Analysis of spatial discrimination in the lumbar spine of normal man

Eliza Leja

Supervisor: Adam Taube
Acknowledgements

I would like to thank my supervisor Adam Taube for guiding me throughout the thesis writing process. I also would like to thank Bo Nyström and Birgitta Schillberg for providing the data and giving the opportunity to be present during the procedure at the spine surgery clinic in Strängnäs. Last, but not least, many thanks to my English coach Tom Homewood.
Abstract

A clinical study was performed in order to determine if healthy test subjects can differentiate between adjacent and separated pairs of vertebrae in the lumbar spine. The variable of interest was number of correctly specified pairs of vertebrae. The test subjects were evaluated in terms of sensitivity and specificity of this test. Bootstrap resampling was applied in the data analysis. The results clearly indicated that the test subjects in this study were able to successfully determine whether a pair of adjacent or separated vertebrae was tested during the procedure.

Key words: lumbar spine, spinal disc, adjacent vertebrae, separated vertebrae, sensitivity, specificity, efficiency.
List of Figures

1.1 Spinal disc structure. [6] ............................................. 4
1.2 Spinal anatomy. [4] .................................................. 5

2.1 Vertebrae in the lumbar spine and coccyx region. [4] [8] ........ 8
2.2 AUC of a discrete classifier. ....................................... 15
2.3 Number of times a pair is correctly specified under the null hypothesis. 18
2.4 Power curves as a function of sample size for standard deviations
   1.73, 2.45 and 3. ...................................................... 20
2.5 Power curves for a one-sided t-test with SD = 1.73, SD = 2.45 and
   SD = 3. ................................................................. 21

3.1 Position of the needles during the procedure. ....................... 22
3.2 Study protocol. ...................................................... 23

4.1 Number of correctly specified pairs. ............................. 24
4.2 Observed sensitivity (r) and specificity (s) in the test procedure. .. 25
4.3 Observed efficiency (r + s). ....................................... 27
4.4 Incorrectly specified pairs. ........................................ 28

5.1 Number of correctly specified pairs. ............................. 29
5.2 30000 resampled means of number of correctly specified pairs of vertebrae. ......................................................... 30
5.3 Observed distribution of number of times correctly specified pair. .. 31
# List of Tables

2.1  *All pairs of adjacent (a) and separated (s) vertebrae A, B, C and D.*  
2.2  *Pairs of adjacent and separated vertebrae A, B, C and D.*  
2.3  *Frequency summary of pair specification for a single test subject.*  
2.4  *Expected frequencies for n pairs of vertebrae.*  
2.5  *Frequency summary of pair specification for a single test subject.*  

5.1  *Frequency summary of the area under the ROC curve for the 18 test subjects.*  

7.1  *Power calculations for deciding the size of the study for sample sizes N = 10 and N = 15.*  
7.2  *Observed number of correctly specified pairs and calculated sensitivity, specificity, and efficiency for 18 test subjects.*  
7.3  *Sensitivity, specificity, efficiency, and area under ROC curve with corresponding 95% confidence interval for 18 test subjects.*  
7.4  *Incorrectly specified pairs for each of the 18 test subjects.*  


Chapter 1

Introduction

1.1 Background

This clinical study is concerned with severe chronic low back pain originating from a vertebral disc. Spinal discs function as ligaments that hold the vertebrae of the spine together and primarily act as shock absorbers between adjacent vertebrae. The discs are constructed of a soft inner core (nucleus pulposus) containing a loose network of fibres suspended in a mucoprotein gel, enclosed by a harder outer layer (annulus fibrosus) composed of concentric sheets of collagen fibres, see Figure 1.1. Due to natural ageing, intervertebral discs degenerate by dehydrating and becoming more stiff. In some cases this degeneration results in pain. Another factor as for example a twisting injury can damage the disc and lead to degeneration. Pain occurring in a damaged vertebral disc can last for a long time. Painful discs may, however, also occur without disc degeneration. [5]

![Spinal disc structure.](image)

*Figure 1.1: Spinal disc structure. [6]*
The spinal column consists of twenty three vertebral discs of which five are located in the low part of the back, i.e. lumbar spine, shown in Figure 1.2 [5]. The lumbar spine area is the focus of this study.

During the past 10 years many studies have been comparing conservative (physiotherapy, exercise, cognitive behaviour therapy) and surgical (arthrodesis) treatments of patients with severe chronic low back pain. In most studies the results of both types of treatments were the same. Arthrodesis suffers from two major problems: patient selection and locating the origin of pain. It is important that the patient in question experiences pain originating from a vertebral disc, and that in turn it can be determined in which disc the pain is occurring, in order to perform arthrodesis. Locating the origin of pain is problematic since an aching spinal disc can not be identified through X-ray or MRI examination. [10]
The purpose of this study is to investigate whether a particular method of examining the lumbar spine region could be applied to patients in order to locate an aching spinal disc. The idea is that if low back pain is indeed originating from a spinal disc, the exact location of the disc could be determined by provoking it through gentle tapping the vertebrae. In conjunction with the underlying problem the first basic question is whether healthy test subjects are able to distinguish pairs of adjacent vertebrae from pairs of separated vertebrae.

1.2 Similar studies

Searching for earlier studies of this specific type of procedure wasn’t very straightforward. This type of method for examining the lumbar spine in order to detect an aching disc doesn’t seem to have been performed before. However, there are many previous studies describing the method of two-point discrimination tests performed on other body areas such as the fingertips, arms etc.

