Manuelle Medizin

Übersichten

Manuelle Medizin 2019 · 57:451–479
https://doi.org/10.1007/s00337-019-00581-5
Published online: 29 October 2019

Jacob Patijn[1,2]

[1] International Academy of Manual/Muscoloskeletal Medicine, IAMMM, Zurich, Switzerland
[2] Department of Translational Neuroscience, School of Mental Health and Neuroscience, Maastricht University, Maastricht, The Netherlands

Reproducibility protocol for diagnostic procedures in Manual/Musculoskeletal Medicine

Edition 2019

Electronic supplementary material

The online version of this article (https://doi.org/10.1007/s00337-019-00581-5) contains further reading material. The online version of the article and further reading are available at http://www.springermedizin.de/manuelle-medizin. The further reading material is available at the end of the article as "Supplementary Material".

Preface to the 2019 edition of the reproducibility protocol

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 completely rewrite the last edition of 2012.

In this new IAMMM protocol, essential changes are made in the different periods of the protocol format. The direct reason for these changes is the fact that nowadays, sample size calculations based on statistical power are advised to estimate study sample sizes. Sample calculations based for studies using kappa coefficients are faced with several difficulties. The kappa coefficient of a diagnostic procedure is not an absolute measure as such. Its value is dependent on many other factors. As a consequence, the same kappa coefficient can have different levels of the overall agreement and the prevalence of the index condition. Furthermore, in a reproducibility study, the kappa coefficient is only related with positive judged diagnostic procedures. In contrast, the overall agreement is an absolute measure and concerns both positive and negative judged diagnostic procedures, reflecting more the daily reality of a clinician. Therefore, in the new protocol we focus in the first instance on calculating a sample size for the overall agreement.

As a consequence, the training period has been extended with a pilot study to evaluate the standardisation of the diagnostic procedure and the overall agreement period has been removed from the protocol.

Since the overall agreement as such is an absolute measure and reflects more the daily practice of a clinician dealing with diagnostic procedures and is most informative for clinicians, the starting point of the protocol is more focused on the overall agreement and in the second instance the kappa coefficient is emphasised. Attention is paid to the comprehensibility and readability of the text compared to the previous edition. Tables and figures have been renewed. 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.

The reproducibility protocol has been elaborated in such a way that it can be used as a kind of "cookbook 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 also be acknowledged in university education.

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

Therefore, in the introduction the original reasons to develop this protocol are again mentioned because they are

still very relevant for present day M/M Medicine.

In this edition, a list of reproducibility studies (exclusively using kappa statistics) 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 who use this protocol will send their comments to the present second Scientific Director of the Academy. In this way, we can continuously improve and update the protocol.

The IAMMM would also like to encourage scientists and/or educationalists, who receive and use this latest edition of the 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.

Table of contents

I. 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 cookbook 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, frequently with many different diagnostic

procedures and many different therapies 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 Fig. 1.

The consequences of this statement are five-fold:

1. Most existing different approaches within M/M Medicine have 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. Furthermore, 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 of 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

Manuelle Medizin 2019 · 57:451–479 https://doi.org/10.1007/s00337-019-00581-5

J. Patijn

Reproducibility protocol for diagnostic procedures in Manual/Musculoskeletal Medicine. Edition 2019

Abstract

The International Academy of Manual/Musculoskeletal Medicine (IAMMM) has completely revised the protocol for reproducibility studies of diagnostic procedures in Manual/Musculoskeletal Medicine (M/M Medicine). The protocol was and is aimed at the practitioners in the field of M/M Medicine. This IAMMM protocol can be used in a very practical way and makes it feasible to perform reproducibility studies equally well in private practices and clinics for M/M Medicine with two or more physicians and as by educational boards of the societies of M/M Medicine. This IAMMM protocol provides practical solutions for sample size calculations in reproducibility studies using kappa statistics. Step by step, many different statistical aspects of reproducibility studies are explained, resulting in a very structured protocol format of how to perform a reproducibility study using overall agreement and the kappa value as statistical outcome.

Keywords

Interobserver agreement · Prevalence of index condition · 0.50-P_{index} method · Sample size · Kappa value

Reproduzierbarkeitsprotokoll für diagnostische Verfahren in Manueller Medizin. Ausgabe 2019

Zusammenfassung

Die International Academy of Manual/Musculoskeletal Medicine (IAMMM) hat das Protokoll für Reproduzierbarkeit von diagnostischen Verfahren in der Manuellen Medizin vollständig revidiert. Das IAMMM-Protokoll war und ist insbesondere für Praktiker auf dem Gebiet der Manuellen Medizin gedacht. Dieses Protokoll kann auf sehr anwendungsorientierte Weise verwendet werden und macht es möglich, Reproduzierbarkeitsstudien gleichermaßen gut in Privatpraxen und Kliniken mit zwei oder mehr Ärzten, wie bei Vorständen einer Gesellschaft für Manuelle Medizin, durchzuführen. Dieses IAMMM-Protokoll bietet praktische Lösungen für Berechnungen der Stichprobengroße in Reproduzierbarkeitsstudien mit Kappa-Statistik. Schritt für Schritt werden sehr viele statistische Aspekte von Reproduzierbarkeitsstudien erklärt, was am Ende in einem sehr strukturierten Protokollformat resultiert, wie man eine Reproduzierbarkeitsstudie unter Verwendung von statistischer Gesamtübereinstimmung und Kappa-Wert-Ergebnisdaten durchführt.

Schlüsselwörter

Vereinbarung zwischen Beobachtern · Prävalenz der Indexbedingung · 0,50-P_{index}-Methode · Probengröße · Kappa-Wert

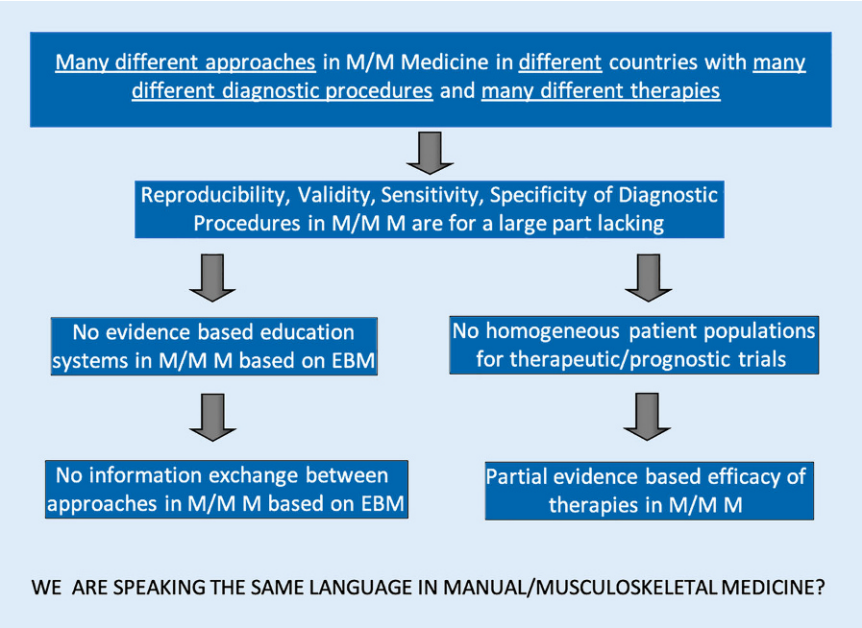

Fig. 1 ▲ Summary of the problem and its consequences for Manual/Musculoskeletal Medicine (M/M M) as defined by the previous Scientific Committee of FIMM (Intenational Federation for Manual Medicine) and adapted by its successor the present International Academy of Manual/Musculoskeletal Medicine (IAMMM). *EBM* evidence-based medicine

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

To provide practitioners in the field of Manual/Musculoskeletal Medicine with the right tools to perform high-quality reproducible studies, the IAMMM has the possibility to organise in cooperation with national societies for Manual/Musculoskeletal Medicine or university departments a 1-day instructional course. In the course, beyond the theoretical explanations of statistics in reproducibility studies, practical training is an essential aspect of the instructional course.

Previous IAMMM International Instructional Courses have been organised in many countries, such as the Czech Republic, Denmark, France, Germany, India, Netherlands, Russia and the United Kingdom.

Detailed information about these courses is available from the IAMMM logistic officer (sjerutte@gmail.com).

The Academy Conference

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, newly developed therapeutic and/or diagnostic algorithms and other new scientific work are presented. In a fruitful discussion between the 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 inexpensive 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. Furthermore, 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 medical research is the fact that different name are used for the same condition. Therefore, we think it important first to provide the reader of this protocol with an overview of the definitions used in this protocol. In clarifying and illustrating the definitions in greater detail, reading becomes much easier. In particular those definitions that are used in reproducibility studies are elaborated in greater detail based on 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 [1, 2]. 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 synonymous to **reproducibility** and accuracy is synonymous to **validity**. Both terms **reproducibility and validity** are generally used, as is in this protocol [3].

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 subject. In the case of a single observer we are

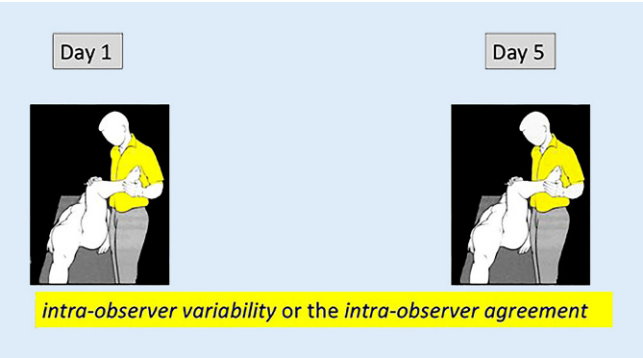

Fig. 2 ▲ Reproducibility with one observer performing the same diagnostic procedure on the same subject on day one and day five. The reproducibility reflects the intra-observer agreement

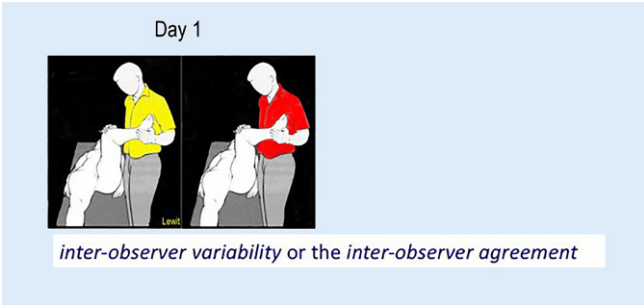

Fig. 3 ▲ Reproducibility with two different observers performing the same diagnostic procedure on the same subject on one occasion (day 1). The reproducibility reflects the inter-observer agreement

Fig. 4 ▲ Patrick Test

dealing with **intra-observer agreement** or intra-observer variability. The same observer uses the same diagnostic procedure in the same subject but on two different occasions (**Fig. 2**).

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 subject at one occasion (**Fig. 3**).

Definition diagnosis versus diagnostic procedure. In essence, the reproducibility of a diagnostic procedure has nothing to do with a diagnosis as such. In medicine, diagnostic procedures are the constituent parts of the 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 our protocol, dichotomous outcomes for the final outcome of a diagnostic procedure such as Yes or No is used. To illustrate this in greater detail we take as example the Patrick Test (**Fig. 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 the hypothesis of the Patrick Test (meant to test the 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 couch on both sides (see *double arrow* in **Fig. 4**). Next, the left/right distances are mutually compared. The observers in advance have agreed that the side with the largest distance has a positive Patrick Test. Whether 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 subjects 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 the final 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 constituent components of the procedure and the use of the same defined final judgement (in our protocol dichotomous outcome, Yes/No). Reproducibility has nothing to do with a genuine diagnosis. This is very important to realise because it explains why no selection procedure for the study population in a reproducibility study is needed. In a reproducibility study with a dichotomous outcome, the diagnostic procedure as such is independent of a final diagnosis because the latter is based on a whole diagnostic arsenal.

In summary, reproducibility is about a fully, detailed and standardised description on how to perform the various components of a diagnostic procedure and a detailed and standardised description on how to measure and interpret the final 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 or criterion standard or a reference test. In a separate protocol we will give a more detailed

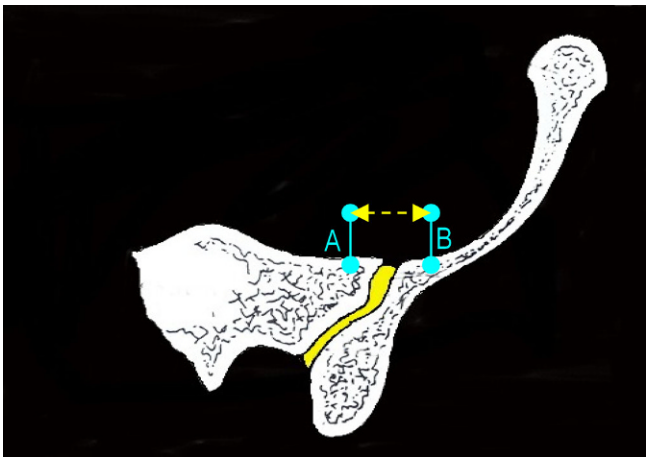

Fig. 5 ▲ A hypothetical x-ray method is used to evaluate the mobility of the SI-joint. Two (*blue*) iron rods are placed on both sides of the SI-joint (*A* and *B*). X-rays are taken before and after a Patrick manoeuvre in the subject. The distance between *A* and *B* is measured before and after a Patrick manoeuvre. The difference between the distances (before and after the Patrick Test) is a measure for mobility

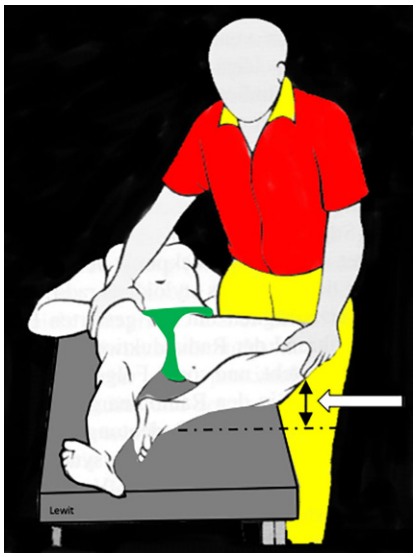

Fig. 6 ▲ Patrick Test. The double black arrow illustrates the distance between the left knee and the examination table (*dotted line*). The subjective left/right difference can be used as an outcome of the diagnostic procedure

introduction to various forms of validity, among which the criterion validity probably is the best know.

To explain this form of validity in greater detail we use again the example of the Patrick Test. As mentioned in "Definition diagnosis versus diagnostic procedure", the hypothesis of the Patrick Test was the testing of the mobility of the SI-joint [4]. Suppose we have already a good reproducibly 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 a reference test that can quantify the mobility of a SI-joint during the Patrick Test manoeuvre.

