will provide a better understanding of how gender interaction between doctor and patient affects quality of skin examination and how gender interaction is affected by training. This knowledge allows for a more educated approach to possible gender bias in health screenings and direction for future mitigation of bias in addition to the methods already in use. The experimental methods used here can be used for future studies both for a more in-depth study of gender interactions in skin examinations, and for an analysis of other types of examinations.

Document Outline

The remaining sections of this thesis will go into the methods used for data collection and analysis, results of those analyses, discussion of the results compared to expectations from literature, and a conclusion including limitations and areas of future study this thesis could lead to. Chapter 2 will cover experimental methods for data collection and data analysis. Chapter 3 will go over the results of the experiment, and the analyses used to address the research questions. Chapter 4 will discuss the results in the context of the literature review and expected outcomes. Chapter 5 will go over the limitations inherent in the experiment and future directions or verify the results and address those limitations.

Chapter 2

METHODS

The following chapter will detail the steps taken to collect participant data during the total body skin exam. All aspects of experimental design including recruitment, and training protocol were conducted by the Penn State School of Medicine. Data was collected and analyzed by researchers at the Human Performance Assessment and Modeling (HPAM) Lab at Penn State in State College.

Hypotheses

To get a baseline of how examiners will perform in this experiment regardless of the gender of their patient, an analysis of the relationship between exam duration and accuracy should be performed. Based on the researchers' logic going into this experiment, it was assumed that a longer exam would mean a more accurate exam as more time would allow examiners to view more of their patient's skin. However, literature on gender bias in healthcare seems to relate gender bias with longer but less accurate services [1,4]. Additionally, sources reviewed for this thesis found a link between gender discordance and a lower quality of care [8,9]. There was also literature suggesting that having protocols in place, like the new examination order developed by Penn State's School of Medicine, may help improve decision making in new or uncomfortable scenarios [15]. The following hypotheses were developed to address the research questions around Total Body Skin Exams (TBSEs).

H1: There will be a positive relationship between exam duration and accuracy of a TBSE

H2: The percent completeness of the TBSE for gender concordant doctor-patient pairs will be higher than for gender discordant pairs.

H3: TBSE duration for gender concordant doctor-patient pairs will be greater than the exam duration for gender discordant pairs

H4: Trained group examiners will not have significantly different TBSE percent completeness between their gender concordant and discordant patients.

H5: Trained group examiners will not have significantly different TBSE durations between their gender concordant and discordant patients.

Participants

To test these hypotheses, 31 participants were recruited from Penn State's School of Medicine, following the IRB approval. Of the 31 total participants, 10 were first-year medical students, eight were second-year medical students, four were third year, seven were fourth year, and two were medical professionals. While none of the participants had any experience with the improved TBSE order, the two medical professionals both had previous experience with skin exams in general. Of the students, none of the first-year students had any experience with skin exams. Second-, third-, and fourth-year students were more likely to have experience with skin exams as they progressed through their schooling, went on rotations, or had more extracurricular experience. To control experience level, experimental groups were constructed to attempt to have equal numbers of each level of experience.

Gender of participants was not able to be controlled for in the same way. The participants that were eligible and available for this study happened to comprise a group that had more women than men. Table 2-1 shows the number of male and female participants in each test group. A full breakdown of subject information can be found in the appendix.

Table 2-1: Number of male and female participants in the control and experimental groups

	Male	Female	Total
Control	3	12	15
Experimental	5	11	16
Total	8	23	31

Data Collection

The participants were randomly allotted into two groups; an experimental group (n=16) and a control group (n=15) (Table 2-1). Before beginning their examinations, those subjects in the experimental group were shown a video demonstrating a specific order in which the skin exam should be completed. Subjects in the control group were not given specific instructions on how to complete the exam. The participants then had one chance to examine each of the two model patients. The participants started the exam when they were ready and stopped the exam when they felt they had thoroughly examined their patient. For this experiment, an abridged exam order was shown to participants that excluded examination of the patients' groin and buttocks. Both patients wore undergarments, shorts, footwear, and gowns. Examiners could ask patients to remove or undo their footwear and gowns for sections of the exam, but patients were never asked to undress any further. Each of the 31 participants examined the same two models (one male and one female) for a total of 62 observations. The order of these two examinations was randomized for each participant to limit the effects of a learning curve.

Each examination was recorded in two ways. First, participants were asked to don a Tobii Glasses eye tracking system (Figure 2-1) which includes both a front facing camera and infrared sensors. The front camera recorded the examiners view during the exam, and the sensors tracked eye movement which provided eye tracking data. These two outputs from the Tobii glasses allowed the gaze of the examiner to be overlaid on the video of their examination. The point of focus was visualized on the video as a red circle. A second camera mounted in the corner of the room provided a third-person view of both the examiner and the model patient. This view was used to further investigate abnormalities in the Tobii footage, and to get a better sense of the movements of the examiner and patient throughout the exam.

Figure 2-1: Tobii glasses showing front facing camera and infrared sensors in the lenses

The footage with the eye tracking data was uploaded into a software package for qualitative data analysis called MAXQDA. In this program, the video was divided into timed segments and labeled based on what body part was being observed and notes detailing any missed sections. Researchers analyzing the data were not told which examiners were part of which experimental group to limit any possible biases in data reporting.

To analyze the above stated hypotheses, data from examiner-patient pairs were separated into four groups. These consisted of a female subject examining a female patient (n=23), a female subject examining a male patient (n=23), a male subject examining a female patient (n=8), and a male subject examining a male patient (n=8). Due to the large gap in male and female participant sample sizes, no statistical analysis will be performed comparing male and female examiners.

From these groups, pairwise t-tests were performed to compare the male and female patients of female examiners. Pairwise t-tests of male and female patients of male examiners were performed separately. These tests revealed if either group of examiners treated their patients any differently based on patient gender. To perform these tests all outliers were removed, and the data was tested for normality. To classify a point as an outlier, it must be three standard deviations away from the mean. Normality can be determined using a Shapiro-wilk test. Once all assumptions are met the test can be performed with $p<0.05$ as significant.

