REPRODUCIBILITY DIAGNOSTIC PROCEDURES IN MANUAL/MUSCULOSKELETAL MEDICINE

EDITION 2019

International Academy of Manual/Musculoskeletal Medicine

Jacob Patijn, MD, PhD
EDITION 2019, PROTOCOL FORMAT FOR DIAGNOSTIC PROCEDURES IN MANUAL/MUSCULOSKELETAL MEDICINE
Jacob Patijn, MD, PhD
Copyright © International Academy of Manual/Musculoskeletal Medicine
Preface to the 2019 Edition of the Reproducibility

The last complete revised edition of this protocol was published by the International Academy of Manual/Musculoskeletal Medicine (IAMMM) in 2012. In the meantime much experience has been gained with the daily practice of this protocol. It has stimulated scientists and practitioners in our field of manual/musculoskeletal medicine (M/M Medicine) to perform reproducibility studies according the format of this protocol. Based on these experiences and the emergence of new data from research in this field, the IAMMM thought it necessary to revise the last edition of 2012.

In this IAMMM protocol, attention is paid to the comprehensibility and readability of the text compared to the previous edition. In this way the protocol is accessible to those practitioners in M/M Medicine, who are less familiar with statistics and in particular with the format of performing reproducibility studies.

A new chapter is dedicated to the calculation of the sample size in reproducibility studies.

The reproducibility protocol has been elaborated in such a way that it can be used as a kind of “Cook Book format“ to perform reproducibility studies with kappa statistics.

The protocol format can be used in a very practical way and it makes it feasible to perform reproducibility studies in private M/M Medicine clinics with two or more physicians and by Educational Boards of the M/M Medicine Societies.

The protocol is used as the syllabus for the International Instructional Course for Reproducibility Studies organised by the IAMMM, and we sincerely hope that the protocol will be acknowledge in University education as well.

As in previous editions of the protocol, this Edition 2019 strongly emphasises the need to perform reproducibility studies in M/M Medicine.

Therefore, in the introduction again the original reasons to develop this protocol are mentioned, because they are still very relevant for present day M/M Medicine.
Also in this 2012 edition, a list of reproducibility studies (exclusively using kappa statistic) of the different region of the locomotion system are published.

The IAMMM is aware that it is a continuous process to keep a protocol like this one updated. We do hope that scientists and educationalists, which use this protocol, will send their comments to the present second Scientific Director of the Academy, In this way, we continuously can improve and update the protocol.

The IAMMM also wants to encourage scientists and/or educationalists, which receive and use this last edition protocol, to disperse it among their colleagues and students. Hereby the protocol becomes accessible to a larger audience of practitioners in the field of M/M Medicine.

Professor Jacob Patijn, MD, PhD
IAMMM Scientific Director

Address for reprints and comments:
Dr. Sjef Rutte, MD, MSc, Logistic officer IAMMM
sjerutte@gmail.com
## CONTENTS

Preface to the 2019 Edition of the Reproducibility

<table>
<thead>
<tr>
<th>I. INTRODUCTION</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>The IAMMM International Instructional Course for Reproducibility Studies</td>
<td>10</td>
</tr>
<tr>
<td>The Academy Conference</td>
<td>10</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>II REPRODUCIBILITY and VALIDITY</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reliability</td>
<td>11</td>
</tr>
<tr>
<td>Reproducibility</td>
<td>12</td>
</tr>
<tr>
<td>Intra-observer agreement</td>
<td>12</td>
</tr>
<tr>
<td>Inter-observer agreement</td>
<td>13</td>
</tr>
<tr>
<td>Diagnosis versus Diagnostic Procedure</td>
<td>13</td>
</tr>
<tr>
<td>Validity</td>
<td>16</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>III REPRODUCIBILITY STUDIES: DATA</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nature Data of reproducibility studies</td>
<td>19</td>
</tr>
<tr>
<td>Qualitative Diagnostic Procedures</td>
<td>19</td>
</tr>
<tr>
<td>Quantitative Diagnostic Procedures</td>
<td>20</td>
</tr>
<tr>
<td>Inappropriate statistics qualitative data</td>
<td>21</td>
</tr>
<tr>
<td>Percent Agreement/Overall Agreement/Observed Agreement</td>
<td>21</td>
</tr>
<tr>
<td>Correlation Coefficients</td>
<td>21</td>
</tr>
<tr>
<td>Appropriate statistics of qualitative data in reproducibility studies</td>
<td>21</td>
</tr>
<tr>
<td>Kappa Statistics</td>
<td>21</td>
</tr>
<tr>
<td>Appropriate statistics of quantitative data in reproducibility studies</td>
<td>22</td>
</tr>
<tr>
<td>Choice of statistics and clinical consequences</td>
<td>22</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>IV REPRODUCIBILITY STUDIES: KAPPA STATISTICS</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Definition Kappa Value</td>
<td>23</td>
</tr>
<tr>
<td>Overall Agreement</td>
<td>25</td>
</tr>
<tr>
<td>Prevalence and Prevalence of the Index Condition</td>
<td>26</td>
</tr>
<tr>
<td>Index Condition</td>
<td>26</td>
</tr>
<tr>
<td>Prevalence of the Index Condition</td>
<td>26</td>
</tr>
<tr>
<td>Prevalence</td>
<td>28</td>
</tr>
<tr>
<td>Expected Agreement by Chance</td>
<td>28</td>
</tr>
<tr>
<td>Calculation Kappa Value</td>
<td>29</td>
</tr>
<tr>
<td>Interpretation Kappa Value: general</td>
<td>31</td>
</tr>
<tr>
<td>Interpretation Kappa Value: dependency of overall agreement</td>
<td>32</td>
</tr>
<tr>
<td>Interpretation Kappa Value: dependency of prevalence of the index condition</td>
<td>34</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>V DEVELOPING REPRODUCIBILITY STUDIES: GENERAL ASPECTS</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nature of the Diagnostic Procedure to be evaluated in a study</td>
<td>37</td>
</tr>
<tr>
<td>Diagnosis</td>
<td>37</td>
</tr>
<tr>
<td>Syndrome</td>
<td>37</td>
</tr>
<tr>
<td>Diagnostic procedure</td>
<td>37</td>
</tr>
<tr>
<td>Number of diagnostic procedures evaluated in reproducibility studies</td>
<td>39</td>
</tr>
<tr>
<td>Too many diagnostic procedures in reproducibility studies</td>
<td>39</td>
</tr>
<tr>
<td>Combinations of a several different diagnostic procedures: mutual dependency</td>
<td>39</td>
</tr>
<tr>
<td>Combinations of a few different tests: mutual dependency of test and final “syndrome diagnosis”</td>
<td>42</td>
</tr>
<tr>
<td>Large number of different diagnostic procedures with a “diagnostic protocol”</td>
<td>45</td>
</tr>
<tr>
<td>Hypothesis of a diagnostic test in reproducibility studies</td>
<td>45</td>
</tr>
<tr>
<td>Number of Observers in reproducibility studies</td>
<td>47</td>
</tr>
<tr>
<td>Characteristics of Observers in reproducibility studies</td>
<td>48</td>
</tr>
<tr>
<td>Number of patients in reproducibility studies</td>
<td>49</td>
</tr>
</tbody>
</table>
## CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>VI</strong></td>
<td>THE PROBLEM OF THE RELATION BETWEEN THE PREVALENCE OF THE INDEX CONDITION AND THE KAPPA VALUE</td>
</tr>
<tr>
<td></td>
<td>Defining the problem of the Prevalence of Index Condition</td>
</tr>
<tr>
<td></td>
<td>Influencing the Prevalence of Index Condition in Advance</td>
</tr>
<tr>
<td><strong>VII</strong></td>
<td>PROTOCOL FORMAT REPRODUCIBILITY STUDIES</td>
</tr>
<tr>
<td></td>
<td>Preparation Phase</td>
</tr>
<tr>
<td></td>
<td>Participating members and Logbook</td>
</tr>
<tr>
<td></td>
<td>Transparency of Responsibility</td>
</tr>
<tr>
<td></td>
<td>Logistics of Reproducibility Studies</td>
</tr>
<tr>
<td></td>
<td>Finance in Reproducibility Studies</td>
</tr>
<tr>
<td></td>
<td>Training Phase</td>
</tr>
<tr>
<td></td>
<td>Observer and Patient Selection</td>
</tr>
<tr>
<td></td>
<td>Selection and Number of Diagnostic procedures</td>
</tr>
<tr>
<td></td>
<td>Agreement about Test Performance</td>
</tr>
<tr>
<td></td>
<td>Agreement about Test Hypothesis</td>
</tr>
<tr>
<td></td>
<td>Agreement about Test Judgement</td>
</tr>
<tr>
<td></td>
<td>Evaluation Form</td>
</tr>
<tr>
<td></td>
<td>Overall Agreement Phase</td>
</tr>
<tr>
<td></td>
<td>Observer and Patient Selection</td>
</tr>
<tr>
<td></td>
<td>Blinding Procedures</td>
</tr>
<tr>
<td></td>
<td>Study Phase</td>
</tr>
<tr>
<td></td>
<td>Observer and Patient Selection</td>
</tr>
<tr>
<td></td>
<td>Blinding Procedures</td>
</tr>
<tr>
<td></td>
<td>Data Phase</td>
</tr>
<tr>
<td></td>
<td>Kappa Calculation</td>
</tr>
<tr>
<td></td>
<td>Publication Phase</td>
</tr>
<tr>
<td></td>
<td>Introduction Section</td>
</tr>
<tr>
<td></td>
<td>Material and Methods Section</td>
</tr>
<tr>
<td></td>
<td>Results Section</td>
</tr>
<tr>
<td></td>
<td>Discussion Section</td>
</tr>
<tr>
<td><strong>VIII</strong></td>
<td>GOLDEN RULES FOR REPRODUCIBILITY STUDIES</td>
</tr>
<tr>
<td><strong>IX</strong></td>
<td>APPENDIX LITERATURE KAPPA STUDIES</td>
</tr>
</tbody>
</table>
Figure 1. Summary of the problem and its consequences for Manual/Musculoskeletal Medicine as defined by the previous Scientific Committee of FIMM and adapted by its successor the present Academy of Manual/Musculoskeletal Medicine.
1. **Introduction**

The IAMMM developed this protocol in a standardised format for reproducibility studies.
In particular, this protocol provides scientists and daily practitioners in our field of M/M Medicine with a practical format, in a more or less cook book form, to perform reproducibility studies. The primary reason for the Academy to develop this kind of protocol is still relevant:

*There are many different approaches (schools) in M/M Medicine in many countries of the M/M Medicine world, with many different diagnostic procedures and many different therapies frequently for the same clinical picture.*

The predecessor of the IAMMM, the Scientific Committee of FIMM, formulated the problem with respect to diagnostic procedures in Manual/Musculoskeletal Medicine (M/M Medicine) and is summarised in figure 1.

**The consequences of this statement are five-fold:**

1. Most existing different approaches within M/M Medicine has no reproducible proven diagnostic procedures in the various regions of the locomotion system. As a consequence the reproducibility, validity, sensitivity and specificity of these diagnostic procedures are largely lacking.

2. Because this lack of good reproducibility, validity, sensitivity and specificity studies of the diagnostic procedures of the different schools in M/M Medicine, mutual comparison of diagnostic procedures of the different approaches in M/M Medicine is impossible. In the present situation, scientific information exchange and fundamental discussions, based on solid scientific results and methods, between these different M/M Medicine approaches is frequently impossible.
3. Each of the different approaches in M/M Medicine has their own education system. Most of the diagnostic procedures taught by these education systems lack good reproducibility. This makes the transferability of the taught diagnostic procedures between the various M/M Medicine education systems questionable. Besides, mutual exchange between education systems of diagnostic procedures is hampered. Since we are living in the age of evidence-based medicine, medical educational systems in general and M/M Medicine in particular have to be based as far as possible on evidence based educational teaching material. Most important, proven reproducible diagnostic procedures are mutually exchangeable and can stimulate discussions between the various approaches in M/M Medicine.

4. In the absence of reproducible diagnostic procedures in M/M Medicine only heterogeneously defined study populations for efficacy trials can be used. Therefore, comparison of results of efficacy trials, with the same therapeutic approach (for instance manipulation), is hardly possible. If the present situation continues, it will slow down of the badly needed process of professionalization of M/M Medicine.

5. Non-reproducible diagnostic procedures of different schools, ill-defined therapeutic approaches and low quality study designs are the main causes for the weak evidence of a proven therapeutic effect of M/M Medicine.