In an article by Michael F. Nolan, two-point discrimination sensitivity was evaluated. The main objective was to examine if there is a difference in test results between healthy, young adult men and women. The sensitivity limits were determined in 11 skin regions: upper-lateral arm, lower-lateral arm, mid-medial arm, mid-posterior arm, mid-lateral forearm, mid-medial forearm, mid-posterior forearm, over 1st dorsal interosseus muscle, thumb, long finger and little finger. During testing each area, the interpoint distance was increased or decreased until the test subjects were consistently able to correctly determine whether they had been touched by one point or two points simultaneously. Also, the smallest distance at which reliable differentiation was made between application of one or two points was noted for each of the test subjects. The results showed that there were no significant differences in two-point discrimination tests between men and women in the different skin areas of upper limb, except for the skin of the medial surface of the forearm. In this test, the results were significantly better ($p < 0.005$) for women than men. [9]

The basics of these tests are similar but not the same as the medical trial in this thesis. In the article mentioned, two spots on the skin are stimulated simultane-
ously whereas in our testing the two vertebrae are tapped one after the other. In the article, the correct responses for each individual were observed, and in our trial the number of correctly specified pairs was of interest. Then the mean values and corresponding mean deviations were calculated. The Student’s $t$-test was used to test for two-point sensitivity differences in skin areas of the upper limb between men and women. This is not the main objective in our study, we are interested in the procedure itself, but the approach is similar.

In another study regarding the lumbar region of the spine, the aim was to investigate whether or not a correlation exists between the changes in patients’ skin sensation, seen in lumbo-sacral disk root pain, and the disease condition during conservative treatment [13]. This treatment included a weeks bed rest, physiotherapy, and medication. The study included 20 right-handed females of age 40-58 years. One of the tests performed was straight leg raising test and the VAS scale was used to evaluate the pain. Also, two-point discrimination test values for L4, L5 and S1 dermatomes of before and after the treatment were analysed, along with changes between the involved and intact lower limbs.

A BASELINE$^R$ plastic two-point discriminator instrument was used to measure the skin sensitivity. Light pressure of two pins was applied simultaneously to the dermatomes in order to measure a test subjects two-point discrimination test values. The results concluded that there were significant ($p < 0.001$) differences between two-point discrimination tests before and after the treatment, and between pain scale values in the involved and intact limbs before and after the treatment. Though, no statistically significant correlation ($r = 0.017$) between two-point discrimination test and straight leg raising tests was observed.
Chapter 2

Study design

2.1 Background

The medical procedure in question is based on the vertebrae bordering the spinal discs in the lumbar spine region. We are interested in examining all possible pairs of vertebrae derived from the following four vertebrae shown in Figure 2.1: A (L3), B (L4), C (L5), and D (S1).

![Diagram of vertebrae in the lumbar spine and coccyx region](image)

*Figure 2.1: Vertebrae in the lumbar spine and coccyx region. [4] [8]*

The pairs are constructed from the three lower vertebrae in the lumbar spine region (L3, L4, L5) and the upper sacral vertebra (S1). Taking into account the order in which the vertebrae are tapped results in sixteen pairs in total.
Four of these pairs represent the same vertebra being tapped twice (AA, BB, CC, DD) and won’t be used in the analysis. The remaining twelve pairs, six of adjacent and six of separated vertebrae, are the combinations of interest. The pairs are being tapped one by one in randomised order and each test subject is supposed to classify whether the two vertebrae in the pair are adjacent or separated. The answers result in two possible outcomes, either correctly or incorrectly specified pairs of vertebrae. The variable of interest is the number of correct answers, i.e. the number of correctly specified pairs for each test subject.

### 2.2 Randomisation

Randomisation is a very important part of conducting clinical trials. Firstly, randomisation will determine the properties of the standard methods used in the statistical analysis, i.e the distribution of the test statistic when the null hypothesis is true [11]. All potential outcomes of the statistical units in the study are known under the null hypothesis and in turn a significance level, or $p$-value, can be calculated to determine how unusual the actual observed statistic is in comparison to all possible values of that particular statistic [12].

Another aim is to avoid bias arising from the use of judgement or systematic arrangements to influence the results. The investigator and supporting staff might influence reporting response to certain test procedures and the test subjects might influence compliance, cooperation, or provision of information. In clinical trials, in addition to randomisation, the technique of blinding is usually employed to avoid the risk of personal bias when comparing treatments. In this case it is a so called
double blinding where neither the person in charge of the procedure, nor the patient have knowledge about the order of tapping the vertebrae pairs. [2]

In this study the test subjects are supposed to specify if a pair of adjacent or separated vertebrae is being tapped. Here, randomisation is concerned with the order of tapping the six pairs of adjacent and six pairs of separated vertebrae, see Table 2.2.

<table>
<thead>
<tr>
<th>Adjacent</th>
<th>Separated</th>
</tr>
</thead>
<tbody>
<tr>
<td>AB</td>
<td>BC</td>
</tr>
</tbody>
</table>

Table 2.2: Pairs of adjacent and separated vertebrae A, B, C and D.

Every pair is assigned a number from 1 to 12. A table of random numbers is then used to create k randomised sequences of these twelve pairs, one for each test subject.

### 2.3 Classification design

#### 2.3.1 Structure

The study design is such that there are two possible outcomes in specifying the pair of vertebrae being tapped, either adjacent ("+") or separated ("-"). For each test subject the variable of interest is the number of correctly specified pairs. There are two errors that can occur in this test procedure: a pair of adjacent vertebrae is classified as separated ("false negatives") and a pair of separated vertebrae is classified as adjacent ("false positives"). A single individual’s test results can be evaluated along the lines of a diagnostic test structure where the observed classifications are summarised in a 2×2 contingency table as in Table 2.3.

There are two measurements, derived from the test results above, to be considered when evaluating a diagnostic test. Using the conventional terminology for medical tests, the following conditional probabilities, denoted r and s, are of interest:
Table 2.3: Frequency summary of pair specification for a single test subject.

