Measuring Clinical Uncertainty and Equipoise by Applying the Agreement Study Methodology to Patient Management Decisions

Robert FAHED
University of Ottawa Faculty of Medicine

Tim E. DARSAUT
Alberta Hospital Edmonton

Behzad FARZIN
Centre de recherche du CHUM

Miguel CHAGNON
Universite de Montreal

Jean Raymond (✉ **jean.raymond@umontreal.ca**)
Centre Hospitalier de L'Universite de Montreal https://orcid.org/0000-0003-1978-4274

Research article

Keywords: equipoise, randomized trials, methodology, agreement, reliability, kappa

Posted Date: February 3rd, 2020

DOI: https://doi.org/10.21203/rs.2.22528/v1

Version of Record: A version of this preprint was published on August 25th, 2020. See the published version at https://doi.org/10.1186/s12874-020-01095-8.

Abstract

Background: Clinical dilemmas in the treatment of patients translate into disagreements in decision-making. Such disagreements can reveal clinical uncertainty that should be addressed through care research. Our goal was to explore the use of reliability study methods to measure the degree of clinical uncertainty and equipoise regarding the use of rival management options prior to the conduct of randomized trials.

Methods: The study design resembles an inter-/intra-observer diagnostic reliability study. A portfolio of a sufficient number of diverse individual patients sharing a similar clinical problem and covering a wide spectrum of clinical presentations can be independently submitted to a variety of clinicians who manage that problem. Clinicians are asked to choose one of the predefined management options that are involved in the clinical dilemma. Intra-rater agreement can be assessed at a later time with a second evaluation.

Results: Descriptive statistics are presented, and results analyzed using kappa statistics. Interpretation of results can be facilitated by providing examples or by translating the results into clinically meaningful summary sentences. Reporting should follow standard guidelines.

Conclusions: Measuring the uncertainty regarding management options for clinical problems may reveal disagreements, provide an empirical foundation for the notion of equipoise, and inform or facilitate the design/conduct of clinical trials to address the clinical dilemma.

Background

Reproducibility is a fundamental scientific property required of any instrument, diagnostic test or prognostic score. [1] The same goes for therapy: outcome-based medical care can be trusted because it proposes interventions that have reproducibly been shown to provide better patient outcomes in a randomized trial.

Treatment decisions or management recommendations made by clinicians are authoritative judgments that have real-life impact on patients. Shouldn't we verify that physician management decisions are reliable? In simple terms, we are more likely to trust the doctor who, when asked the same question twice, provides the same answer both times.

It seems natural to think that recommendations made by doctors are inevitably variable and ungeneralizable, for working on a case-by-case basis and following a complex and ineffable reasoning process, they take into account the unique histories, characteristics and circumstances of the particular patient. [2] Upon reflection however, this in itself does not make clinicians' recommendations fundamentally different from other clinical judgments such as diagnosis which equally concerns unique individuals: the clinician's verdict (the output of the process) comes often down to allocating the patient one of a few categories, whether the judgment concerns diagnosis (disease present/absent) or management options (do not treat / treat medically / treat surgically). No matter the process underlying the clinical decision in question, if it leads to contradictory judgments or courses of action when the same patient is presented to the same or to different clinicians more than once, then the process is unreliable. The similarities are such that the reliability of clinical recommendations can be assessed the same way the reliability of a diagnostic imaging test is studied (Fig. 1).

Just as variability in making diagnoses should be studied, rather than minimized or eliminated through consensus sessions, [3] disagreements regarding clinical decisions need not be resolved through multidisciplinary meetings, Delphi processes [4] or practice guidelines unsupported by evidence.[5] Variability and inconsistency in clinical decision making can be informative: the uncertainty can reveal gaps in medical knowledge or identify suboptimal practices that could be improved. Because neurovascular interventions are image-guided and crucially depend on particular anatomical features of the patient's lesion, we very naturally came to apply the methodology typically used in studying the reliability of imaging diagnoses to clinical decision making.[6–8] Measuring the uncertainty in making clinical recommendations can be a preliminary step to the design or conduct of trials to address the uncovered uncertainty. [9–13] The goal of this paper is to expose how the methodology of reliability/agreement studies can be adapted to assess the repeatability of clinical management decisions and measure clinical uncertainty and equipoise.

Methods

The methodology we propose is inspired from the Guidelines for Reporting Reliability and Agreement Studies (GRRAS) originally designed for diagnosis/score/measurements. [1] The method presupposes that a dilemma concerning the use of at least 2 different options in the management of some patients affected with a similar clinical problem has been identified.

Two preliminary remarks are in order: first, it is important to mark the difference with a survey of opinions of preferred treatments: More than surveying whether clinicians agree in principle or in theory, a reliability study is an empirical investigation that tests the reproducibility of judgments made in practice for a series of particular cases. [4, 14]

Second, real clinical decisions are made once and then acted upon, while studies which assess reproducibility require the independent repetition of the same question concerning the same patient (or the same sort of patients) more than once. Although the study involves real clinicians and real patients, the context of the study is artificial, for decisions do not affect patient care. The experimental set up can be made to somewhat resemble clinical practice, however this may not always be possible, or even desirable (as we will see with the problem of prevalence below).