At present, it is still the opinion the IAMMM to create the best possible conditions for exchange of scientific information between the various schools in M/M Medicine. This information exchange must be based on results of solid scientific work. By comparing the results of good reproducibility studies, performed by different schools, a fundamental discussion can start. The main aim of this discussion is not to conclude which school has the best diagnostic procedure in a particular area of the locomotion system, but to define a set of validated diagnostic procedures which can be adopted by the different schools and become transferable to regular medicine. The Academy wants to provide the Societies for M/M Medicine with standardised scientific protocols for future studies.
The IAMMM International Instructional Course for Reproducibility Studies

The IAMMM Conferences

The best forum to create a discussion platform for the different schools in M/M Medicine is the Academy Conference organised by the IAMMM every year. This 2-day conference is organised every second year. At this Academy Conference, preliminary results of studies, proposals for research protocols, new developed therapeutic and/or diagnostic algorithms and other new scientific work are presented. In a fruitful discussion between audience and presenters many ideas can be exchanged based on solid scientific work, without interference of “school politics”.

The Scientific Director of the IAMMM, emphasises that good reproducibility of diagnostic procedures in M/M still has the first priority. These kinds of studies are easy and cheap to perform and form the best base for mutual discussion between schools in M/M Medicine. Co-operation with universities and active involvement of the Societies for M/M Medicine is indispensable and crucial for the future work of the IAMMM.
II. RELIABILITY: REPRODUCIBILITY AND VALIDITY

Definitions
Before performing a reproducibility study, it is essential that one becomes familiar with the nomenclature used in this kind of study. Besides, it is of utmost importance that the difference between reproducibility and validity is well understood. One of the major problems in medicine and also in research is the fact that different names are used for the same condition. Therefore we think it important first to provide the reader with an overview of the definitions used in this protocol. In clarifying and illustrating the definitions in greater detail, the reading becomes much easier. In particular those definitions that are used in the reproducibility studies are elaborated in greater detail based on the experience from previous reproducibility studies.

1. Reliability
In the scientific medical literature the term reliability is frequently (mis)-used in relation to the evaluation of diagnostic procedures. Reliability reflects the overall consistency of a measure. A diagnostic procedure is said to have a high reliability if it produces similar results under consistent conditions. Reliability comprises how well two persons use and interpret the same diagnostic procedure. However, reliability does not automatically imply validity.

Therefore Reliability must be subdivided into Precision and Accuracy. Precision is the same as Reproducibility and Accuracy is the same as Validity. Both terms Reproducibility and Validity are generally used, as is in this protocol.

{BinAbRahman:2015vy} {Dhillon:2003vk}
1.1 Reproducibility

Definitions
Reproducibility of a diagnostic procedure reflects the extent of agreement of a single person (observer) or different observers using the same diagnostic procedure in the same patient. In the case of a single observer we are dealing with intra-observer agreement or intra-observer variability. The same observer uses the same diagnostic procedure in the same patient but on two different occasions. (See figure 2)

![Figure 2](image1.png)

*intra-observer variability or the intra-observer agreement*

In the case of (two) different observers, we are dealing with inter-observer agreement or inter-observer variability. The (two) different observers use the same diagnostic procedure in the same patient at one occasion. (See figure 3)
Definition Diagnosis versus Diagnostic Procedure

In essence, the reproducibility of a diagnostic procedure or has nothing to do with a diagnosis as such. In medicine, diagnostic procedures are the constituent parts of a whole diagnostic arsenal that finally can lead to a particular diagnosis. Reproducibility of a diagnostic procedure reflects how well observers have standardised the whole diagnostic procedure as such and its final judgement. In the protocol, dichotomous outcomes for the final outcome of a diagnostic procedure such as Yes or No. To illustrate this in greater detail we take as example the Patrick Test (figure 4). In M/M Medicine the Patrick Test is frequently used to evaluate the mobility of the sacroiliac joint (SI-joint). However, in reproducibility studies, one has to separate this hypothesis of the Patrick Test (mobility of the SI-joint) from the Patrick Test as a diagnostic procedure as such. For instance, observers subjectively estimate the distance between the knee and the cough on both sides. (See double arrow in figure 4) Next, the distances are mutually compared and observers has agreed that the side with the largest distance has a positive Patrick Test.

If a positive found Patrick Test reflects a decreased mobility of the SI-joint at the same side is not proven yet. This concerns the validity of the Patrick Test. Validity will be explained later.
In this example the Patrick Test can be one of the constituent diagnostic procedures of a whole diagnostic arsenal for instance in patients with low back and leg pain. Based on a whole diagnostic arsenal, medical history and neurological examination included, the final diagnosis can be a lumbar radicular compression of the L5 root left. This is a genuine diagnosis in the sense that aetiology and prognosis are known. The Patrick Test as such in this example of a diagnostic arsenal does not provide you with final the diagnosis of a lumbar radicular compression of the L5 root left. The Patrick Test in this case is one of the many performed diagnostic procedures. Therefore, the reproducibility of a particular diagnostic procedure only has to do with standardisation of the various components of the procedure and the use of the same defined final judgement (in our protocol dichotomous outcome, Yes/No) and nothing to do with a diagnosis. This is very important to realise, because it explains why no selection procedure is needed for the study population of a reproducibility study because the diagnostic procedure that is evaluated in a reproducibility studies with a dichotomous outcome are independent of a final diagnosis that is based on a whole diagnostic arsenal.
In summary, reproducibility is about a fully, detailed and standardized description on how to perform the diagnostic procedure and a detailed and standardised description on how to measure and interpret the outcome of the diagnostic procedure, to ultimately reproduce their findings.
1.2 Validity

Definition
Validity measures the extent to which the diagnostic procedure actually does test what it is supposed to test. More precisely, validity is determined by measuring how a diagnostic procedure performs against the gold, criterion standard or reference test. In a separate protocol we will give a more detailed introduction to various forms of validity, among which the criterion validity probably is the best known.

To explain this form of validity in greater detail we use again the example of the Patrick Test. As mentioned in the previous chapter, the hypothesis of the Patrick Test was the testing of the mobility of the SI-joint.

Suppose we have already a good reproducible proven Patrick Test. The hypothesis of the Patrick Test was the evaluation of the presence or absence of mobility of the SI-joint. To evaluate the validity of the Patrick Test, we need a gold standard or reference test that can quantify the mobility of a SI-joint during the Patrick Test manoeuvre.

Figure 5. A hypothetical x-ray method is used to evaluate the mobility of the SI-joint. Two (blue) iron rods are placed on both sides of the SI-joint (A and B). X-rays are taken before and after a Patrick manoeuvre in the patient. The distance between A and B is measured before and after a Patrick manoeuvre. The difference between the distances (before and after the Patrick Test) is a measure for mobility.
Suppose we can use the stereophotogrammetry method with two iron rods (A and B, see figure 5) placed on both sides of the SI-joints during a Patrick Test manoeuvre. The distance between these two iron rods is measured on X-rays taken before and at the end stage of a Patrick Test manoeuvre. In case there is a difference in the measured distances, then the Patrick manoeuvre has introduced a movement in the SI-joint. Suppose we find a difference in the distances between the rods A and B before and at the end stage of the Patrick manoeuvre, and suppose this method has proven be reproducible. Than we have a gold standard or reference test for the Patrick Test. Subsequently, we now have to perform a validity study in which the Patrick Test and an X-ray SI-joint stereophotogrammetry (the gold standard) are simultaneously performed in the same group of patients. By comparing the results of the Patrick Test (figure 6) and the X-ray SI-joint stereophotogrammetry (figure 5) the validity of the Patrick Test can be estimated. In M/M Medicine many diagnostic procedures have been developed, each which its own hypothesis. However, we have to realise that gold standards and reference tests are lacking in the vast majority of these diagnostic procedures.
In all above-mentioned examples we have used the Patrick Test. As stated earlier, observers judge the Patrick Test by simply and subjectively estimating the distance of between left and right knee and the cough. (See double arrow in figure 7). In this case one has to develop a quantitative method, which measures this distance between the knee and the cough. This quantitative method can be used as the so-called reference test to estimate the validity of the Patrick Test with respect to a range of motion.

Figure 7. Patrick Test. The double black arrow illustrates the distance between the left knee and the examination table. The left/right difference can be used as an outcome of the diagnostic procedure.
III. REPRODUCIBILITY STUDIES: Nature of Diagnostic Procedures

1. Nature of Data in Reproducibility Studies
Before starting a reproducibility study one has first to realise what kind of diagnostic procedure we are dealing with. The nature of the data of a reproducibility study dictates the kind of statistics we have to apply. In general we have two kinds of diagnostic procedures: Qualitative and Quantitative.

1.2 Qualitative Diagnostic Procedures
Qualitative diagnostic procedures are the most used diagnostic procedures in M/M Medicine and are characterised by subjective interpretation of the observer with respect to the result of a performed diagnostic procedure (end feel, motion restriction, resistance). In qualitative diagnostic procedures, the outcome of the diagnostic procedure (the interpretation of the diagnostic procedure or outcome) can be divided in two kinds of data: nominal data and ordinal data. Outcomes of diagnostic procedures with nominal data refer to existence or absence of a particular feature and have a dichotomous character reflecting a contrast. Also the contrast between male and female in gender as outcome is a typical example of such a nominal data. Typical diagnostic procedures in M/M Medicine in which the outcome of the diagnostic produces nominal data are for example diagnostic procedures which evaluate “end feel” (abnormal Yes or No), pain provocation test (pain Yes or No) under different conditions (provoked by observer or provoked by movements of the patient) and range of motion (restricted Yes or No). For reproducibility studies with these kinds of data kappa statistics are most appropriate (see chapter IV).

If the outcome of a diagnostic procedure has different categories with a natural order we are dealing with ordinal data (good, better, best). An example is the outcome of such a diagnostic procedure, which evaluates the measure of range of motion, and is divided into minimal, moderate and severe restriction. Other examples are the character of end feel subdivided into normal end feel, soft end feel, and hard end feel. In this case weighted kappa statistics are indicated. This kind of ordinal data is also used in standard x-ray off the cervical spine in which the severity of the degenerative changes of the cervical spine are subdivided in categories. (Kellgren:1957tj){Cote:1997ta} In figure 15A there is a very high overall agreement of 0.99, the whole kappa/ P_index curve is located above the 0.60 kappa cut off level. In 98% the kappa value is ≥ 0.6. In case a low
overall agreement, the kappa/ $P_{\text{index}}$ curve will ultimately drop under the 0.60 kappa cut off level line and now

However, we have to wonder whether such a subdivision in ordinal outcome data, both in M/M Medicine and radiology, has any clinical consequence for the patient we see in our daily practice.

One has to consider if such a subjective subdivision of a diagnostic procedure outcome (normal, moderate, severe) has consequences for one’s therapeutic intervention or reflects a certain clinical condition of the patient. In M/M Medicine, in most of the cases there is no argument to use such a subjective subdivision of outcome of diagnostic procedures. Only in circumstances, in which one wants to use a diagnostic procedure to evaluate its outcome during a period of time, subjective subdivision is indicated. However, outcomes of diagnostic procedures, with subjective subdivision, are quite difficult to make reproducible. In particular, the problem is how to standardise this subjective subdivision of a diagnostic procedure. Besides, a gold standard or reference test is necessary to estimate the validity. In this case it is advisable to use a quantitative method with a device, which measures in detail the outcome of the diagnostic procedure.

1.3 Quantitative Diagnostics Procedures

In quantitative diagnostic procedures, mostly measured with a certain kind of device, findings are quantified in degrees, millimetres, kg etc. and are mentioned interval or continuous data. A good example is measurement of joint motion of the finger in degrees by goniometry.

First of all one has to evaluate the reproducibility of the device (test/retest). In this test/retest procedure, the systematic measurement failure can be estimated based on the dispersion of the data values.

For interval or continuous data the appropriate statistics are Intraclass Correlation and Paired T-test (two tailed). In case of several different interval data ANOVA analysis is indicated.

Secondly, a gold standard is needed to measure the validity of the method.

Thirdly, for these kinds of quantitative procedures with devices, normative values are needed. A study of the method in normal subjects is needed to estimate the effect of gender and age. Quantitative diagnostic procedures can serve as gold standards for qualitative diagnostic procedures.
2. **Inappropriate statistics of qualitative data in reproducibility studies**
   Frequently, inappropriate statistics are applied to measure the reproducibility of a diagnostic procedure. The main flaw is that agreement is often confused with trend or association, which is the assessment of the predictability of one variable from another. Hereunder the flaws of several statistical methods in reproducibility studies are listed.

2.1 **Percentage of Agreement/Overall Agreement/Observed Agreement** ($P_{obs}$).
   In reproducibility studies, using dichotomous outcome data, just mentioning the *Percent Agreement/Overall Agreement* gives no real information about the reproducibility of a diagnostic procedure. *Percent Agreement/Overall Agreement* is the ratio of the number of subjects in which the observers agree to the total number of observations. The main problem is that the *Percent Agreement/Overall Agreement* does not take into account of the agreement that is expected to occur solely by chance alone. In our protocol we use the term *Observed Agreement* ($P_{obs}$). This will be further explained in chapter IV.