<table>
<thead>
<tr>
<th>Test outcome</th>
<th>True condition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adjacent</td>
<td>Separated</td>
</tr>
<tr>
<td>+</td>
<td>True positive</td>
</tr>
<tr>
<td>-</td>
<td>False negative</td>
</tr>
</tbody>
</table>

\[ r = \text{Prob}(+ \text{ Outcome} | \text{Adjacent}) = \text{sensitivity} \]

\[ s = \text{Prob}(- \text{ Outcome} | \text{Separate}) = \text{specificity} \]

Sensitivity \((r)\) is defined as the probability of a test subject correctly specifying a pair of adjacent vertebrae, and specificity \((s)\) denotes the probability of correctly specifying a pair of separated vertebrae [1]. These proportions are obtained in the following way:

\[ r = \frac{\text{true positive frequency}}{\text{true positive frequency} + \text{false negative frequency}} \quad (2.1) \]

and

\[ s = \frac{\text{true negative frequency}}{\text{true negative frequency} + \text{false positive frequency}} \quad (2.2) \]

The test values of sensitivity and specificity vary between 0% and 100%. For example, in a test with 80% sensitivity, the test subject is able to correctly specify 80% of the adjacent vertebrae pairs, meanwhile 20% of adjacent vertebrae pairs are incorrectly classified. [1]

We now introduce the following notations for a single test subject:

\[ n = \text{number of pairs tapped} \]

\[ p = \text{proportion of adjacent pairs} \]

\[ q = \text{proportion of separated pairs} \]

and we have that \(p + q = 1\). The observed number of
correctly specified pairs of adjacent vertebrae is the product of the \(n\) pairs in the study, proportion \(p\), and sensitivity \(r\). The product of \(n\) pairs, proportion \(q\), and specificity \(s\) gives the observed number of correctly specified separated vertebrae pairs for a single test subject.

<table>
<thead>
<tr>
<th>Test outcome</th>
<th>True condition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adjacent</td>
<td>Separated</td>
</tr>
<tr>
<td>+ (rnp)</td>
<td>((1 - s)np)</td>
</tr>
<tr>
<td>- ((1 - r)np)</td>
<td>(snq)</td>
</tr>
<tr>
<td>Total</td>
<td></td>
</tr>
<tr>
<td>(np)</td>
<td>(nq)</td>
</tr>
</tbody>
</table>

Table 2.4: Expected frequencies for \(n\) pairs of vertebrae.

### 2.3.2 Evaluation

If the probability of obtaining the outcome ”+” is the same regardless of whether a pair is adjacent or separated, it means that

\[
sensitivity = 
= \text{Prob}(+ | \text{Adjacent}) = \text{Prob}(+ | \text{Separate}) = 1 - \text{Prob}(- | \text{Separate}) = \]

= 1 - specificity.

Therefore, for a completely useless test, i.e a test without any prognostic value, it is true that

\[
r + s = 1. \tag{2.3}
\]

This is the null hypothesis in our analysis. In other words, a test subject is unable to distinguish between pairs of adjacent and pairs of separated vertebrae. On the other hand, if the test procedure is maximally effective, meaning that the test always results in correct decisions, it would imply that

\[
\text{Prob}(+ | \text{Adjacent}) = \text{Prob}(- | \text{Separate}) = 1.
\]
For such a test we have that

\[ r + s = 2. \quad (2.4) \]

If there are as many adjacent vertebrae pairs as separated, i.e. \( p = q = 0.5 \), the total number of correctly specified pairs for each individual is defined as

\[ rnp + snq = n(0.5r + 0.5s) = \frac{n(r + s)}{2}. \quad (2.5) \]

In the actual trial we will obtain the following absolute frequencies \( x_a \) and \( x_s \) for the number of correctly classified adjacent and separated pairs of vertebrae respectively, as seen in Table 2.5.

<table>
<thead>
<tr>
<th>Test outcome</th>
<th>True Condition</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Adjacent</td>
<td>Separated</td>
<td>Total</td>
</tr>
<tr>
<td>+</td>
<td>( x_a )</td>
<td>( 6 - x_s )</td>
<td>( x_a + 6 - x_s )</td>
</tr>
<tr>
<td>-</td>
<td>( 6 - x_a )</td>
<td>( x_s )</td>
<td>( 6 - x_a + x_s )</td>
</tr>
<tr>
<td>Total</td>
<td>6</td>
<td>6</td>
<td><strong>12</strong></td>
</tr>
</tbody>
</table>

*Table 2.5: Frequency summary of pair specification for a single test subject.*

It then follows that

\[ E(x_a/6) = r \quad \text{and} \quad E(x_s/6) = s \]

and the corresponding estimators are defined as

\[ \hat{r} = x_a/6 \quad \text{and} \quad \hat{s} = x_s/6. \]

As a measurement \( \hat{e} \) of how good a test subject is in classifying the actual adjacent or separated pairs of vertebrae, we use

\[ \hat{e} = \hat{r} + \hat{s} = (x_a + x_s)/6. \]

We define this measurement as *efficiency* and under the null hypothesis we have that
\[ E(x_a + x_s) = 6. \]

If a test subject lacks the ability to distinguish adjacent vertebrae pairs from separated, then

\[ E(\hat{e}) = 1 \]

or

\[ E(x_a + x_s) = 6, \]

and for a test subject who always correctly specifies all pairs of vertebrae it would mean that

\[ E(\hat{e}) = 2 \]

or

\[ E(x_a + x_s) = 12. \]

The ideal for the test is to be both highly sensitive and specific. The relationship between sensitivity and specificity is described by receiver operator characteristic (ROC) curves. In a ROC curve, sensitivity is plotted as a function of the false positive rate (1 - Specificity) for all possible cut-off points [1]. Since the test is based on a discrete classifier, the ROC curve for a single test subject will have the following shape, see Figure 2.2. [3]

The area under the curve (AUC) measures the accuracy of the test, i.e. how well a test subject can discriminate between pairs of adjacent and separated vertebrae. The further up to the left the curve is, the better the test, i.e. the closer AUC is to 1. The dotted diagonal line represents zero discrimination ability with an AUC of 0.5 [1]. The traditional academic point system can be used as a guide for classifying the test accuracy [14]:

- 0.9 - 1 = excellent
- 0.8 - 0.9 = good
• 0.7 - 0.8 = fair
• 0.6 - 0.7 = poor
• 0.5 - 0.6 = fail.