Suppose we can use the stereophotogrammetry method with two iron rods (A and B, ◘ **Fig. 5**) placed on both sides of the SI-joints [5] 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, movement in the SI-joint is elicited. 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 we have duplicated the results of this method (proven to be reproducible), in this instance we have a good gold standard to evaluate the validity of the Patrick Test. Subsequently, we now have to perform a validity study in which the Patrick Test and an X-ray SI-joint stereophotogrammetry (as gold standard) are simultaneously performed in the same group of subjects. By comparing the results of the Patrick Test (◘ **Fig. 6**) and the X-ray SI-joint stereophotogrammetry (◘ **Fig. 5**) the validity of the Patrick Test can be estimated. In M/M Medicine, many diagnostic procedures have been developed, each with its own hypothesis. However, we have to realise that gold standards or 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 between the couch on the left and right side (see double arrow in ◘ **Fig. 6**). In this case one has to develop a quantitative method, which measures this distance between the knee and the couch. 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.

III. Reproducibility studies: nature of diagnostic procedures

1. Nature of data in reproducibility studies

Before starting a reproducibility study, first we have 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 etc.). In qualitative diagnostic procedures, the outcome of the diagnostic procedure can be divided into 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 or binary character reflecting a contrast. Also, the contrast between male and female in studies with gender as outcome is a good example of such nominal data. Other typical diagnostic procedures in M/M Medicine, in which the outcomes of the diagnostic produce are nominal data, are for ex-

ample diagnostic procedures that evaluate "end feel" (abnormal Yes or No), pain provocation diagnostic procedure (pain Yes or No) under different conditions (provoked by observer or provoked by movements of the subject) and range of motion (restricted Yes or No). For reproducibility studies evaluating these kinds of data, kappa statistics are indicated (see " IV. Reproducibility studies: kappa statistics").

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-rays of the cervical spine in which the severity of the degenerative changes of the cervical spine are subdivided into categories [6, 7].

However, we have to question whether this kind of clinical subdivision into ordinal outcome data, both in M/M Medicine and radiology, has any sense at all. We have to consider whether subjective subdivisions of a diagnostic procedure outcome (for instance normal, moderate, severe) have consequences for the diagnostic and therapeutic indications of the subject. In M/M Medicine, in most of the cases there are no solid arguments to use such a subjective subdivision of outcome in diagnostic procedures. Only in circumstances in which one wants to use a diagnostic procedure to evaluate its outcome during a period of time can subjective subdivision be 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 of the subdivision. 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 recorded as 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, analysis of variance (ANOVA) 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 (P_o)/observed agreement (P_o)

In reproducibility studies, using dichotomous outcome data, just mentioning one of the synonymous terms *percent agreement or overall agreement or observed agreement* does not provide the entire information about the reproducibility of a particular diagnostic procedure. In our protocol we use the term *observed agreement* ($\mathbf{P_o}$). $\mathbf{P_o}$ in a reproducibility study is the ratio of the number of subjects in which the observers agree to the total number of observations. The main problem is that in reproducibility studies using a dichotomous outcome, the $\mathbf{P_o}$ does not take into account of the agreement that is expected to occur solely by chance alone. This will be further elaborated in "IV. Reproducibility studies: kappa statistics".

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 [χ^2] and Phi) or do not account for systematic observer bias (Pearson's product moment correlation, rank order correlation) [8, 9].

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 data. This statistical method will be extensively explained in "IV. Reproducibility studies: kappa statistics".

3.2 Appropriate statistics in quantitative data reproducibility studies

To evaluate the reproducibility of repetitive measurements with quantitative/ continuous data (that may be interval or ratio data), the paired samples t-test and/or the intraclass correlation coefficient (ICC) 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).

The ICC is the statistical method of choice for the reproducibility of observers for continuous data (cm, mm, etc.). The calculated factor R in this statistical pro-

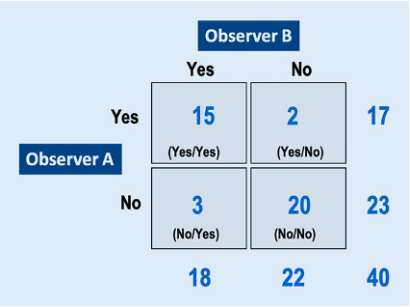

Fig. 7 ▲ The results of a reproducibility study with 40 subjects and two observers *A* and *B* presented in a 2 × 2 contingency table (see text)

cedure is 1 if the ratings are identical for all pairs, but less than or equal to 0 in the absence of reproducibility.

3.3 Choice of statistics and clinical consequences

In reproducibility studies, the choice of statistics is not only dependent on the measurement level of the collected data (nominal, ordinal, interval, ratio). It also depends on the type of clinical decision concluded from the findings of the reproducibility study.

Suppose the reproducibility of 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 can be quantified by 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, the kappa coefficient is 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 "III. Reproducibility studies: nature of diagnostic procedures", most diagnostic procedures in daily practice of M/M Medicine produce an outcome of the diagnostic procedure that is nominal, and often even dichotomous (Yes/No, Present/Absent, Normal/Abnormal).

For these kinds of dichotomous data, kappa statistics are appropriate. In this section, the different kappa coefficients are explained and illustrated with the results of previous reproducibility studies to highlight different aspects, problems and flaws of this statistical method. Frequently used terms in kappa statistics are defined and explained. This section is essential to understand the reproducibility protocol elaborated in "V. Developing reproducibility studies: general aspects". Although many formulas will be shown in this section for illustration, all these formulas will be integrated in a spreadsheet (see "VII. Protocol format reproducibility study", section "Statistics period") for automatic calculation of the kappa coefficient and related data of a reproducibility study. You do not have to remember these formulas.

1. Definition of the kappa coefficient

The kappa coefficient is a measure of inter- or intra-observer agreement (see "II. Reliability: Reproducibility and validity", section "Reproducibility") corrected for agreement occurring by chance.

Why do kappa statistics correct for the chance? If you perform a diagnostic procedure on a subject 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 coefficient (see "IV. Reproducibility studies: kappa statistics", section "Calculation of the kappa coefficient").

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 subjects. The outcome possibility of the diagnostic procedure was: Positive (Yes) or Negative (No). Both observers A and B examined the 40 subjects with the Patrick Test and recorded their findings. By combining the results of both observers per subject 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 subjects the Patrick Test positive (Yes/Yes), **2.** Both Observer A and Observer B judge in the same subjects 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 subjects (Yes/No), **4.** Observer A judges the Patrick Test negative while Observer B judges the Patrick Test positive in the same subjects (No/Yes). The results of these four categories can be depicted in a so-called 2 × 2 contingency table (**Fig. 7**). In the rows and the columns are the total numbers of subjects in which the Observer A and Observer B judge the Patrick Test positive or negative.

Observer A judged 17 Patrick Tests as positive and 23 as negative. Observer B judged 18 Patrick Tests as positive and 22 as negative. By adding per observer both figures the end result is 40.

Based on the data from this 2 × 2 contingency table, different important aspects of the kappa statistics can be calculated. As shown later some of these aspects will influence the final kappa coefficient.

2. Overall agreement

2.1 Definition

Under the "III. Reproducibility studies: nature of diagnostic procedures", section "Percentage agreement/overall agreement (P_o)/observed agreement (P_o)", the synonymous terms percentage of agreement/overall agreement ($\mathbf{P_o}$)/observed agreement ($\mathbf{P_{obs}}$) were introduced in reproducibility studies, the overall agreement reflects the percentage of subjects 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 ($\mathbf{P_o}$). This means that the overall agreement $\mathbf{P_o}$ is calculated by the sum of the number of subjects in which both observers judge the diagnostic procedure positive and negative, divided by the total number of subjects in the study. In **Fig. 8**, a similar 2 × 2 contingency

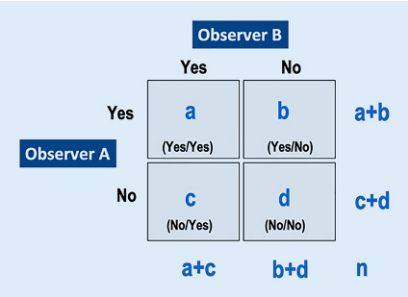

Fig. 8 ▲ The results of a theoretical reproducibility study with *n* subjects and two observers *A* and *B* presented in a 2 × 2 contingency table (see text)

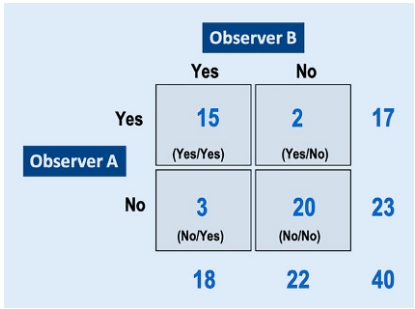

Fig. 9 ▲ The results of a reproducibility study with 40 subjects and two observers *A* and *B* presented in a 2 × 2 contingency table (see text)

table is shown as in **Fig. 7** but now based on a theoretical reproducibility study.

The formula for the overall agreement or observed agreement $\mathbf{P_o}$ based on the data of **Fig. 8** is:

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

Based on the data of the 2 × 2 contingency of the reproducibility study shown in **Fig. 7** the observed agreement $\mathbf{P_o}$ is calculated as follows:

$$\mathbf{P_o} = (15 + 20)/40 = 0.88\,.$$

This $\mathbf{P_o}$ will later be inserted in the final formula to calculate the final kappa coefficient (see "IV. Reproducibility studies: kappa statistics", section "Calculation of the kappa coefficient").

As will be explained later, the overall agreement is very important in a reproducibility study—because it influences strongly the magnitude of a kappa coefficient (see "IV. Reproducibility studies: kappa statistics", section "Interpretation of kappa coefficient: dependency of the prevalence of the index condition P_{index}").

2.2 Prevalence and prevalence of the index condition

Three new statistical concepts, used in reproducibility studies, are introduced: index condition, prevalence and prevalence of the index condition.

2.3 Index condition definition

The **index condition** is synonymous with the positive judged diagnostic procedure of a subject participating in a reproducibility study. In **Figs. 7 and 8** 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.

2.4 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 2 × 2 contingency table we have three boxes with a number of subjects with a positive diagnostic procedure: the box with Yes/Yes, the box with Yes/No, the box with No/Yes. In the example of **Fig. 8**, these boxes are filled out with a, b and c and in **Fig. 9** with 15, 2 and 3. To calculate the **prevalence of the index condition** $\mathbf{P_{index}}$ we need a special formula. Based on a theoretical 2 × 2 contingency table shown in **Fig. 8**, the formula for the prevalence of the index condition $\mathbf{P_{index}}$ is:

$$\mathbf{P_{index}} = \frac{[\mathbf{a} + (\mathbf{b} + \mathbf{c})/2]}{\mathbf{n}}.$$

Based on the 2 × 2 contingency table of the reproducibility study shown in **Fig. 10** the formula for the prevalence of the index condition $\mathbf{P_{index}}$ is:

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

Both observed agreement ($\mathbf{P_o}$) and prevalence of the index condition ($\mathbf{P_{index}}$) are important in a reproducibility study. Their values are decisive for the magnitude of the final kappa coefficient in a reproducibility study (see "VI. The relation between the kappa coefficient and the prevalence the index condition P_o", section "Defining the P_{index} problem").

2.5 Prevalence

Prevalence is a statistical concept referring to the number of subjects with a positive diagnostic procedure that are present in a study sample. 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 abovementioned is not feasible in reproducibility studies with a dichotomous outcome of the diagnostic procedure. Only the prevalence of the index condition ($\mathbf{P_{index}}$) can be calculated.

However, it is possible to calculate the prevalence of the positive judged per observer. In **Fig. 9**, 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 ($\mathbf{P_{index}}$) in the example of **Fig. 9** is **0.44**.

3. Expected agreement by chance

As stated before, the kappa coefficient is a measure for inter-observer agreement or intra-observer agreement corrected for agreement occurring by chance. Because, if you perform a diagnostic procedure in a subject 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 $\mathbf{P_c}$. This expected agreement by chance $\mathbf{P_c}$ will integrated in the final formula to estimate a kappa coefficient (see "IV. Reproducibility studies: kappa statistics", section "Calculation of the kappa coefficient").

The formula for the expected agreement by chance $\mathbf{P_c}$ [10] based on the theoretical reproducibility study shown in **Fig. 8** is:

$$\mathbf{P_c} = \frac{\mathbf{a} + \mathbf{b}}{\mathbf{n}} \mathbf{x} \frac{\mathbf{a} + \mathbf{c}}{\mathbf{n}} + \frac{\mathbf{c} + \mathbf{d}}{\mathbf{n}} \mathbf{x} \frac{\mathbf{b} + \mathbf{d}}{\mathbf{n}}$$

Table 1 Diagram according to Landis and Koch [11] to interpret the kappa coefficient of a reproducibility study. Strength of agreement is the same as reproducibility

Kappa value	Strength of agreement
–0.20 to 0.00	Absence
0.00–0.20	Slight
0.21–0.40	Fair
0.41–0.60	Moderate or Good
0.61–0.80	Substantial
0.81–1.00	Almost perfect

The expected agreement by chance $\mathbf{P_c}$ will be used for the final formula to estimate a kappa coefficient (see "IV. Reproducibility studies: kappa statistics", section "Calculation of the kappa coefficient").

Based on the 2 × 2 contingency table of the reproducibility study shown in **Fig. 9**, the expected agreement $\mathbf{P_c}$ can be calculated as:

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

4. Calculation of the kappa coefficient

To calculate the kappa coefficient, we need the observed agreement $\mathbf{P_o}$ elaborated in "Definition kappa coefficient" of this section and the expected agreement by chance $\mathbf{P_c}$ of section "Calculation of the kappa coefficient" of this section to be inserted in the formula for the kappa coefficient $\boldsymbol{\kappa}$:

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

When we apply the kappa formula on the data of the reproducibility study as shown in **Fig. 9**, the expected agreement by chance $\mathbf{P_c}$ will be 0.51, the observed agreement $\mathbf{P_o}$ is 0.88. Inserting these figures in the kappa formula leads to:

$$\boldsymbol{\kappa} = \frac{0.88 - 0.51}{1 - 0.51} = 0.55$$

The prevalence of the index condition $\mathbf{P_{index}}$ in this study is 0.44 (see above) with an observed agreement $\mathbf{P_o}$ of 0.88 (see above).

5. Interpretation of kappa coefficient: general

The kappa coefficient, 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 coefficient of a reproducibility study. The most widely used is the scheme of Landis and Koch [11]. They stated that kappa coefficients above 0.60 represent good to almost perfect agreement beyond chance between two observers. In contrast, kappa coefficients of 0.40 or less represent absence to fair agreement beyond chance. Kappa coefficients between 0.40 and 0.60 reflect a fair to good agreement beyond chance (**Table 1**).