2.2 **Correlation coefficients**
   In many reproducibility studies correlation and association measures are used to evaluate the reproducibility of clinical data. The problem is that some do not have the ability to distinguish a trend towards agreement from disagreement (Chi-Square ($\chi^2$) and Phi) or do not account for systematic observer bias (Pearson’s product moment correlation, Rank order correlation).{Bland:2003dl}{Tammemaggi:1995va}

3. **Appropriate statistics of qualitative data in reproducibility studies**

3.1 **Kappa statistics**
   Kappa statistics are the statistics of choice for evaluating intra- and/or inter-observer reproducibility for ordinal and nominal (dichotomous) data. This statistical method will be extensively explained in chapter IV.
3.2 Appropriate statistics in quantitative data reproducibility studies
To evaluate the reproducibility of repetitive measurements with quantitative data (interval or continuous data), the Paired T-test and/or Intraclass Correlation is indicated. This kind of statistic is used in cases of test/retest procedures when a device is used to quantify a clinical finding (range of motion).
One-way analysis of variance Intraclass Coefficient (ICC) is the statistical method of choice for the reproducibility of observers for interval data (cm, mm, etc). The calculated factor R in this statistical procedure is 1 if there are identical ratings, less than 0 in absence of reproducibility. A limitation of the ICC is that it provides no information about the magnitude of disagreement between observers.

3.3 Choice of statistics and clinical consequences
In reproducibility studies, the choice of statistics is not only dependent on the character of the collected data (nominal, ordinal, interval). It also depends on the type of clinical decision concluded from the findings of the reproducibility study.
Suppose the reproducibility of a leg length inequality has to be evaluated. The results of this reproducibility study have to be used to decide whether or not a heel lift is indicated to correct leg length inequality. In this case reproducibility has to be analysed by ANOVA ICC statistics for interval data. In contrast, if results of this reproducibility study have to be used to decide which side (left or right) has to be adjusted, kappa statistics are indicated for nominal data.

*In summary, in reproducibility studies of any kind, the nature of the collected data (nominal, ordinal, interval or continuous) and the final clinical purpose of the reproducibility study as such, are decisive for the applied statistical method.*
IV. REPRODUCIBILITY STUDIES: Kappa statistics

As mentioned already in chapter III, most diagnostic procedures in daily practice of M/M Medicine have an outcome of the diagnostic procedure with nominal data with a dichotomous character (Yes or No). For these kinds of dichotomous data, kappa statistics are the most appropriate. In this chapter the kappa statistics are explained and illustrated with the results of previous reproducibility studies to highlight different aspects of problems and flaws of this statistical method. Frequently used terms in kappa statistics are defined and explained. This chapter is essential to understand the reproducibility protocol elaborated in chapter V. Although many formulas will be shown in this chapter for illustration, all these formulas will be integrated in a spreadsheet (see chapter VII, paragraph 5) for automatic calculation of the kappa value and related data of a reproducibility study. You don't have to remember these formulas.

1. Definition Kappa Value

Kappa Value is a measure for inter- or intra-observer agreement (see chapter II, paragraph 1.1) corrected for the chance. Why do the kappa statistics correct for the chance? If you perform a diagnostic procedure on a patient with a dichotomous outcome Yes or No, just by chance (50%) you can judge a diagnostic procedure positive. In the kappa statistics this chance can be calculated with a formula. The result of this calculation is integrated in the final formula to estimate the kappa value (see chapter IV, paragraph 4).

To illustrate the kappa statistics in detail, we use an example of a hypothetical reproducibility study in which two observers A and B perform the Patrick Test in 40 patients. The outcome possibility of the diagnostic procedure was: Positive (Yes) or Negative (No). Both observers A and B examined the 40 patients with the Patrick Test and recorded their findings. By combining the results of both observers per patient at the end of the study four categories of results between observers are possible: 1. Both Observer A and Observer B judge in the same patients the Patrick Test positive (Yes/Yes), 2. Both Observer A and Observer B judge in the same patients the Patrick Test negative (No/No), 3. Observer A judges the Patrick Test positive, while Observer B judges the Patrick Test negative in the same patients (Yes/No), 4. Observer A judges the Patrick Test negative while Observer B judges the Patrick Test positive in the same patients (No/Yes). The results of these four categories can be depicted in a so-called 2x2 contingency table (see figure 8). In the rows and the columns are the
total numbers of patients in which the Observer A and Observer B judge the Patrick Test positive or negative. Observer A judged 17 Patrick Tests positive and 23 negative. Observer B judged 18 Patrick Tests positive and 22 negative. By adding per observer both figures the end result is 40.

Figure 8. The results of a reproducibility study with 40 patients and two observers A and B presented in a 2x2 contingency table. (see text).

Based on the data from this 2x2 contingency table different important aspects of the kappa statistics can be calculated. As shown later some of these aspects will influence the final kappa value.
Overall Agreement

Definition

In paragraph 2.1 the terms Percentage of Agreement/Overall Agreement/Observed Agreement ($P_o$) were introduced. In reproducibility studies, the Overall Agreement reflects the percentage of patients in which the observers agree about the outcome or judgement of the diagnostic procedure. In kappa statistics and in our protocol Overall Agreement is also named the Observed Agreement ($P_o$). This means that the Overall Agreement $P_o$ is calculated by the sum of the number of patients in which both observers judge the diagnostic procedure positive and negative, divided by the total number of patients of the study. In figure 9 below, a similar 2x2 contingency table is shown as in figure 8 but now based on a theoretical reproducibility study.

![Observer B](Observer A)

<table>
<thead>
<tr>
<th></th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td>(Yes/Yes)</td>
<td>(Yes/No)</td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>c</td>
<td>d</td>
</tr>
<tr>
<td>(No/Yes)</td>
<td>(No/No)</td>
<td></td>
</tr>
<tr>
<td>a+b</td>
<td>c+d</td>
<td></td>
</tr>
<tr>
<td>a+c</td>
<td>b+d</td>
<td></td>
</tr>
<tr>
<td>n</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 9. The results of a theoretical reproducibility study with n patients and two observers A and B presented in an 2x2 contingency table. (see text).

The formula for the Overall Agreement or Observed Agreement $P_o$ based on the data of figure 9 is:

$$P_o = \frac{a + d}{n}$$

Based on the data of the 2x2 contingency of the reproducibility study shown in figure 8 the Observed Agreement $P_o$ calculated as follows:

$$P_o = (15 + 20)/40 = 0.88.$$  

This $P_o$ will later be inserted in the final formula to calculate the final kappa value (see chapter IV, paragraph 4).
As will be explained later, the overall agreement is very important in a reproducibility study - because it influences strongly the magnitude of a kappa value (see chapter IV, paragraph 7).

2. Prevalence of the Prevalence of the Index Condition

Three new, used in reproducibility studies, are introduced: Index Condition, Prevalence and Prevalence of the index condition.

2.1 Index Condition

Definition
The index condition is synonymous with the positive judged diagnostic procedure of a patient participating in a reproducibility study. In figure 8 and 9 the index condition is illustrated by the “YES” In reproducibility studies with diagnostic procedure and a dichotomous outcome (final judgement), a positive judged diagnostic procedure by observers is referred to as the index condition.

2.2 Prevalence of the index condition

Definition
The prevalence of the index condition in reproducibility studies reflects the frequency of positive judged diagnostic procedures in the study population by both observers. In the 2x2 contingency table we have three boxes with a number of patients with a positive diagnostic procedure: the box with Yes/Yes, the box with Yes/No, the box with No/Yes). In the example of figure 9, these boxes are filled out wit a, b and c and in figure 11 with 15, 2 and 3. To calculate the prevalence of the index condition $P_{\text{index}}$ we need a special formula. Based on a theoretical 2x2 contingency table shown in figure 10, the formula for the prevalence of the index condition $P_{\text{index}}$ is:

$$P_{\text{index}} = \frac{[a + (b+c)/2]}{n}$$
Based on the 2x2 contingency table of the reproducibility study shown in figure 11 the formula for the prevalence of the index condition $P_{\text{index}}$ is:

$$P_{\text{index}} = \frac{[15 + (2+3)/2]}{40} = 0.44$$

Both Observed Agreement ($P_o$) and prevalence of the index condition ($P_{\text{index}}$) are important in a reproducibility study. Their values are decisive for the magnitude of the final kappa value in a reproducibility study (see chapter VI, paragraph 1).
2.3 **Prevalence**

Prevalence is a statistical concept referring to the number of subjects with a positive diagnostic procedure that are present in a study population. Because two observers examine the same subject in a reproducibility study, each examined subject can have both a positive and a negative judged diagnostic procedure. Therefore, a prevalence in the sense of the above-mentioned is not feasible in reproducibility studies with a dichotomous outcomes of the diagnostic procedure. Only the prevalence of the index condition \( P_{\text{index}} \) can be calculated. However, it is possible to calculate the prevalence of the positive judged per observer. In figure 11, the prevalence of the positive judged diagnostic procedure by observer A is: \((15 + 2)/40 = 0.43\) and for observer B \((15 + 3)/40 = 0.45\).

The prevalence of the index condition \( P_{\text{index}} \) in the example of figure 8 is 0.44.

3. **Expected Agreement by Chance**

As stated before, the kappa value is a measure for inter-observer agreement or intra-observer agreement corrected for chance. Because, if you perform a diagnostic procedure in a patient with the dichotomous outcome (Yes or No), you just by chance can judge a diagnostic procedure positive or negative. Therefore we have to calculate the expected agreement by chance \( P_c \). This expected agreement by chance \( P_c \) will integrated in the final formula to estimate a kappa value (see chapter IV, paragraph 4).

The formula for the expected agreement by chance \( P_c \) based on the theoretical reproducibility study shown in figure 12, is:

\[
P_c = \frac{a + b}{n} \times \frac{a + c}{n} + \frac{c + d}{n} \times \frac{b + d}{n}
\]
Figure 12. The results of a theoretical reproducibility with n patients and two observers A and B presented in a 2x2 contingency table. (see text).

The expected agreement by chance $P_c$ will be used for the final formula to estimate a kappa value (see chapter IV, paragraph 4).

Based on the 2x2 contingency table of the reproducibility study shown in figure 11 (page 28), the expected agreement $P_c$ can be calculated as:

$$P_c = \frac{17}{40} \times \frac{18}{40} + \frac{23}{40} \times \frac{22}{40} = 0.51$$

4. Calculation Kappa Value

To calculate the kappa value we need the observed agreement $P_o$ elaborated in paragraph 1 of this chapter and the expected agreement by chance $P_c$ of paragraph 4 of this chapter to be inserted in the formula for the kappa value $\kappa$:

$$\kappa = \frac{P_o - P_c}{1 - P_c}$$

When we apply the kappa formula on the data of the reproducibility study as shown in figure 13, the expected agreement by chance $P_c$ will be 0.51, the observed agreement $P_o$ is 0.88. Inserting these figures in the kappa formula leads to:

$$\kappa = \frac{0.88 - 0.51}{1 - 0.51} = 0.75$$
The prevalence of the index condition $P_{\text{index}}$ in this study is 0.44 (see page 28) with an Observed agreement $P_\text{o}$ of 0.88. (see page 26)

![Contingency Table](see text)

Figure 13. The results of a reproducibility study with 40 patients and two observers A and B presented in a 2x2 contingency table. (see text).
5. **Interpretation of Kappa Value**

The kappa value, as a measure for intra-observer or inter-observer agreement, can be either negative or positive. It can range between \(-1\) and \(+1\). Several schemes are available to interpret the kappa value of a reproducibility study. The most widely used is the scheme of Landis and Koch{Landis:1977vpa}. They stated that kappa values above 0.60 represent good to almost perfect agreement beyond chance between two observers. In contrast, kappa values of 0.40 or less represent absence to fair agreement beyond chance. Kappa values between 0.40 and 0.60 reflect a fair to good agreement beyond chance (see table 1).

However, the standards for strength of agreement provided by Landis and Koch is just an agreement about the in kappa interpretation. The same kappa value can be based on different values of the overall agreement \((P_o)\). A very low or negative kappa value can be the result of a very high or low \(P_{\text{index}}\) and does reflect the quality of the agreement between two observers about a diagnostic procedure. This will be further explained in paragraph 6 and 7 of this chapter.