Investigators are interested in assessing the repeatability of clinical management decisions within and between clinicians on particular patients (Fig. 1). Thus there are 3 components (3 dimensions) to each decision: Each decision D is one of Y choices which are made by one clinician X on one patient Z. Each component (Y, X, or Z) is determined by the experimental design (detailed in the methodology section of the report): 1) Decisions are one of a pre-specified number of categories (spectrum of 'management choices', which corresponds to the diagnostic categories of diagnostic studies); 2) Decisions concern an individual belonging to a heterogeneous collection of particular patients affected by the same problem or disease under investigation (spectrum of patients); and 3) Decisions are repeatedly made by a single (intra-rater) or multiple (inter-rater) clinicians of various backgrounds, practices and expertise (spectrum of raters or clinicians). The study team chooses the management categories (the subject of the clinical dilemma) that will be offered as choices and they assemble a collection of patients and of clinician responders. While each decision, clinician and patient are unique, the study must compare decisions to evaluate and summarize their repeatability when they concern the same individual. The agreement study involves preparing a portfolio of patients that is then independently submitted to several physicians. The severity of the reliability test, the subsequent interpretation and the future generalizability of the results all depend on the number and variety of individuals included in the experiment.

Spectrum of Patients

What kind of patients should be included in the study? The classic method to select patients that has the theoretical advantage of allowing statistical inference from the selected individuals to a population is to proceed with random sampling from that population. However, such populations are rarely available in reality. Furthermore, for pragmatic reasons, the number of patients to be studied must be limited, and a small 'representative' sample may not include the types and proportions of patients that are necessary to properly test reliability (of diagnoses or of management decisions). Attempts by the study team to duplicate the frequencies naturally found in medical practice in their constructed portfolio can create serious imbalance in the answers obtained. As with diagnostic tests, the statistical indices that will be used to summarize results are sensitive to prevalence (or frequency of decisions). [15, 16] If the object of the diagnostic reliability study is a rare disease, for example, the portfolio cannot include the proportion of patients naturally affected (say 1/1000); the same goes for management categories (such as invasive surgery). Finally, we must remember that here, we are not interested in capturing an index which estimates the distribution of a disease or characteristic in a population, nor in finding out which management option is most frequently used by a population of doctors, but the goal of the study is to rigorously test whether the clinical judgments made are repeatable, one patient at a time, no matter the circumstances, clinicians or patients. Thus, while the portfolio must include a diversity of patients, and it may be constructed to resemble a clinical series, it does not have to be 'representative' of a theoretical or specific population. The challenge is more akin to testing an experimental apparatus in a laboratory with specimens of a known composition (positive and negative controls), prior to using the apparatus to explore specimens of unknown composition. Just as the reliability of a balance is not rigorously tested by weighing the same object 10 times, or by weighing objects of very similar weights, but by testing it with a wide range of weights, the reliability of clinical judgments must be tested with a diversity of particular patients, ideally covering a wide range of possible clinical encounters, along various spectra (age, size, location, duration of symptoms etc..), whether they concern diagnostic verdicts or therapeutic decisions. In practice, the portfolio will typically be artificially constructed to include prototypical patients selected by members of the study team (who are familiar with the clinical dilemma) to be 'positive' and 'negative controls’ for the various diagnostic or decision categories, to make sure they will be represented in the final decisions, as well as a substantial proportion of less typical or ‘grey zone’ cases.

The amount of information which should be provided for each patient included in the portfolio is a difficult question. To minimize the chance that clinicians might disagree based simply on different interpretation of the information provided, we believe it should be limited to the essential, for the purpose of the study is not to identify all potential reasons to disagree on a particular patient, but to measure the clinical uncertainty that remains even when extraneous reasons for potential disagreement are minimized.

While each patient included in the study is a concrete particular, sometimes uniquely identified by their radiograph or angiogram, for example,[6–8, 17] the patient can always be grouped (at the time of clinical decisions or at the time of analyses,) with other patients in a number of conceptual generalizations (or subgroups) that, according to some background knowledge pertinent to the clinical dilemma being studied, can influence clinical decisions. Investigators may be interested in exploring which patient or disease characteristic is associated with which decision. Patient or disease characteristics that will be included in each particular clinical vignette of the portfolio are generalizations (sometimes each with its own spectrum that may influence decisions. These may or may not be ‘reasons for decisions’ or ‘reasons for actions’, and they may be weighted differently by different clinicians. Investigators interested in exploring such details should ensure they include a sufficient number of particular patients with and without the characteristics of interest in the portfolio.

Like the baseline characteristics included in the registration form of a clinical trial, the information must be made available for each patient and expressed in a standardized fashion. These baseline characteristics are summarized in a descriptive Table of patients included in the study (Table 1).

Table 1
Reporting patient management agreement studies (inspired from GRRAS)

TITLE AND ABSTRACT	**1. Identify in title and/or abstract the clinical dilemma for which uncertainty and intra-inter physician agreement was investigated**
INTRODUCTION	2. Name and describe the subject of interest explicitly: what disease(s), what available management options, what clinical dilemmas are being considered
	3. Specify the patients that are confronted with uncertainty
	4. Specify the clinicians involved in making clinical decisions or recommendations
	5. Describe what is already known about reliability/agreement and provide a rationale for the study.
METHODS	6. Explain how the number of patients and clinicians was chosen.
	7. Describe how patients and clinicians were selected.
	8. Describe the experimental setting (e.g time interval between sessions, availability of clinical information, blinding…)
	9. State whether judgments were made independently.
	10. Describe the statistical analyses
RESULTS	11. State the actual number of raters and subjects that were included, and the number of replicated judgments which were collected.
	12. Describe the characteristics of clinicians (training, experience) and patients (any clinical data judged relevant to the study question).
	13. Report estimates of reliability and agreement including measures of statistical uncertainty.
DISCUSSION	Discuss the practical relevance of results.
AUXILIARY MATERIAL	Provide detailed results if possible (e.g online).

The source of patients included in the study should be mentioned in the study report. Patients may be selected from the data base of a registry or of a clinical trial. In such cases, the selection criteria of the trial should be mentioned. The exact selection of cases will of course impact results; the series of cases can be provided in extensio at the time of publication.