The following expression for AUC in terms of sensitivity and specificity is derived:

\[
AUC = 1 - \frac{r(1-s)}{2} - \frac{s(1-r)}{2}(1-s)(1-r) = \\
= 1 - \frac{r-rs+s-rs}{2} - (1-r-s+rs) = \\
= \frac{2-r+rs-s+rs-2+2r+2s-2rs}{2} = \frac{r+s}{2}.
\]

This expression has been derived earlier in (2.5), meaning that \((r+s)/2\) represents the proportion of all correctly specified pairs of vertebrae. Multiplying the AUC with the number of pairs tapped gives the number of correctly specified pairs for a single test subject.
2.4 Distribution assumptions

The medical procedure is built on examining twelve pairs of vertebrae in the lumbar region of the back. The clinical study includes \( k \) test subjects and for each one of them a number of correct answers is registered. This means that the study will result in a series of \( k \) independent observations of sensitivity \((r_1, \ldots, r_k)\), specificity \((s_1, \ldots, s_k)\), and efficiency \((e_1, \ldots, e_k)\). In order to perform power calculations, construct confidence intervals, etc., it is of interest to study the properties of the population from which these observations have been derived. In particular, some idea about the variability of the population is needed. There are three cases being considered:

- taking into account the order of tapping the vertebrae
- not taking into account the order of tapping the vertebrae
- variance based on the range of the values

Looking closer at all pairs of vertebrae, there are six unique pairs among the twelve defined when the order of tapping the vertebrae is not taken into account. Each pair is then tapped twice by changing the order of tapping each vertebra within the pair. Tapping \( n \) pairs of vertebrae can be described in terms of performing \( n \) trials where each trial results in a success (correctly specified pair) or a failure (incorrectly specified pair). Taking into account the order of knocking on each vertebrae would mean considering \( n = 12 \) different pairs of vertebrae. If indeed the classification of a pair can be assumed independent from pair to pair, then under the null hypothesis the number of correct answers for every test subject is considered to follow a binomial distribution based on \( n = 12 \) trials and success probability \( p = 0.5 \),

\[
X \sim \text{Bin}(p = 0.5; n = 12),
\]

with expectation

\[
E(X) = np = 12 \cdot 0.5 = 6
\]

and variance

\[
V(X) = npq = 12 \cdot 0.5 \cdot 0.5 = 3 \quad \text{with} \quad SD = \sqrt{V(X)} = 1.73.
\]
However, if twelve independent pairs can not be assumed, the number of correctly specified pairs follows a different model. When the order of tapping each vertebrae is not taken into account, there are no longer twelve unique independent pairs under consideration. This is a distribution assumption where some of the vertebrae pairs are classified the same way when tapped two times, meanwhile the other pairs are not. An extreme case arises when every pair is either correctly or incorrectly specified both times. Then the number of correct answers is defined as $X = 2Y$ where $Y$ is modelled as an independent binomially distributed variable, based on $n = 6$ trials and success probability $p = 0.5$,

$$Y \sim \text{Bin}(p = 0.5; n = 6)$$ (2.10)

with expectation

$$E(Y) = np = 6 \cdot 0.5 = 3$$ (2.11)

and variance

$$V(Y) = npq = 6 \cdot 0.5 \cdot 0.5 = 1.5.$$. (2.12)

This further implies an expectation

$$E(X) = E(2Y) = 2 \cdot np = 2 \cdot 6 \cdot 0.5 = 6$$ (2.13)

and variance

$$V(X) = V(2Y) = 4V(Y) = 4 \cdot 1.5 = 6 \text{ with } SD = \sqrt{V(X)} = 2.45$$ (2.14)

of the number of correctly specified pairs under the null hypothesis.

In the case of twelve independent pairs there are three possible outcomes for correctly specifying the same pair being tapped twice. The outcome for each vertebrae pair ranges between zero and two, meaning that the same pair can be correctly specified zero, one, or two times. In the extreme case, when only six independent pairs are assumed, tapping the vertebrae of the same pair in different order will always result in the same pair being specified. This means that three pairs are
always twice correctly specified and three pairs are twice incorrectly specified under the null hypothesis. The distribution of how many times a pair is correctly specified under the null hypothesis is shown in Figure 2.3.

![Figure 2.3: Number of times a pair is correctly specified under the null hypothesis.](image)

(a) Twelve independent pairs tapped.  
(b) Six independent pairs tapped twice.

In the third distribution case, a very rough approximation of the standard deviation is defined. Basically, the data range under the null hypothesis, 12 - 0 = 12, is divided in four equal parts and as a result a standard deviation equal to 3 is considered, along with an expectation of \( E(X) = 6 \) correctly specified pairs.

### 2.5 Power calculations

From the description of the study design we know that \( X \) is a variable that ranges between the values 0 and 12. According to the previously described statistical model, the standard deviation is either \( SD(X) = 1.73 \) or \( SD(X) = 2.45 \). In the worst case, a standard deviation of \( SD(X) = 3 \) is defined. This would mean considering an optimistic, pessimistic, or very pessimistic distribution assumption for the number of correctly specified pairs.

The main objective of the study is to examine the ability of spatial discrimination
for a group of $N$ test objects. In the planning stage of the trial the size of the study group was investigated in order to obtain a certain level of statistical power. However, the problem can’t be approached in the usual way, as is the case when comparing control and treatment groups in a clinical trial. When determining the size of this study, some assumptions are needed regarding the underlying distribution and alternative hypothesis.

For a single test subject, a measurement of interest was defined as the number of correctly specified pairs with an expectation of $E(X) = 6$ under the null hypothesis. For the group as a whole, the mean of the number of correct answers for all the test subjects combined, i.e. $M = \Sigma X/N$ will be considered as a test variable.

When performing power calculations based on hypothesis tests, the underlying alternative hypothesis should be concerned with the smallest effect that can still be reliably detected [7]. We are looking for the probability of rejecting a false null hypothesis. The number of test subjects used in the study should be chosen so that rejecting the null hypothesis of the individuals not being able to spatially discriminate between adjacent and separated vertebrae pairs has a large probability when the null hypothesis is false. In the planning stage it seemed reasonable to assume that under the alternative hypothesis $r + s = 1.4$, i.e. that sensitivity and specificity combined vary around 70%. Then, under the alternative hypothesis, the number of correctly specified pairs is expected to be

$$E(X) = 6(r + s) = 6 \cdot 1.4 = 8.4.$$  

(2.15)

As the sample size grows towards infinity, the expected distribution of the sample mean will approach normality regardless of the distribution shape of the individual observations [7]. Therefore, determining the power of the test in this study was based on one-sided hypothesis $t$-test at a significance level of $\alpha = 0.01$. The resulting power curves as a function of sample size in the three distribution assumption cases are plotted in Figure 2.4.
It can be seen that a reasonably high statistical power can be attained already with
a study size of 10 test subjects under the alternative hypothesis of $E(X) = 8.4$.
The test results in a power of 98%, 78%, and 58% for standard deviations 1.73,
2.45, and 3 respectively. Choosing $N = 15$ individuals increases the statistical
power to 99%, 93%, and 78%. These calculations are summarized in Table 7.1 in
Appendix.