<table>
<thead>
<tr>
<th>Kappa value</th>
<th>Strength of agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.20 - 0.00</td>
<td>Absence</td>
</tr>
<tr>
<td>0.00 - 0.20</td>
<td>Slight</td>
</tr>
<tr>
<td>0.21 - 0.40</td>
<td>Fair</td>
</tr>
<tr>
<td>0.41 - 0.60</td>
<td>Moderate or Good</td>
</tr>
<tr>
<td>0.61 - 0.80</td>
<td>Substantial</td>
</tr>
<tr>
<td>0.81 - 1.00</td>
<td>Almost perfect</td>
</tr>
</tbody>
</table>

Table 1. Diagram according to Landis and Koch{Landis:1977vpa} to interpret the kappa value of a reproducibility study. Strength of agreement is the same as reproducibility.
6. **Interpretation of Kappa Value: dependency of the overall agreement**

As already mentioned in paragraph 1 of this chapter, the overall agreement is a very important factor to interpret the kappa value of a reproducibility study.

In reproducibility studies, the overall agreement $P_o$ reflects the percentage of patients in which the observers agree about the outcome or judgement of the diagnostic procedure. More precisely, it reflects the total percentage in which both observers agree about positive and negative found diagnostic procedure in the same patients.

In the example of figure 14, the overall agreement $P_o$, is calculated by the sum of the number of patients in which both observers judge the diagnostic procedure positive and in which both observers judge the diagnostic procedure negative, divided by the total number of patients of the study. In our example the overall agreement $P_o = \frac{15 + 20}{40} = 0.88$.

![Figure 14. The results of a reproducibility study with 40 patients and two observers A and B presented in a 2x2 contingency table. (see text).](image)

Many published reproducibility studies in M/M Medicine, presenting too low kappa values without mentioning the overall agreement data. (period). In figure 15A and 15B the relation between the kappa/ $P_{index}$ curve and the overall agreement is illustrated, with curve of two different overall agreements $P_o$.

In figure 15A there is a very high overall agreement of 0.99, the whole kappa/ $P_{index}$ curve is located above the 0.60 kappa cut off level. In 98% the kappa value is $\geq 0.6$. In case a low overall agreement, the kappa/ $P_{index}$ curve will ultimately drop under the 0.60 kappa cut off level line and now kappa value will $\geq 0.6$.(figure 15 B) Because a part of the kappa/ $P_{index}$ curve drops below the zero line, negative kappa values can occur.
For instance, if the overall agreement $P_o$ decreases from 0.98 to 0.79, the kappa/ $P_{\text{index}}$ curve slowly shift downwards and become finally located under the zero line. The percentage with kappa values $\geq 0.6$ will decrease from 99% to 0%. In figure 16, all kappa/ $P_{\text{index}}$ curves of the $P_o$ interval 0.98 – 0.79 are depicted.
7. **Interpretation of Kappa Value: dependency of the prevalence of the index condition $P_{index}$**

As already mentioned in paragraph 2.2 of this chapter, the prevalence of the index condition $P_{index}$ is a very important factor. Not only how to interpret a kappa value of a reproducibility study, but the $P_{index}$ importantly can influence the level of the kappa value.

In reproducibility studies, the prevalence of the index condition $P_{index}$ is synonymous with the frequency of all positive judged diagnostic procedures (the index condition) by the observers. The $P_{index}$ has to be calculated by a special formula. (see paragraph 2.2). The relation between the kappa value and the $P_{index}$ is illustrated in figure 13.

![Figure 16. Relation between kappa value and prevalence of the index condition. The dotted line is the cut off level of 0.60. The kappa/prevalence index curves with an overall agreement $P_{obs}$ between 0.97 and 0.79.](image)

![Figure 17. Relation between kappa value and prevalence index of the index condition $P_{index}$. The dotted line is the cut off level of 0.60. The blue dots indicate low kappa values in case of a low (left blue dot) or a high $P_{index}$ value (right blue dot).](image)

The kappa/ $P_{index}$ curve in figure 17 illustrates that if the $P_{index}$ is low (0.2) or high (0.9), the matching kappa values will be under the kappa cut off line of 0.6. This means that in this study population of a reproducibility study, there are too few (low $P_{index}$) or too many (high $P_{index}$) positive found diagnostic procedures. As stated in paragraph 5, the standards for strength of agreement provided by Landis and Koch (Landis:1977vpa) (see table 1 below) was just an accordance about the in kappa interpretation and without a scientific base. The same kappa value can be based on different
values of the $P_o$ and $P_{index}$. A very low or negative kappa value can be the result of a very high or low $P_{index}$ and does reflect the quality of the agreement between two observers about a diagnostic procedure.

<table>
<thead>
<tr>
<th>Kappa value</th>
<th>Strength of agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.20 - 0.00</td>
<td>Absence</td>
</tr>
<tr>
<td>0.00 - 0.20</td>
<td>Slight</td>
</tr>
<tr>
<td>0.21 - 0.40</td>
<td>Fair</td>
</tr>
<tr>
<td>0.41 - 0.60</td>
<td>Moderate or Good</td>
</tr>
<tr>
<td>0.61 - 0.80</td>
<td>Substantial</td>
</tr>
<tr>
<td>0.81 - 1.00</td>
<td>Almost perfect</td>
</tr>
</tbody>
</table>

Table 1. Diagram according to Landis and Koch (Landis:1977vpa) to interpret the kappa value of a reproducibility study. Strength of agreement is the same as reproducibility.

The relation between the kappa value and $P_o$ and $P_{index}$ has consequences for the interpretation of kappa values in published reproducibility studies. As can see in figure 18, the same kappa value can be located on the left or right side of different kappa/ $P_{index}$ curves. Each kappa/ $P_{index}$ curve has its own $P_o$ value, ranging from an overall agreement of 0.97 till 0.79. The kappa value 0.4 can be due to a low overall agreement $P_o$, of which the top of the kappa/ $P_{index}$ curves is on or below the cut off line of 0.6 (inner curves near the left red arrow). Besides, the kappa value 0.4 can also be due to a too high or too low $P_{index}$, of which the top of the kappa/ $P_{index}$ curves is above the cut off line of 0.6 (outer curves near the right red arrow), because these kappa/ $P_{index}$ curves have a high $P_o$ value.

The same is partly true for reproducibility studies finding kappa values $\geq$ 0.6 without mentioning data about $P_o$ and $P_{index}$. Although authors conclude, based on the scheme of Landis and Koch (table 1), a very good reproducibility of the diagnostic procedure and subsequently advice to use this diagnostic procedure in daily practice, they were just lucky that the prevalence of the index condition $P_{index}$ was not too high or too low. The $P_{index}$ is always calculated after completing the study and therefore is not known in advance.

8. Interpretation of Kappa Value: bias

Bias is the case when observers produce different patterns of ratings or outcomes (Cicchetti:1990ww). No systematic pattern of scoring trends should be present by any observer. If a solid training phase is incorporated in the study protocol, bias should not be a
In the IAMMM protocol a well-defined training phase is incorporated.

Concluding, published reproducibility studies, that does not mention the values of the $P_o$ and $P_{\text{index}}$, and in which authors concluded an absence of clinical value because of a low found kappa value and using the standards for strength of agreement provided by Landis and Koch \cite{Landis:1977vp}, have to be rejected.

How to deal with the relation between kappa value and $P_o$ and $P_{\text{index}}$ in reproducibility studies will be elaborated later (see chapter VI, paragraph 2).

Figure 18. Relation between kappa value and prevalence of the index condition. The dotted line is the cut off level of 0.60. The kappa/prevalence index curves with an overall agreement $P_{\text{obs}}$ between 0.97 and 0.79. Red dots are all kappa values of 0.4. (see text)
V. DEVELOPING REPRODUCIBILITY STUDIES: General Aspects

1. What is the Nature of the Diagnostic Procedure(s) to be evaluated?
The first step, before starting a reproducibility study of a diagnostic procedure(s) in M/M Medicine, is to be clear about the nature of the diagnostic procedure(s) to be evaluated. In reproducibility studies and in daily medical practice, it is essential to realise the difference between a diagnosis, a syndrome and relation with diagnostic procedures.

Diagnosis
In a genuine diagnosis, by definition the aetiology and prognosis of the disease are known, for instance bacterial meningitis.

Syndrome
A syndrome is a combination of signs and symptoms that appear together in a high frequency in a certain population, for instance a sacroiliac syndrome, low back pain. The aetiology however is unknown or diverse.

Diagnostic procedure
In (M/M) medicine, diagnostic procedures are the constituent parts of a whole diagnostic arsenal that finally can lead to a particular diagnosis or syndrome.
A diagnostic procedure is a procedure, performed by a clinician, to identify and/or objectify in a qualitative (subjective) manner a clinical symptom of the patient. In both genuine diagnosis and syndromes, diagnostic procedures are needed. Seldomly, we can rely one single diagnostic procedure to make a diagnosis or define a syndrome.
For example the single finding of the absence of an Achilles tendon reflex does not constitute a lumbar radicular syndrome. Only the additional combination of findings of radiating pain, sensory deficit, motor deficit and a positive Lasègue are necessary to make the conclusion of a lumbar radicular syndrome. Since we are dealing with a syndrome, the aetiology can be as well as an intervertebral disc prolaps as a tumour in the intervertebral foramen, both with lumbar nerve root involvement. In our daily practice we are dealing with many non-specific clinical conditions, for instance low back pain. Since in low back pain 85% of the aetiology is lacking, we have to rely on diagnostic procedures to form principally syndromes of low back pain.
Also in our educational systems, many tests are thought to the students as a “diagnostic” procedure. For instance, diagnostic procedure for restricted passive cervical rotation. The students just learn how to perform the whole procedure of passive cervical rotation (setting of the hand, applied force etc.). Such a restriction can have many reasons and it therefore gives no information about a particular diagnosis or syndrome. Therefore, a combination of diagnostic procedures has to be performed, which all together point in the same direction towards a particular clinical syndrome or diagnosis. In summary, before starting a reproducibility study of a diagnostic procedure(s) in M/M Medicine observers have to agree about its nature and have to realise that:

a. **A single diagnostic procedure is never related to a particular diagnosis or syndrome.**

   *In a reproducibility study of a single diagnostic procedure, just the reproducibility of the execution of the whole performance of the diagnostic procedure and the judgement of the observers is evaluated (for instance a positive or negative judged Patrick Test).*

b. **Different diagnostic procedures are related to a particular syndrome.**

   *In a reproducibility study of a set of diagnostic procedures, just the reproducibility of the combination of the different diagnostic procedures in relation to a “Syndrome” is evaluated (for instance the absence (no) or presence (yes) of a sacroiliac syndrome). In this case the different diagnostic procedures must be mutually independent for the observers (see paragraph 2.2).*

c. **Several diagnostic procedures are related to a particular diagnosis.**

   *In a reproducibility study of a set of diagnostic procedures, the reproducibility of the combination(s) of the different diagnostic procedures in relation to a diagnosis are evaluated (for instance the absence (no) or presence (yes) of international criteria for rheumatoid arthritis of a knee). In this case the different diagnostic procedures must also be mutually independent for the observers (see paragraph 2.2).*
2. Number of diagnostic procedures evaluated in a reproducibility study

2.1 Too many diagnostic procedures
Reproducibility studies in non-specific clinical conditions, for low back pain, sometimes show the evaluation of a large number of diagnostic procedures at the same time, for instance all diagnostic procedures in the lumbar region. In these kinds of reproducibility studies, many of the diagnostic procedures at the end show low kappa values and subsequently it is concluded that this diagnostic procedures have no clinical value.

As already explained in paragraphs IV 5 and IV 6, the prevalence of the index condition $P_{\text{index}}$ and overall agreement $P_{\text{o}}$ influence greatly the final kappa value of a study. Since data of $P_{\text{o}}$ and $P_{\text{index}}$ are frequently lacking in studies evaluating many diagnostic procedures at the same time, a definite conclusion about the reproducibility of the diagnostic procedures with low kappa values cannot be drawn.

The largest flaw of this kind of reproducibility studies with many diagnostic procedures is the absence of a training period. As a consequence and very predictable a low overall agreement $P_{\text{o}}$ is obtained for many diagnostic procedures and therefore a low kappa value.

2.2 Combinations of a few different diagnostic procedures: mutual dependency
As mentioned already in paragraph 1 of this chapter, reproducibility studies can also evaluate a combination of diagnostic procedures in relation to the existence of a particular syndrome or diagnosis. In M/M Medicine, a combination of diagnostic procedures is frequently used in relation to sacroiliac syndromes. It was also stated in paragraph 1 that the individual diagnostic procedures in this combination have to be mutually independent. How to evaluate the mutual dependency of the diagnostic procedures evaluated in reproducibility studies?

Mutual Dependency of Diagnostic procedures
Kappa statistics are normally used for the agreement between observers — the inter-observer reproducibility - as illustrated in figure 19. In this example of a 2 x 2 contingency table we have two observers A and B.
Figure 19. The results of a reproducibility study with 40 patients and two observers A and B presented in a 2x2 contingency table.

The same kappa statistics can be used to determine the mutual dependency between diagnostic procedures used in a study. Instead of the two observers A and B, we now use per observer (A) a set of two tests (I and II). The agreement and disagreement between Test I and Test II of the examined patients is likewise estimated in a 2x2 contingency table (figure 20).