Illustrative example: Mechanical thrombectomy for acute stroke

Mechanical thrombectomy has revolutionized the management of acute stroke from large vessel occlusion, but uncertainty remains regarding indications. Several randomized trials have demonstrated the benefits of thrombectomy, but they used various eligibility criteria, such as age cutoffs, time of presentation, thrombus location, or extent of cerebral infarction. To study management decisions made by clinicians, 41 patients with acute ischemic stroke were assembled in a portfolio, which included approximately 1/3 of “positive control” patients (“typical” thrombectomy indications according to published guidelines), 1/3 of “negative controls” (thrombectomy definitely not indicated, for there was no large-vessel occlusion) and 1/3 of “grey zone” cases (patients not sufficiently studied in past trials, such as patients > 80, with minor symptoms (NIHSS score < 6) or with a large infarct). The information provided for each patient was limited to that which is routinely transmitted for making a decision: age, gender, symptoms and their severity according to the National Institute of Health Stroke Scale (NIHSS, a scale ranging from 0 to 42 that summarizes the extent of neurological deficits in stroke patients), time of symptom onset and key images of the brain at risk of infarction (Fig. 2).

Spectrum of Clinicians

The study of the reliability of clinical decisions should include numerous clinicians of various backgrounds and experiences, from all specialties involved in the various management options pertinent to the dilemma under study, as each specialty shares a body of knowledge and beliefs (and frequently a preference for the treatment it usually performs). What renders a scale or a treatment recommendation reliable, is that judgments are repeatable even when made by clinicians of various backgrounds and experience in diverse patients. The questionnaire will collect some baseline information on participating clinicians, and the characteristics of the clinicians involved in the study can be detailed in a table. Results can also be analyzed separately for some subgroups of clinicians (for each specialty, or for experienced or ‘senior’ clinicians). Of course, clinicians from various specialties may have diverging opinions, but even colleagues with the same background working in the same center and exposed to similar experiences may not make the same treatment recommendation for the same patient.[17] Again, this is no different from an evaluation of a diagnostic test. The goal of the study is not to find which treatment is most popular in some population of specialists, nor to try to identify ‘the right treatment’ by polling opinions. Thus it is not necessary for clinicians to be a representative sample of one specialty or another (although they may be). In the thrombectomy example, the study included 60 vascular neurologists who routinely managed acute ischemic stroke patients and 20 interventional neuroradiologists (who perform thrombectomy) working in French stroke centers. There were 49 junior (0–10 years) and 37 senior clinicians (> 10 years of experience).[18] Participants responding to the survey are asked to seriously consider each case as if it were a momentous clinical decision, but respondents should be reassured they will not be judged; they should not be afraid of being “wrong”, because unlike an accuracy study, there is no gold standard with which to evaluate performance.

The problem is more delicate with intra-rater studies. These may be very informative, but they are rarely performed. [18–20] Better agreement can be expected when the same clinician responds twice to the same series of cases (typically weeks apart in patients presented in a different order to assure independence between judgments), but the risk here is that the clinician may reveal their own inconsistencies in decision-making. In the case of diagnostic tests, poor intra-rater agreement (across multiple raters) is evidence of the lack of reliability of the score/measurement/ diagnostic categories, and a strong indication that the scale or categories should be modified.[20] We see no reason to conclude differently with management decisions: when asked the same question twice, a clinician’s inconsistencies in recommending opposing options to the same patient only reasserts a high degree of uncertainty regarding the clinical dilemma being examined. Participating in such intra-rater studies can be a humbling experience, but one that can convince the participant that a clinical trial may be in order.

Management categories

For each case, clinicians are independently asked which predefined option they would recommend or carry out. Choices are readily made when the questionnaire is conceived at the time of the design of a randomized controlled trial (RCT): the options are the 2 treatments being compared. Particular attention should be paid to the wording of questions, as ambiguities can affect the reliability of responses. Of course, agreement will be less frequent when the number of possible options is increased: it is more difficult to agree on the use of various treatments (“would you use A, B, C, or conservative management?”), than agreeing on: “would you treat this patient with A? [Yes/No].” Categorical responses can sometimes be dichotomized at the time of analyses. [8, 17]

Results will of course depend on the way the questions are formulated, and the best way to conceive the questionnaire will depend on the particular object of the study.

In the thrombectomy example the question was: “Do you perform/refer the patient for thrombectomy?” (Yes/No). There was no other option (no third answer such as “I don’t know”). The questionnaire should be given a test run with a few ‘test patients’ on a few ‘test clinicians’ before proceeding with the real study, as the wording of the questions included may need to be modified when problems with the first iterations are encountered.

Additional questions

The investigators may ask, for each decision, the level of confidence of participants.[6, 8, 17] If the questionnaire is prepared as a preliminary step in the design of a RCT, participants can also be asked the direct question: would you propose, to this patient, participation in a trial that randomly allocates treatments A and B. [7, 8, 17, 18] In the aforementioned study, participants were asked: Would you propose a trial comparing standard medical therapy with or without thrombectomy for this patient?

Statistical power and analyses

The number of cases and clinicians necessary to judge reliability with sufficient rigor and power depends on several parameters. [21] They must be predefined and justified in the study protocol. The number of patients to be studied is typically limited for pragmatic reasons. The larger the number of cases to be studied, the smaller the number of clinicians will be willing to participate. We have found that for simple questions with a binary outcome, as a rule of thumb a minimal number of ten raters reviewing 30–50 patients is necessary for the study to be informative. [6–8, 20]

There are many statistical approaches to measure reliability and agreement, depending on the type of data (categorical, ordinal, continuous), the sampling method and on the treatment of errors. [1] Reliability in treatment recommendations (categories) is most frequently analyzed using kappa-like statistics. There are several types of kappa statistics, and a discussion of the appropriate use of one or the other is beyond the scope of this article. A statistician should be involved in the design of the study early on.