Chapter 3

RESULTS

This section contains the results of 62 skin exam trials as well as the results of the pairwise t-tests used to test this thesis's hypotheses. Full results tables are available in the appendix. While the first three hypotheses applied to both male and female examiners, male examiner sample size was too small to break into control and trained groups. So only the female examiners were used for the last two hypotheses.

Relationship Between Exam Duration and Exam Accuracy

The first hypothesis of this thesis is that examination duration and accuracy will have a positive relationship. This would mean that longer exams would be associated with more accurate exams. To test this hypothesis, exam duration was taken as the independent variable and was plotted against exam accuracy as a dependent variable. A scatter plot was created that could then be analyzed for its fit to different types of graphical relationships.

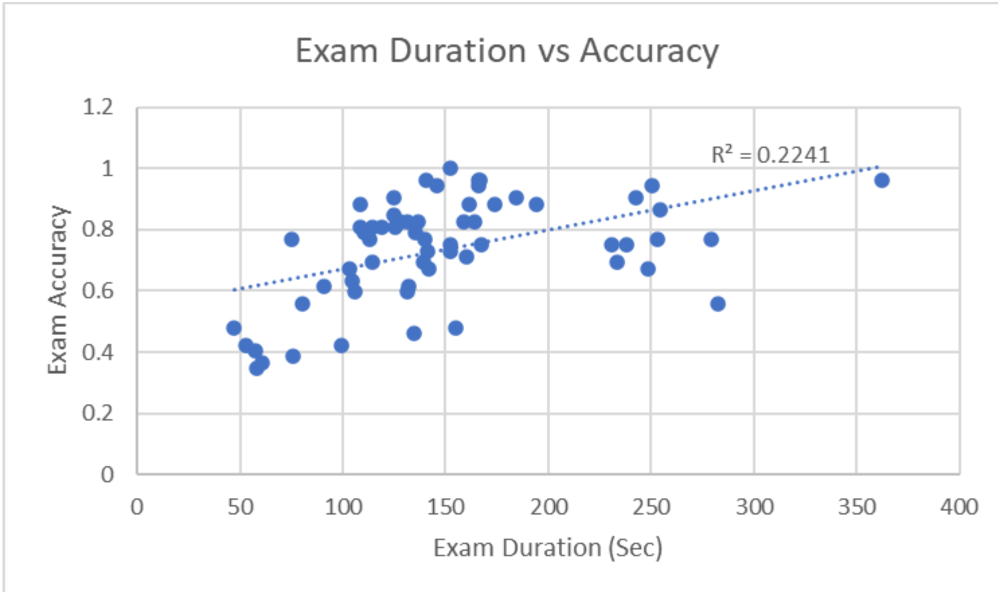

Figure 3-1: Scatter plot of all 62 examinations with exam duration vs accuracy

Figure 3-1 shows a scatter plot of all 62 examinations graphed together. There seems to be a positive relationship between duration and accuracy in general, but it doesn't appear to be very strong (R squared = 0.2241), and there are clusters of points that may indicate other variable effects. To further analyze this relationship, the exams were separated by gender of the examiner and graphed on two different plots.

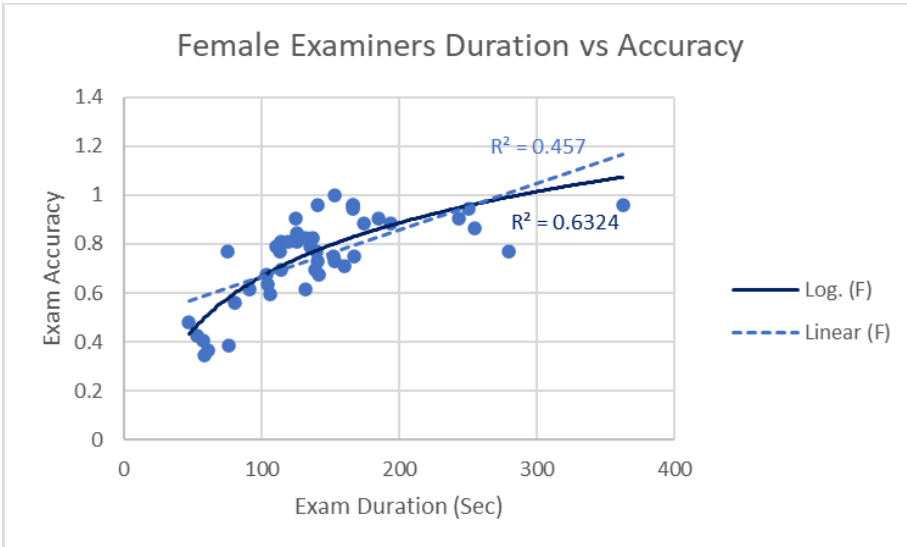

Figure 3-2: Exam duration versus accuracy for female examiners

In Figure 3-2, only the female examiners were plotted to see if a stronger relationship might be found if the variability of examiner gender was removed. In this plot both a linear trendline and a logarithmic trendline were compared. Since no examiner could go above 100 percent accuracy, there will naturally be an asymptote at y=1 on this plot. Here the linear model accounts for about 45% of the variability in exam accuracy, while the log model accounts for 63% of the variability. Both plots indicate positive relationships between exam duration and exam accuracy in female examiners.