Figure 20. A 2x2 contingency table showing the agreements and disagreement between Test I and Test II of the examined patients of observer A to estimate the mutual dependency between Test I and Test II.
Based on the data in the boxes of figure 20, a kappa value can be calculated. A kappa value of $\geq 0.40$ means that there is a probability that Test I and Test II (figure 20) are mutually dependent diagnostic procedures for observer A. Also in this case the values of the $P_o$ and $P_{index}$ are necessary for proper interpretation of the kappa value.

The reason for such a mutual dependence of tests is the fact that observer A judges Test I positive, he subsequently and unconsciously judges Test II also positive. In a previous reproducibility study, the problem of evaluating too many tests in the same reproducibility study was illustrated. {VanDeursen:1993va} In this study, three observers were involved and used 6 SI-Tests to make a final conclusion of the presence of an SI-Joint dysfunction. The data of two observers A and B are used as an example. In the reproducibility study 6 SI-Joint tests (I, II, III, IV, V and VI) were used. Based on the 6 SI-joint tests observers A and B had to judge whether the examined patients have the SI-Joint Dysfunction Syndrome yes or no. Instead of showing an separate 2x2 contingency table for each observer A and B of all possible combinations of two SI-Test, the data are summarised in one single table (Table 2).

<table>
<thead>
<tr>
<th>Tests</th>
<th>I</th>
<th>II</th>
<th>III</th>
<th>IV</th>
<th>V</th>
<th>VI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Obsver</td>
<td>A</td>
<td>B</td>
<td>A</td>
<td>B</td>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>I</td>
<td>B</td>
<td>+0.61</td>
<td>+0.33</td>
<td>-0.12</td>
<td>+0.39</td>
<td>+0.19</td>
</tr>
<tr>
<td>II</td>
<td>A</td>
<td>-0.09</td>
<td>+0.02</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>III</td>
<td>A</td>
<td>+0.25</td>
<td>-0.01</td>
<td>+0.34</td>
<td>+0.17</td>
<td>+0.34</td>
</tr>
<tr>
<td>IV</td>
<td>A</td>
<td>+0.34</td>
<td>-0.29</td>
<td>+0.25</td>
<td>-0.05</td>
<td>+0.15</td>
</tr>
<tr>
<td>V</td>
<td>A</td>
<td>+0.61</td>
<td>-0.22</td>
<td>+0.18</td>
<td>+0.43</td>
<td>+0.89</td>
</tr>
<tr>
<td>VI</td>
<td>A</td>
<td>+0.61</td>
<td>-0.22</td>
<td>+0.18</td>
<td>+0.43</td>
<td>+0.89</td>
</tr>
</tbody>
</table>

Table 2. Kappa values between pairs of tests I to VI, subdivided per observer A and B.
(see text)

In the most upper row (yellow), 6 SI-Tests I to VI are listed. In the far left column, these tests I to VI are also listed from top to bottom but now in black. In the second left column, the observers A and B within each SI-Test
row are listed. In the next columns to the right the kappa values for each observer A and B per SI-Test I to VI are shown. These kappa values are calculated based on the principles used in the 2x2 contingency table presented in figure 20.

Table 2 has to be read in the following way. If one wants to look for a mutual dependency between test V and test VI of observer A, the first step is follow the yellow dashed line with arrow to the right, starting from observer A in left upper square till you reach the square under the yellow SI-Test V at the top row. Next, from this position, follow the vertical column of this test V downwards (black dashed line with arrow downwards), till you reach the horizontal row corresponding with test VI of observer A. The kappa value you will find in this case is +0.89 (see square right lower corner of the table). The kappa value +0.52, depicted beneath that of +0.89 illustrates the same relation between test V and VI but now for observer A. Both kappa values 0.89 and 0.52 are above 0.40 and when using standards for strength of agreement provided by Landis and Koch, both kappa values demonstrates a possible mutual dependency of the tests.

2.3 Combinations of a few different tests: mutual dependency of test and final “syndrome diagnosis”
In M/M Medicine in general and the SI-joint dysfunction syndromes in particular, reproducibility studies use combination of diagnostic procedures to make a final judgement about the existence of a clinical sign or syndrome. We use again the example of a study mentioned in paragraph 2.2 of this chapter. Observers A and B had to judge, based on six SI-Tests I to VI, the existence of a SI-Joint dysfunction syndrome - yes or no.{vanDeursen:1999wl}
To evaluate which of the six SI-Tests (I to VI) the observers (unconsciously) have used for their final judgement of SI-Joint dysfunction syndrome, kappa statistics can be applied again.
The data for estimation of the mutual dependency between a single SI-Test and the final judgement of a SI-dysfunction syndrome (Syndrome Diagnosis) of Observer A are presented in a 2x2 contingency table of figure 21.
Instead of showing an separate 2x2 contingency table for each observer A and B for all SI-Tests and the final conclusion of a SI-dysfunction syndrome, the complete data are summarised in one single table (Table 3).

![Contingency Table](image)

**Figure 21.** A 2x2 contingency table showing the agreements and disagreement between Test I and a final “Syndrome Diagnosis” (SI-dysfunction syndrome) of the examined patients of observer A, to estimate the mutual dependency between Test I and a final “Syndrome Diagnosis”.

**Table 3.** Kappa values of relation between separate SI-Tests I to VI and final judgement of the existence of a SI-dysfunction Syndrome, per observer A and B. (see text)
The kappa values in the dashed boxes in the right column are above 0.40 and when using standards for strength of agreement provided by Landis and Koch, that both observers mainly use (unconsciously) SI-Test V and VI for their final judgement. Observer A additionally uses probably SI-Test IV too for his final judgement. The other SI-Tests I to IV are hardly involved in the final judgement of observer A and B.

Another flow of reproducibility studies evaluating different diagnostic procedures for the same clinical phenomenon, for instance different SI-Test for SI-Joint dysfunction, is almost never clear in what “functional system” the different diagnostic procedures are located. Pain provocation SI-Tests and motion pattern SI-Tests (Vorlauf phenomenon) are located in two different systems resp. nociceptive system and postural system. The outcome of the diagnostic procedure have to be in the same functional system as the diagnostic procedure as such. For instance SI-joint provocation tests with the outcome of a numeric pain score.

Reproducibility studies, using combinations of several diagnostic procedures to make a final judgement of the existence of a clinical sign or syndrome and not evaluating the mutual dependency of the diagnostic procedures and/or mutual dependency of diagnostic procedures and final judgement, have no clinical value. The same is true for reproducibility studies, evaluating a set of diagnostic procedures for a single diagnosis, advocate the use of a minimal number of positive diagnostic procedures to confirm the final judgement, for instance, you 3 out of 6 diagnostic procedures.

Realising the amount of work of reproducibility studies evaluating, the many diagnostic procedures that have been developed in the six decades in M/M Medicine it is advisable to evaluate only one diagnostic procedure in a reproducibility study. Secondly, developing a new diagnostic procedure for M/M Medicine, it is advisable to perform a reproducibility study before publishing the new diagnostic procedure.
2.4 Large number of different diagnostic procedures in a “diagnostic protocol”

In some reproducibility studies, the observer(s) have to classify a patient within a particular system using a diagnostic protocol with a large number of diagnostic procedures. A well-known example is the McKenzie System that distinguishes several different syndromes for instance for low back pain. However, frequently the single diagnostic procedures were not evaluated with respect to their reproducibility properties. Although, observers may agree to a large extent about their final judgement to classify a patient with the diagnostic protocol, it is unclear what diagnostic procedure(s) or combination of diagnostic procedures the observers used for their conclusion. This kind of reproducibility studies using a diagnostic protocol with a large number of diagnostic procedures has to incorporate only reproducible proven diagnostic procedures. Besides the mutual dependency of the diagnostic procedures and with the final judgement has to evaluated as part of the reproducible study. as illustrated in paragraph in the previous paragraphs V.2.2 and V.2.3).

3. Hypothesis of the diagnostic procedure in a reproducibility study

The hypothesis of a diagnostic procedure as such can influence the final result of a reproducibility study. More precisely, there is a relation between the extent of agreement (read kappa value) and the supposed hypothesis by the observers participating in the study. In general, hypothesis means what the observers assume what their diagnostic procedure really is supposed to test. In case of a simple hypothesis such as range of motion there is no problem. The problem arises when observers just adapt the hypothesis of a diagnostic procedure from their textbooks or what they were taught in their M/M Medicine courses. A well-known example is the mobility of the sacroiliac joint (SI-joint). In M/M Medicine, a vast number of SI-joint Tests have been developed, all supposedly testing the mobility of the SI-Joint. Looking carefully and critically at all these different SI-Joint Tests, we have to question whether all these diagnostic procedures evaluate the same aspect of the SI-Joint mobility. All the more, because all these different SI-Joint Tests differ substantially in their performance.
Although it has been proven in cadaver studies that mobility of the SI-joint exists, it is impossible, even for the most experienced observer, to test manually the mobility of the SI-Joint. Nevertheless in many reproducibility studies involving SI-Joint Tests, this incorrect hypothesis is still the starting point. In a reproducibility study, incorrect hypothesis as such can influence the observer agreement and consequently the final kappa value of the study. Because it is essential to understand the effect of the hypothesis, two examples from previous performed reproducibility studies are presented to illustrate this phenomenon. In a former reproducibility study 3 observers (A,B,C), wanted to evaluate the reproducibility of hypo-mobility of the SI-joint, based on 6 SI-Tests (I to VI). Their hypothesis of the used SI-Tests was that all these diagnostic procedures could demonstrate the presence or absence of mobility of a SI-joint. The three well-experienced observers (all were M/M Medicine course leaders) adapted the hypotheses of the 6 used SI-Tests from literature. In table 4 the kappa results are listed between observers (A↔B, A↔C, B↔C), per SI-Test (I to VI) and with respect to their final judgement of the absence or presence of a SI-joint hypo-mobility (SI-dysfunction syndrome diagnosis).

<table>
<thead>
<tr>
<th>Observers</th>
<th>SI-Test</th>
<th>I</th>
<th>II</th>
<th>III</th>
<th>IV</th>
<th>V</th>
<th>VI</th>
<th>SI-Diagnosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>A↔B</td>
<td>0.11</td>
<td>-0.08</td>
<td>-0.05</td>
<td>+0.29</td>
<td>-0.16</td>
<td>-0.05</td>
<td>-0.05</td>
<td>-0.05</td>
</tr>
<tr>
<td>A↔C</td>
<td>+0.08</td>
<td>+0.10</td>
<td>+0.38</td>
<td>+0.20</td>
<td>+0.06</td>
<td>+0.14</td>
<td>+0.14</td>
<td>+0.14</td>
</tr>
<tr>
<td>B↔C</td>
<td>+0.03</td>
<td>-0.16</td>
<td>-0.23</td>
<td>+0.05</td>
<td>+0.13</td>
<td>-0.09</td>
<td>-0.09</td>
<td>-0.09</td>
</tr>
</tbody>
</table>

Table 4. Kappa values of the inter-observer agreement (A↔B, A↔C, B↔C) in a reproducibility study performed by three observers (A, B, C) using 6 SI-Tests (I to VI) to make the final judgement about the mobility of the SI-joint. (see text)

Note all kappa values between pairs of observers are below 0.60 both for the individual SI-Tests I to VI and for the final judgement of the absence or presence of a SI-joint hypo-mobility.
In a second reproducibility study{Patijn:2000dq}, the same two observers (A, B) from the previous study mentioned above, wanted to evaluate the reproducibility of the SI-joint dysfunction based on 3 SI-joint tests (test I, test II, test III from the above-mentioned first study). Observers first renounce their previous hypothesis of the used three SI-Tests. Namely, that all these three tests could determine the extent of the SI-mobility. Secondly, by very precisely looking at all aspects of the performance of the diagnostic test procedures and their judgement, the observers A and B concluded by mutual deliberation that all three SI-Tests measured increased muscle tone of different muscle groups related to the lumbo-sacral-hip complex. Because no structural abnormalities were found, a SI-joint dysfunction was assumed. Observers argued that increased muscle tone led to motion restriction and resistance at the end of the passive performed procedure. Based on these 3 SI-Tests, the observers had to judge whether or not a SI-joint dysfunction existed. In figure 22, the 2x2 contingency table of this study is presented together with the kappa value, prevalence index and overall agreement.