Analyses can be repeated for various subgroups of patients or clinicians. For example, in case of an agreement study involving physicians from various specialties, it can be useful to study the degree of agreement within each specialty, to show that disagreements are not explained by various training or backgrounds. [8, 18] In the example of thrombectomy decisions, (dis)agreements were independent from physicians' background (Fig. 3).

Similarly, if it is known that some patient characteristic is commonly used to select one option rather than the other, agreement for patients sharing that characteristic can be analyzed. It should be noted that subgroup analyses reduce the number of observations; there may not be a sufficient variety and number of cases to adequately assess the reliability of decisions regarding that particular characteristic; confidence intervals are irremediably wider and results should be interpreted with caution.

Results

The report should be transparent and follow standardized guidelines (Table 1). [1]

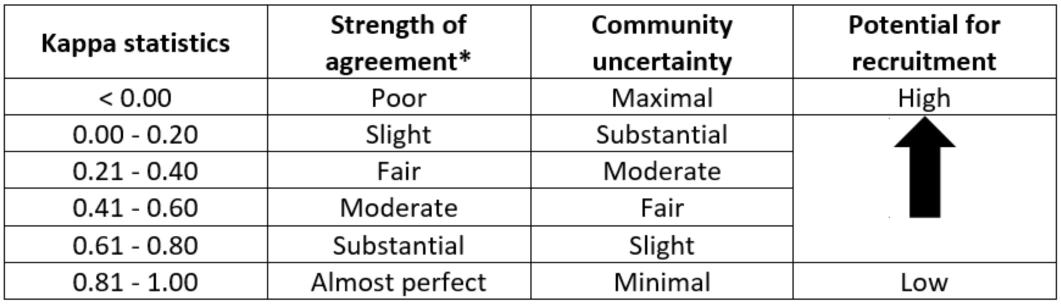
The results section normally includes descriptive statistics regarding the total number of management decisions, summarized in tables or figures. Comparing decision categories made by various subgroups of clinicians on various subgroups of patients may sometimes be of interest. Returning to our example, the 86 physicians opted for thrombectomy in 61.2% of cases (56.9% of cases among the 60 neurologists, and 71.1% among the 26 interventional neuroradiologists). Thrombectomy decisions varied among neurologists (between 29.3% and 87.8%) and among interventional neuroradiologists (between 36.6% and 97.6% of decisions). (Fig. 3A). Even when the clinicians' individual rates of thrombectomy decisions were similar, clinicians did not necessarily agree on which particular patients should be treated. For example, two physicians opting for thrombectomy in 80% of patients can still disagree on 0% (i.e they agree on every cases)) to 40% of cases (i.e none of the 20% of patients with a decision against thrombectomy are the same between the two physicians). It is possible to show the distribution of decisions for each case: in the present example, the bar graph (Fig. 3B) shows that if a majority of raters (> 90%) agreed on thrombectomy in some cases (top and bottom parts of the histogram), there was wide disagreement in many others. Graphs such as A and B do not permit a clear appreciation of the overall extent of agreement: the first shows a wide range of thrombectomy decisions, but does not show to what extent physicians disagreed on particular patients. Panel (B) shows how many clinicians 'voted' for thrombectomy for each patient, but a summary of overall agreement or reliability cannot be shown with this form of presentation.

The most important results concerning the repeatability of management decisions are typically expressed using indices (such as kappa values) summarized in tables or figures to allow a rapid appreciation of the overall results and simplify comparisons between subgroups. Unfortunately, such indices often have little meaning to clinicians. While a scale of interpretation can be provided (such as Landis and Koch [22]), interpretation can be facilitated by translating results into clinically meaningful sentences: Calculations of kappa values (Fig. 3C) showed that interrater agreement for thrombectomy was below the “substantial” level (defined as 0.6 for Landis and Koch). [22] At least 1/3 of physicians disagreed on thrombectomy decisions in more than 1/3 of cases (15/41 cases). Providing particular examples at both extremes of the spectrum (cases with near-perfect agreement and cases with maximal disagreement, when they occur) may also help illustrate the results of the study.

Discussion

The studies we propose are designed to transparently identify and measure the clinical uncertainty (a concept we prefer to ‘equipoise’) involved in the management of specific clinical problems, not to provide a ‘truth’ based on expert opinions. The identification and estimation of such uncertainties can serve many purposes: first, we believe the clinical community, clinicians and patients alike, should be aware that diverse options are actually being proposed for the management of similar patients, if only to make alternative options available. Second, clinicians may be reassured to realize that the uncertainty is not due to some personal failure, they are not alone in being uncertain. This step may encourage members of the community to get organized and prepare for the work that needs to be done: to accept the uncertainty revealed by the study, and proceed with the clinical research that addresses that uncertainty. A way to care for individual patients in their best medical interest in the meantime is still needed. This can be done within the context of a care trial, as detailed elsewhere.[23] Third, such studies can be a prelude to the design or conduct of such care trials. [11–13] When a reliability study is designed with a trial in view, it can provide empirical evidence of Freedman’s notion of ‘clinical equipoise’, when ‘an honest professional disagreement among expert clinicians about the preferred treatment’ exists, [24] a result that may reassure clinicians, patients and ethics committees. While in our view no such equipoise condition is necessary when evidence regarding what to do is lacking, randomized allocation to 2 different options may become impossible if all clinicians agree on one option for a particular group of patients (an unlikely event). The range of kappa values could be interpreted within a scale of uncertainty that could indicate RCT feasibility (likelihood of recruitment), as suggested in Table 2. Of course, the predictive value of any such scale on the recruitment actually achieved would need to be empirically verified.