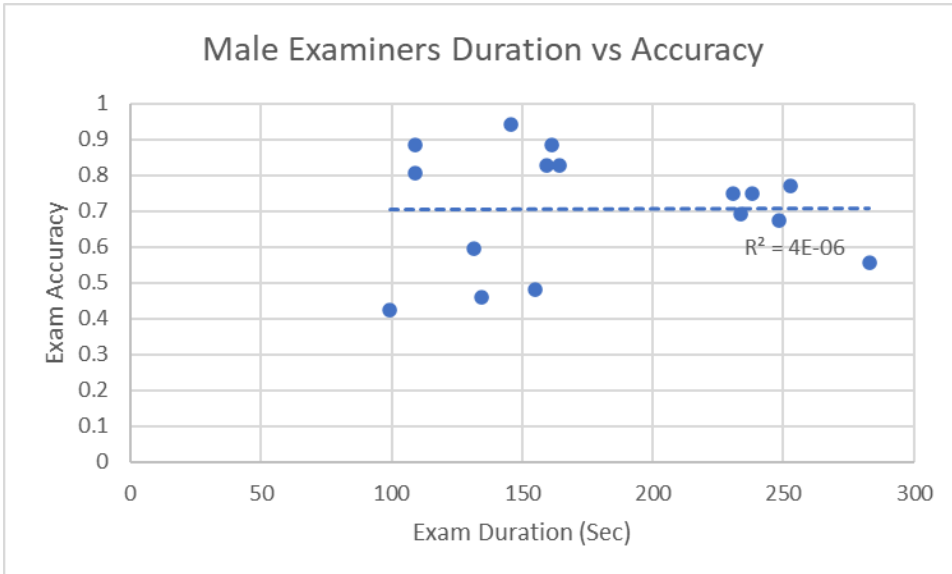

Figure 3-3: Exam duration vs accuracy for male examiners

Figure 3-3 is the plot created when only male examiner trials are plotted on the duration vs accuracy graph. Here there is no positive relationship between exam duration and accuracy in male examiners. The trendline had no positive slope and accounted for effectively 0% of the variability in the data set. These points are nearly randomly scattered in this plot and cannot be described with a positive linear model.

Exam Completeness for Gender Concordant and Discordant Dyads

The second research hypothesis was that gender concordant pairs will complete a higher percent of the TBSE than gender discordant pairs. In order to test this hypothesis, a pairwise t-test was used to observe the difference in percent exam completion between male and female patients of female examiners, and male and female patients of male examiners. For this hypothesis, male patient results were subtracted from female patient results to calculate the mean difference.

Table 3-1: Summary statistics and analysis of total exam completeness for female examiners

	Female Patient	Male Patient
N	23	23
Average Percent Complete	0.73	0.76
Standard Deviation	0.164	0.184
Mean Diff.	-0.037	
Pairwise t-test	0.01337 < 0.05	
Power	0.09	

The averages in Table 2-1 indicate that male patients of a female doctor received a more complete exam on average than female patients of a female examiner. This is also shown by the negative mean difference. The p-value was smaller than 0.05, so the null hypothesis is rejected. Power, however, is very low.

Another analysis was conducted to determine if a similar significant relationship could be explained by a difference in the examination of one or more body segments, but the significant difference in exam percent completion was only found between male and female patients of a female examiner at the total exam level.

Table 3-2: Male examiner total exam completeness and statistics

	Female Patient	Male Patient
N	8	8
Average Time (sec)	.704	.711
Standard Deviation (sec)	.19	.13
Mean Diff. (sec)	-0.00721	
Pairwise t-test	0.7966 > 0.05	
Power	0.03*	

*Sample size is too small for reliable power calculation

Table 3-2 shows summary statistics and the result of a pairwise t-test in the comparison of male and female patients of a male examiner. The female patient has a lower average exam completeness, but also has a higher standard deviation. The p-value was calculated at 0.7966, so the null hypothesis that the mean difference is zero cannot be rejected. The power for this analysis is also very low. No segment level analysis was performed for the male examiner group because the power level was already too small.

Exam Duration for gender concordant and discordant dyads

The third hypothesis was that gender concordant pairs would perform faster examinations than the gender discordant pairs. An initial test for normality showed that the data for total exam time was not normally distributed. Participant number 1 had the highest total exam time of any participant and the difference between their male and female patient times was more than 3 standard deviations away from the mean. Once this outlier point was removed the total exam time was shown to be normally distributed. Participant 1's data was analyzed for exam time bringing the sample size down to 22. However, in data sets that had no outliers and were shown to be normally distributed, participant 1 was kept.

Table 3-3: Female examiner total exam time summary statistics and statistical analysis

	Female Patient	Male Patient
N	22	22
Average Time (sec)	142.67	126.60
Standard Deviation (sec)	55.71	50.46
Mean Diff. (sec)	16.07	

Pairwise t-test	P = 0.009944 < 0.05
Power	0.17

Table 3-3 shows the summary statistics from the female examiner. The average total examination time was longer for female patients of a female examiner than for male patients. This is also shown by the mean difference of the two. A pairwise t-test returns a significant result, and the null hypothesis of a zero mean difference is rejected. The power of this analysis was reported at 0.17. To determine if this significant relationship could be caused by time differences on specific segments, the total exam time was broken into segment exam times and more pairwise t-tests were performed. Additionally, total exam time was broken into transition time, and effective exam time (excluding transition times).

Table 3-4: Time analysis of individual body segments

	Scalp	Left arm	back	Right anterior leg	Effective exam time
Mean difference (sec)	9.41	1.2	2.38	-2.27	11.19
Pairwise p-value	0.000413	0.0328	0.0252	0.014	0.0086
Norm p-value	**0.0089**	0.4064	**0.04844**	0.484	0.5983

Table 3-4 shows the individual items on the checklist that had significantly different times between male and female patients. The mean differences are all positive except for right anterior leg exam times. This indicates that more time was spent on male patients than on female patients for the right anterior leg. All segments listed have p-values less than 0.05 indicating a significant relationship. For the Shapiro wilks normality test, the difference in back and scalp examination times had a p-values less than 0.05. This indicates that the null hypothesis of a normal distribution is rejected. In this case, the normality of the data set can be analyzed by eye.