Plotting the power curves as a function of the true mean difference for sample sizes
$N = 10$ and $N = 15$, it seems appropriate to choose a sample size of at least 15
test subjects in order to detect the desired effect size, see Figure 2.5.
Figure 2.5: Power curves for a one-sided t-test with $SD = 1.73$, $SD = 2.45$ and $SD = 3$. 
Chapter 3

Methodology

3.1 Practical procedure and data collection

The procedure takes place in an X-ray room and is supposed to last 10-15 minutes. Before starting the actual trial, the test subject is given a chance to get familiar with the procedure in order to provide his/her answers. The doctor goes over all possible pairs of vertebrae in an unspecified order and informs the test subject whether an adjacent or separated pair of vertebrae is being tapped. Then, whenever the test subject is ready, the procedure can begin.

The test subject lies prone during the entire procedure. In preparation for the procedure, the skin of the lower lumbar is sterile washed. Then, by using fluroscopy, four short injection needles are applied to the three lower vertebrae in the lumbar and the upper sacral vertebra, see Figure 3.1.

![Figure 3.1: Position of the needles during the procedure.](image)
Each needle is firmly attached to the posterior spinal process of the vertebra in question by gently tapping the top of each needle. The pain experienced during application of the needles can be likened to venipuncture for blood sampling. A tiny amount of local anaesthetic, 0.1 ml Xylocain 10 mg/ml, is injected to numb the posteriorium surrounding the tip of the needle. Then lightly striking the needles won’t be experienced as pain, only as a slight shock sensation.

The twelve pairs of vertebrae are tapped in the randomised order specified for each test subject, unknown to the doctor. After tapping each pair of vertebrae, the test subject’s answer of whether the two vertebrae tapped are adjacent (A) or separated (S) is noted in the protocol by an assistant, see Figure 3.2.

<table>
<thead>
<tr>
<th>Test subject No:</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>pair of vertebrae</td>
<td>BA</td>
<td>BD</td>
<td>DC</td>
<td>AD</td>
<td>AB</td>
<td>AC</td>
<td>CA</td>
<td>CB</td>
<td>CD</td>
<td>DB</td>
<td>DA</td>
<td>BC</td>
</tr>
</tbody>
</table>
| classification | A  | S  | A  | ...| ...| ...| ...| ...| ...| ...| ...| ...

Figure 3.2: Study protocol.

### 3.2 Statistical methods

In the planning stage of this experiment some underlying assumptions and hypothesis regarding the variable of interest needed to be stated. The power calculations in preparation of this study in section 2.5 are based on the normality assumption. Also, the theoretical distribution of correctly specified pairs under the null hypothesis is assumed to be symmetrically distributed around its hypothesized mean. Therefore, the classic $t$-test would seem to be the appropriate method for testing our null hypothesis. However, after examining the data, this method was not suitable for analysing the data. Instead, bootstrap resampling was used in hypothesis testing and obtaining desired confidence intervals [15]. The calculations were performed using the statistical software R.
Chapter 4

Results

4.1 Data description

The clinical trial resulted in subjecting 18 individuals to this particular medical procedure in order to test their ability of distinguishing adjacent vertebrae pairs from separated. The number of correctly specified pairs (mean = 10.17, SD = 1.42) was obtained for each test subject and the distribution of the observed outcomes is shown in Figure 4.1.

![Figure 4.1: Number of correctly specified pairs.](image-url)
All test subjects were able to correctly specify more than eight pairs of vertebrae, except for one individual with six correct answers. It was also observed that only one of the test subjects was able to correctly determine all twelve pairs. More than half (56%) of the individuals correctly specified 11 pairs of vertebrae. In five (28%) cases the procedure resulted in 9 correct answers and another single individual correctly determined 10 pairs of vertebrae.

Ten of the 18 test subjects were chosen among the staff at the clinic and therefore had prior medical knowledge (mean = 10, SD = 1.054), whereas 8 test subjects had none (mean = 10.38, SD = 1.85). No statistically significant ($p = 0.629$) difference between these groups with regard to their ability to discriminate between adjacent and separated vertebrae pairs was observed.

### 4.2 Sensitivity and specificity

As previously stated, sensitivity and specificity measure the proportion of correctly specified adjacent and separated vertebrae pairs respectively. The distribution of the observed sensitivity (mean = 0.81, SD = 0.16) and specificity (mean = 0.89, SD = 0.17) for the 18 test subjects is not symmetrical and clearly non-normal, see Figure 4.2.

![Figure 4.2: Observed sensitivity ($r$) and specificity ($s$) in the test procedure.](image)
The mass of the distribution is concentrated around the higher values for both sensitivity and specificity. This is not unexpected since these measures are derived from correctly specified adjacent and separated pairs of vertebrae respectively. No values lower than 0.5 (11%) were observed for sensitivity, see Table 7.2. The highest observed frequency (39%) was a sensitivity of 0.83. In four (22%) cases a sensitivity of 0.67 was calculated, and for five (28%) other test subjects a value of 1 was obtained. No statistically significant \( p = 0.102 \) difference between the medical (mean = 0.75, SD = 0.18) and non-medical (mean = 0.87, SD = 0.12) groups was observed.

As for specificity, only in one (6%) case the value of 0.33 was observed with the remaining values ranging between 0.67 (6%) and 1 (55%) with the highest observed frequency. In six (33%) cases specificity resulted in a value of 0.83. The difference between the groups with medical knowledge (mean = 0.92, SD = 0.09) and without (mean = 0.85, SD = 0.24) was not statistically significant \( p = 0.518 \).