Table 2 : Index of community uncertainty and potential for trial recruitment



Kappa statistics	Strength of agreement*	Community uncertainty	Potential for recruitment
< 0.00	Poor	Maximal	High
0.00 - 0.20	Slight	Substantial	⬆
0.21 - 0.40	Fair	Moderate	
0.41 - 0.60	Moderate	Fair	
0.61 - 0.80	Substantial	Slight	
0.81 - 1.00	Almost perfect	Minimal	Low

RCTs remain poorly accepted by patients and the medical community alike. [25–28] One obstacle is the notion that by participating in a RCT, the clinician abdicates and the patient exchanges a personalized decision for randomized allocation, a method whose sole purpose is to decrease bias and provide generalizable knowledge. This idea ignores the benefits randomized allocation can play in balancing the risks of receiving an inferior treatment. [29, 30] If medical care should always be individualized, this does not mean that the doctor always knows what to do. But the notion of personalized care has encouraged patients to expect a definite answer. We believe patients confronted with such dilemmas, as established by low agreement in these types of studies, are better cared for within the context of a care trial. The fact that equally respected clinicians would have chosen the rival treatment option could reassure clinicians and patients that the treatment that is randomly allocated is a treatment the patient could have received had they sought the opinion of a different expert.

The methodology we propose can undoubtedly be improved with experience as other investigators explore the best way to adapt it to other areas of medical care.

Conclusion

Reliability studies of clinicians' recommendations can reveal and measure clinical uncertainty regarding the best treatment, increase open-mindedness regarding the possibility of alternative options, provide an empirical foundation for the notion of equipoise, and inform or facilitate the design/conduct of clinical trials to address the clinical dilemma. Such studies may show the necessity to change the way we practice, from unrepeatable, unverifiable decisions, to a more prudent and systematic approach that takes uncertainty into account. When no one really knows what to do, integrating research methods to clinical care may be in the best medical interest of individual patients. Collecting empirical evidence regarding variability in treatment recommendations may, in the future, become an important component of a science of clinical practice.

Abbreviations

GRRAS
Guidelines for Reporting Reliability and Agreement Studies

NIHSS
National Institute of Health Stroke Scale
RCT
Randomized controlled trial

Declarations

Ethics approval and consent to participate: NA

Consent for publication: NA

Availability of data and material: NA

Competing interests: None

Funding: None

Authors' contributions:

Robert FAHED: writing, figures

Tim DARSAUT: supervision, writing

Behzad FARZIN: writing

Miguel Chagnon: writing, statistical analyses

Jean RAYMOND: supervision, writing, final approval.

Acknowledgements: None

References

[1] Kottner J, Audige L, Brorson S, Donner A, Gajewski BJ, Hrobjartsson A, et al. Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. J Clin Epidemiol. 2011;64:96-106.

[2] Pellegrino ED. The Anatomy of Clinical Judgments. Clinical Judgment: A Critical Appraisal: Springer, Dordrecht; 1977. p. 169-94.

[3] Bankier AA, Levine D, Halpern EF, Kressel HY. Consensus interpretation in imaging research: is there a better way? Radiology. 2010;257:14-7.

[4] Etminan N, Brown RD, Jr., Beseoglu K, Juvela S, Raymond J, Morita A, et al. The unruptured intracranial aneurysm treatment score: a multidisciplinary consensus. Neurology. 2015;85:881-9.

[5] Cenzato M, Boccardi E, Beghi E, Vajkoczy P, Szikora I, Motti E, et al. European consensus conference on unruptured brain AVMs treatment (Supported by EANS, ESMINT, EGKS, and SINCH). Acta neurochirurgica. 2017;159:1059-64.

[6] Darsaut TE, Fahed R, Macdonald RL, Arthur AS, Kalani Y, Arikan F, et al. Surgical or endovascular management of ruptured intracranial aneurysms: an agreement study. J Neurosurg. 2018;Jul 13:1-7. doi: 10.3171/2018.1.JNS172645. [Epub ahead of print].

[7] Darsaut TE, Gentric JC, McDougall CM, Gevry G, Roy D, Weill A, et al. Uncertainty and agreement regarding the role of flow diversion in the management of difficult aneurysms. AJNR American journal of neuroradiology. 2015;36:930-6.

[8] Fahed R, Batista AL, Darsaut TE, Gentric JC, Ducroux C, Chaalala C, et al. The Treatment of Brain Arteriovenous Malformation Study (TOBAS): A preliminary inter- and intra-rater agreement study on patient management. Journal of neuroradiology. 2017;44:247-53.

[9] Khoury NN, Darsaut TE, Ghostine J, Deschaintre Y, Daneault N, Durocher A, et al. Endovascular thrombectomy and medical therapy versus medical therapy alone in acute stroke: A randomized care trial. Journal of neuroradiology. 2017;44:198-202.

[10] Darsaut TE, Findlay JM, Magro E, Kotowski M, Roy D, Weill A, et al. Surgical clipping or endovascular coiling for unruptured intracranial aneurysms: a pragmatic randomised trial. Journal of neurology, neurosurgery, and psychiatry. 2017;88:663-8.

[11] Darsaut TE, Jack AS, Kerr RS, Raymond J. International Subarachnoid Aneurysm Trial - ISAT part II: study protocol for a randomized controlled trial. Trials. 2013;14:156.

[12] Darsaut TE, Magro E, Gentric JC, Batista AL, Chaalala C, Roberge D, et al. Treatment of Brain AVMs (TOBAS): study protocol for a pragmatic randomized controlled trial. Trials. 2015;16:497.

[13] Raymond J, Gentric JC, Darsaut TE, Iancu D, Chagnon M, Weill A, et al. Flow diversion in the treatment of aneurysms: a randomized care trial and registry. Journal of neurosurgery. 2017;127:454-62.

[14] Cockroft KM, Chang KE, Lehman EB, Harbaugh RE. AVM Management Equipoise Survey: physician opinions regarding the management of brain arteriovenous malformations. J Neurointerv Surg. 2014;6:748-53.