Analysis of Normality

T-tests are robust in the sense that a slight deviation from a normally distributed data set may still return valuable results. Since the Shapiro Wilks test for normality returned a p-value close to 0.05 for the back exam, a visual analysis of the data can be used to determine if the set is close enough to normal to still use the t-test results. The p value of the scalp was too far from 0.05 to warrant a visual check.

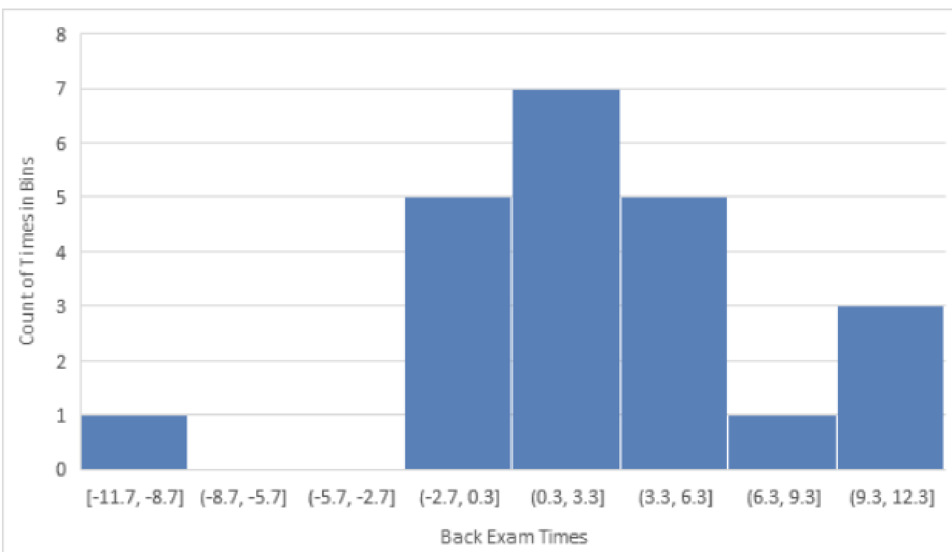

Figure 3-4: Distribution of back examination times

The distribution in Figure 3-4 shows that the data maintains a bell shape with a slight skew left. Since the distribution is unimodal and frequency tapers on either side a t-test can still be used with the caveat that ideal normality was not achieved. This allows the results from Table 3-4 to be maintained. So, it can be said that female examiners spent significantly more effective exam time on their female patients than their male patients, specifically on the left arm and back. However, female examiners were also found to spend significantly more time on male right anterior legs, than they did for their female patients.

Table 3-5: Male examiner total exam times and statistics

	Female Patient	Male Patient
N	8	8
Average Time (sec)	185.79	178.99
Standard Deviation (sec)	75.56	55.26
Mean Diff. (sec)	6.8	
Pairwise t-test	0.4753 > 0.05	
Power	0.04*	

*Sample size is too small for reliable power calculation

Table 3-5 shows summary statistics and the result of a pairwise t-test in the comparison of male and female patients of a male examiner. The female patient has a longer average exam time, but it also has a higher standard deviation. The null hypothesis that the mean difference is zero cannot be rejected. The power for this analysis is much lower than for the female examiners due to the smaller sample size so an analysis at the segment level was not performed for male examiners.

Exam Completeness for Female control and experimental groups

Of the female examiners recruited for this study, 11 were in the trained group and 12 were in the control group. The following tables contain data from the female examiner trials separated by experimental group.

The analyses from this data set will be pairwise t-tests comparing the male and female patients of a female examiner in the control group and a separate comparison of the male and female patients of a female examiner in the experimental group. The fourth hypothesis posits that the difference in percent exam completion found in the analysis of the second hypothesis may still be found in the control group, but that gender concordant and discordant pairs in the trained group will have no significant difference between them in the metric of percent exam completion.

Table 3-6: Summary table for female control and trained groups' percent exam completion

	Control Group		Trained Group	
	Female Patient	Male Patient	Female Patient	Male Patient
N	12	12	11	11
Average	0.729	0.752	0.726	0.775
Std. Dev.	0.14	0.17	0.189	0.198

Table 3-6 shows the average and standard deviation of percent exam completion for the male and female patients of female examiners in the control and trained groups. In both experimental groups, the average male percent completion is higher than for female patients, but the standard deviations create a large margin of error.

Figure 3-5 demonstrates the distributions of total exam completeness of male and female patients examined by control group and trained female examiners. The box plots all reach up to near 100% completion, but the lower extremes of the trained group box plots are more complete than those of the control group. The outliers shown here are not large enough to be removed.

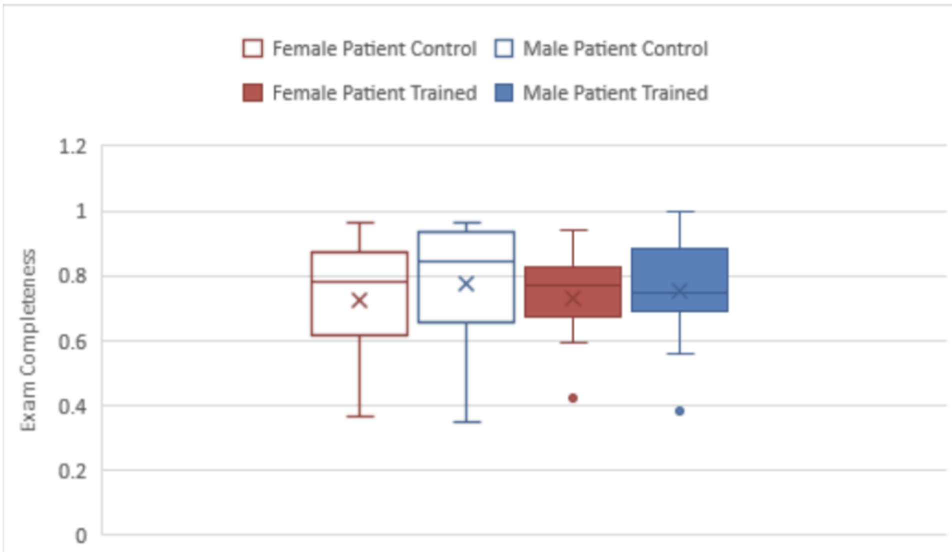

Figure 3-5: Female examiner box plots of total exam completeness by experimental group