<table>
<thead>
<tr>
<th>SI-joint dysfunction</th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observer A</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>38</td>
<td>0</td>
</tr>
<tr>
<td>No</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Prevalence: 0.85
Overall Agreement: 0.98
Kappa Value: 0.7

Note, that the kappa value has risen to 0.70 just by changing the hypothesis of three SI-Tests (I, II, III) used in this reproducibility study. In the first study (see table 4) the kappa values of SI-Tests I, II and III were resp. 0.11, -0.08 and -0.05.
Whatever diagnostic procedure is selected for a reproducibility study, step by step the whole diagnostic procedure and its final judgement has to be analysed for observers to agree about what they think the diagnostic procedure really is. Based on this agreement, the observers can define a more plausible hypothesis for the diagnostic procedure, which can completely contradict the hypothesis stated in the literature. Therefore, before analysing the diagnostic procedure, sometimes the original described diagnostic procedure in the literature has to be renounced.

4. **Characteristics and number of observers to be involved in a study**

4.1 **Number of Observers**

In published reproducibility studies, the number of observers participating in the study varies from 2 to sometimes 10. Because of a better clinical application of a diagnostic procedure, some authors advocate the use of more than two observers in a reproducibility study. Authors simply argue that the more observers agree about a diagnostic procedure, the better are the reproducibility properties of that diagnostic procedure. However, this assumption is based on a serious logical error. Reproducibility studies are primarily meant to provide us with information about all the aspects of the reproducibility properties of a diagnostic procedure. This means that the number of observers in essence has no relation to the reproducibility properties of a diagnostic procedure as such in a reproducibility study. Before starting a reproducibility study, the two observers have to agree about all the details of the performance of the diagnostic procedure and its final judgement.

As will be explained in the reproducibility protocol format (see chapter VII) this agreement is acquired by introducing a training phase in the protocol format of the study. If in a reproducibility study several observers are used, who have not passed the training phase of the protocol, the final low kappa values reflects more the personal interpretation or their comprehension of the non-trained observers instead of the reproducibility properties of the evaluated diagnostic procedure.
Therefore, only two observers are needed in a reproducibility study, if only the reproducibility property of a diagnostic procedure have to be evaluated.

If a reproducibility study is meant to evaluate the effect of education on several participating observers by implementing in the study protocol several training phases, more than two observers can be used to participate in the study. {Degenhardt:2005wu}

4.2 Characteristics of Observers

In many reproducibility studies, observers with different levels of skills are involved. These levels are used as a predictive or explanatory factor for the level of the kappa values found by the different observers involved in the study. For using observers with different levels of skills in reproducibility studies, the same objections count as for the idea to use more than 2 of observers in a study.

Reproducibility studies are primarily meant to provide us with information about all the aspects of the reproducibility properties of a diagnostic procedure. This means that level of skills of the observers in essence have no relation to the reproducibility properties of a diagnostic procedure as such. Before starting a reproducibility study, the observers have to be independent from their personal skills, and to agree about all the details of the performance of the diagnostic procedure and its judgement in the trainings phase.

As will be explained in the reproducibility protocol format (see chapter VII) this agreement is acquired by introducing a training phase in the protocol format of the study, in case only the reproducibility properties of a diagnostic procedure has to be evaluated. If in a reproducibility study several observers are used, who have not passed a training phase of the protocol, the final obtained kappa value reflects for instance more the personal interpretation of the well-experienced observer and the comprehension of the evaluated test of the less experienced student instead of the reproducibility properties of the evaluated diagnostic procedure.
Over the years of their profession, well-experienced practitioners in M/M Medicine have unconsciously developed their own personal interpretation about the performance and about the judgement of a diagnostic procedure. As a consequence, their diagnostic procedure may differ from the original described test in literature. In students, a lack of experience with the diagnostic procedure may play a role and influence the final kappa value. It is emphasised in our protocol format (see chapter VII, paragraph 2) to implement a training phase, for each observer irrespective the level of skill. Only than a standardisation of the performance and judgement of a diagnostic procedure is guaranteed.

5. **Number of patients to be involved in a reproducibility study**

In previous editions, the number of 40 patients was arbitrary and was seen as a “statistical minimum” to perform these kinds of studies. Also from a practical point of view a rounded number of 40 was chosen to make these kinds of reproducibility studies easy and cheap to perform. In general, it was stated that for simple reproducibility studies, 40 subjects were sufficient. Nowadays, power calculations are advised to estimate the study sample size. However, power calculations for sample size are only possible in case a null hypothesis have been formulated such as in randomised controlled trials (RCT). Kappa statistics were designed for descriptive purposes and as a basis for statistical inference, but kappa statistics are typically not used as a null hypothesis–testing statistic. {Donner:2010hk}{Tooth:2004gh} Different other approaches have been developed, but were mainly meant for multiple observers with a dichotomous outcome variable. {Donner:2010hk}

Indirectly related with the problem of the sample size of a reproducibility is the fact that kappa value of a diagnostic procedure is not an absolute measure. Its value is dependent of the prevalence of the index condition \( P_{\text{index}} \) and the overall agreement \( P_o \) (see paragraph IV 6 and 7). This means that the same kappa value can have different levels of the \( P_o \) and the \( P_{\text{index}} \). In a reproducibility study, the kappa value is only related with the positive judged diagnostic procedures. In figure 23 the squares with a (yes/yes), b (yes/no) and c (no/yes) are decisive for the final kappa value. The overall agreement \( P_o \) concerns both the positive and negative judged diagnostic procedures, depicted in the squares a (yes/yes) and b (no/no).
Figure 23. Example a theoretical reproducibility with n patients and two observers A and B presented in a 2x2 contingency table. (see text).

The overall agreement $P_o$ is an absolute measurement and reflects the daily practice of a clinician dealing with diagnostic procedures. As a clinician one wants to know how reproducible his diagnostic procedure is, both for a positive and negative final judgements of his diagnostic procedure. The overall agreement $P_o$ is the most appropriate measure for conveying the relevant information in a 2x2 table and is most informative for clinicians. {deVet:2013ks} However, when using dichotomous outcomes, we have to realize that the overall agreement $P_o$ is not corrected for the chance. In the previous IAMMM protocols the level of the $P_o$ was chosen to guarantee always kappa values $\geq 0.6$. If the $P_o$ of a reproducibility study is 0.79, no kappa value $\geq 0.6$ can be obtained.

Figure 24. Relation between kappa value and prevalence of the index condition. The dotted line is the cut off level of 0.60. The kappa/prevalence index curves with a too low overall agreement $P_o$ of 0.79 is located beneath the cut off level line of 0.60. The kappa/prevalence index curves with overall agreement $P_o$ of 0.97 is located far above the cut off level line of 0.60.
Dependent of the level of the $P_o$ the percentage of kappa values $\geq 0.60$ will rise ($P_o 0.80 - \kappa \geq 0.60$ in 25%, $P_o 0.82$ in 45%, $P_o 0.83$ in 52%), at the same time the kappa/prevalence index curves will shift upwards when the $P_o$ increases (see figure 24).

Since the overall agreement $P_o$ is an absolute measurements and reflects more the daily reality of a clinician, calculating the sample size of a study population, we first focus on the overall agreement. The question arises: what, from clinical point of view, is an acceptable level for the $P_o$? Important is the fact that the results of reproducibility studies may not change clinical practice in a way that leads to poorer patient outcomes.\cite{McHugh:2012up} A precise cut off level for the $P_o$ is not available in the literature. This is the reason that many texts in literature only recommend acceptable levels of overall agreement range from 0.75 op to 0.90 \cite{McHugh:2012up,Birkimer:1979vv,Graham:vb,Hartmann:1977tw}.

From statistical point of view a level of the $P_o$ of 0.80 or higher can guarantee a kappa value above the cut off level of 0.60. In case of a $P_o$ of 0.80 25% of the possible kappa values is larger than 0.60. A $P_o$ of 0.83 provides 50% of the possible kappa value larger than 0.60.\cite{figure 25}
VI RELATION BETWEEN THE KAPPA VALUE AND THE PREVALENCE OF THE INDEX CONDITION $P_{\text{index}}$

1. Defining the $P_{\text{index}}$ Problem

As already mentioned and elaborated in Chapter IV, paragraph 6, one cannot correctly interpret the kappa value of a reproducibility study without knowing the prevalence of the index condition $P_{\text{index}}$. However, one of the major disadvantages of kappa statistics in reproducibility studies is the fact that the $P_{\text{index}}$ is not known in advance. Only after completion of the study can the $P_{\text{index}}$ be calculated. In all reproducibility study there is always a risk that the $P_{\text{index}}$ is far too high (too many positive judged tests) or far too low (too few positive judged tests). In figure 20 the relation between the kappa value and the $P_{\text{index}}$, is shown again.
Figure 20. Relation between kappa value and prevalence index of the index condition $P_{\text{index}}$. The dotted line is the cut off level of 0.60. The blue dots indicate low kappa values (arrows) in case of a low (left blue dot) or a high $P_{\text{index}}$ (right blue dot of the index condition.
Both blue dots in figure 21 illustrate the risk that, after completion of the reproducibility study, the kappa value appeared to be too low (far under the cut off level of 0.60) as a consequence of a too high $P_{\text{index}}$ value (too high frequency of positive judged tests in the study population) or too low $P_{\text{index}}$ value (too low frequency of positive judged tests in the study population).

Theoretically, a $P_{\text{index}}$ value of 0.50 is preferable because with that value the kappa value will be located at the top of the kappa/prevalence index curve. (see figure 21)
Figure 21. Relation between kappa value and $P_{\text{index}}$. The broken line is the cut off level of 0.60. The blue dot indicate a kappa value in case of a $P_{\text{index}}$ of 0.50 (red figure).

As can be seen in the kappa/$P_{\text{index}}$ curve of figure 21, the inter-observer reproducibility level is well over 0.80, because a large part of the kappa/$P_{\text{index}}$ curve is above the cut off level of a kappa value of 0.6 – the broken horizontal line.
2. **Influencing the $P_{\text{index}}$ in advance: the 0.50-$P_{\text{index}}$ method**

In a fictitious reproducibility study, two observers A and B wanted to evaluate the reproducibility of test I. They used a study population of 40 patients. Both observers A and B examined first 20 patients with Test I (see figure 22).

![Flow diagram of a reproducibility study](image)

Figure 22. Flow diagram of a reproducibility study with two observers A and B. Both observers perform Test I in their own patients ($n_1$ and $n_2$). Both observers send their patients to each other (arrows). Each observers examine a total of 40 patients.

Subsequently observer A sent his examined population $n_1$ to observer B. Observer B did the same by sending his population $n_2$ to observer A.

At the end, both observers have examined 40 patients with Test I. It is essential that both observers in advance agreed that mutual communication as well as communication with the examined patient was not allowed. By this blinding procedure, there is a 50% chance that observer B will have a positive or a negative test when he examines every patient sent to him from observer A.
However, as stated before, the number of positive tests by both observers is not known in advance. More precisely, there is always a risk in this study format to get a too high or too low $P_{\text{index}}$, which can result in an unwanted low kappa value as a measure for inter-observer agreement.

The 0.50-$P_{\text{index}}$ method: As explained in paragraph 1 of this chapter, a $P_{\text{index}}$ of 0.50 is preferable because the kappa value will be located at the top of the kappa/ $P_{\text{index}}$ curve (see figure 21). We can take the same fictitious reproducibility study from figure 22, in which two observers A and B wanted to evaluate the reproducibility of test I.

Figure 23. Flow diagram of a reproducibility study using the 0.50-$P_{\text{index}}$ method. Both observers A and B perform Test I in their own patients. Both observers send 10 positive and 10 negative tests to each other. The overall agreement in advance was 85%, (see text).

Suppose that they, before starting the reproducibility study with the 40 patients, managed to obtain an overall agreement of, let us say, 85%.
Instead of sending just 20 patients to each other as illustrated in figure 22, both observers A and B now send 10 patients with a positive Test I and 10 patients with a negative Test I to each other (see figure 23). By this way observer B receives 20 patients from observer A and vice versa. This is what we call the 0.50- \( P_{\text{index}} \) method, Since the whole procedure was blinded (mutual communication as well as communication with the examined patient was not allowed) again every patient sent by observer A had 50% chance for observer B to have positive or negative test. According to the obtained 85% overall agreement between observers, observer B will agree in 8.5 (0.85x10) of the 10 patients with a positive Test I and will disagree in 1.5 (0.15x10) of the 10 patients with a positive Test I. The same is true for the patients sent by observer B to observer A. The end result is that observers agree in 17 patients with a positive or negative Test I and disagree in 6 patients. The results of this fictitious reproducibility study are presented in a 2x2 contingency table (figure 24).

![Contingency Table](image)

**Figure 24.** A 2x2 contingency table showing the agreements and disagreement between observer A and B about the existence of a positive Test I, using the Pervalence-0.50-method.

This theoretical format of the 0.50- \( P_{\text{index}} \) method has been proven in
a real reproducibility study (Patijn, J Orth Med, 2004).
In that study, two observers P and E evaluated the reproducibility of
the passive hip flexion tests. After a training phase and overall
agreement phase, they obtained an overall agreement level of 88.2%.
In the final kappa phase, using the 0.50-P_index method in 40 patients,
the obtained prevalence of the index condition appeared to be 0.44 (fig. 25), which is near the ideal value of 0.50.