[15] Feinstein AR, Cicchetti DV. High agreement but low kappa: I. The problems of two paradoxes. Journal of clinical epidemiology. 1990;43:543-9.

[16] Cicchetti DV, Feinstein AR. High agreement but low kappa: II. Resolving the paradoxes. Journal of clinical epidemiology. 1990;43:551-8.

[17] Darsaut TE, Estrade L, Jamali S, Bojanowski MW, Chagnon M, Raymond J. Uncertainty and agreement in the management of unruptured intracranial aneurysms. Journal of neurosurgery. 2014;120:618-23.

[18] Ducroux C, Fahed R, Khoury N, Gevry G, Kalsoum E, Labeyrie M, et al. Intravenous thrombolysis and thrombectomy decisions in acute ischemic stroke: an interrater and intrarater agreement study. Rev Neurol (Paris). 2018;In press.

[19] Farzin B, Gentric JC, Pham M, Tremblay-Paquet S, Brosseau L, Roy C, et al. Agreement studies in radiology research. Diagn Interv Imaging. 2017;98:227-33.

[20] Farzin B, Fahed R, Guilbert F, Poppe AY, Daneault N, Durocher AP, et al. Early CT changes in patients admitted for thrombectomy: Intrarater and interrater agreement. Neurology. 2016;87:249-56.

[21] Donner A, Rotondi MA. Sample size requirements for interval estimation of the kappa statistic for interobserver agreement studies with a binary outcome and multiple raters. Int J Biostat. 2010;6:Article 31.

[22] Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33:159-74.

[23] Raymond J, Darsaut TE, Altman DG. Pragmatic trials can be designed as optimal medical care: principles and methods of care trials. J Clin Epidemiol. 2014;67:1150-6.

[24] Freedman B. Equipoise and the ethics of clinical research. N Engl J Med. 1987;317:141-5.

[25] Fiorella D, Mocco J, Athur A, Siddiqui A, Heck D, Albuquerque F, et al. Randomized controlled trials for everything? J Neurointerv Surg. 2015;7:861-3.

[26] Mansouri A, Cooper B, Shin SM, Kondziolka D. Randomized controlled trials and neurosurgery: the ideal fit or should alternative methodologies be considered? J Neurosurg. 2016;124:558-68.

[27] Kelley M, James C, Alessi Kraft S, Korngiebel D, Wijangco I, Rosenthal E, et al. Patient Perspectives on the Learning Health System: The Importance of Trust and Shared Decision Making. The American journal of bioethics : AJOB. 2015;15:4-17.

[28] Robinson EJ, Kerr CE, Stevens AJ, Lilford RJ, Braunholtz DA, Edwards SJ, et al. Lay public's understanding of equipoise and randomisation in randomised controlled trials. Health technology assessment (Winchester, England). 2005;9:1-192, iii-iv.

[29] Raymond J, Fahed R, Darsaut TE. Randomize the first patient. Journal of neuroradiology. 2017;44:291-4.

[30] Fahed R, Darsaut TE, Raymond J. The introduction of innovations in neurovascular care: Patient selection and randomized allocation. World Neurosurg. 2018. Oct;118:e99-e104. doi: 10.1016/j.wneu.2018.06.127. Epub 2018 Jun 23.