Table 3-7: Summary statistics and analysis of total exam completeness for female examiners separated by experimental group

	Control Group		Trained Group	
N	12		11	
Patient	Female	Male	Female	Male
Average completeness	.726	.775	.729	.752
Standard Deviation	0.189	0.198	0.14	0.17
Mean Diff	-0.049		-0.023	
Pairwise t-test	.0485		.1545	
Power	.09		.05	

Table 3-7 shows summary statistics of the completeness results as well as the results of the pairwise t-tests. In both groups male patients had more complete exams on average, but not by much. This is also indicated in the negative mean difference. The control group has a larger magnitude mean difference and larger standard deviations than the trained group. With a p-value equal to 0.0487, the control group is shown to have a significant difference between male and female patients of a female examiner. There is no such relationship in the trained group. This relationship is significant in the control group but not in the trained group. Power is lower here than it has been in previous analyses. When completeness was analyzed at the body segment level there were no significant relationships found.

Exam Duration for female control and trained groups

The last hypothesis of this thesis is that trained groups of female examiners will have no significant difference between their male and female patients on the metric of exam duration. There may still be a significant relationship between the male and female patients of the control group of female examiners. To test this hypothesis pairwise t-tests were performed between the male and female patients of female examiners in the control group with separate t-tests for the patients in the trained group.

Table 3-8: Summary table for female control and trained groups' total exam duration

	Control Group		Trained Group	
	Female Patient	Male Patient	Female Patient	Male Patient
N	12	12	11	11
Average	153.24	130.49	132.11	122.72
Std. Dev.	72.37	66.71	32.17	29.41

Table 3-8 shows the summary statistics of the female and male patients in the control and trained groups. In both experimental groups, the female patients had longer exams on average. The difference in average exam time between the male and female patients is larger in the control group than in the trained group. However, the control group also has larger standard deviations than the trained group.

Figure 3-6 demonstrates the distribution of total exam times for all male and female patients of female examiners in the control and trained groups. While all four have low end exam times around a minute, control group exam times range up to over 250 seconds. Training group exam times only go up to around 175 seconds. The outlier in the group of male patients of a trained examiner is not large enough to remove from the set.

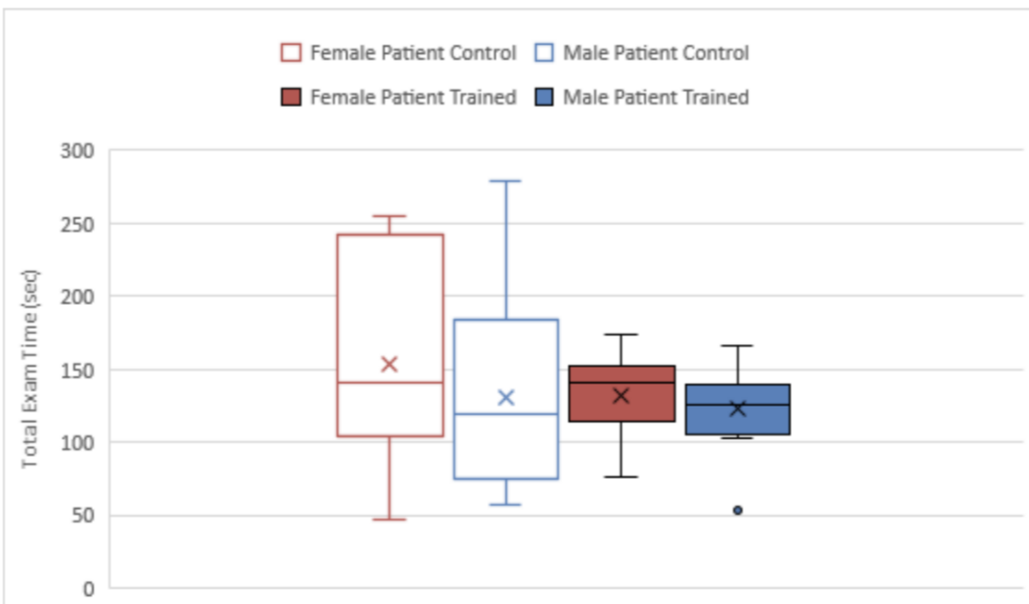

Figure 3-6: Female examiner box plots separated by patient gender and experimental group

Table 3-9: Female examiner summary statistics and analysis separated by experimental group

	Control Group		Trained Group	
N	12		11	
Patient	Female	Male	Female	Male

Average Time(sec)	153.24	130.49	132.11	122.72
Standard Deviation (sec)	72.37	66.71	32.17	29.41
Mean Diff (sec)	22.74		9.39	
Pairwise t-test	P = 0.0487 < 0.05		P = 0.1607 > 0.05	
Power	0.12		0.11	

Table 3-9 shows some summary statistics as well as the results of the pairwise t-tests. In both groups exam time for female patients is longer than for male patients. Relatedly both mean differences are positive. The control group has a larger mean difference and larger standard deviations than the trained group. With a p-value equal to 0.0487, the control group is shown to have a significant difference between male and female patients of a female examiner. There is no such relationship in the trained group

Chapter 4

DISCUSSION

Hypotheses

H1: There will be a positive relationship between exam duration and accuracy of a TBSE

H2: The percent completeness of the TBSE for gender concordant doctor-patient pairs will be higher than for gender discordant pairs.

H3: TBSE duration for gender concordant doctor-patient pairs will be greater than the exam duration for gender discordant pairs

H4: Trained group examiners will not have significantly different TBSE percent completeness between their gender concordant and discordant patients.

H5: Trained group examiners will not have significantly different TBSE durations between their gender concordant and discordant patients.