As expected the overall agreement remained stable. A kappa value of 0.74 was obtained.

**In summary, the 0.50-P_index method has been proven to be feasible and to solve one of the main drawbacks of kappa statistics used in reproducibility studies. However, the method cannot be used if more than one test is involved in a reproducibility study!**
As shown in figure 26, an entire reproducibility study can be subdivided into six different periods or phases, which successively have to be passed through. Each phase is characterised by different components which are essential for that particular phase. The presented protocol format is developed for two observers evaluating one single diagnostic test at a time. The arguments for this decision is elaborated in chapter V, paragraph 2.

Figure 26. Flow chart of planning in different phases of a reproducibility study.
1. **PREPARATION PHASE**

In the preparation phase, mainly agreements about the study conditions and logistics of the reproducibility study are concluded. In this phase, when necessary, consent of a local medical ethical committee has to be obtained.

1.1 **Study Conditions: Participating Members and Logbook**
First of all one has to form a research group with members who are going to participate in the reproducibility study. Provide the members with this protocol. Discuss the main outline of the purpose of the study. An important issue of this phase is the introduction of a logbook to be used during the whole study. In this logbook all the agreements between participating members are recorded. In case of disagreement, the observers can always check in the logbook which previous agreements were made.

1.2 **Study Conditions: Transparency of Responsibility**
Secondly, an important aspect in this phase is the nomination of one person who has the final responsibility for the whole reproducibility study.
This person in particular is responsible for updating the logbook. Also in this preparation phase the sequence of authorship for the publication has to be fixed. The responsible person decides in cooperation with the group who is doing what during the study, for instance developing an evaluation form for the study. The responsible person can be one of the observers, performing the diagnostic test, or a chosen independent and participating colleague.

1.3 Logistics Study
An important issue is how to arrange the logistics of a reproducibility study. The best circumstances are when both observers work in the same outdoor clinic or institute and have consulting hours at the same time. In the different phases of the protocol, except the training phase, it is easy to recruit patients for the study. This is essential, because both in the overall agreement phase (see paragraph 3) and the study phase (see paragraph 4) observers have to send patients to each other. Special arrangements have to be made for the training phase, in which both observers have to examine at least 10 patients in detail to agree about the test performance and its final judgement. It is advisable to reserve special time for the training phase. Both the overall agreement and the study phase can easily be performed during the regular consulting hours of an outpatient clinic when both observers at the same time see their patients.

1.4 Finance
In essence a reproducibility study costs no extra finance outside the time spent by the observers in the different phases of the protocol. Therefore, these kinds of studies are feasible for all kinds of clinics of M/M Medicine. No extra support from statistical experts is needed, because an “excel” file is added in this protocol which automatically calculates all the necessary data and kappa values when basic data are filled into the spreadsheet (see chapter VII, paragraph 5).
1.5 Approval by the Local Committee of Ethics.
The final research protocol inclusive a copy of the written Information to the patients etc. must be forwarded to the Committee of Ethics.

Before doing so, be careful that all the obligations related to this process is fulfilled

2. TRAINING PHASE

2.1 Observers and Patient Selection
The training phase is a very essential phase of the reproducibility study. In this phase, the basis is created for both a successful outcome of the overall agreement phase and for the final results of the reproducibility study; namely a good kappa value. In general only two observers (see Chapter V, Paragraph 4.1) are needed and 10 different patients. These patients can be randomly selected from the observers’ own patient files or from the planned daily outpatient clinic programme. No particular exclusion or inclusion criteria are necessary for these 10 patients. Each observer successively chooses a patient - who will have a positive or negative test.

Figure 28. Trainings Phase of the Reproducibility Protocol.
2.2 Selection and number of Tests
Observers must agree upon which tests they wish to evaluate in the reproducibility study. Of course, a literature study is mandatory to decide whether the tests have been evaluated before. The aspect of the number of tests has been discussed in chapter V.2. In general, it is advisable to evaluate one test at a time. In case of more than one test to be evaluated at the same time, the 0.50-prevalence index method can only be applied for one of the tests (see Chapter VI, paragraph 2).

1. Agreement about Test Performance
After agreement upon which test to evaluate in the study, a preliminary description of the test performance takes place. It is advisable to have the original description of the test, and based on that, to make the first description of the test(s). This description must be very detailed and specific, taken both the observer (examiner) and the patient into account. How are the two of them positioned? How are the hands of the observer placed on the patient? Etc. Etc.
Then training of the two observers starts.
For this training phase, special time (in the form of sessions) has to be reserved. Besides training on patients it is advisable for the observers to start the first session with the performance of the test on each other. Subsequently both observers have to examine three or four patients in the following sessions. This examination can be done in two to three sessions (3 x 4 hours). Both observers have to examine the same patient in detail. Per patient one observer performs the test to be evaluated for reproducibility, while the other observer observes. Next the roles of the observers change. The examining observer informs the other observer about all the details of his performance and judgement of the test. The following elements of the test procedure have to be discussed in detail for later standardisation of the entire test performance:

1. position of the patient
2. position of the observer
3. position or placement of the left and right hand and or fingers
4. direction of the passive or active motion
5. anatomical land mark for the directed motion

During the examination of the 10 patients, the observers by consensus have to agree about all the details of the performance of the test, and this must be recorded in the logbook.

*It is advisable for both observers to train the test by performing routinely the test on all their outpatient clinic patients.*

1. **Agreement about Test Hypothesis**
In chapter V.3 we discussed that the hypothesis of a diagnostic test as such can influence the final result of a reproducibility study. Most of the tests mentioned in textbooks
and course syllabuses are based on unproven hypotheses. Therefore, in reproducibility studies observers have to forget these hypotheses. Based on the detailed performance of the test, observers have to agree about what the test actually tests. For instance the passive hip flexion adduction test is supposed to test the mobility of the SI-joint (textbook hypothesis). Looking closely at the performance of the test, it is much more plausible that the range of motion of the passive hip flexion adduction is dependent on the muscle tone of different muscle groups related to the lumbo-sacral-hip complex (working hypothesis).

The same is true for pain provoking tests. Some pain tests are supposed to identify a particular structure as the source for the pain, for instance the SI-joint. However in all kinds of SI-joint pain provoking tests, many more anatomical structures than the SI-joint can be the source for pain. The best working hypothesis in this kind of test is that different structures functionally related to the SI-joint can be the cause of the pain. In general most tests in M/M Medicine are related to range of motion and don’t give rise to problems of defining the working hypothesis. In other cases, a working hypothesis of the test can be defined by the observers, by carefully looking at all the details of the performance of the test under search.

2.5 Agreement about Test Judgement
Both the agreements of the two observers about the performance and the hypothesis of the test are decisive for the final judgement of the test. The observers have to look carefully on how they normally use the test in their daily practice and in particular how they judge the positive or negative result. Frequently, and with very experienced practitioners, this judgement happens almost semi-automatically. The participating observers have to be very careful to look how they judge a test in daily practice and mutually compare these judgements. Sometimes, a semi-quantitative method is necessary. For instance in a
reproducibility study of the passive hip flexion adduction test the left/right difference can be semi-quantified, using the number of fingers to measure distance between the chest and the knee (Patijn, J Orth Med, 2004). In figure 29 it is presented how to semi-quantify a difference in range of motion of the passive hip flexion adduction test. The black hand represents the number of fingers between the chest and the knee in the right hip flexion test. The same procedure is repeated on the left side.

Figure 29. Semi-quantification by number of fingers of the passive hip flexion adduction tests.
The numbers of fingers of both sides is estimated. A left/right difference of more than one finger is decisive and the side with the largest number of fingers is the most restricted side and labelled as a positive test. Dependent on the test, observers have to agree about how they judge a test to be positive or negative, and whether it is necessary to semi-quantify this judgement. In general it is advisable to look how observers have to use the test in their daily practice. The decision about the judgement of the test is recorded in the logbook.

2.6 Evaluation Form.
Based on the results of the performance and judgement discussions, an
form is developed which has to be used in the study (see figure 30)

REPRODUCIBILITY EVALUATION FORM
PASSIVE HIP FLEXION
ADDUCTION TEST

OVERALL AGREEMENT PHASE ■ STUDY PHASE □

Patient Registration Number □□□□□ Male ■ / Female □

Aim:
Number patients: 20 (10 per observer)
Number Observers: 2
Inclusion Criteria: Pain Locomotion System
Exclusion Criteria: None
Selectie: Consecutive
Logistics: Direct performance of test
Blinding: No mutual communication between observers and patient and observer

Hypothesis Test:

Performance test:

Sem-Quantification:

Judgement Test:

<table>
<thead>
<tr>
<th>Test</th>
<th>Left positive</th>
<th>Right positive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Passive Hip Flexion Adduction</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

evaluation form used in the overall agreement and study phase
This evaluation form is used in the overall agreement period as well as in the study period. In this form a brief overview of the study is mentioned. But more importantly, the details about the patient, the performance, the semi-quantification and the judgement of the test are recorded. At the bottom of the form the positive side of the test can be registered.

3. OVERALL AGREEMENT PHASE

The main aim of the overall agreement phase is to obtain an overall agreement $P_0$ percentage higher than 80% (see chapter IV, paragraph 8).

11. Two Observers
12. Selection of 20 patients
13. Blinding Procedures

Overall Agreement Phase
($> 80\%$)

Figure 31. Overall Agreement Phase of the Reproducibility Protocol.

3.1 Observers and Patients Selection
The same two observers of the training phase perform the overall agreement phase. Because the observers have to send patients to each other, it is preferable that both observers work at the same clinic, so that the program of the clinic can be organised with respect to the study.
Each observer has to select at random 10 patients out of his policlinic programme.
The process can be done as follows:
Forthy evaluation forms are made, as each observer examines $10 + 10$ patients, as well as 40 forms for patient data.
In principle, all patients with pain of the locomotion system are suitable for inclusion in the overall agreement phase. An exception is
the case of the test which is related to a particular disease entity with special inclusion and exclusion criteria. Each observer performs the test in all the patients of his policlinic programme and select patients for the study in which he is convinced that the test is positive or negative.
In case we examine a lumbar test, both a negative or a positive lumbar test in a patient with say headache and or neck pain can be used in the study. It is not necessary to select patients with complaints in the same region in which the test under search is located. In each examined patient, the observers fill out one evaluation form with the results and one form with the patient data. When all 20 patients have been examined by both observes, the evaluation forms are collected and the number of patients in which observers agree (Yes/Yes and No/No) are calculated according a 2x2 contingency table is shown as in figure 33.

![Figure 32] The results of a theoretical reproducibility with n patients and two observers A and B presented in a 2x2 contingency table. (see text).

The formula for the overall agreement or observed agreement $P_o$ based on the data of figure 33 is:

$$P_o = \frac{a + d}{n}$$
In the case of a too low overall agreement, i.e. a percentage < 80%, the observers must go back to the training phase.
In the second training phase, 10 new patients must be examined in all details with respect to the performance and the judgement of the test. The recorded agreements of the logbook now become very essential. Observers have to look very carefully at all details, in particular, the semi-quantification, which can lead to problems in interpretation. If the problems are located, they are recorded in the logbook. The evaluation form is adapted and a second overall agreement has to be performed.
If again no overall agreement of > 80% is obtained the observers have again to go back to a new training phase and a subsequent third overall agreement phase. If still no substantial overall agreement percentage is obtained the observers have to discuss the continuation of the reproducibility study. Furthermore they have to wonder whether the test under search is suitable for education purposes because of a disputed transferability illustrated by the repeatedly found low overall agreement.
Publishing this kind of results is very valuable for education systems in M/M M.

3.2 Blinding Procedures.
Both in the overall agreement phase and in the following study phase adequate blinding procedures are essential. Except for pain evaluating tests, no communication between observer and patient during the performance of the procedure is allowed. Of course the observer, who has selected a patient for the study, has to inform the patient about the aim of the study, in accordance with the written information for patients. However, the observer who receives a patient for the study from his colleague is not allowed to communicate with the patient about the study at all. No communication is allowed between observers about the examined patients included in the overall agreement phase and the following study phase of the reproducibility study. The best way is to install an
independent person who collects the filled out evaluation forms directly after the examination of the patient. This independent person can also calculate the overall agreement percentages. If a substantial overall agreement percentage has been obtained, the preparations for the study phase with the $0.50 - P_{index}$ method can be made.
Proceeding to the study phase indicates that the overall agreement is >80%, and provided that nothing changes the overall agreement will be constant. In the study phase, evaluating one single test, the 0.50-P$_{\text{index}}$ method is used (see chapter VI, paragraph 2). If more than one test is evaluated at the same time, there is always the risk of a very low or very high P$_{\text{index}}$ resulting in an unwanted low kappa value (see chapter IV, paragraph 6).