Figures

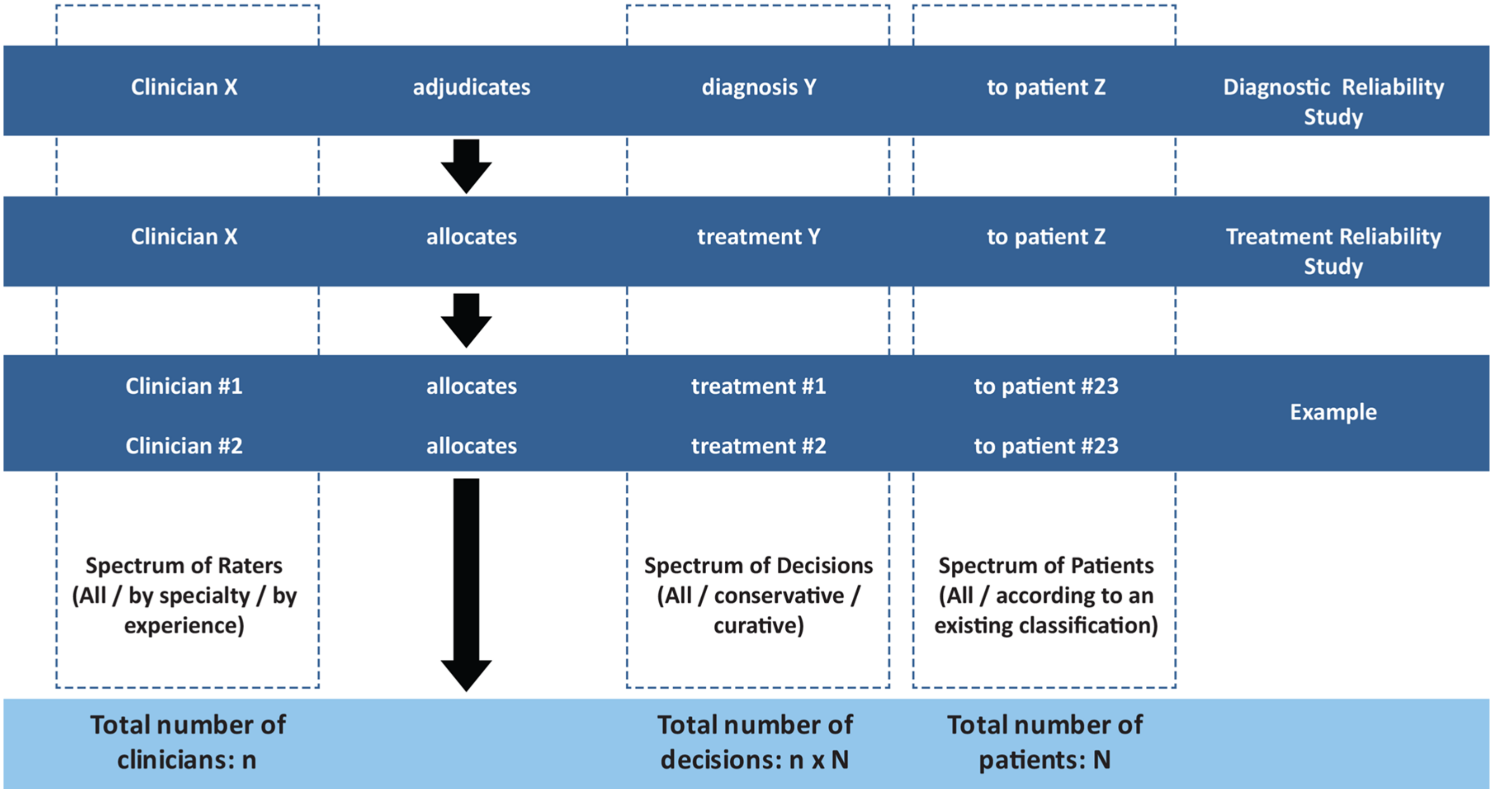
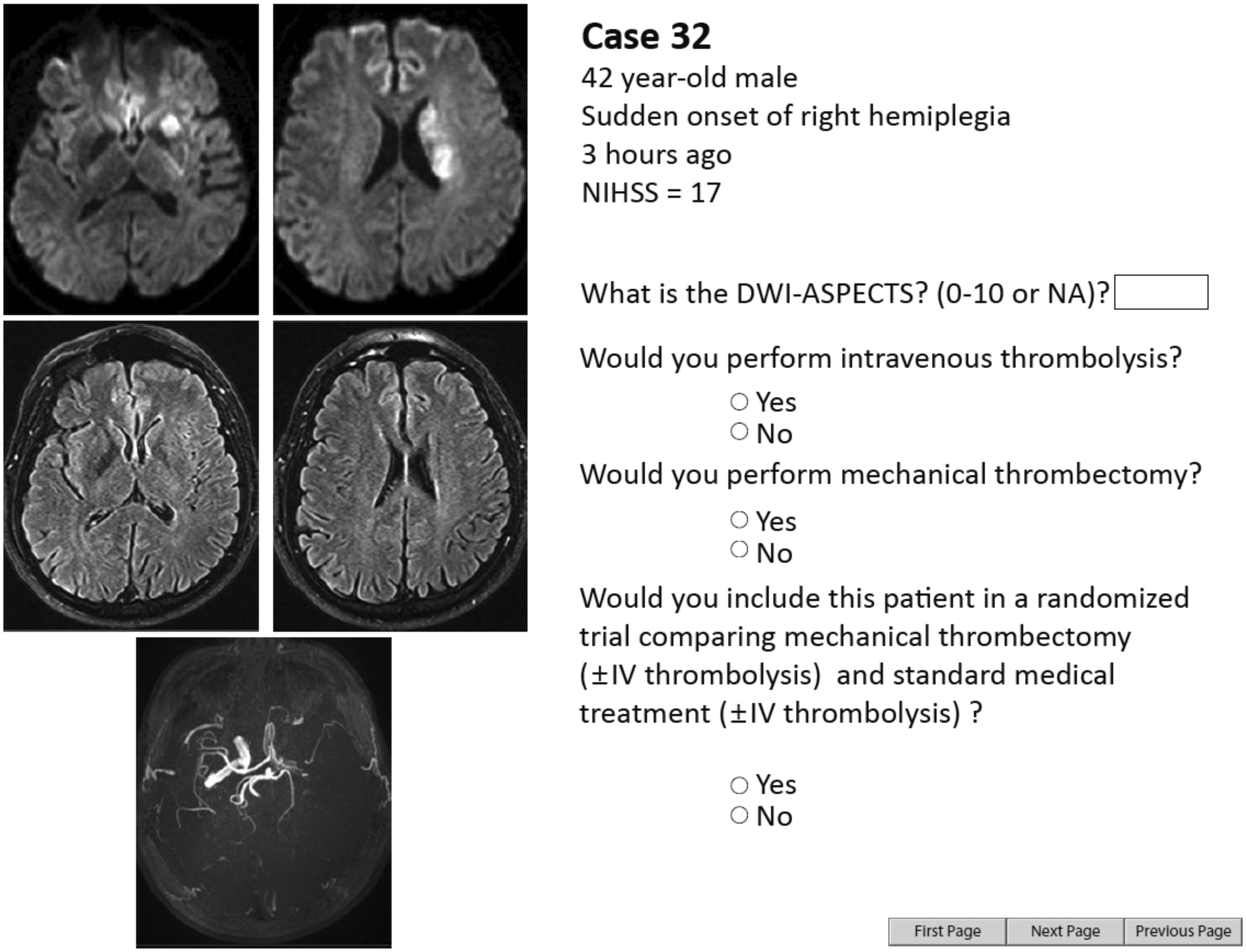

Clinician X
adjudicates
diagnosis Y
to patient Z
Diagnostic Reliability Study
Clinician X
allocates
treatment Y
to patient Z
Treatment Reliability Study
Clinician #1
allocates
treatment #1
to patient #23
Clinician #2
allocates
treatment #2
to patient #23
Example
Spectrum of Raters (All / by specialty / by experience)
Spectrum of Decisions (All / conservative / curative)
Spectrum of Patients (All / according to an existing classification)
Total number of clinicians: n
Total number of decisions: n x N
Total number of patients: N

Figure 1

Reliability studies - Reliability studies of diagnostic tests assess the agreement among X clinicians for the diagnosis Y for each of the Z patients included in the study. The reliability studies of treatment decisions we propose use a similar methodology to study the agreement for management options. After asking X clinicians to choose one of the Y management options proposed for each of the Z patients, we can measure the agreement/uncertainty.