Hypothesis one is true for female examiners as a positive relationship was found between exam duration and accuracy but was not found to be true for male examiners. Hypothesis two was not true in any of the groups as the only significant difference found showed a higher completion rate for male patients of female examiners. Hypothesis three was found to be true for female examiners, but no significant difference in exam duration was found between male and female patients of male examiners. Both hypothesis four and five were true for female examiners. The control groups showed differences in exam duration and accuracy between their male and female patients, but no significant difference was found in the female examiners who had been trained with the new protocol.

Graphical Analysis

In figures 3-1 to 3-3, all 62 examinations analyzed in this thesis were plotted together to compare exam duration and accuracy. It was expected that there would be a positive relationship between these two metrics. This means that examiners who took longer to perform their exams were expected to have more accurate exams.

While a positive relationship was seen in the first figure, only about 22% of the variability in the data set was explained by the linear model. Additionally, the scatter plot seemed to be resolved into a few clusters of points. This led to two new plots being created to observe the relationship between exam duration and accuracy in male and female examiners separately.

Figure 3-2 contained all results from the 46 female examiner trials (male and female patients) plotted on the same axes as before. In this new plot, the linear model now accounted for almost 46% of variability in the data set. A logarithmic model was also added to this plot. Since it would be impossible to go past 100 percent accuracy no matter how much time examiners spent, there would naturally be an asymptote at y=1. A logarithmic trendline could replicate this asymptote and possibly be a better fit. Indeed, the log model accounted for 63% of the variability. Bothe the linear and log plots demonstrate that positive models do a good job representing the relationship between exam duration and accuracy in female examiners which matches the initial expectations for these skin exams.

Figure 3-3 is a plot of the 16 male examiner trials. Interestingly, there was no positive relationship seen in this plot and the linear trendline fit to these data accounted for almost none of the variability in exam accuracy. This goes against what was expected for these skin exams, but it is unclear why this might be the case. The male examiners did not have more experience than the female examiners, and they had similar average percent exam completion. The male examiners' average exam durations seemed to be shorter than their female counter parts, but it is unlikely that

that would cause such a drastically different scatter plot. It is possible that a positive relationship still exists between exam duration and accuracy, but that the small sample size of male examiners does not allow us to fully view that relationship. More experiments would need to be done with more male examiners to know for sure what type of relationship exists here.

Exam Times

The average total exam durations for all examiner groups were higher for female patients than for male patients. While this relationship was not significant for the male examiners (Table 3-5), female examiners did take significantly longer examining their female patients than their male patients (Table 3-3).

One explanation for this difference could be that the model female patient in this study provided a slightly different exam experience for the subjects than the male patient. The male patient was mostly bald which made for much quicker scalp exams. For the female patient, an effective scalp exam would require examiners to part through her hair. Additionally, while both model patients kept their shorts on for the entire exam, the female patient was also wearing a bra. Some subjects carried on with the exam only looking at skin that was already exposed. Other subjects would move parts of the bra strap around the top of the chest and the back to expose more skin for examination. This would increase their total exam time. In the lower extremities, somewhat of an opposite effect was noted. The male patient model was wearing socks for the exam and the female patient was wearing sandals. The sandals could easily slip off while the socks took a little more time. These challenges are reflected in the increased exam times for scalp, back, and lower extremities.

In Table 3-4 a breakdown of significant exam time findings at the segment level is shown. A significant increase in time is shown for female patients in areas of the scalp, back, and left arm with the left arm having the lowest mean difference and the scalp having the highest. More research needs to be done into the scalp relationship as the data set did not turn out to be normally distributed. On the other side, there was shown to be a significant increase in male exam time of the right anterior leg.

It is worthwhile to note that effective exam time (total time – transition time) was found to be significantly longer for female patients of a female examiner than for male patients as well, which accounts for the significant relationship found at the total exam time. However, a significant relationship was not found for transition time, so while moving bra straps and removing socks may increase a subject's transition time, it did not factor into the significant over all exam time findings.

It is possible that the reason these relationships exists for female examiners and not for male examiners is that male examiners did not feel comfortable performing the more intimate aspects of the skin exam on female patients leading to less of a difference in time between their male and female patients [8]. Conversely it could be possible that female examiners were more particular about their exams with their female patients than the male examiners or with their male patients [8].

Based on the literature, there was no mention of female doctors treating patients differently based on gender and male doctors having no difference between their patients. Infact, it was found in some cases, that female patients of male doctors had the worst outcomes [9]. So, it is surprising that the female examiners would be the only ones showing a difference in exam time between their patients.

The difference in exam time see in female participants was predicting in the original hypotheses though it was expected for male and female participants. Hypothesis three stated that exam duration would be longer for gender concordant doctor-patient pairs. This hypothesis was based on the assumption in the first hypothesis that there would be a positive relationship between exam duration and accuracy. A longer more accurate skin exam would be ideal as it would have the best chance of catching a skin abnormality, and gender concordant pairs were expected to receive the best examinations [4,8,9]. So, if exam accuracy in female patients of female examiners remained comparable to or better than in male patients, then these findings would match what was found in the literature. To better understand the quality of examination given by these participants, exam duration needs to be analyzed in conjunction with exam completeness.

Exam Completion

In the completion metric, male examiners appear to have no difference between male and female patients (Table 3-2). However female examiners performed significantly more complete exams on male patients than on female patients (Table 3-1). That the female examiners performed more complete exams on average for their male patients is surprising because it is the opposite of what was expected from the background literature [8,9]. It is difficult to do any follow up analysis from this result because the significant relationship does not show up during any segment exams. However, the lower exam completion rates in female patients of female examiners coupled with the same group having longer examination times does match up with a different pattern discussing in the literature.