However, the standards for strength of agreement provided by Landis and Koch is just an agreement about the kappa interpretation.

The same kappa coefficient can be based on different values of the overall agreement ($\mathbf{P_o}$). A very low or negative kappa coefficient can be the result of a very high or low $\mathbf{P_{index}}$ and does reflect the quality of the agreement between two observers about a diagnostic procedure.

This will be further explained below in sections "Interpretation of kappa coefficient: dependency of the overall agreement" and "Interpretation of kappa coefficient: dependency of the prevalence of the index condition P_{index}" of this section.

6. Interpretation of kappa coefficient: dependency of the overall agreement

As already mentioned in section "Definition of the kappa coefficient" of this section, the overall agreement is a very important factor to interpret the kappa coefficient of a reproducibility study. In reproducibility studies, the overall agreement $\mathbf{P_o}$ reflects the proportion of subjects in which the observers agree about the outcome or judgement of the diagnostic procedure. More precisely, it reflects the total proportion in which both observers agree about positive and negative found diagnostic procedure in the same subjects.

In the example of **Fig. 14**, the overall agreement $\mathbf{P_o}$ is calculated by the sum of the number of subjects 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 subjects of the study. In our example the overall agreement $\mathbf{P_o} = (15 + 20) / 40 = 0.88$.

Many published reproducibility studies in M/M Medicine show low kappa coefficients without mentioning the overall agreement data.

In **Fig. 10** the relation between the kappa/P_{index} curve and the overall agreement is illustrated, with curves for two different overall agreements $\mathbf{P_o}$.

In **Fig. 10a** 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 coefficient is ≥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 the kappa coefficient will be <0.6 (**Fig. 10b**). Because a part of the kappa/P_{index} curve drops below the zero-line, negative kappa coefficients can occur.

For instance, if the overall agreement Po decreases from 0.98 to 0.79, the kappa/P_{index} curve slowly shift downwards and become finally located under the zero line. The percentage with kappa coefficients kappa coefficients ≥0.6 will decrease from 99 to 0%. In **Fig. 11**, all kappa/P_{index} curves of the $\mathbf{P_o}$ interval 0.98–0.79 are depicted.

7. Interpretation of kappa coefficient: dependency of the prevalence of the index condition P_{index}

As already mentioned in section "Prevalence and the prevalence of the index condition" of this section, the prevalence of the index condition $\mathbf{P_{index}}$ is a very important factor. Not only how to interpret a kappa coefficient of a reproducibility study, but the $\mathbf{P_{index}}$ importantly can influence the level of the kappa coefficient.

In reproducibility studies, the prevalence of the index condition P_{index} is synonymous with the frequency of all positive judged diagnostic procedures (the

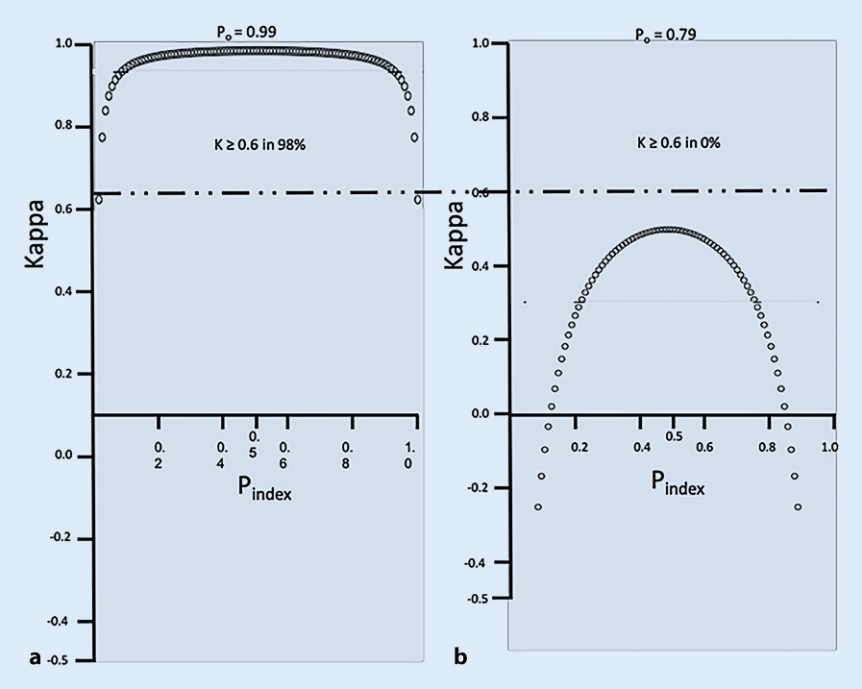

Fig. 10 ▲ Relation between kappa coefficient and prevalence of the index condition. The *dotted/dashed horizontal line* is the cut off level of 0.60. **a** The kappa/prevalence index curve with an overall agreement P_o of 0.97. The part of the curve above the horizontal line represent 98% kappa coefficients larger than 0.60. **b** The kappa/prevalence index curve with an overall agreement P_o of 0.77. The curve does not cross the horizontal line, illustrating that 0% of the kappa coefficient exceed 0.60

index condition) by the observers. The P_{index} has to be calculated using a formula (see section "Prevalence of the index condition definition"). The relation between the kappa coefficient and the P_{index} is illustrated in ◘ **Fig. 9**.

The kappa/P_{index} curve in ◘ **Fig. 12** illustrates that if the P_{index} is low (0.2) or high (0.9), the matching kappa coefficients will be under the kappa cut off line of 0.6. This means that in this study sample, there are too few (low P_{index}) or too many (high P_{index}) positively judged diagnostic procedures. As stated in the section "Interpretation of the kappa coefficient", the standards for strength of agreement provided by Landis and Koch [11] (◘ **Table 1**) was just an accordance about the kappa interpretation and without a scientific base. The same kappa coefficient can be based on different values of the $\mathbf{P_o}$ and $\mathbf{P_{index}}$. A very low or negative kappa coefficient can be the result of a very high or low $\mathbf{P_{index}}$ and does reflect the quality of the agreement between two observers about a diagnostic procedure.

The relation between the kappa coefficient and $\mathbf{P_o}$ and $\mathbf{P_{index}}$ has consequences for the interpretation of kappa coefficients in published reproducibility studies. As can be seen in ◘ **Fig. 13**, the same kappa coefficient can be located on the left or right side of different kappa/P_{index} curves. Each kappa/P_{index} curve has its own $\mathbf{P_o}$ value, ranging from an overall agreement of 0.97 till 0.79. The kappa coefficient 0.4 can be due to a low overall agreement $\mathbf{P_o}$, 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 coefficient 0.4 can also be due to a too high or too low $\mathbf{P_{index}}$, 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 $\mathbf{P_o}$ value.

The same is partly true for reproducibility studies finding kappa coefficients ≥0.6 without mentioning data about $\mathbf{P_o}$ and $\mathbf{P_{index}}$. 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 $\mathbf{P_{index}}$ was not too high or too low. The $\mathbf{P_{index}}$ is always calculated after completing the study and therefore is not known in advance.

8. Interpretation of kappa coefficient: bias

Bias can be present when observers produce different patterns of ratings or outcomes [12]. 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 problem [13]. In the IAMMM protocol a well-defined training phase is incorporated.

Concluding, published reproducibility studies that do not mention the values of the $\mathbf{P_o}$ and $\mathbf{P_{index}}$ and in which authors concluded an absence of clinical value because of a low observed kappa coefficient and using the standards for strength of agreement provided by Landis and Koch [14] have to be interpreted with caution.

How to deal with the relation between kappa coefficient and $\mathbf{P_o}$ and $\mathbf{P_{index}}$ in reproducibility studies will be elaborated later ("The problem of the relation between the prevalence of the index condition and the kappa coefficient", section "Influencing the prevalence of index condition in advance").

V. Developing reproducibility studies: general aspects

1. Nature of the diagnostic procedure to be evaluated in a study

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.

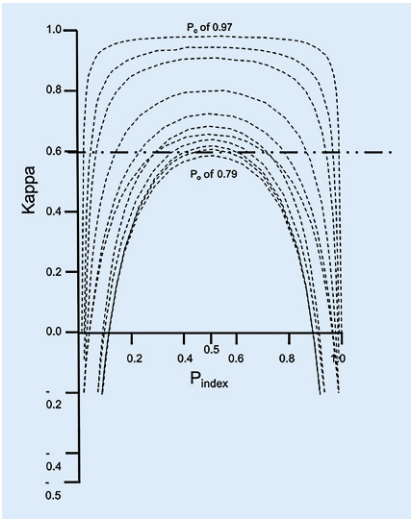

Fig. 11 ▲ Relation between kappa coefficient and prevalence of the index condition. The *dotted line* is the cut off level of 0.60. The kappa/prevalence index curves with an overall agreement P_o between 0.97 and 0.79

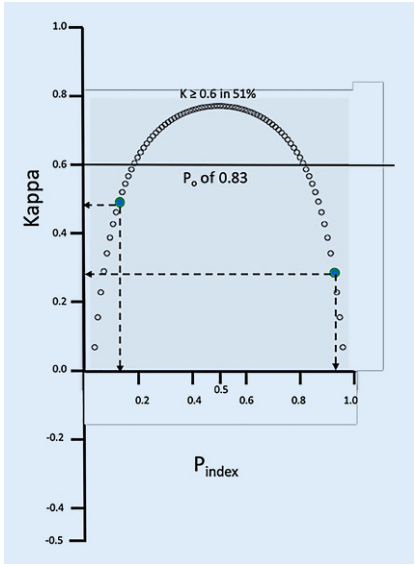

Fig. 12 ▲ Relation between kappa coefficient and prevalence index of the index condition P_{index}. The horizontal line is the cut off level of 0.60. The part of the curve above this line represent 51% of the kappa coefficients ≥0.60. The blue dots indicate low kappa coefficients (*horizontal arrows*) in case of a low P_{index} (*left blue dot*) or a high P_{index} (*right blue dot*)

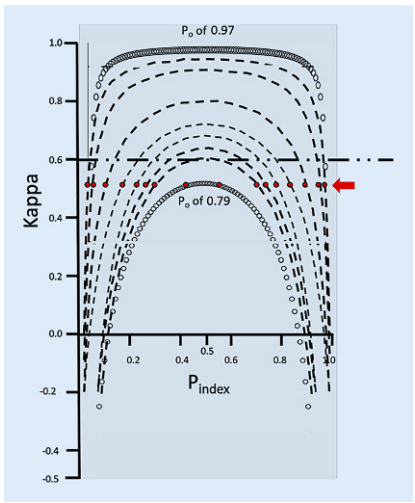

Fig. 13 ▲ Relation between kappa coefficient and prevalence of the index condition. The *horizontal dotted/dashed line* is the cut off level of 0.60. The kappa/prevalence index curves with an overall agreement P_o between 0.97 and 0.79. *Red dots* represent all kappa with the same coefficient of 0.4, but on different kappa/prevalence index curves (see text)

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 subject. In both genuine diagnoses and syndromes, diagnostic procedures are needed. We can rarely 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. 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 prolapse 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 principally rely on diagnostic procedures to form syndromes of low back pain.

Also, in our educational systems, many diagnostic procedures are taught 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 as such.

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 section "Combinations of a few different diagnostic procedures: mutual dependency").
c. **Several diagnostic procedures are related to a particular diagnosis.**

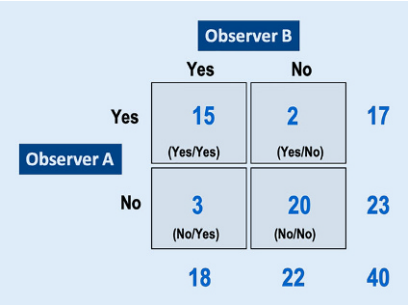

Fig. 14 ▲ The results of a reproducibility study with 40 subjects and two observers *A* and *B* presented in a 2 × 2 contingency table (see text)

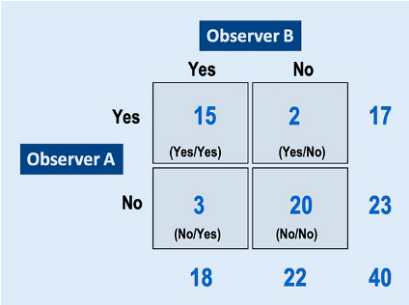

Fig. 15 ▲ The results of a reproducibility study with 40 subjects and two observers *A* and *B* presented in a 2 × 2 contingency table

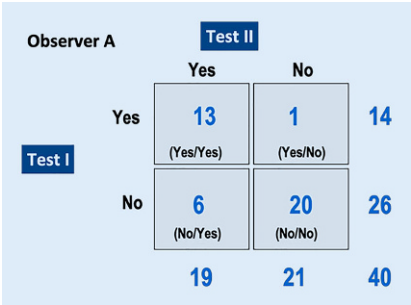

Fig. 16 ▲ A 2 × 2 contingency table showing the agreements and disagreement between Test I and Test II of the examined subjects of observer *A* to estimate the mutual dependency between Test I and Test II

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 section "Combinations of a few different diagnostic procedures: mutual dependency"*).*

2. Number of diagnostic procedures evaluated in reproducibility studies

2.1 Too many diagnostic procedures

Reproducibility studies in non-specific clinical conditions, for low back pain, sometimes evaluate 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 coefficients and subsequently it is concluded that these diagnostic procedures have no clinical value.

As already explained in the sections "Interpretation of the kappa coefficient" and "Interpretation of kappa coefficient: dependency of the overall agreement" (under "IV. Reproducibility studies: kappa statistics"), the prevalence of the index condition $\mathbf{P_{index}}$ and overall agreement $\mathbf{P_o}$ influence greatly the final kappa coefficient of a study. Since data of $\mathbf{P_o}$ and $\mathbf{P_{index}}$ 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 coefficients 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 $\mathbf{P_o}$ is obtained for many diagnostic procedures and therefore a low kappa coefficient.

2.2 Combinations of a few different diagnostic procedures: mutual dependency

As mentioned already in the section " Nature of the diagnostic procedure to be evaluated in a study", 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 investigated 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 **Fig. 15**. In this example of a 2 × 2 contingency table we have two observers A and B.

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 diagnostic procedures (Test I and Test II). The agreement and disagreement between Test I and Test II of the examined subjects is likewise estimated in a 2 × 2 contingency table (**Fig. 16**).

Based on the data in the boxes of **Fig. 16**, a kappa coefficient can be calculated. A kappa coefficient of ≥0.40 means that that there is a probability that Test I and Test II (**Fig. 16**) are mutually dependent diagnostic procedures for observer A. Also, in this case the values of the $\mathbf{P_o}$ and $\mathbf{P_{index}}$ are necessary for proper interpretation of the kappa coefficient.

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 diagnostic procedures in the same reproducibility study was illustrated [15]. 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 subjects have the SI-joint dysfunction syndrome yes or no.