Case 32

42 year-old male
Sudden onset of right hemiplegia
3 hours ago
NIHSS = 17

What is the DWI-ASPECTS? (0-10 or NA)? □

Would you perform intravenous thrombolysis?

○ Yes
○ No

Would you perform mechanical thrombectomy?

○ Yes
○ No

Would you include this patient in a randomized trial comparing mechanical thrombectomy (±IV thrombolysis) and standard medical treatment (±IV thrombolysis) ?

○ Yes
○ No

Figure 2

The portfolio - Example from the electronic portfolio used for the thrombectomy agreement study. Each page displayed a clinical vignette with basic clinical information (age, gender, NIHSS score, etc…) and a few selected brain imaging slices. For each patient, raters were asked whether they would perform mechanical thrombectomy (yes/no). Other questions were also asked for further analyses on other parameters (agreement for intravenous thrombolysis, etc…).

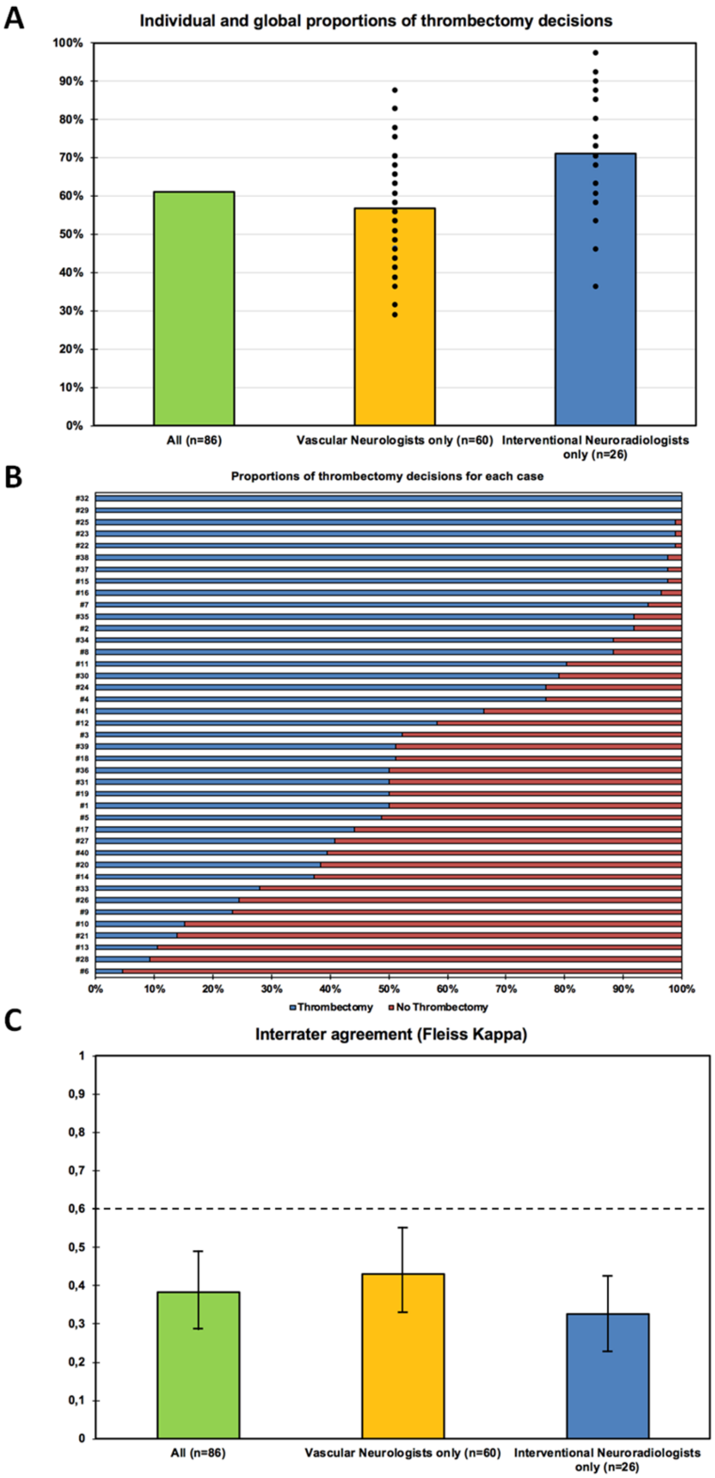


Figure 3

Thrombectomy decisions - Legend: Panel A shows the proportion (%) of decisions to perform thrombectomy (in percentages) for all raters and among each specialty. Black dots represent the individual results of each of the 86 clinicians. The bar graphs show similar proportions of decisions between neurologists and interventional neuroradiologists (INRs), but they hide individual discrepancies among physicians, shown here by black dots, revealing a wide range of decisions. Panel B shows, for each patient, the proportions (%) of thrombectomy decisions. This panel better illustrates the spectrum of results in various patients, as it shows that some cases had almost unanimous decisions for (complete/almost complete blue bar at the top) or against thrombectomy (complete/almost complete red bar at the bottom part). However, a significant proportion of cases (in the middle) reveal wide disagreements. None of these panels can give an overall idea of the degree of agreement in the study. Panel C shows the levels of agreements (through kappa values) in a bar graph. It shows that thrombectomy decision lacks reliability (i.e kappa value is below 0.6) for all raters and also within each subspecialty (vascular neurologists and interventional neuroradiologists).