Gender bias was found to cause diagnoses in women to take longer than men, and to be less accurate than for men [4]. This thesis assumed that longer less accurate diagnoses for women, may also translate to longer less accurate examinations. If this were the case, then these experiments would seem to identify a very real difference in how men and women are examined, possibly caused by gender bias. It was expected that since this was a more intimate exam that those participants examining a member of their own gender would be more comfortable with, and more sensitive to the patient's issues [8]. However, it is possible that gender bias could still negatively affect the examination quality of female patients even from female examiners [7].

Without performing surveys to better understand possible innate gender biases from the participants, it cannot be said what decisions were impacted by gender bias. The results here do show it as a possibility, however. Future studies may benefit from including an initial gender bias survey in data collection.

Control vs Trained Groups

Since the male examiner group was too small to split up, control and trained groups can only be compared in female examiners. An initial look at the box plots (Figure 3-5, 3-6) and standard deviations (Table 3-7, 3-9) seems to show that trained examiners are more consistent in their treatment of male and female patients than the control group. Standard deviations for total exam time are larger in the control group than the trained group and the box plots of the trained group cover a smaller range of times than the control group. This relationship is somewhat present in the exam completion data, but to a lesser extent.

Female examiners in the control group took significantly longer to examine female patients than male patients and performed significantly less complete exams on female patients. However, those in the trained group, took the same amount of time on both patients and exam completion rates were not shown to be different between male and female patients.

From the literature, there was evidence that employing a protocol can in some cases help improve confidence in decision making especially for inexperienced examiners [15]. While it cannot be said for sure whether gender bias caused female patients of female examiners to have lower quality exams, the combination of longer less accurate exams is certainly not ideal. The result of having no significant difference between male and female patients in the trained group, where a difference once existed is a plus, and seems to corroborate some of the literature findings.

Chapter 5

CONCLUSION

From the 62 total body skin examination (TSBE) trials that came out of the Penn State College of Medicine, five hypotheses were posited and analyzed within this thesis. The first hypothesis, that there would be a positive relationship between TBSE duration and accuracy appeared to be true for female examiners but could not be accepted for male examiners. The second hypothesis that gender concordant doctor-patient pairs would perform more complete exams was never reflected in the data. Male examiners showed no difference in exam completion between their male and female patients, and female examiners had worse completion rates for patients of their same gender than with male patients. This reduced accuracy for female patients of female examiners coupled with the slower exam time points to a possible bias for male patients during skin exams. However, female examiners in the trained group seemed to perform more consistent exams between their male and female patients and no significant difference was found in exam time or accuracy.

More tests would need to be performed to establish the presence of actual innate gender bias, but the use of a protocol such as the one developed in this experiment shows promising potential for keeping TBSE quality consistent between patients of different genders. Future tests designed to explore the use of protocols in known environments of gender bias could open them up to mitigate gender bias in other fields of medicine allowing more people to experience equitable healthcare.

Limitations and Future Directions

There were some limitations in this study that weakened the results such as sample size limitations, effectiveness of the training program, lack of examination quality metrics, and representation limitations. Small sample sizes limit the power of statistical analyses and put more importance on the data being normally distributed. More subjects would increase the power and if there was a big enough sample size might make the statistical analyses more robust. The low sample size also makes the difference between the male and female examiner group sizes more apparent. If there were more male examiners to match the female examiner numbers, then statistical comparisons between the two would be easier to perform.

The training program, while it was developed to optimize the skin examination process, can only be effective if the participants learn it properly. In this study, subjects in the experimental group were only shown a video of the proper procedure before they tried their hand at it themselves. If the subjects were given a chance to practice for the exam, then the results may have been more telling. For this to be implemented the control group would also have to be given some chances to practice giving a skin exam.

A limitation in the analysis of this training program is that examiner performance was not recorded past a list of examined body segments. From this data it is not possible to analyze how well each segment was examined as the only metric available was time. In future studies, patients could be assessed prior to the experiment to note what skin abnormalities warrant further investigation. Then when participants perform the skin exam, the number of abnormalities noted better reveals the quality of the skin exam.

As far as representation goes, this analysis of the study did not take race, ethnicity, or gender outside of the gender binary into account. The race and ethnicity of the participants was not collected. But there could very well be differences in gender interactions depending on the culture a participant was raised in. Some cultures may be more or less comfortable performing certain types of exams if there is gender discordance. This analysis also did not consider racial interactions between the doctor and the patient. While the focus was on gender bias, racial bias could look very similar and may be hard to distinguish.

More research needs to be done to account for the small sample sizes of this study. If it is true that use of a training program helps keep examiners more consistent, then that could be very useful in the face of gender bias. However, with such low power there is a high chance for errors. In the future, the same, or a similar study could be performed with a large sample size where all subjects get time to practice performing the skin exam. Most of the participants were students with little to no experience. Giving them this extra practice would help to better see the effects of the training program and not just the effects of little experience. Having a large sample size would also limit the importance of normality and increase power, allowing any findings to be taken more seriously.

There are a lot of promising results coming out of this study. Even with the small sample size here there seems to be a major improvement in the performance of skin exams in gender concordant and discordant pairings.

REFERENCES

1. Jed Writing Team, "Understanding Discrimination and Bias", n.d., The Jed Foundation. Accessed https://jedfoundation.org/resource/understanding-discrimination-and bias/#:~:text=Discrimination%20is%20the%20unjust%20treatment,us%2C%20we%20treat%20them%20differently.
2. Wex Definitions Team. "Gender Bias". Legal Information Institute. June 2020. https://www.law.cornell.edu/wex/gender_bias#:~:text=Gender%20bias%20refers%20to%20a,real%20or%20perceived%20gender%20identity.
3. Verdonk, P., Benschop, Y.W.M., de Haes, H.C.J.M. *et al.* From gender bias to gender awareness in medical education. *Adv in Health Sci Educ* **14,** 135–152 (2009). https://doi.org/10.1007/s10459-008-9100-z
4. J. Leonard "Gender Bias in Medical Diagnosis," Medical News Today, Jun. 2021. Accessed on: Mar. 2, 2022 [Online] Available: https://www.medicalnewstoday.com/articles/gender-bias-in-medical-diagnosis
5. Album, D., & Westin, S. (2008). Do diseases have a prestige hierarchy? A survey among physicians and medical students. *Social Science and Medicine, 66,* 182–188.
6. Kiss, A., & Meryn, S. (2001). Effect of sex and gender on psychosocial aspects of prostate and breast cancer. *British Medical Journal, 323,* 1055–1058.
7. J. Schmittdiel, et al. "Effect of Physicians and Patient Gender Concordance on Patient Satisfaction and Preventive Care Practices" Journal of General Internal Medicine, 2000. Accessed on Mar. 10, 2022. [Online]. Available: https://doi.org/10.1046/j.1525-1497.2000.91156.x