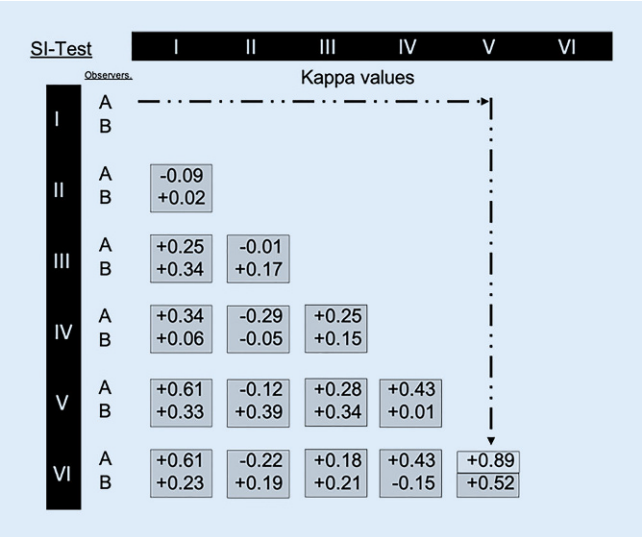

Fig. 17 ◀ Kappa coefficients between pairs of SI-Tests I to VI, subdivided per observer *A* and *B* (see text)

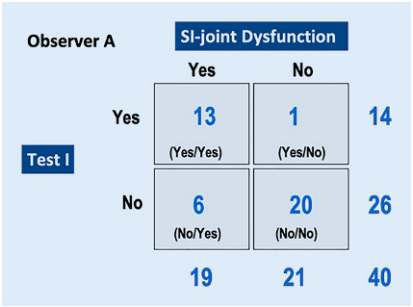

Fig. 18 ▲ A 2 × 2 contingency table showing the agreements and disagreement between Test I and the final "syndrome diagnosis" (SI-joint dysfunction syndrome) of the examined subjects of observer *A*, to estimate the mutual dependency between Test I and this final "syndrome diagnosis"

Instead of showing a separate 2 × 2 contingency table for each observer A and B of all possible combinations of two SI-Tests, the data are summarised in one single table (**Fig. 17**).

In the most upper black row, 6 SI-Tests I to VI are listed. In the far-left black 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 coefficients for each observer A and B per SI-Test I to VI are shown. These kappa coefficients are calculated based on the principles used in the 2 × 2 contingency table presented in **Fig. 16**.

Fig. 17 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 to follow the black dashed/dotted line with arrow to the right, starting from observer A in left upper square till you reach the square under number V of the SI-tests at the top of the 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 coefficient you will find in this case is +0.89 (see square right lower corner of the table in the figure). The kappa coefficient +0.52, depicted beneath that of +0.89 illustrates the same relation between Test V and VI but now for observer A. Both kappa coefficients 0.89 and 0.52 are above 0.40 and when using standards for strength of agreement provided by Landis and Koch, both kappa coefficients demonstrates a possible mutual dependency of the diagnostic procedures.

2.3 Combinations of a few different diagnostic procedures: mutual dependency of diagnostic procedure 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 section "Combinations of a few different diagnostic procedures: mutual dependency" of this section. 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 [16].

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-joint dysfunction syndrome (= syndrome diagnosis) of Observer A are presented in a 2 × 2 contingency table of **Fig. 18**.

Instead of showing a separate 2 × 2 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 (**Fig. 19**).

The kappa coefficients 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 also probably uses SI-Test IV for his final judgement. The other SI-Tests I to IV are hardly involved in the final judgement of observer A and B.

Another flaw of reproducibility studies, evaluating different diagnostic procedures for the same clinical phenomenon (for instance different SI-Tests for SI-joint dysfunction), is the fact that it is almost never clear to what "functional system" the different diagnostic procedures are related. Pain provocation SI-Tests and motion pattern SI-Tests (Vorlauf phenomenon) are located in two different functional systems: nociceptive system and postural system. The outcome of the diagnostic procedures have to be in the same functional system as the diagnostic procedure as such.

For instance, SI-joint provocation diagnostic procedures with the outcome of a numeric pain score.

Reproducibility studies, using combinations of several diagnostic procedures

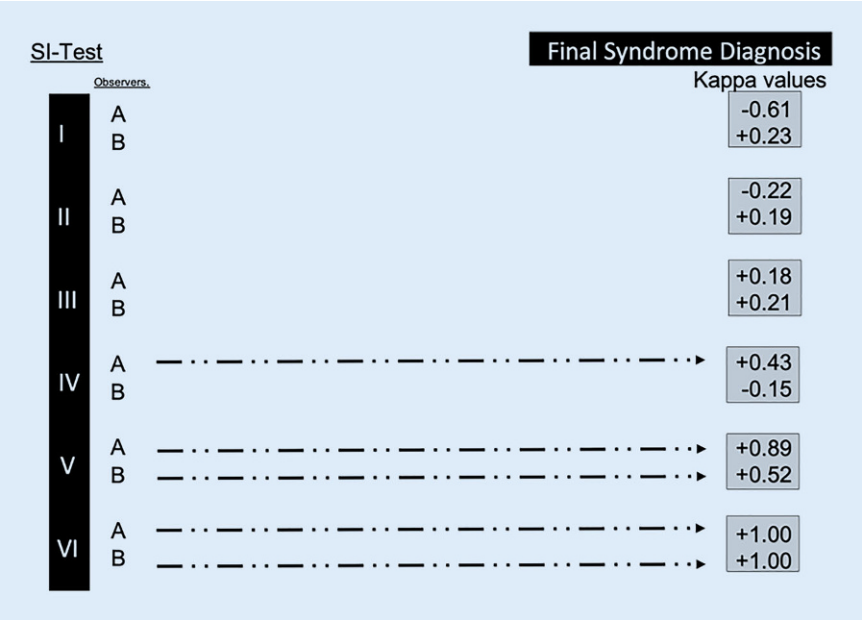

Fig. 19 ▲ Kappa coefficients of relation between separate SI-Tests I to VI and final judgement of the existence of a SI-dysfunction syndrome, per observer *A* and *B* (see text)

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, use 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 subject 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 [17–19]. 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 subject with the diagnostic protocol, it is unclear what diagnostic procedure(s) or combination of diagnostic procedures the observers used for their conclusion. These kinds of reproducibility studies using a diagnostic protocol with a large number of diagnostic procedures have to incorporate only reproducibly proven diagnostic procedures. Besides the mutual dependency of the diagnostic procedures used, their dependency with the final judgement has to be evaluated in the statistical analysis of the reproducible study as illustrated in sections "Combinations of a few diagnostic procedures: mutual dependency" and "Combinations of a few different diagnostic procedures: mutual dependency of diagnostic procedure and final 'syndrome diagnosis' " (under "V. Developing reproducibility studies: general aspects").

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 coefficient) 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, especially 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 [20–22], 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, an incorrect hypothesis as such can influence the observer agreement and consequently the final kappa coefficient 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) [15, 23]. 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.

SI-Test	I	II	III	IV	V	VI	SI-Dysfunction
Observer Couples	kappa values						
A <-> B	+0.11	-0.08	-0.05	+0.29	-0.16	-0.05	-0.05
A <-> C	+0.08	+0.10	+0.38	+0.20	+0.06	+0.14	+0.14
B <-> C	+0.03	-0.16	-0.23	+0.05	+0.13	-0.09	-0.09

Fig. 20 ▲ Kappa coefficients 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)

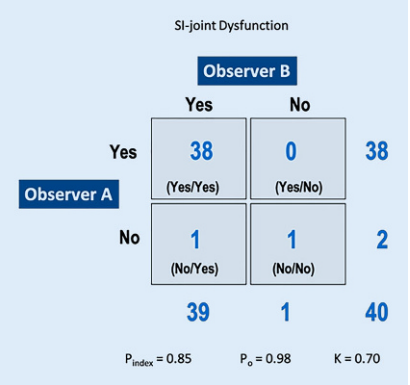

Fig. 21 ▲ A 2 × 2 contingency table showing the agreements and disagreement between observer *A* and *B* about the existence of a SI-joint dysfunction based on three SI-Tests (see text)

The three well-experienced observers (all were M/M Medicine course leaders) adapted the hypotheses of the 6 used SI-Tests from literature. In ◘ **Fig. 20** 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).

Note all kappa coefficients 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 [24], 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 diagnostic procedures could determine the extent of the SI-mobility. Secondly, by very precisely looking at all aspects of the performance of the diagnostic procedures and their judgement, 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 SI-joint dysfunction existed. In ◘ **Fig. 21**, the 2 × 2 contingency table of this study is presented together with the kappa coefficient, prevalence of index condition and overall agreement.

Note that the kappa coefficient 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 ◘ **Fig. 20**) the kappa coefficients of SI-Tests I, II and III were 0.11, –0.08 and –0.05 respectively.

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 tests.

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 originally 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 the reproducibility properties of that diagnostic procedure are.

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 "VII. Protocol format reproducibility study") this agreement is acquired by introducing a training phase in the protocol format of the study. If in a reproducibility study several observers who have not passed the training phase of the protocol are used, the final low kappa coefficients reflect more the personal interpretation or the 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 several training phases in the study protocol, more than two observers can be used to participate in the study [25].

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 coefficients 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 relationship with the reproducibility properties of a diagnostic procedure as such. Before starting a reproducibility study, the observers have, independently from their personal skills, to agree about all the details of the performance of the diagnostic procedure and its final judgement in the training phase.

As will be explained in the reproducibility protocol format (see "VII. Protocol format reproducibility study"), 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 have to be evaluated. If in a reproducibility study several observers who have not passed a training phase of the protocol are used, the final obtained kappa coefficient reflects for instance more the personal interpretation of the well-experienced observer and the comprehension of the evaluated diagnostic procedure 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 originally described diagnostic procedure in literature. In students, a lack of experience with the diagnostic procedure may play a role and influence the final kappa coefficient. It is emphasised in our protocol format (see "VII. Protocol format reproducibility study", section "Training period") to implement a training phase for each observer irrespective the level of skill. Only then is standardisation of the performance and judgement of a diagnostic procedure guaranteed.

5. Number of subjects to be involved in a reproducibility study

In previous editions of his protocol, a total of 40 subjects in study phase was arbitrarily chosen as a "statistical minimum" to perform these kinds of studies. From a practical point of view, a rounded number of 40 was chosen to make these kinds of reproducibility studies relatively easy and cheap to perform. In general, it was advised by a statistician that for simple reproducibility studies with dichotomous outcome and using kappa statistics, 40 subjects were sufficient. Nowadays, sample size calculations based on statistical power are advised to estimate study sample sizes. However, such calculations for sample size are only possible in case a null hypothesis can be formulated such as in randomised controlled trials (RCT). Because kappa statistics is not generally recommended for null hypothesis testing, sample size calculations based on power are not strictly relevant in reproducibility studies with a dichotomous outcome [13]. Instead and more important are the size and stability of the estimates determined by the width of the confidence intervals. 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 [13, 26]. Other approaches have been developed, but were mainly meant for multiple observers with a dichotomous outcome variable [26].Very important to realize and strongly related with the problem of the sample size of a reproducibility is the fact that the kappa coefficient of a diagnostic procedure is not an absolute measure as such. Its value is always dependent on the prevalence of the index condition $\mathbf{P_{index}}$ and to a lesser degree on the overall agreement $\mathbf{P_o}$ (see "IV. Reproducibility studies: kappa statistics", sections "Interpretation of kappa coefficient: dependency of the overall agreement" and "Interpretation of kappa coefficient: dependency of the prevalence of the index condition P_{index}"). This means that the same kappa coefficient can have different levels of the $\mathbf{P_o}$ and the $\mathbf{P_{index}}$. Furthermore, in a reproducibility study, the kappa coefficient and $\mathbf{P_{index}}$ is only related with the positive judged diagnostic procedures. In **Fig. 22**, the squares with data **a** (yes/yes), **b** (yes/no) and **c** (no/yes) are exclusively decisive for the final kappa coefficient. The data **d** (no/no), illustrating the diagnostic procedures observers also agree about, is not included in the kappa coefficient as measure for the reproducibility of a diagnostic procedure. Only the overall agreement $\mathbf{P_o}$ concerns both the positive and negative judged diagnostic procedures observers agree about and are depicted in the squares **a** (yes/yes) and **b** (no/no).

The overall agreement $\mathbf{P_o}$ as such is an absolute measure and reflects more the daily practice of a clinician dealing with diagnostic procedures. As a clinician, one wants to know how reproducible the diagnostic procedure is, both for a positive and negative final judgement of the diagnostic procedure. The overall agreement $\mathbf{P_o}$ is the most appropriate measure for conveying the relevant information in a 2×2 table and is most informative for clinicians [27]. However, when using dichotomous outcomes, we have to realise that the overall agreement $\mathbf{P_o}$ is not corrected for the chance. In the previous IAMMM protocols a minimum level of the $\mathbf{P_o}$ was chosen to guarantee kappa coefficients ≥0.6 in the final study [28]. If the $\mathbf{P_o}$ of a reproducibility study is 0.79, no kappa coefficient ≥0.6 can be obtained.

Dependent on the level the percentage of kappa coefficients ≥0.60 will rise ($\mathbf{P_o} = 0.80$: $\kappa \geq 0.60$ in 25%, $\mathbf{P_o} = 0.82$: $\kappa \geq 0.60$ in 45%, $\mathbf{P_o} = 0.83$: $\kappa \geq 0.60$ in 52%). When the $\mathbf{P_o}$ increases, the kappa/prevalence index curves will simultaneously shift upwards (see **Fig. 23**).