### 4.3 Efficiency

In this clinical study the efficiency was defined as the sum of sensitivity and specificity and is considered to be a measurement of how well an individual can distinguish between pairs of adjacent and separated vertebrae. It denotes the proportion of correctly specified pairs of vertebrae in total for each test subject. Figure 4.3 is displaying the distribution of efficiency (mean = 1.69, SD = 0.24) for the 18 test subjects.

The lowest observed value is an efficiency of 1 (6%) with all other observed values higher than that. The largest observed frequency for efficiency is of the value 1.83 in ten (55%) cases. For only one (6%) individual the test resulted in an efficiency of 2, and for another one (6%) a value of 1.67 was calculated. Five (27%) individuals showed a test efficiency of 1.50. No statistically significant \( p = 0.628 \) difference between the medical (mean = 1.67, SD = 0.17) and non-medical (mean = 1.73, SD = 0.31) groups was observed.
4.4 Incorrectly specified pairs of vertebrae

Just out of curiosity it is interesting to see whether a specific pair of vertebrae is repeatedly incorrectly specified. The bar chart in Figure 4.4 shows all the incorrectly specified pairs and their absolute frequencies. The numbers in each bar represent the corresponding relative frequencies. This is just a description and no further analysis of these observations will be carried out.

Of the 34 incorrectly specified pairs 22 were supposed to be adjacent and 12 separated. It seems that it was more difficult to determine two vertebrae located next to each other than apart. The values of interest are the higher frequencies. For instance, 6 (33%) individuals were unable to specify AB as a pair of adjacent vertebrae and at the same time no BA pairs were misspecified. Also a high number of pairs BC and CB were incorrectly determined, 6 (33%) and 4 (22%) respectively. One individual misspecified both BC and CB, and therefore in 9 (50%) unique cases the test subjects were unable to determine that vertebrae B and C are adjacent. The specific pairs CD and DC were wrongly determined 1 (6%) and correspondingly 5 (28%) times. In one case an individual incorrectly specified both pairs which means that in 5 (28%) unique cases the test subjects determined that vertebrae C and D are separated when in fact they were not.
When it comes to pairs of separated vertebrae, the highest observed frequency of misspecified pairs was CA with 5 (28%) pairs of separated vertebrae specified as adjacent. The same pair of vertebrae in reverse order, AC, was misspecified only 1 time where this particular test subject also incorrectly specified the pair CA. As a result, in 5 (22%) individual cases the vertebrae C and A were determined to be adjacent. Only one test subject determined DA to be a pair of adjacent vertebrae. Lastly, two individuals incorrectly specified the location of vertebrae B and D relative to each other both times leaving only 2 (0.11%) unique situations where individuals were unable to classify it as a pair of separated vertebrae. All of these observed frequencies for each test subject are shown in Table 7.4 in Appendix.
Chapter 5

Analysis

5.1 Correctly specified pairs

Since the number of correctly specified pairs can take values between 0 and 12, Figure 5.1 clearly shows that the distribution of the observed number of correct answers does not come from a population with a mean of $\mu = 6$, it is significantly higher than that. Under the null hypothesis we would expect values both below and above 6 correctly specified vertebrae pairs. In this case there are no observations smaller than 6 and almost all of them are larger than 8.

![Figure 5.1: Number of correctly specified pairs.](image)
The basic one-sample t-test rejects the null hypothesis of $\mu = 6$ ($p$-value < 0.000) at the 1% significance level. The mean is significantly larger than 6 correctly specified pairs ($p$-value < 0.000) with the corresponding two-sided 99% confidence interval of [9.19 ; 11.14]. However, Looking at the observed data, a one-sample t-test is not the best method for testing whether the data comes from a population with $\mu = 6$. The data is clearly non-symmetrical and violates the normality assumption. Therefore, bootstrap resampling is applied in this analysis for handling the observed data.

By plotting the number of correctly specified vertebrae pairs, it is obvious that the null hypothesis of a population mean $\mu = 6$ is rejected ($p = 0.00$) at the 1% significance level. The test subjects in this study are clearly able to discriminate between pairs of adjacent and separated vertebrae. The corresponding 99% adjusted bootstrap percentile (BCa) confidence interval of the mean number of correctly specified pairs is estimated to be [8.89 ; 10.78]. The bootstrap resampled distribution of the sample mean is shown in Figure 5.2. The estimated mean (10.17) of our test statistic is denoted by the dotted line. It is easy to see that a mean of six correctly specified pairs is highly unlikely to be observed.
An interesting observation in the data is that the odd numbers of correctly specified vertebrae pairs have the highest observed frequency, either 9 or 11 correctly specified pairs. It would seem reasonable to assume that once a pair being knocked on is correctly specified, the test subjects would be able to correctly specify the same vertebrae pair the second time when knocked on each vertebrae in reversed order.

Looking closer at the distribution of twice specified pairs, it is very skewed in Figure 5.3.

![Figure 5.3: Observed distribution of number of times correctly specified pair.](image)

The observed high proportion of twice correctly specified pairs suggests that the test subjects in this study are able to tell the difference between adjacent and separated pairs of vertebrae. Also, a proportion of 22% of pairs were once correctly specified and 4% were misspecified twice. The observed SD = 1.43 is lower than all three of the hypothesised in section 2.4. Therefore, it would seem that the study procedure follows a model based on the first distribution assumption stated in Section 2.4, i.e. the pairs tapped are independent. They could be seen as independent because the vertebrae in the pair are tapped one at the time, not
simultaneously. In order to correctly classify a pair two times in a row, the test subject must have good sense of the individual vertebrae location. Mistaking the position of the first, or second, or both vertebrae would result in a false classification. Tapping a certain pair in reversed order means tapping the second vertebrae from before first. For example, this vertebrae could have been correctly identified the first time but not the second, and vice versa. Therefore it is not certain that tapping the same pair in reversed order will result in the same classification. This was observed in few cases where a certain pair and its reverse were tapped right after one another.