8. G. Zaharias, L. Piterman, M. Liddell, “Doctors and Patients: Gender Interation in the Consultation,” Acad Med. Feb. 2004, Accessed: Mar. 10, 2022. [Online]. Available doi: : 10.1097/00001888-200402000-00011. PMID: 14744716.

9. C. J. D. Wallis, et al. “Association of Surgeon-Patient Sex Concordance with Postoperative Outcomes,” JAMA Surg. Feb. 2022 1;157(2):146-156. Accessed Mar. 14, 2022. [Online]. Available doi: 10.1001/jamasurg.2021.6339. PMID: 34878511; PMCID: PMC8655669.

10. M. F. Helm, “Optimizing the total-body skin exam: An Observational Cohort Study,” J Am Acad Dermatol, Nov. 2019. Accessed Mar. 14, 2022. [Online]. Available: https://doi.org/10.1016/j.jaad.2019.02.028

11. A. K. Day, et al. “The Skin Cancer and Sun Knowledge (SCSK) Scale: Validity, Reliability, and Relationship to Sun-Related Behaviors Among Young Western Adults”. Health Education & Behavior. 2014;41(4):440-448. Accessed Mar. 2, 2022 [Online]. Available doi: 10.1177/1090198114529129

12. J.L. Davis, et al. “Gender differences in cancer screening beliefs, behaviors, and willingness to participate: implications for health promotion,” Am J Mens Health, May 2012. Accessed Mar. 8, 2022. [Online]. Available doi: 10.1177/1557988311425853. Epub 2011 Nov 8.

13. K. D. Bertakis, “The influence of gender on the doctor-patient interation,” Patient Educ Couns. 2009 Sep;76(3):356-60. Accessed: Mar. 14, 2022. [Online]. Available doi: 10.1016/j.pec.2009.07.022. Epub 2009 Aug 3. PMID: 19647968

14. Jasmine R Marcelin, Dawd S Siraj, Robert Victor, Shaila Kotadia, Yvonne A Maldonado, The Impact of Unconscious Bias in Healthcare: How to Recognize and Mitigate It, *The Journal of Infectious Diseases*, Volume 220, Issue Supplement_2, 15 September 2019, Pages S62–S73, https://doi.org/10.1093/infdis/jiz214

15. Nibbelink, C. W., & Brewer, B. B. (2018). Decision-making in nursing practice: An integrative literature review. Journal of Clinical Nursing, 27(5-6), 917–928. doi:10.1111/jocn.14151

16. A. C. Schieber, et al. "Do gender differences affect the doctor–patient interaction during consultations in general practice? Results from the INTERMEDE study, Family Practice, vol. 31, iss. 6, pg. 706–713, Dec. 2014, Accessed Mar. 13, 2022. [Online]. Available: https://doi.org/10.1093/fampra/cmu057

Appendix

FULL DATA SETS

Table A-1: Female examiner times and percent completeness

			Total Exam Time		Total Exam Completeness	
Participant No.	Experimental Group	Experience Level	Female Patient	Male Patient	Female Patient	Male Patient
1	Control	3rd Year	362.4	166.4	96%	96%
2	Trained	2nd Year	152.6	166.4	100%	94%
5	Control	MD	141	75.3	73%	77%
6	Trained	1st Year	140.7	136.7	96%	83%
8	Trained	1st Year	135.3	112.9	79%	77%
9	Control	2nd Year	167.2	132.2	75%	62%
10	Trained	4th Year	174	125.2	88%	85%
11	Control	4th Year	128.1	110.5	83%	79%
12	Control	1st Year	58.3	57.3	35%	40%
13	Trained	2nd Year	152.6	139.3	73%	69%
14	Trained	2nd Year	160.5	141.9	71%	67%
15	Control	1st Year	166.7	131.7	96%	83%
17	Control	2nd Year	250.7	279.3	94%	77%
18	Control	4th Year	124.8	119.1	90%	81%
19	Trained	2nd Year	114.5	103.4	69%	67%
20	Control	2nd Year	254.4	194.1	87%	88%
21	Control	1st Year	47.1	60.7	48%	37%
22	Trained	1st Year	80.6	106.1	56%	60%
24	Control	4th Year	104.5	90.7	63%	62%
26	Trained	4th Year	152.4	139.6	75%	77%

28	Trained	3rd Year	114.2	125.4	81%	81%
29	Control	4th Year	242.8	184.5	90%	90%
31	Trained	1st Year	75.8	53	38%	42%

Table A-2: Male examiner times and percent completeness

			Total Exam Time (sec)		Total Exam Completeness	
Participants	Experimental Group	Experience Level	Female Patient	Male Patient	Female Patient	Male Patient
3	Trained	4th Year	108.8	108.8	88%	81%
4	Trained	MD	252.8	230.6	77%	75%
7	Control	2nd Year	282.8	248.3	75%	69%
16	Trained	2nd Year	282.8	248.3	56%	67%
23	Control	1st Year	154.8	131.6	48%	60%
25	Trained	3rd Year	159.2	164.3	83%	83%
27	Trained	1st Year	145.6	161.3	94%	88%
30	Control	3rd Year	99.5	134.4	42%	46%