Since the overall agreement $\mathbf{P_o}$ is an absolute measure (in contrast to the kappa coefficient as a relative measure [27]) reflecting more the daily reality of a clinician, we focus in first instance on calculating a sample size for the overall agreement $\mathbf{P_o}$. The first question we have to answer is what, from clinical point of

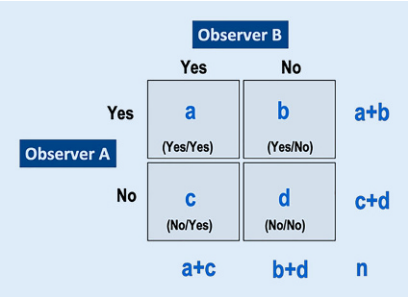

Fig. 22 ▲ Example a theoretical reproducibility with *n* subjects and two observers *A* and *B* presented in a 2 × 2 contingency table (see text)

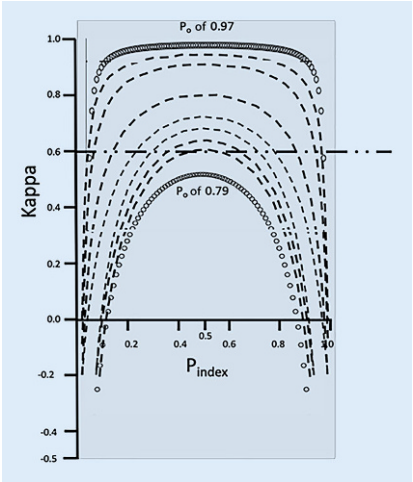

Fig. 23 ▲ Relation between kappa coefficient and prevalence of the index condition. The *horizontal dotted/dashed line* is the cut off level of 0.60. The kappa/prevalence index curves with a too low overall agreement P_o of 0.79 is located beneath the cut off level line of 0.60. The kappa/prevalence index curves with overall agreement P_o of 0.97 is located far above the cut off level line of 0.60

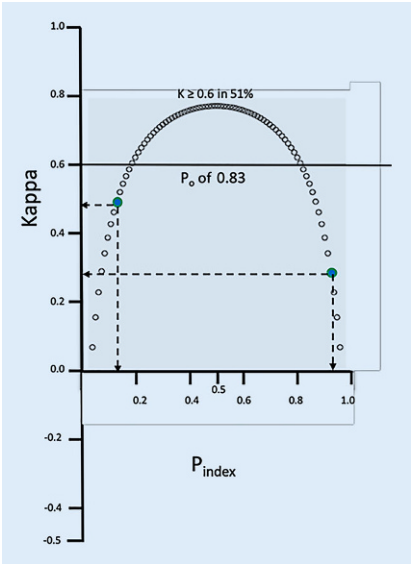

Fig. 24 ▲ Relation between kappa coefficient and prevalence index of the index condition P_{index}. The horizontal line is the cut off level of 0.60. The *blue dots* indicate low kappa coefficients (*horizontal arrows*) in case of a low P_{index} (*left blue dot*) or a high P_{index} (*right blue dot*)

view, is a clinically acceptable level for $\mathbf{P_o}$? One important aspect is the fact that results of reproducibility studies may not change clinical practice in a way that leads to non-ethical poorer subject outcomes [10]. However, being acceptable and/or ethical are both relative issues because we also accept in RCTs, showing a significant effect of a treatment, a non-responder's percentage of 33% or higher. This an everyday reality in medicine that we have to accept and the same is true for the fact that a $\mathbf{P_o}$ of 1.00 of a diagnostic procedure is in most cases not feasible in M/M Medicine. Therefore, you can not find a precise acceptable cut off level for $\mathbf{P_o}$ in the literature. This is the very reason why many texts in the literature only recommend acceptable levels of overall agreement $\mathbf{P_o}$ that range from 0.80 up to 0.90 [10, 29–31].

From clinical point of view, we have to realise that in M/M Medicine, diagnostic procedures are for a large part physical examination procedures. They are the constituent parts of a whole diagnostic arsenal that finally can lead to a particular diagnosis or syndrome diagnosis. One isolated diagnostic procedure in M/M Medicine as such never results in a poorer subject outcome. In M/M Medicine, we rarely have to rely one single diagnostic procedure to make a diagnosis or define a syndrome diagnosis.

In our daily M/M Medicine practice, we deal for a large part with many non-specific clinical conditions with the consequence that we have to rely a number of many different diagnostic procedures. Therefore, in M/M Medicine, a combination of diagnostic procedures has to be performed, which all together point in the same direction towards a particular clinical syndrome, syndrome diagnosis or differential diagnosis. Abovementioned clinical arguments about an acceptable level for $\mathbf{P_o}$ make a minimum value of 0.80 for $\mathbf{P_o}$ acceptable.

From statistical point of view a $\mathbf{P_o}$ level of 0.80 or higher can guarantee a kappa coefficient ≥0.60. In case $\mathbf{P_o} = 0.80$, 25% of the kappa coefficients are ≥0.60. With a $\mathbf{P_o}$ of 0.83, this percentage rises to 50% (**Fig. 24**).

Combining the two points of view, a $\mathbf{P_o}$ of 0.83 is advisable in reproducibility studies with dichotomous outcomes.

To estimate a sample size for a particular $\mathbf{P_o}$ the formula shown below.

$$\text{Sample Size } \mathbf{n} = \frac{\mathbf{P_o}\,(1 - \mathbf{P_o})\,1.96}{\mathbf{ME}^2}$$

ME stands for margin of error and is used as meaning the radius of a confidence interval. A confidence interval (CI) is a type of interval estimate, calculated from the statistics of the observed data that might contain the true value of an unknown population parameter. In other words, confidence interval means a quantification of the level of confidence that a particular parameter lies in the interval.

Depending on the levels of $\mathbf{P_o}$ and the margin of error ME, the sample size will differ. In **Table 2** the sample sizes (*n*) are shown for a ME of 0.1.

From **Table 2** it can be concluded that with a constant ME of 0.1, a rise in the $\mathbf{P_o}$ level results in a decrease of the sample size. However, the choice of a particular ME value is arbitrary and depends on how the digits after the decimal point are rounded. In **Table 2** an ME of 0.1 is used with only one digit after the decimal point. In **Table 3** the sample sizes are shown for the different ME values and now with three digits after the decimal point. In essence all these ME values become 0.1 after rounding to one digit after the decimal point is performed. With a constant $\mathbf{P_o}$ level of 0.83 different sample sizes are acquired due to only slight changes in the ME level. Note that a slight increase of the ME level from 0.100 to 0.116 decreases the sample size from 54 subjects to the 40 subjects. This number of subjects was used in the previous IAMMM protocols.

In the study period the 0.50-$\mathbf{P_{index}}$ method is used to tackle the problem of mutual dependency of the kappa coefficient and the $\mathbf{P_{index}}$ (see "VI. The relation between the kappa coefficient and the

Table 2 Sample size estimation of an overall agreement reproducibility study with a margin of error of 0.1 based on different values of P_o. CI min and CI max mark the boundaries of the confidence interval (CI)

Margin of error (ME)	P_o	1–P_o	Sample size	CI min P_o	CI max P_o
0.1	0.78	0.22	66	0.7	0.9
0.1	0.79	0.21	64	0.7	0.9
0.1	0.80	0.20	61	0.7	0.9
0.1	0.81	0.19	59	0.7	0.9
0.1	0.82	0.18	57	0.7	0.9
0.1	0.83	0.17	54	0.7	0.9
0.1	0.84	0.16	52	0.7	0.9
0.1	0.85	0.15	49	0.8	1.0
0.1	0.86	0.14	46	0.8	1.0
0.1	0.87	0.13	43	0.8	1.0
0.1	0.88	0.12	41	0.8	1.0
0.1	0.89	0.11	38	0.8	1.0

Table 3 Sample size estimation of an overall agreement reproducibility study with different margins of error (ME) and a constant P_o of 0.83. CI min and CI max mark the boundaries of the confidence interval (CI)

Margin of error (ME)	Z-score[2]	P_o	1-P_o	Sample size	CI min P_o	CI max P_o
0.125	3.8416	0.83	0.17	**35**	0.71	**0.96**
0.124	3.8416	0.83	0.17	**35**	0.71	**0.95**
0.123	3.8416	0.83	0.17	**36**	0.71	**0.95**
0.122	3.8416	0.83	0.17	**36**	0.71	**0.95**
0.121	3.8416	0.83	0.17	**37**	0.71	**0.95**
0.120	3.8416	0.83	0.17	**38**	0.71	**0.95**
0.119	3.8416	0.83	0.17	**38**	0.71	**0.95**
0.116	3.8416	0.83	0.17	***40***	0.71	**0.95**
0.111	3.8416	0.83	0.17	***44***	0.72	**0.94**
0.106	3.8416	0.83	0.17	***48***	0.72	**0.94**
0.102	3.8416	0.83	0.17	***52***	0.73	**0.93**
0.098	3.8416	0.83	0.17	***56***	0.73	**0.93**

prevalence of the index condition P_{index}", section "Influencing the P_{index} in advance: the 0.5-P_{index} method"). To apply the 0.50-$\mathbf{P_{index}}$ method, it must be possible to divide the sample size for the study population by 4. Therefore, only samples sizes 40, 44, 48 and 52 can be used in reproducibility studies with a dichotomous outcome and a $\mathbf{P_o}$ level of 0.83. If one regularly performs these kinds of studies, a sample size of 52 is advised. In other cases, the original used sample size of 40 is sufficient.

Based on an overall agreement study (see protocol) with a calculated sample size a **Po** can be acquired with a proper CI interval. Subsequently and based on the same data the kappa (κ) coefficient can be calculated as well. In this way, a reproducibility study with a dichotomous outcome provides you with two measures for the same interobserver agreement: an absolute measure $\mathbf{P_o}$ and a relative measure κ.

VI. Relation between the kappa coefficient and the prevalence of the index condition P_{index}

1. Defining the P_{index} problem

As already mentioned before and elaborated in "IV. Reproducibility studies: kappa statistics", section "Interpretation of kappa coefficient: dependency of the overall agreement", one cannot correctly interpret the kappa coefficient of a reproducibility study without knowing the prevalence of the index condition $\mathbf{P_{index}}$. However, one of the major disadvantages of kappa statistics in reproducibility studies is the fact that the $\mathbf{P_{index}}$ is not known in advance. Only after completion of the study can the $\mathbf{P_{index}}$ be calculated. In all reproducibility studies using kappa statistics, there is always a risk that eventually the $\mathbf{P_{index}}$ will be far too high (too many positive judged diagnostic procedures) or far too low (too few positive judged diagnostic procedures). In **Fig. 24** the relation between the kappa coefficient and the $\mathbf{P_{index}}$, is shown again.

In the example of **Fig. 25**, the blue dots illustrate the risk that, after completion of the reproducibility study, the kappa coefficient appeared to be too low (far under the cut off level of 0.60) as a consequence of a too high $\mathbf{P_{index}}$ value (too high frequency of positive judged diagnostic procedures in the study population) or too low $\mathbf{P_{index}}$ value (too low frequency of positive judged diagnostic procedures in the study population).

Theoretically in this example, a $\mathbf{P_{index}}$ value of 0.50 is preferable because with that value, the kappa coefficient will be located at the top of the kappa/prevalence index curve (**Fig. 25**).

As can be seen in **Fig. 25**, the kappa/P_{index} curve with a $\mathbf{P_o}$ of 0.83 is for a large part above the cut off of 0.6. In reproducibility studies with a $\mathbf{P_o} \geq 0.80$, the kappa coefficient corresponding with a $\mathbf{P_{index}}$ of 0.50 is by definition >0.60. Therefore, we had to evaluate in advance the influence of the $\mathbf{P_{index}}$ in such a way that eventually after completion of the reproducibility study the $\mathbf{P_{index}}$ will be around 0.50.

2. Influencing the P_{index} in advance: the 0.50-P_{index} method

In a fictitious reproducibility study with a dichotomous outcome (yes/no), two observers A and B wanted to evaluate the reproducibility of diagnostic procedure (Test I). They used a study population of total 40 subjects. Each observer examines his own 20 subjects with Test I

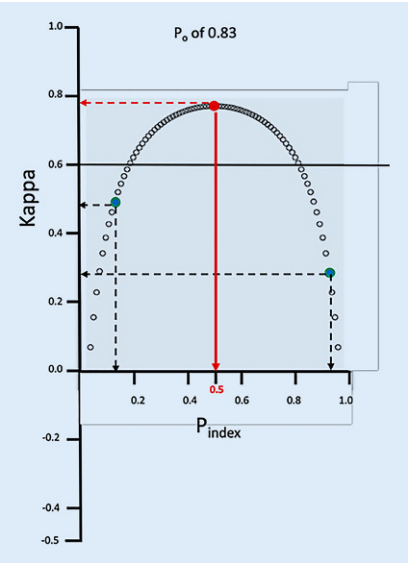

Fig. 25 ▲ Relation between kappa coefficient and prevalence index of the index condition P_{index}. The *horizontal line* is the cut off level of 0.60. The *blue dots* indicate low kappa coefficients (*horizontal arrows*) in case of a low P_{index} (*left blue dot*) or a high P_{index} (*right blue dot*). The *red dot* is the kappa coefficient with a P_{index} of 0.50 and a kappa coefficient above the cut off level of 0.60

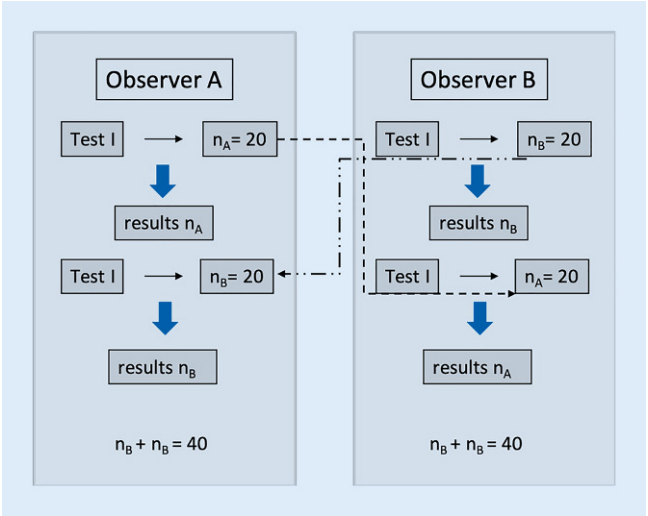

Fig. 26 ◀ Flow diagram of a reproducibility study with two observers *A* and *B*. Both observers perform Test I in their own subjects (n_A and n_B). Both observers send their subjects to each other (*dotted arrows*). Each observer examines a total of 40 subject (see text)

(**Fig. 26**). Both observers were blinded for the results of his colleague observer. Both mutual communication and communication with the examined subject was not allowed.

Subsequently, each observer sent his 20 examined subjects to his colleague observer. At the end, both observers have examined 40 subjects with Test I. Due to the blinding procedure, no bias can occur because there is always a 50% chance that an observer will have a positive or a negative test result when he examines every subject sent to him from his colleague observer. However, as stated before, the number of positive judged diagnostic procedures by both observers is not known in advance. More precisely, there is always a risk in this study format to obtain a $\mathbf{P_{index}}$ that is too high or too low which can result in an undesirably low kappa coefficient as a measure for inter-observer agreement. As explained in "Defining the $\mathbf{P_{index}}$ problem" of this section, a $\mathbf{P_{index}}$ of 0.50 is preferable because the kappa coefficient will be located at the top of the kappa/P_{index} curve (**Fig. 25**).

We can take a fictitious reproducibility study, in which two observers A and B wanted to evaluate the reproducibility of a diagnostic procedure (Test I) in a study population of 100 subjects. Suppose they have trained the whole diagnostic procedure of Test I and evaluated the standardisation of their diagnostic procedure (final judgement included) in a small pilot study (see "VII. Protocol format reproducibility study", section "Training period") and succeeded to have an estimated $\mathbf{P_o}$ of 0.85.