5.2 Sensitivity and specificity

In the previous section it was clear that sensitivity and specificity in this trial are significantly larger than 0.5 ($p_r = 0.00$ and $p_s = 0.00$). The corresponding 99% adjusted bootstrap percentile (BCa) confidence intervals for the mean of sensitivity and specificity are $[0.69 ; 0.89]$ and $[0.70 ; 0.95]$. Further, these two measurements can be evaluated by examining the ROC curves for the 18 tests subjects. Instead of plotting 18 graphs, the calculated areas under the curves and corresponding 95% confidence intervals are summarized in Table 7.3 in the Appendix section. The frequency distribution of the area under the ROC curves is presented in Table 5.1.

<table>
<thead>
<tr>
<th>Area under ROC</th>
<th>LL</th>
<th>UL</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>0.16</td>
<td>0.84</td>
<td>1</td>
</tr>
<tr>
<td>0.76</td>
<td>0.49</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>0.83</td>
<td>0.67</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>0.88</td>
<td>0.72</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0.93</td>
<td>0.79</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

*Table 5.1: Frequency summary of the area under the ROC curve for the 18 test subjects.*
For each of the 18 test subjects the same test was performed where they had to specify the pair of vertebrae that was being tapped, either adjacent or separated. The accuracy of this test depends on how well the test subjects are able to specify the two types of pairs. The area under the ROC curve measures this accuracy [14]. In only one (6%) case the area was 0.5 which is considered to be a poor test. In three (17%) cases the area under the ROC curve was observed to be "fair" with a value of 0.76. Also one (6%) value of 0.88 and 1 were observed, representing a "good" and "excellent" test result respectively. For ten (56%) of the test subjects the area under the ROC curve resulted in a value of 0.93 which is considered to be an "excellent" test. In two (11%) other cases a "good" test result of 0.83 was recorded. Overall, this can be considered as a good test because most of the values of the area under the ROC curve are larger than 0.8. This would imply that the test subjects are able to successfully distinguish between pairs of adjacent and separated vertebrae.

5.3 Efficiency

Concluding that sensitivity and specificity are significantly larger than 0.5 would further imply that efficiency, the previously defined measurement for how well a test subject is able to correctly specify pairs of adjacent and separated vertebrae in total, is greater than the hypothesised value of $r + s = 1$. This is clearly the case in Figure 4.3 in Section 4.3. The smallest observed value of efficiency is 1 and all the other values are larger than that. The observed efficiency comes from a population with the mean significantly ($p = 0.00$) larger than 1, thereby rejecting the null hypothesis. The 99% adjusted bootstrap percentile (BCa) confidence interval of the mean of efficiency is estimated to be [1.50 ; 1.79].
Chapter 6

Conclusions

A clinical study consisting of 18 test subjects was performed in order to investigate whether healthy individuals possess the ability to discriminate between pairs of adjacent and separated vertebrae. The variable of interest is the number of correctly specified vertebrae pairs. The method of bootstrap resampling was applied to the observed data in hypothesis testing and obtaining confidence intervals for the parameter of interest, i.e. the mean of correctly specified pairs. The results clearly showed that the number of correctly specified vertebrae pairs comes from a population with a mean significantly greater than 6, indicating that individuals in this test procedure were able correctly classify pairs of adjacent and separated vertebrae.

Further, analysing the results in terms of sensitivity and specificity, i.e. the proportion of correctly specified adjacent and separated vertebrae pairs respectively, the null hypothesis of a completely worthless test was rejected. This was clearly seen by examining the distribution of the observed data. The mean of both sensitivity and specificity was observed to be significantly larger than 0.5, therefore implying the mean of efficiency being significantly larger than the hypothesised value of 1. Also, when reviewing the area under the ROC curves of the test for each individual, in 15 of the 18 cases an AUC larger than 0.8 was observed, corresponding to "good" and "excellent" test procedure results.

Among the 18 test subjects in this study, 10 of them had previous medical knowledge meanwhile the others did not. No significant group differences were observed with regard to correctly specified vertebrae pairs, sensitivity, specificity, and efficiency at the 1% significance level.
To summarise, all of the null hypothesis defining a completely worthless test where the individuals lack the ability of spatial discrimination between adjacent and separated vertebrae pairs were rejected. The results clearly show that the test subjects are able to distinguish pairs of adjacent vertebrae from separated.
Bibliography


Chapter 7

Appendix

Power calculations

<table>
<thead>
<tr>
<th>Dispersion assumption</th>
<th>Null hypothesis: (r + s) = 1</th>
<th>Alternative hypothesis: (r + s) = 1.4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>E(M)</td>
<td>SE(M)</td>
</tr>
<tr>
<td>N = 10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SD = 1.73</td>
<td>6</td>
<td>0.55</td>
</tr>
<tr>
<td>SD = 2.45</td>
<td>6</td>
<td>0.77</td>
</tr>
<tr>
<td>SD = 3.00</td>
<td>6</td>
<td>0.95</td>
</tr>
<tr>
<td>N = 15</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SD = 1.73</td>
<td>6</td>
<td>0.45</td>
</tr>
<tr>
<td>SD = 2.45</td>
<td>6</td>
<td>0.63</td>
</tr>
<tr>
<td>SD = 3.00</td>
<td>6</td>
<td>0.77</td>
</tr>
</tbody>
</table>