Figure 33. Study Phase of the Reproducibility Protocol.

4.1 Observers and Patients Selection

The same two observers from the overall agreement phase perform the study phase. Because the observers have to send patients to each other, it is preferable that both observers work at the same clinic, so that the program of the clinic can be organised with respect to the study. Each observer selects 20 patients from his policlinic programme, making 40 patients in total. This means that 80 evaluation forms have to be made in advance (40 per observer).

In principle all patients with pain related to the locomotion system are suitable for inclusion to the study phase. Consequently, the observer performs the test on all the patients of his policlinic programme, except in case the test is related to a particular disease entity with special inclusion and exclusion criteria. He chooses patients in whom he is convinced that the test is positive or negative. This means that negative or positive lumbar tests in patients with
headache and or neck pain can also be used for the study. Each observer sends 10 patients with a positive test and 10 patients with a negative test to the other observer according to the scheme presented in figure 34. The total study population will be 40 patients.

Figure 34. Flow diagram of a reproducibility study using the 0.50-P$_{index}$ method. First both observers A and B perform Test I in their own patients and select 10 patients with a positive and 10 patients with a negative test. Subsequently they send their own population of 20 patients to each other (resp. n$_A$ and n$_B$). Observer A will now examine the population of observer B (n$_B$ = 20) and Observer B will examine the population of observer A (n$_A$ = 20). In total both observers will see 40 patients (n$_A$ + n$_B$).

When the 40 patients have been examined by both observers, the results from each patient’s evaluations forms are collected and added to the 2x2 contingency table. The number of patients in which observers agree (Yes/Yes = a and No/No = d) and disagree (Yes/No = b and No/Yes = c) are calculated as shown in figure 35.
4.2 **Blinding Procedures.**

Adequate blinding procedure is just as essential in the study phase, as in the overall agreement phase, and the procedure is identical except for the number of patients to examine. This means that, except for pain evaluating tests, no communication is allowed between observer and patient during the performance of the procedure. Of course the observer, who has selected his patient for the study, has to inform his patient about the aim of the study, etc (informed consent!). However, the observer who receives a patient for the study from his colleague is not allowed to communicate with the patient about the study at all. No communication is allowed between observers about the examined patients included in the study phase of the reproducibility study. The best way is to install an independent person who collects the filled out evaluation forms directly after the examination of the patient.
5. DATA PHASE

The data distilled from the evaluation form is arranged according the 2x2 contingency table as shown in figure 36. The figures a, b, c and d are inserted in the calculation table of the spreadsheet (see below).

17. Data recording/analysis

Figure 36. Data Phase of the Reproducibility Protocol.

1. Kappa calculation

In any spreadsheet the kappa value can automatically be calculated by inserting the figures into the formula shown in chapter IV. Only the figures corresponding to position a, b, c and d have to be inserted in the first four cells: for example, resp. 0, 0, 1 and 39 in figure 38. The columns indicated with capitals A to V (black, see figure 38) of a spreadsheet represent the different components to calculate the overall agreement, the P_index and the final kappa value. The labels of the columns are shown in the first row (blue, see figure 38).

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
<th>K</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>a</td>
<td>b</td>
<td>c</td>
<td>d</td>
<td>n</td>
<td>a+b</td>
<td>a+c</td>
<td>c+d</td>
<td>b+d</td>
<td>a+d</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>39</td>
<td>40</td>
<td>0</td>
<td>1</td>
<td>40</td>
<td>39</td>
<td>39</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>L</th>
<th>M</th>
<th>N</th>
<th>O</th>
<th>P</th>
<th>Q</th>
<th>R</th>
<th>S</th>
<th>T</th>
<th>U</th>
<th>V</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Po</td>
<td>(a+b)/n</td>
<td>(a+c)/n</td>
<td>(c+d)/n</td>
<td>(b+d)/n</td>
<td>MxN</td>
<td>OxP</td>
<td>Pc</td>
<td>Po-Pc</td>
<td>1-Pc</td>
</tr>
<tr>
<td>2</td>
<td>0,98</td>
<td>0,00</td>
<td>0,03</td>
<td>1,00</td>
<td>0,98</td>
<td>0</td>
<td>0,975</td>
<td>0,98</td>
<td>0,00</td>
<td>0,03</td>
</tr>
</tbody>
</table>

Figure 37. Spreadsheet with columns A to V and row 1 and 2 to calculate the kappa value. Labels of columns are blue. In row 2 the values of a,b,c and d are inserted resp. 0,0,1 and 39.

83
In the second row formulas have to be inserted from cell E2 to cell V2. These formulas are listed in table 4.
Table 5 Formulas to use in the spreadsheet in columns E to V in row 2 to automatically calculate the kappa value. Labels of columns are blue. All formulas start with an = mark.

<table>
<thead>
<tr>
<th>Column</th>
<th>label</th>
<th>Cell</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>E</td>
<td>n</td>
<td>Total</td>
<td>E2 =A2+B2+C2+D2</td>
</tr>
<tr>
<td>F</td>
<td>a+b</td>
<td>F2</td>
<td>=A2+B2</td>
</tr>
<tr>
<td>G</td>
<td>a+c</td>
<td>G2</td>
<td>=A2+C2</td>
</tr>
<tr>
<td>H</td>
<td>c+d</td>
<td>H2</td>
<td>=C2+D2</td>
</tr>
<tr>
<td>I</td>
<td>b+d</td>
<td>I2</td>
<td>=B2+D2</td>
</tr>
<tr>
<td>J</td>
<td>a+d</td>
<td>J2</td>
<td>=A2+D2</td>
</tr>
<tr>
<td>K</td>
<td>( P_{\text{index}} )</td>
<td>Prevalence index cond</td>
<td>K2 =A2/E2+B2/2<em>E2+C2/2</em>E2</td>
</tr>
<tr>
<td>L</td>
<td>( P_{o} )</td>
<td>Overall Agreement</td>
<td>L2 =J2/E2</td>
</tr>
<tr>
<td>M</td>
<td>(a+b)/n</td>
<td>M2</td>
<td>=F2/E2</td>
</tr>
<tr>
<td>N</td>
<td>(a+c)/n</td>
<td>N2</td>
<td>=G2/E2</td>
</tr>
<tr>
<td>O</td>
<td>(c+d)/n</td>
<td>O2</td>
<td>=H2/E2</td>
</tr>
<tr>
<td>P</td>
<td>(b+d)/n</td>
<td>P2</td>
<td>=I2/E2</td>
</tr>
<tr>
<td>Q</td>
<td>M x N</td>
<td>Q2</td>
<td>=M2*N2</td>
</tr>
<tr>
<td>R</td>
<td>O x P</td>
<td>R2</td>
<td>=O2*P2</td>
</tr>
<tr>
<td>S</td>
<td>( P_{c} )</td>
<td>Expected Chance Agreement</td>
<td>S2 =Q2+R2</td>
</tr>
<tr>
<td>T</td>
<td>( P_{0}.P_{c} )</td>
<td>T2</td>
<td>=L2-S2</td>
</tr>
<tr>
<td>U</td>
<td>1-.( P_{c} )</td>
<td>U2</td>
<td>=1-S2</td>
</tr>
<tr>
<td>V</td>
<td>kappa</td>
<td>V2</td>
<td>=T2/U2</td>
</tr>
</tbody>
</table>

By copying the entire row 2 to the next row, more kappa values within the study can be calculated such as the mutual dependency of tests when more than one test is evaluated (see chapter V, paragraph 2.2). Or, if more tests are used to make one final diagnosis (see chapter V, paragraph 2.3). However, in both cases the 0.50- \( P_{\text{index}} \) method cannot be used.
6. PUBLICATION PHASE

INTERNATIONAL ACADEMY OF

In the publication of the results of a reproducibility study it is essential that the reader is provided with adequate information on what grounds the authors based their final conclusion.

In rough outlines the format of a publication consists of an Introduction, a section with Material and Methods, a section with Results and finally the discussion.

1. Introduction Section

In the introduction the literature is reviewed about the diagnostic procedures, their hypotheses and their reliability. Previous articles are mentioned – if any –, and possible problems with their methodology can be used as supplementary arguments for performing the study. Additionally, at the end of the introduction the reasons and aims of the present study are mentioned.

2. Material and Methods Section

In this section the most relevant data of the reproducibility study are recorded. First of all, under the heading Materials in this section all the characteristics of the patients have to be mentioned such as:

1. Data about the source population (hospital, outdoor clinic, special clinic, etc) and/or whether it is a particular complaint group, syndrome or diagnosis group. In case of normal subjects (students, staff) data about recruitment procedures must be mentioned.

2. Data about how the patients were selected from the source population on entrance, consecutive, every other, non-selective sample procedure etc.)

3. Data about inclusion criteria for entrance of the patients in the study

4. Data about exclusion criteria of the patients

5. Demographic data (gender, age of patients or normal subjects) that are recorded in the study.

6. Number of patients or normal subjects in different phases of the study.
Under the heading **Methods** in this section all the characteristics of the study format have to be mentioned such as:

1. If in the **Preparation Phase** data about the participating members of
the study are mentioned such as an independent observer, ratification by ethical committee, informed consent, financial support, and the use of a logbook to register the consensus procedures.

1. **If a Training Phase** was used in the study format.
2. Detailed data about the performance of the diagnostic test. It is not sufficient simply to refer to a test from the literature.
3. Detailed data about the hypothesis of the test of the observers.
4. Detailed data about the judgement of the test(s) and/or semi-quantification.
5. Detailed data about the conditions for a final “diagnosis” in case of several tests.
6. **Number of tests**
7. Data about the characteristics of the observers (experience etc.)
8. **If an Overall Agreement Phase** was used in the study format.
9. **If in the Study Phase** the $0.50 - P_{\text{index}}$ method was used

At the end of the Material and Method Section data has to be provided about the statistical methods used in the study. In most of the cases it will be the kappa method.

1. If other statistic methods were used, the reason why has to presented.
2. If more tests were evaluated, their mutual dependency has to be analysed (see chapter V, paragraph 2.2).
3. In the case of a final “diagnosis” based on several tests, the mutual dependency between a single test and the final diagnosis has to be analysed (see chapter V, paragraph 2.3).

1. **Results Section**
In this section only results of the study have to be presented and no interpretations of results.

1. Data about the demographic characteristics of the population (gender, age of patients or normal subjects).
2. Data about the prevalence(s) of the index condition and not just frequencies (see chapter IV, paragraph 2.2).

3. Data about the overall agreement of the test(s) (see chapter IV, paragraph 1).

4. Data about mutual dependencies of tests per observer in table format (see table 2)

5. Data about mutual dependencies, final diagnosis and individual tests per observer in table format (see table 4)

6. Presentation of the raw data in 2x2 contingency tables, as illustrated in figure 38, together with the results of the kappa value.

<table>
<thead>
<tr>
<th>Test I positive</th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observer A</td>
<td>17</td>
<td>3</td>
</tr>
<tr>
<td>Observer B</td>
<td>3</td>
<td>17</td>
</tr>
</tbody>
</table>

P_{index}: 0.50

Overall Agreement P_0: 0.85

Kappa Value K: 0.7
1. **Discussion Section**

In this section results of the study are discussed by comparing it with the results from literature (if available). Formulate a clear conclusion with its consequences for daily practice. Essential in the discussion section is to mention the flaws of the study and the recommendation to improve future studies. In case the reproducibility is good to excellent is particularly important to mention the next steps to be taken: study on validity, sensitivity and specificity, on the predictive value of a positive and a negative test result, and likelihood ratio.
In Figure 39 the scheme is presented again to show the different aspects and phases of a reproducibility study on which Golden rules for a reproducibility study can be based. Reproducibility studies are easy to perform and not restricted to large institutes like universities. Private practices or other institutes with two or more practitioners in M/M Medicine are very suitable for these kinds of studies.

Figure 39. Flow chart of planning in different phases of a reproducibility study.
Rule 1: Create a clear logistic and responsibility structure for the reproducibility study in the preparation phase. One single person must be responsible for the entire process of the whole study including the logbook.

Rule 2: Use a logbook for the study

Rule 3: Always include a training phase in the study
Full agreement must be achieved about all details of the test(s) under search.

Rule 4: Always include an overall agreement phase in the study
This period is essential to achieve a substantial overall agreement > 0.80. Lower values result by definition to low kappa values.

Rule 5: Always repeat the training phase in case of a too low overall agreement

Rule 6: By preference evaluate in a reproducibility study only one test

Rule 7: Always describe in detail the diagnostic test under search

Rule 8: Always present raw data in 2x2 contingency tables

Rule 9: Always present the kappa value together with the values of the overall agreement and the P index.