Instead of sending just 20 subjects to each other as illustrated in **Fig. 26**, both observers A and B now send 25 subjects with a positive Test I and 25 subjects with a negative Test I to each other (**Fig. 27**). By this way observer B receives 50 subjects from observer A and vice versa. Since the whole procedure was blinded (mutual communication as well as communication with the examined subject was not allowed) again every subject sent by observer A had 50% chance for observer B to have a positive or negative judged Test I. Based on the $\mathbf{P_o}$ of 0.85 of the pilot study, observer B will agree in 21 of 25 ($0.85 \times 25 = 21.25$) subjects with observer A having a positive judged Test I and will disagree in 4 of 25 ($0.15 \times 25 = 3.75$) subjects with a positive judged Test I. The same can be calculated for the negative judged tests. The same situation exists for the subjects sent by observer B to observer A. The end result is that the observers agree in 42 subjects with a positive or negative Test I and disagree in 16 subjects. The data of this fictitious reproducibility study can be presented in a 2×2 contingency table (**Fig. 28**). Based on the theoretical data (figures in the squares in parentheses) shown in the 2×2 contingency table of **Fig. 28** the $\mathbf{P_o}$, $\mathbf{P_{index}}$ and kappa coefficient can be calculated. The overall agreement $\mathbf{P_o} = 0.85$ and $\mathbf{P_{index}} = 0.50$ with a kappa coefficient of 0.70. This fictitious reproducibility study, in which 2 observers sent 50% positive and 50 negative judged tests to each other under strict blinding circumstances, is the typical example of what is called the 0.50-P_{index} method.

This theoretical format of the 0.50-P_{index} method has been proven in a first reproducibility study [32] and subsequent studies using this method [33–37]. In that first study, two observers P and E evaluated the reproducibility of the Passive Hip Flexion Test. After a training phase, they obtained an overall agreement $\mathbf{P_o}$ level of 0.88 in the pilot study. In the final study phase, using the 0.50-P_{index} method in 40 subjects, the obtained a prevalence of the index condition $\mathbf{P_{index}}$ of 0.44 (**Fig. 29**), which is near the ideal $\mathbf{P_{index}}$ value of 0.50.

As expected, the overall agreement $\mathbf{P_o}$ of 0.88 remained stable compared to the value of the pilot study. A kappa coefficient 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 with dichotomous outcomes and two observers.

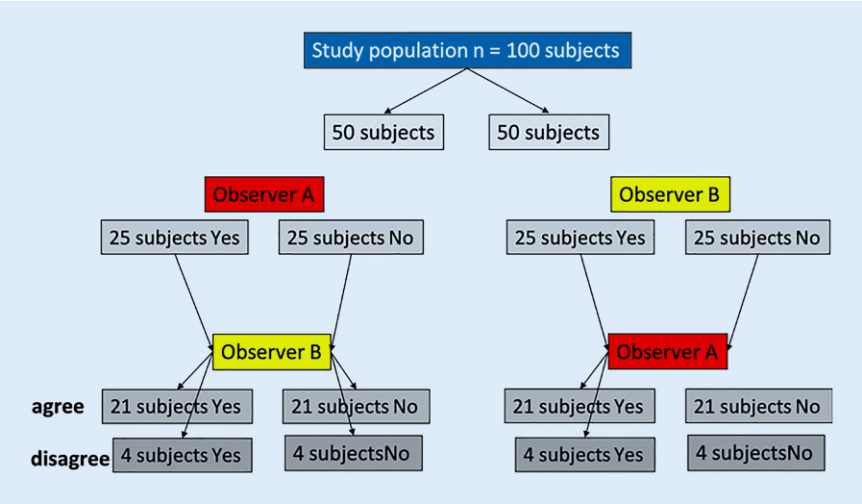

Fig. 27 ▲ Flow diagram of a reproducibility study using the 0.50-P_{index} method. Both observers *A* and *B* perform Test I in their own subjects. Both observers send 25 subjects with a positive judged Test I (yes) and 25 subjects with a negative judged Test I to each other. Based on the overall agreement of the pilot study performed in advance of 0.85, the number of subjects they agree and disagree about can be calculated (see text)

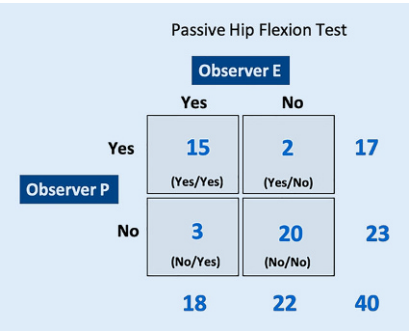

Fig. 29 ▲ A 2 × 2 contingency table showing the agreements and disagreement between observer *P* and *E* about the existence of a positive or negative judged Passive Hip Flexion Test, using the 0.50-P_{index} method (see text)

VII. Protocol format reproducibility study

As shown in ◘ **Fig. 30**, an entire reproducibility study can be subdivided into five different periods, which have to be successively completed. 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 procedure at a time with a dichotomous outcome.

The arguments for this decision are elaborated in "V. Developing reproducibility studies: general aspects", section "Number of diagnostic procedures evaluated in a reproducibility study".

1. Logistic period

In the logistic phase, mainly agreements about the study conditions, participating members and logistics of the reproducibility study are concluded (◘ **Fig. 31**).

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 IAMMM 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 agreed. The responsible person decides in cooperation with the participating members who is doing what during the study, for instance developing an evaluation form for the study.

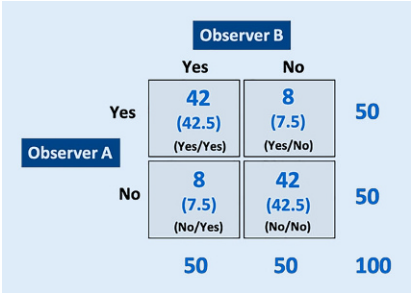

Fig. 28 ▲ The 2 × 2 contingency table of the fictitious reproducibility study with a dichotomous outcome (Yes/No) in 100 subjects showing the agreements and disagreement after using the 0.50-P_{index} method resulting in a P_{index} of 0.50 (see text)

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 outpatient 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 subjects for the study. This is essential because both in the training phase (see section "Training period") and the study phase (see section "Study period") observers have to send subjects to each other. Special arrangements have to be made for an aspect of the training phase, in which both observers have to examine 10 subjects in detail to agree about the diagnostic procedure performance and its final judgement. It is advisable to reserve special time for the training phase. Both the training and the study phase can easily be performed during the regular consulting hours of an outpatient clinic when both observers see their subjects at the same time.

1.4 Finance

In essence a reproducibility study requires no extra financing 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

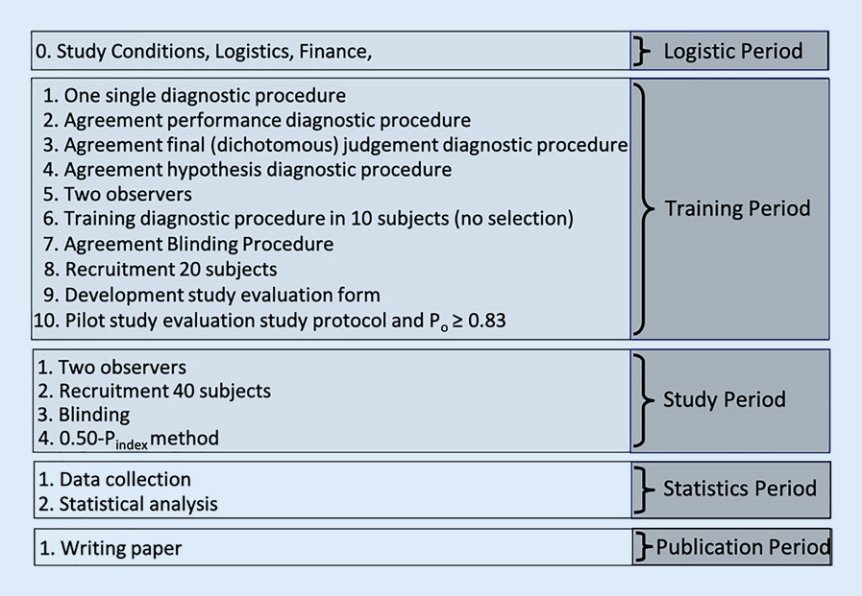

Fig. 30 ▲ Flow chart of planning in different periods of a reproducibility study

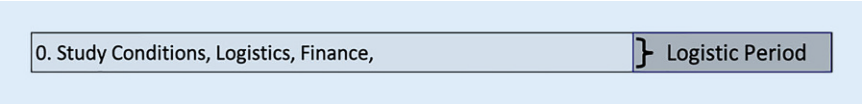

Fig. 31 ▲ Logistic phase of the reproducibility protocol

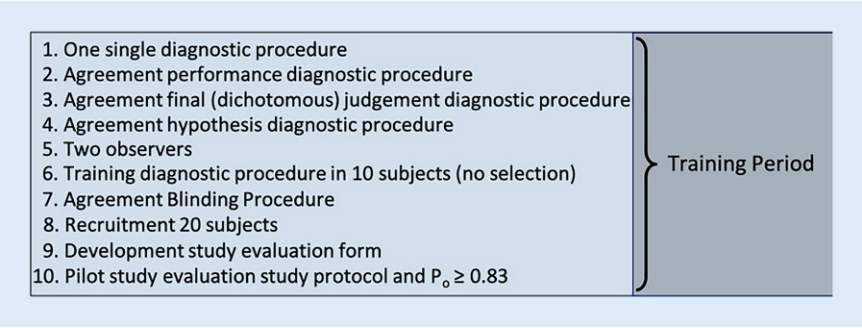

Fig. 32 ▲ Training period of the reproducibility protocol

because an "excel" file is added in this protocol which automatically calculates all the necessary results (**P_o**, **P_{index}**, **kappa**) for publication (see "VII. Protocol format reproducibility study", section "Statistics period").

1.5 Approval by the local ethical committee

In this phase, depending on the local ethical regulation of patient participation in studies, consent the local medical ethical committee of your hospital or university about the ethical aspects of your study and the conditions for an informed consent must be obtained.

In some universities, the final research protocol, including a copy of the written information about the study to the subjects etc., must be forwarded to the ethical committee.

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

2. Training period

2.1 Observers and subject recruitment

The training phase is a very essential phase of the reproducibility study. In this phase, the basis is created for a successful outcome of the reproducibility study. The first 6 items of the training period are mainly dedicated to the mutual agreement between observers about all detailed aspects of the performance and final judgements of the diagnostic procedure they want to evaluate in the reproducibility study. The best way to train the diagnostic procedure is when both observers participate in a session in which they perform the diagnostic procedure on the same subject. Comparing each other's performance of the diagnostic procedure and the definition of their final judgement have to result for a univocal performance of the diagnostic procedure (**Fig. 32**).

In general, only two observers (see "V. Developing reproducibility studies: general aspects", section "Number of observers") are needed and maximal 10 different subjects. These subjects can be randomly recruited from the observers' own patient record files or medical staff. No particular exclusion or inclusion criteria are necessary for these 10 subjects. One has to realise that in a reproducibility study of a diagnostic procedure, just the reproducibility of the execution of the whole performance of the diagnostic procedure and its final judgement by the observers are evaluated (for instance a positive or negative judged Patrick Test) and not the existence of a diagnosis!

Each observer can choose a subject—who will have a positive or negative outcome of the diagnostic procedure.

2.2 Selection of diagnostic procedure(s) to be evaluated

Observers must agree upon which diagnostic procedure they wish to evaluate in the reproducibility study. Before a definite choice, the literature has to be studied to determine whether the diagnostic procedure have been evaluated before. The aspect of the number of diagnostic procedures to be evaluated has been discussed in "V. Developing repro-

Fig. 33 ▲ Semi-quantification by number of fingers of the Passive Hip Flexion/Adduction Test (see text)

ducibility studies: general aspects", section "Number of diagnostic procedures evaluated in a reproducibility study". In general, it is advisable to evaluate one diagnostic procedure at a time. In case of more than one diagnostic procedure is evaluated in the same reproducibility study, the 0.50-P_{index} method cannot be applied anymore (see "VI. The relation between the kappa coefficient and the prevalence of the index condition P_{index}", section "Influencing the P_{index} in advance: the 0.50-P_{index} method"). In this case there is a risk that many diagnostic procedures finally show low kappa coefficients because of too low or too high $\mathbf{P_{index}}$ values (see "IV. Reproducibility studies: kappa statistics", section "Interpretation of kappa coefficient: dependency of the prevalence of the index condition P_{index}").

2.3 Mutual agreement about performance of diagnostic procedure

After agreement upon which diagnostic procedure will be evaluated in the study, a preliminary description of the performance of the diagnostic procedure takes place. It is advisable to have the original description of the diagnostic procedure from literature. Subsequently, the observers make the first own description of the diagnostic procedure. This description must be very detailed and specific, taking both the observer (examiner) and the subject into account. Next, the two observers start their session in which they examine 10 subjects. Comparing each other's performance of the diagnostic procedure and the definition of their final judgement have to result in a univocal performance of the diagnostic procedure.

It is advisable for the observers to start in the first session with the performance of the diagnostic procedure on each other. The following elements of the diagnostic procedure have to be discussed in detail for later standardization of the entire performance of the diagnostic procedure:

1. position of the subject
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
6. description final judgement

The observers by consensus have to agree about all the details of the performance of the diagnostic procedure. This consensus has to be recorded in the logbook.

It is advisable for both observers also to train the agreed performance of diagnostic procedure by routinely using this diagnostic procedure in their daily practice.

2.4 Agreement about hypothesis of diagnostic procedure

In "V. Developing reproducibility studies: general aspects", section "Hypothesis of the diagnostic procedure in a reproducibility study", we already discussed that the hypothesis of a diagnostic procedure as such can influence the final result of a reproducibility study. Most of the diagnostic procedures mentioned in textbooks and M/M Medicine course syllabuses are based on unproven hypotheses. Therefore, in reproducibility studies observers should ignore these unproven hypotheses. Based on the detailed performance of the diagnostic procedure, observers have to agree about what the diagnostic procedure actually tests. For instance, the passive hip flexion adduction diagnostic procedure is supposed to test the mobility of the SI-joint (textbook hypothesis). Looking closely at the performance of the diagnostic procedure, 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 diagnostic procedures that include provoking pain. Some of these pain-provoking diagnostic procedures are supposed to identify a particular anatomical structure as the source for the pain, for instance the SI-joint. However, in all kinds of SI-joint pain provoking diagnostic procedures, many different anatomical structures outside the SI-joint can be the source for pain. In this case, the best working hypothesis is that different anatomical structures, functionally related to the SI-joint, can be the cause of the pain. In general, most diagnostic procedures in M/M Medicine are related to range of motion and do not give rise to problems of defining the working hypothesis. In other cases, a working hypothesis of the diagnostic procedure can be defined by the observers by carefully looking at all the details of the performance of the diagnostic procedure being examined.