Table 7.1: Power calculations for deciding the size of the study for sample sizes $N = 10$ and $N = 15$. 
<table>
<thead>
<tr>
<th>TS</th>
<th>Group</th>
<th>Adjacent</th>
<th>Separated</th>
<th>Total</th>
<th>Sensitivity (r)</th>
<th>Specificity (s)</th>
<th>Efficiency (r+s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>4</td>
<td>5</td>
<td>9</td>
<td>0.67</td>
<td>0.83</td>
<td>1.50</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>5</td>
<td>6</td>
<td>11</td>
<td>0.83</td>
<td>1.00</td>
<td>1.83</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>4</td>
<td>5</td>
<td>9</td>
<td>0.67</td>
<td>0.83</td>
<td>1.50</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>5</td>
<td>6</td>
<td>11</td>
<td>0.83</td>
<td>1.00</td>
<td>1.83</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>5</td>
<td>6</td>
<td>11</td>
<td>0.83</td>
<td>1.00</td>
<td>1.83</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>3</td>
<td>6</td>
<td>9</td>
<td>0.50</td>
<td>1.00</td>
<td>1.50</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>6</td>
<td>5</td>
<td>11</td>
<td>1.00</td>
<td>0.83</td>
<td>1.83</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>5</td>
<td>6</td>
<td>11</td>
<td>0.83</td>
<td>1.00</td>
<td>1.83</td>
</tr>
<tr>
<td>9</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>6</td>
<td>0.67</td>
<td>0.33</td>
<td>1.00</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>5</td>
<td>6</td>
<td>11</td>
<td>0.83</td>
<td>1.00</td>
<td>1.83</td>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td>6</td>
<td>5</td>
<td>11</td>
<td>1.00</td>
<td>0.83</td>
<td>1.83</td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td>4</td>
<td>5</td>
<td>9</td>
<td>0.67</td>
<td>0.83</td>
<td>1.50</td>
</tr>
<tr>
<td>13</td>
<td>2</td>
<td>6</td>
<td>4</td>
<td>10</td>
<td>1.00</td>
<td>0.67</td>
<td>1.67</td>
</tr>
<tr>
<td>14</td>
<td>1</td>
<td>3</td>
<td>6</td>
<td>9</td>
<td>0.50</td>
<td>1.00</td>
<td>1.50</td>
</tr>
<tr>
<td>15</td>
<td>2</td>
<td>5</td>
<td>6</td>
<td>11</td>
<td>0.83</td>
<td>1.00</td>
<td>1.83</td>
</tr>
<tr>
<td>16</td>
<td>2</td>
<td>5</td>
<td>6</td>
<td>11</td>
<td>0.83</td>
<td>1.00</td>
<td>1.83</td>
</tr>
<tr>
<td>17</td>
<td>2</td>
<td>6</td>
<td>6</td>
<td>12</td>
<td>1.00</td>
<td>1.00</td>
<td>2.00</td>
</tr>
<tr>
<td>18</td>
<td>2</td>
<td>6</td>
<td>5</td>
<td>11</td>
<td>1.00</td>
<td>0.83</td>
<td>1.83</td>
</tr>
</tbody>
</table>

Group = 1: medical  
Group = 2: non-medical

*Table 7.2: Observed number of correctly specified pairs and calculated sensitivity, specificity, and efficiency for 18 test subjects.*
<table>
<thead>
<tr>
<th>TS</th>
<th>Group</th>
<th>Sensitivity (r)</th>
<th>Specificity (s)</th>
<th>Efficiency (r+s)</th>
<th>AUC (ROC)</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>LL</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0.67</td>
<td>0.83</td>
<td>1.50</td>
<td>0.757</td>
<td>0.491</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0.83</td>
<td>1.00</td>
<td>1.83</td>
<td>0.929</td>
<td>0.789</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0.67</td>
<td>0.83</td>
<td>1.50</td>
<td>0.757</td>
<td>0.491</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>0.83</td>
<td>1.00</td>
<td>1.83</td>
<td>0.929</td>
<td>0.789</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>0.83</td>
<td>1.00</td>
<td>1.83</td>
<td>0.929</td>
<td>0.789</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>0.50</td>
<td>1.00</td>
<td>1.50</td>
<td>0.833</td>
<td>0.670</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>1.00</td>
<td>0.83</td>
<td>1.83</td>
<td>0.929</td>
<td>0.789</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>0.83</td>
<td>1.00</td>
<td>1.83</td>
<td>0.929</td>
<td>0.789</td>
</tr>
<tr>
<td>9</td>
<td>2</td>
<td>0.67</td>
<td>0.33</td>
<td>1.00</td>
<td>0.500</td>
<td>0.162</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>0.83</td>
<td>1.00</td>
<td>1.83</td>
<td>0.929</td>
<td>0.789</td>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td>1.00</td>
<td>0.83</td>
<td>1.83</td>
<td>0.929</td>
<td>0.789</td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td>0.67</td>
<td>0.83</td>
<td>1.50</td>
<td>0.757</td>
<td>0.491</td>
</tr>
<tr>
<td>13</td>
<td>2</td>
<td>1.00</td>
<td>0.67</td>
<td>1.67</td>
<td>0.875</td>
<td>0.715</td>
</tr>
<tr>
<td>14</td>
<td>1</td>
<td>0.50</td>
<td>1.00</td>
<td>1.50</td>
<td>0.833</td>
<td>0.670</td>
</tr>
<tr>
<td>15</td>
<td>2</td>
<td>0.83</td>
<td>1.00</td>
<td>1.83</td>
<td>0.929</td>
<td>0.789</td>
</tr>
<tr>
<td>16</td>
<td>2</td>
<td>0.83</td>
<td>1.00</td>
<td>1.83</td>
<td>0.929</td>
<td>0.789</td>
</tr>
<tr>
<td>17</td>
<td>2</td>
<td>1.00</td>
<td>1.00</td>
<td>2.00</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>18</td>
<td>2</td>
<td>1.00</td>
<td>0.83</td>
<td>1.83</td>
<td>0.929</td>
<td>0.789</td>
</tr>
</tbody>
</table>

Group = 1: medical
Group = 2: non-medical

Table 7.3: Sensitivity, specificity, efficiency, and area under ROC curve with corresponding 95% confidence interval for 18 test subjects.
<table>
<thead>
<tr>
<th>TS</th>
<th>Group</th>
<th>AB</th>
<th>BC</th>
<th>CD</th>
<th>CB</th>
<th>DC</th>
<th>AC</th>
<th>BD</th>
<th>CA</th>
<th>DB</th>
<th>DA</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>√</td>
<td>√</td>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>√</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>√</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>√</td>
<td>√</td>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-</td>
</tr>
<tr>
<td>9</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>√</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-</td>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-</td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td>√</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-</td>
</tr>
<tr>
<td>13</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-</td>
</tr>
<tr>
<td>14</td>
<td>1</td>
<td>√</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-</td>
</tr>
<tr>
<td>15</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-</td>
</tr>
<tr>
<td>16</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-</td>
</tr>
<tr>
<td>17</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-</td>
</tr>
<tr>
<td>18</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>√</td>
</tr>
</tbody>
</table>

Table 7.4: Incorrectly specified pairs for each of the 18 test subjects.