

Research Methodology and Biostatistics Series II - Quality of Medical Research

Abhaya Indrayan^{1*}

¹Department of Clinical Research, Max Super Speciality Hospital, Saket, New Delhi

Correspondence:

Abhaya Indrayan

Email: Abhaya.Indrayan@maxhealthcare.com

Abstract:

Medical research has mushroomed, and concerns are expressed regarding the quality of such voluminous research output. Among several factors, the quality can be substantially improved by effective control of uncertainties that severely afflict the reliability and validity of all empirical research.

Uncertainty is an endemic affliction of human activity, and medical research is particularly susceptible. Two primary types of medical uncertainties are aleatory and epistemic. The former is inherent due to biological, environmental, and other natural variations, and the latter arises mostly from knowledge gaps.

Biostatistical methods are equipped to handle sampling fluctuations that are a source of a major part of the aleatory uncertainties. Tools such as scoring systems, aetiology diagrams, and expert systems can help to reduce some types of epistemic uncertainties. Proper choice of research tools remains crucial for controlling most epistemics.

Key words: Aleatory Uncertainties, Epistemic Uncertainties, Reliability, Validity

Introduction

In the first article of this series in the previous issue, the selection of the problem for research was discussed. After selecting a topic, it is important to plan and conduct a study in a manner that high quality results are obtained. Quality in this case implies believable results that could be confidently implemented to improve health.

When research endeavours do not succeed in reaching a beneficial result, they leave a lesson for us. This does not render the research useless since research is for the unknown and unexpected results are expected. But there are several other “research” that do not contribute anything to our knowledge. This happens because the research questions are hazy, the methodology is sloppy, the results are vague, or the reporting is unclear. Perhaps “such space-occupying lesions” have proliferated more in medical research quagmire than in any

other discipline. Quality has taken a back seat amidst the rush for quantity.

The basic ingredient of quality research is that it should be well-intentioned and should be performed with the care and sincerity it deserves by using an appropriate methodology, and honestly reported.

The most important aspect of research that ensures quality is not the result but the methodology. When a relevant question is examined with an appropriate methodology, high quality research is ensured, irrespective of the result. The methodology includes choosing the appropriate setting (clinic, laboratory, and community) that can elicit the correct answer to the research question, studying the right type of subjects or patients, selecting an unbiased sample of the subjects, and conducting the study on sample size required for adequate reliability of the estimates or power to detect a medically important effect. This also

includes a suitable design for the selection/allocation of the patients keeping in view the possible confounding factors, using the data collection tools with established validity and reliability, ensuring that the data obtained are as correct and complete as possible, and using the appropriate method of statistical analysis considering the type of data and the research objectives. All these aspects will be discussed in detail in the subsequent articles in this series.

Besides explaining the meaning of the quality of research, this article gives details of various kinds of uncertainties and their sources that afflict quality. The tools for controlling and assessing the magnitude of the uncertainties, particularly the reliability and validity, are also discussed. These two concepts are often confused. For further details of planning, conducting, and reporting good quality research, see Indrayan et al.¹

Uncertainties in medical research

Medical uncertainties can be considered as the greatest challenge for quality research. Uncertainty is an endemic affliction to all human activities and science is no exception. The most apparent reason for enormous uncertainties in medical research results is the profound variation in this setup, called the aleatory uncertainties. The other is the limitation of knowledge that hampers medical research despite immense advances in recent times. These are called epistemic uncertainties. The sources of these two types of uncertainties can be identified as follows.