2.5 Agreement about the final judgement of diagnostic procedure

Both the agreements of the two observers about the performance and the hypothesis of the diagnostic procedure are decisive for the final judgement of the diagnostic procedure. The observers have to look carefully on how they normally use the diagnostic procedure in their daily practice and in particular how they judge a positive or negative result. Frequently, and with very experienced practitioners, this judgement happens almost automatically. The participating observers have to be very careful to look how they judge a diagnostic procedure 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 was semi-quantified, using the number of fingers to measure distance between the chest and the knee [32].

REPRODUCIBILITY EVALUATION FORM

Diagnostic Procedure: Passive Hip Flexion/Adduction Test

Pilot Study ■ **Study Period □**

Subject Registration Number □□□□□ Male■ /Female□

Aim:

Number subjects: 20 (10 per observer)

Number Observers: 2
Inclusion Criteria: None
Exclusion Criteria: None
Selection: Consecutive
Blinding: No mutual communication between observers and between subject and observer

***Hypothesis Diagnostic Procedure*:**

Performance Diagnostic Procedure:

Semi-Quantification Outcome Diagnostic Procedure (optional):

Judgement Diagnostic Procedure:

Diagnostic Procedure	Left positive	Right positive
Passive Hip Flexion Adduction Test		

Fig. 34 ▲ Example of reproducibility evaluation form used in the pilot study or study phase

1. One single diagnostic procedure
2. Agreement performance diagnostic procedure
3. Agreement final (dichotomous) judgement diagnostic procedure
4. Agreement hypothesis diagnostic procedure
5. Two observers
6. Training diagnostic procedure in 10 subjects (no selection)
7. Agreement Blinding Procedure
8. Recruitment 20 subjects
9. Development study evaluation form
10. Pilot study evaluation study protocol and $P_o \geq 0.83$

} Training Period

Fig. 35 ▲ Training period of the reproducibility protocol

In ◘ **Fig. 33** this semi-quantification of Passive Hip Flexion/Adduction Test is shown. The black hand represents the number of fingers of the observer between the chest and the knee of the patient. The same procedure is repeated on the left side.

The numbers of fingers of both sides were estimated. A left/right difference of more than one finger was decisive and it was agreed that the side with the largest number of fingers is the most restricted side and labelled as a positive Passive Hip Flexion/Adduction Test. Depending on the diagnostic procedure, observers have to agree about how they judge a diagnostic procedure 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 diagnostic procedure in their daily practice. The decision about the judgement of the diagnostic procedure is recorded in the logbook.

2.6 Agreement about the blinding procedure

Both in the pilot study and study period, adequate blinding procedures are essential. Except for diagnostic procedures that evaluate pain, no communication between observers and between observer and subject during the performance of the whole diagnostic procedure is allowed.

Of course, each observer, who has selected a subject for the study, has to inform the subject about the aim of the study, in accordance with the guidelines of the local ethical committee.

2.7 Study evaluation form

Based on the results of the performance and judgement discussions, an evaluation form is developed which has to be used in the study (see ◘ **Fig. 34**).

This evaluation form is used both in a pilot study and in the study period. In this form a brief overview of the study is mentioned. But more importantly, the details about the subject, the performance, the semi-quantification and the judgement of the diagnostic procedure are recorded. At the bottom of the form the positive judged side of the diagnostic procedure can be recorded.

After the mutual agreement of all aspects of the diagnostic procedure, the whole procedure has to be evaluated with respect to the standardization. In other words, do both observers keep the agreed performance of the diagnostic procedure in detail in a standardized manner as filled out in the evaluation form (◘ **Fig. 35** items 9 and 10).

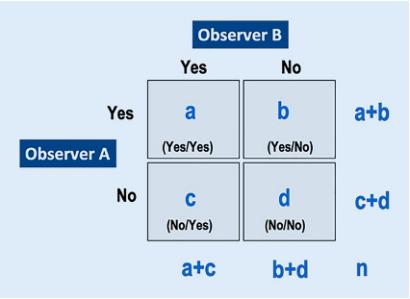

Fig. 36 ▲ The results of a theoretical reproducibility with *n* subjects and two observers *A* and *B* presented in a 2 × 2 contingency table (see text)

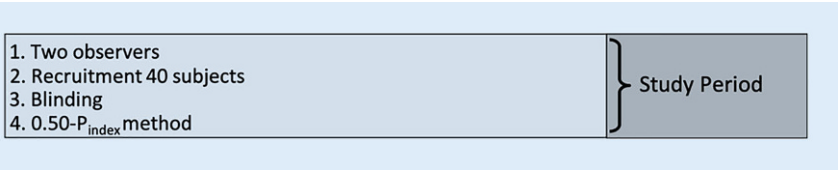

Fig. 37 ▲ Study period of the reproducibility protocol

Both observers are scheduled for their outpatient clinic at the same time and place.

In advance, 40 evaluation forms are made (20 forms for each observer).

For the pilot study each of the two observers has to recruit during his outpatient clinic programs 10 subjects out of his outpatient programme.

In principle, all subjects with pain of the locomotion system are suitable for recruitment for the pilot study.

Each observer performs the diagnostic procedure in all the subjects of his policlinic programme and selects subjects for the study in which he is convinced that the diagnostic procedure is positive or negative.

As explained earlier, it is not necessary to select subjects with complaints in the same region in which the diagnostic procedure under examination is located. One has to realise that in a reproducibility study of a diagnostic procedure, just the reproducibility of the execution of the whole performance of the diagnostic procedure and its final judgement by the observers are evaluated (for instance a positive or negative judged Patrick Test) and not the existence of a diagnosis!

For each examined subject, the observers fill out one evaluation form with the results and one form with only the subject's registration number. The latter form is used by colleague observer who examines the patient as a second observer. When all 20 subjects have been examined by both observers, the evaluation forms are collected and the number of subjects in which observers agree (Yes/Yes and No/No) are calculated according a 2 × 2 contingency table is shown as in Fig. 36.

The formula for $\mathbf{P_o}$ based on the data of Fig. 36 is:

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

In the case the overall agreement is too low, i.e. a $P_o < 0.80$, the observers must again go through a new training phase.

In the renewed training phase, 10 new subjects must be examined in all details with respect to the performance and the judgement of the diagnostic procedure. 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 problem(s) is identified, an adaptation is made and recorded in the logbook. The evaluation form is adapted and a pilot study has to be performed.

If again no $\mathbf{P_o} \geq 0.80$ is obtained, the observers have again to go back to a second renewed training period. If still no substantial $\mathbf{P_o}$ value ≥0.80 is obtained, the observers have to discuss the continuation of the reproducibility study.

Furthermore, they have to consider whether the diagnostic procedure under examination is suitable for educational purposes because of a disputed transferability illustrated by the repeatedly found $\mathbf{P_o} < 0.80$. Although the evaluation of the diagnostic procedure can have a negative result for its diagnostic value, the publication of these kinds of results is very valuable for education systems in M/M Medicine.

4. Study period with 0.50-P$_{index}$ method

Proceeding to the study period indicates that the $\mathbf{P_o}$ value is ≥0.80 (advisable 0.83, see section "Number of subjects to be involved in a reproducibility study") and provided that the overall agreement will be constant. In the study period, evaluating one single diagnostic procedure, the 0.50-P$_{index}$ method is used (see "VI The relation between the kappa coefficient and the prevalence of the index condition P$_{index}$", section "Influencing the P$_{index}$ in advance: the 0.50-P$_{index}$ method"). If more than one diagnostic procedure is evaluated in the same reproducibility study, there is always the risk of a very low or very high P$_{index}$ resulting in an unwanted low kappa coefficient (see "IV. Reproducibility studies: kappa statistics", section "Interpretation of kappa coefficient: dependency of the overall agreement") (Fig. 37).

4.1 Observers and subjects recruitment

The same two observers from the training period perform the study period. Because the observers have to send subjects to each other, it is preferable that both observers work at the same clinic, so that the programme of the clinic can be organised with respect to the study.

Each observer selects 20 subjects from his policlinic programme, making 40 subjects in total. This means that 80 evaluation forms have to be made in advance (40 per observer).

In principle all subjects with pain related to the locomotion system are suitable for inclusion to the study period. Consequently, the observer performs the diagnostic procedure on all the subjects of his policlinic programme, except in case the diagnostic procedure is related to a particular disease entity with special inclusion and exclusion criteria. He chooses subjects in whom he is convinced that the diagnostic procedure is positive or negative.

This means that negative or positive lumbar diagnostic procedures in subjects with headache and or neck pain can also be used for the study. Each observer

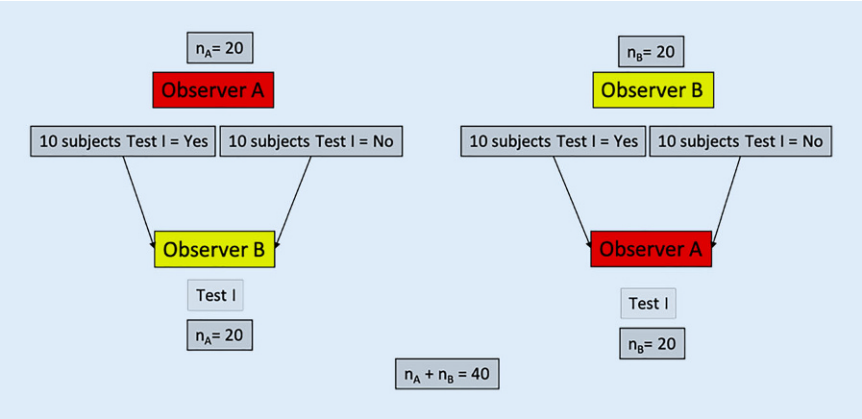

Fig. 38 ▲ Flow diagram of a reproducibility study using the 0.50-P_{index} method. First both observers *A* and *B* perform Test I in their own 20 subjects and select 10 subjects with a positive and 10 subjects with a negative diagnostic procedure (Test I). Subsequently, they send their own population of 20 subjects (n_A and n_B) to each other. 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 subjects ($n_A + n_B$)

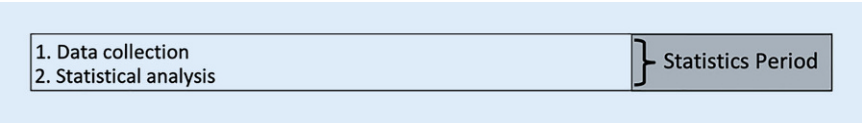

Fig. 40 ▲ Statistic period of the reproducibility protocol

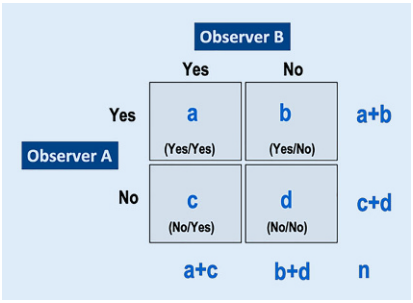

Fig. 39 ▲ Example a theoretical reproducibility with *n* subjects and two observers *A* and *B* presented in a 2 × 2 contingency table (see text)

sends 10 subjects with a positive test and 10 subjects with a negative test to the other observer according to the scheme presented in ◘ **Fig. 38**. The total study population will be 40 subjects.

When the 40 subjects have been examined by both observes, the results from each subject's evaluation forms are collected by an independent person and added to the 2 × 2 contingency table. The number of subjects 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 ◘ **Fig. 39**.

4.2 Blinding procedures

An adequate blinding procedure is as essential in the study period, as in pilot study of the training period. The blinding procedure is identical except for the number of subjects to examine. This means that, except for pain evaluating diagnostic procedures, no communication is allowed between observers and between subject and observer during the performance of the diagnostic procedure. Of course, each observer who has selected a subject for the study has to inform the subject about the aim of the study in accordance with the guidelines of the local ethical committee. However, the observer who receives a subject for the study from his colleague is not allowed to communicate with the subject about the study at all. No communication is allowed between observers about the examined subjects included in the study period of the reproducibility study. The best way is to have an independent person collect the completed evaluation forms directly after the examination of the subject (◘ **Fig. 40**).

5. Statistics period

The independent person collects the filled-out evaluation forms directly after the examination of the subject by the observers. The data obtained from the evaluation forms are arranged according the 2 × 2 contingency table (a, b, c and d) as shown in ◘ **Fig. 39**. Subsequently, the data has to be analyzed and the $\mathbf{P_o}$, $\mathbf{P_{index}}$ and **kappa** have to be calculated. For this purpose, a spreadsheet that automatically calculates all desired results ($\mathbf{P_o}$, $\mathbf{P_{index}}$ **kappa**, confidence intervals [CI]) can be downloaded from the IAMMM website. In this spreadsheet the data a, b, c and d from the 2 × 2 made contingency table can are inserted in the first four columns of the spreadsheet under the labels a, b, c and d of the first row. All desired results $\mathbf{P_o}$ with CI, $\mathbf{P_{index}}$ and **kappa** with CI are calculated, the P_c included.

The same spreadsheet can be used for other calculations, for instance the mutual dependency of diagnostic procedures when more than one diagnostic procedure is evaluated (see "V. Developing reproducibility studies: general aspects", "Combinations of a few different diagnostic procedures: mutual dependency").

By copying entire row 2 to a next row, other data can be filled out under labels a, b, c and d.

6. Publication period

When writing a paper for 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 describing the materials and methods, another section with results and finally the discussion section.

6.1 Introduction section

In the introduction, the literature is reviewed regarding 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.

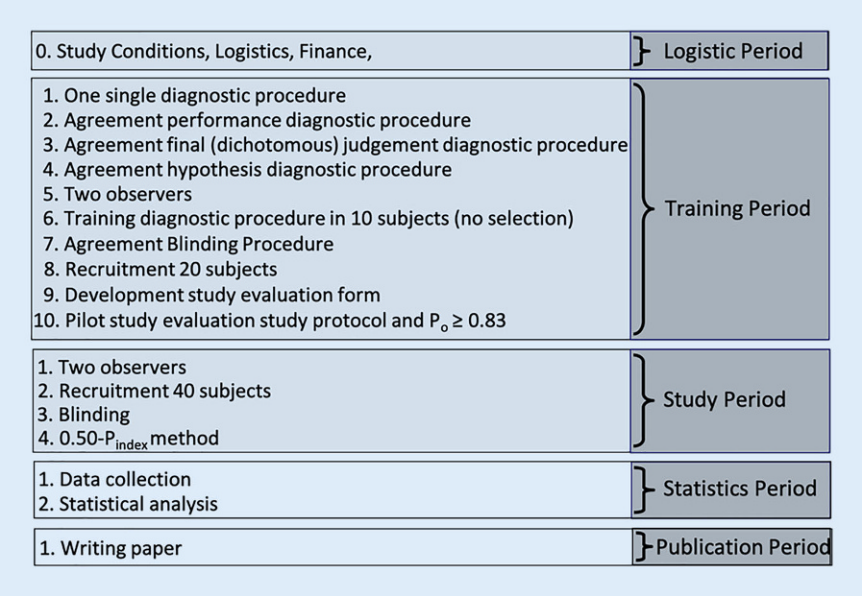

Fig. 41 ▲ Flow chart of planning in different phases of a reproducibility study

Additionally, at the end of the introduction the reasons and aims of the present study are mentioned.

6.2 Materials and methods section

In this section, only the results of the study have to be presented and no interpretation of results. First of all, under the heading **Materials** in this section all the characteristics of the subjects have to be mentioned such as:

1. Data about the source population (hospital, outpatient 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 subjects were selected from the source population (on entrance, consecutive, every other, non-selective sample procedure etc.)
3. Data about exclusion criteria of the subjects
4. Demographic data (gender, age of subjects or normal subjects) that are recorded in the study.
5. Number of subjects or normal subjects in different phases of the study.
6. Under the heading **Methods** in this section all the characteristics of the study format have to be mentioned such as:
   - If in the **Preparation Phase** data about the participating members of the study is 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.
   - If a **Training Phase** was used in the study format
   - Detailed data about the performance of the diagnostic procedure. It is not sufficient to simply refer to a diagnostic procedure from the literature.
   - Detailed data about the hypothesis of the diagnostic procedure of the observers.
   - Detailed data about the judgement of the diagnostic procedure(s) and/or semi-quantification.
   - Detailed data about the conditions for a final "diagnosis" in case of several diagnostic procedures.
   - Number of diagnostic procedures.
   - Data about the characteristics of the observers (experience etc.)
   - If a Pilot study for evaluation of the protocol was used in the study format.
   - If in the **Study Period** the 0.50-P_{index} method was used.
   - At the end of the Materials 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.
   - If other statistical methods were used, the reason why has to presented.
   - If in case of more than one diagnostic procedure was evaluated, the analysis of their mutual dependency has to be presented/described (see "V. Developing reproducibility studies: general aspects", "Combinations of a few different diagnostic procedures: mutual dependency").
   - In the case of a final "diagnosis" based on several diagnostic procedures, the mutual dependency between a single diagnostic procedure and the final diagnosis has to be analysed (see "V. Developing reproducibility studies: general aspects", "Combinations of a few different diagnostic procedures: mutual dependency of diagnostic procedure and final 'syndrome diagnosis'").

6.3 Results section

In this section only the 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 subjects or normal subjects).
2. Data about the $\mathbf{P_o}$ of the diagnostic procedure(s) with confidence intervals, margin of error (see "IV. Reproducibility studies: kappa statistics", section "Definition of the kappa coefficient").
3. Data about the $\mathbf{P_{index}}$ and not just frequencies of the participating observers (see "IV. Reproducibility studies: kappa statistics", section "Prevalence and the prevalence of the index condition").
4. Data about the **kappa coefficient** of the diagnostic procedure(s) with confidence intervals.
5. Data about mutual dependencies of diagnostic procedures per observer in table format (see **Fig. 17**).
6. Data about mutual dependencies, final diagnosis and individual diag-

nostic procedures per observer in table format (see Fig. 19).
7. Presentation of the raw data in 2 × 2 contingency tables.

6.4 Discussion section

In this section results of the study are discussed and compared with the results from other reproducibility studies in the literature. In the introduction section these studies have already been discussed. Explain why the present study has different results compared to those studies from literature.

Formulate a clear conclusion with its consequences for daily practice of M/M Medicine. Essential in the discussion section is to mention the strong and weak aspects of your study and make recommendations to improve for future studies. In case the reproducibility shows good results it is 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.

VIII. Golden rules for reproducibility studies

In Fig. 41 the scheme is presented again to show the different aspects and periods 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.

Rule 1: Create a clear logistic and responsibility structure for the reproducibility study in the preparation phase. A 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 an extensive training period in the study protocol. Between observers, full agreement must be achieved about all details of the test(s) under examination.

Rule 4: Always include a pilot study in the training period to evaluate the standardization of the performance of the diagnostic procedure. It is also essential that this pilot study achieve a $\mathbf{P_o} \geq 0.80$. By definition, lower $\mathbf{P_o}$ values result in low kappa coefficients.

Rule 5: Always repeat the pilot study of 20 subjects in the training phase in case of a $\mathbf{P_o} < 0.80$.

Rule 6: Preferred is the evaluation of only one diagnostic procedure in a reproducibility study.

Rule 7:

Rule 8: Always present raw data in 2 × 2 contingency tables.

Rule 9: Always present the $\mathbf{P_o}$ together with the values of the $\mathbf{P_{index}}$ and **kappa coefficient**.

Corresponding address

Prof. Dr. Jacob Patijn, MD, PhD
International Academy of Manual/ Muscoloskeletal Medicine, IAMMM
Zurich, Switzerland
jacob.patijn@gmail.com

Acknowledgements. The development of the first protocol editions was done in close cooperation with a second editor, Dr. Lars Remvig. The IAMMM wants to thank him for his important previous contributions.

In addition, we want to thank Dr. Sjef Rutte, MD, for editorial advice and critical comments about the readability of the protocol.

Finally, we want to thank Sander van Kuijk, PhD, clinical epidemiologist, Department of Clinical Epidemiology and Medical Technology Assessment, Maastricht University, Netherlands. His contribution was crucial for the development of the proper sample size for reproducibility studies.

Compliance with ethical guidelines

Conflict of interest J. Patijn declares that he has no competing interests.

For this article no studies with human participants or animals were performed by any of the authors. All studies performed were in accordance with the ethical standards indicated in each case.

References

1. Bin Ab Rahman J (2015) Brief guidelines for methods and statistics in medical research. Springer, Berlin
2. Dhillon BS (2003) Series on Industrial & System Engineering – Vol. 2, Human Reliability and Error in Medical System. World Scientific, New Jersey, London, Singapore, Hongkong, ISBN 981-238-359-X
3. Gopalakrishna G, Langendam MW, Scholten RJPM, Bossuyt PMM, Leeflang MMG (2016) Defining the clinical pathway in cochrane diagnostic test accuracy reviews. BMC Med Res Methodol 16(1):153
4. Lewit K (2010) Manipulative therapy: musculoskeletal medicine. Churchill Livingstone Elsevier, Edinburgh, London, New York, Oxford, Philadelphia, St. Louis, Sydney, Toronto, ISBN: 9-780-7020-3056-7
5. Colachis SC, Worden RE, Bechtol CO, Strohm BR (1963) Movement of the sacroiliac joint in the adult male: a preliminary report. Arch Phys Med Rehabil 44:490–498
6. Kellgren JH, Lawrence JS (1957) Radiological assessment of osteo-arthrosis. Ann Rheum Dis 16(4):494–502
7. Côté P, Cassidy JD, Yong-Hing K, Sibley J, Loewy J (1997) Apophysial joint degeneration, disc degeneration, and sagittal curve of the cervical spine. Can they be measured reliably on radiographs? Spine 22(8):859–864
8. Bland JM, Altman DG (2003) Applying the right statistics: analyses of measurement studies. Ultrasound Obstet Gynecol 22(1):85–93
9. Tammemaggi MC, Frank JW, Leblanc M, Artsob H, Streiner DL (1995) Methodological Reproducibility—A comparative study of various indices of reproducibility applied to the reproducibility of ELISA serologic tests for Lyme disease. J Clin Epidemiol 48(9):1123–1132
10. McHugh ML (2012) Interrater reliability: the kappa statistic. Biochem Med 22(3):276–282
11. Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. Biometrics 33(1):159–174
12. Cicchetti DV, Feinstein AR (1990) High agreement but low kappa: II. Resolving the paradoxes. J Clin Epidemiol 43(6):551–558
13. Tooth LR, Ottenbacher KJ (2004) The κ statistic in rehabilitation research: an examination. Arch Phys Med Rehabil 85(8):1371–1376
14. Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. Biometrics 33(1):159–174
15. Van Deursen L, Patijn J (1993) Aufwertung der Ligamentaren Kreuzschmerzen. Manuelle Medizin 31:43–46
16. van Deursen LL, Patijn J, Durinck JR, Brouwer R, van Erven-Sommers JR, Vortman BJ (1999) Sitting and low back pain: the positive effect of rotary dynamic stimuli during prolonged sitting. Eur Spine J 8(3):187–193
17. Kilpikoski S, Airaksinen O, Kankaanpää M, Leminen P, Videman T, Alen M (2002) Interexaminer reliability of low back pain assessment using the McKenzie method. Spine 27(8):E207–E214

18. Laslett M, Young SB, Aprill CN, McDonald B (2003) Diagnosing painful sacroiliac joints: a validity study of a mcKenzie evaluation and sacroiliac provocation tests. Aust J Physiother 49(2):89–97
19. Donahue MS, Riddle DL, Sullivan MS (1996) Intertester reliability of a modified version of McKenzie's lateral shift assessments obtained on patients with low back pain. Phys Ther 76(7):706–716 (discussion717–discussion726)
20. Vleeming A, Van Wingerden JP, Dijkstra PF, Stoeckart R, Snijders CJ, Stijnen T (1992) Mobility in the sacroiliac joints in the elderly: a kinematic and radiological study. Clin Biomech 7(3):170–176
21. Kissling RO, Jacob HA (1996) The mobility of the sacroiliac joint in healthy subjects. Bull Hosp Jt Dis 54(3):158–164
22. Jacob HAC, Kissling RO (1995) The mobility of the sacroiliac joints in healthy volunteers between 20 and 50 years of age. Clin Biomech 10(7):352–361
23. Van Deursen L, Patijn J, Ockhuysen A (1990) The value of some clinical tests of the sacroiliac joint. J Man Med 5:96–99
24. Patijn J, Brouwer R, Lennep LV, Deursen LV (2000) The diagnostic value of sacroiliac tests in patients with non-specific low back pain. J Orthop Med 22(1):10–15
25. Degenhardt BF, Snider KT, Snider EJ, Johnson JC (2005) Interobserver reliability of osteopathic palpatory diagnostic tests of the lumbar spine: improvements from consensus training. J Am Osteopath Assoc 105(10):465–473
26. Donner A, Rotondi MA (2010) Sample size requirements for interval estimation of the kappa statistic for interobserver agreement studies with a binary outcome and multiple raters. Int J Biostat 6(1):Article31
27. de Vet HCW, Mokkink LB, Terwee CB, Hoekstra OS, Knol DL (2013) Clinicians are right not to like Cohen's κ. BMJ 346:f2125
28. Patijn J, Remvig L (2012) Reproducibility and validity: protocol formats for diagnostic procedures in manual musculoskeletal medicine, International Academy of Manual Musculoskeletal Medicine, available on website iammm.net
29. Birkimer JC, Brown JH (1979) Back to basics: percentage agreement measures are adequate, but there are easier ways. J Appl Behav Anal 12(4):535–543
30. Graham M, Milanowski A, Miller J (2012) Measuring and promoting inter-rater agreement of teacher and principal performance ratings. U.S. Department of Education, contract number ED-06-CO-0110, Center for Educator Compensation Reform (CECR), Westat
31. Hartmann DP (1977) Considerations in the choice of inter-observer reliability estimates. J Appl Behav Anal 10:103–116
32. Patijn J, Patijn J, Pragt E, Pragt E, Ruud B, Brouwer R (2005) Reproducibility studies in Manual/Musculoskeletal Medicine: a new method for kappa independence from prevalence. J Orthop Med 27(1):11–16
33. Juul-Kristensen B, Røgind H, Jensen DV, Remvig L (2007) Inter-examiner reproducibility of tests and criteria for generalized joint hypermobility and benign joint hypermobility syndrome. Rheumatology 46(12):1835–1841
34. Vind M, Bogh SB, Larsen CM, Knudsen HK, Sogaard K, Juul-Kristensen B (2011) Inter-examiner reproducibility of clinical tests and criteria used to identify subacromial impingement syndrome. BMJ Open 1(1):e42
35. Junge T, Jespersen E, Wedderkopp N, Juul-Kristensen B (2013) Inter-tester reproducibility and inter-method agreement of two variations of the Beighton test for determining Generalised Joint Hypermobility in primary school children. BMC Pediatr 13:214
36. Remvig L, Duhn PH, Ullman S et al (2009) Skin extensibility and consistency in patients with Ehlers-Danlos syndrome and benign joint hypermobility syndrome. Scand J Rheumatol 38(3):227–230
37. Remvig L, Duhn PH, Ullman S et al (2010) Skin signs in Ehlers-Danlos syndrome: clinical tests and para-clinical methods. Scand J Rheumatol 39(6):511–517

Wilfried-Lorenz-Versorgungsforschungspreis für Arbeit zur Koinzidenz von rheumatoider Arthritis und Diabetes

Das Deutsche Netzwerk Versorgungsforschung (DNVF) e.V. vergibt den Wilfried-Lorenz-Versorgungsforschungspreis für eine Untersuchung zur Koinzidenz von Diabetes und rheumatoider Arthritis. Die 20-köpfige Jury wählte im Gutachterverfahren die Arbeit der Preisträgerin Dr. Katinka Albrecht aufgrund des versorgungsrelevanten Themas und der guten Einbindung in die vorhandene Evidenz aus.

Versicherte mit rheumatoider Arthritis mit und ohne Diabetes mellitus wurden zu ihrer rheumatologischen Versorgung und Krankheitsbelastung befragt. 20 Prozent der 2.500 befragten Personen mit rheumatoider Arthritis hatten auch eine Diabetes-mellitus-Diagnose. Diabetes kam häufiger vor bei Männern, bei älteren Menschen, bei gleichzeitig bestehender Adipositas und bei Personen mit einem niedrigen Haushaltseinkommen. Von Diabetes und Arthritis Betroffene wurden seltener von Rheumatologen behandelt und seltener mit Antirheumatika versorgt als Personen mit Arthritis, die keine zusätzliche Diabetes-Diagnose hatten. Sie waren häufiger im Krankenhaus und hatten weitaus häufiger weitere kardiovaskuläre Begleiterkrankungen, Nierenerkrankungen sowie Depressionsdiagnosen.

Die Studie bestätigt, dass Diabetes eine häufige und relevante Begleiterkrankung der rheumatoiden Arthritis ist. Die Betroffenen haben eine hohe Wahrscheinlichkeit für weitere Begleiterkrankungen und benötigen eine gute fachärztliche Versorgung. Die Krankheitskontrolle der Arthritis mit spezifischer antirheumatischer Therapie ist bei Patienten mit gleichzeitig bestehendem Diabetes umso wichtiger, um weitere Folgeschäden zu vermeiden.

Diana Alchanow,
Deutsches Netzwerk
Versorgungsforschung (DNVF) e.V.