Aleatory uncertainties (Inherent variations)

- Biological — nonmodifiable (age, gender, heredity or genetic make-up, birth order, height, etc.)
- Biological — modifiable (anthropological, physiological, biochemical)
- Socio-economic (income, education, and occupation) that can affect personal hygiene, nutrition, and self-esteem.
- Cultural, behavioural, and psychological (mental status, family system, faith in prayers, sexual practices, addictions, personality traits, tension-anxiety-stress, etc.)
- Observers, instruments, and laboratories—inherent variation in measurements
- Environmental (climate, dust, mosquitoes, flies, pollution, sanitation, water supply, infection load, quality and quantity of health facilities, family and societal support, communication, traffic, laws and their enforcement, etc.)
- Multifactorial (lifestyle, hygiene, knowledge-attitude-practices, susceptibility, utilisation of health services, etc.; importantly, sampling fluctuations)

Epistemic uncertainties (External)

- Universal ignorance about appropriate treatment, cause-effect, etc., for certain ailments, or lack of consensus among experts (e.g., restoring full health of a leukemic patient)

- Nonavailability of data and inadequate knowledge (risk factors for Alzheimer's disease)
- Individual (patient and physician) and societal biases, including biased samples, and suppression of facts (sexual offences, knife injury)
- Not being able to consider all the factors because they are far too many, or because they are not correctly stipulated (factors for longevity)
- Chance that can not be explained, other than inherent variation (differential health of identical twins brought up in the same environment)
- Non availability of appropriate instrument or facility for any particular measurement because it is too expensive or for some other reason, and thus inability to obtain the required information (non-availability of CT scan in a peripheral hospital for neurological problems)
- Incompetency, memory lapse, biasedness, carelessness, lack of validity of observers, instruments, and laboratories, etc. (recall bias, incomplete history, not getting proper answers from the patients, uncalibrated instruments, poor laboratory)
- Inadequate design, wrong analysis, or sloppy interpretation of the results
- Noncompliance of the regimen or nonresponse by the patients

The uncertainty around a result is much more than what is made out by the conventional statistical confidence interval (CI), since CI is only for fluctuations from sample to sample. Consideration of other aleatory uncertainties may provide an enormously large uncertainty interval when variations in the response and instruments are considered. The epistemic uncertainties put a further question mark on the validity of this interval. Many such uncertainties go unnoticed and uncared for, leading to low-quality results in some cases. Let us examine how to handle these uncertainties.

Controlling uncertainties

Aleatory uncertainties are controlled by using appropriate statistical methods for designing, choosing a sample of an adequate size that represents the target population, collecting the right data and its proper collation, and their appropriate analysis. These will be discussed in subsequent articles in this series.

Epistemic uncertainties are relatively difficult to handle. However contradictory it may sound, statistical tools are available that can help in reducing epistemic uncertainties in certain situations. Primary among them are clinimetrics, aetiology diagrams, and artificial intelligence-based expert systems. For other situations, lateral thinking may be required so that new explanations can be conjectured and tested.

Clinimetrics includes two related tools – scoring systems and mathematical models. Scoring systems reduce a multivariate

entity to a univariate quantity—thus increasing the comprehension and utility. They combine demographic and clinical features, laboratory investigations, and radiological findings into one score that makes it easy to decide about the most probable diagnosis (e.g., for acute appendicitis),² gradation of severity (e.g., APACHE score),³ or likely prognosis (e.g., for Hodgkins lymphoma).⁴ Mathematical models also have similar features. Aetiology diagrams, such as the one proposed for myocardial infarction,⁵ help to not miss any aspect of the disease at the time of evaluation of patients. Artificial intelligence-based expert systems, such as for gynaecological diagnostics,⁶ are fast coming up that claim to fill up knowledge gaps. Remember, though, that these tools are only aids, the ultimate decision remains with the clinicians.

Assessing uncertainties

A solar eclipse can be predicted centuries in advance. This is a deterministic phenomenon. Medicine is not so fortunate. Never enough is known about biological systems to predict with that accuracy. Reasoning is the tool of choice in an uncertain environment. Going from qualitative notions such as the possibility of the presence or absence of disease to quantitative notions such as 0.7 probability of disease, statistics is extremely useful in considering the in-between notions of plausibility. Probability is the tool of choice to assess the uncertainties.

The concepts of reliability and validity

Reliability and validity delineate uncertainties in a variety of settings. Reliability is reproducibility in identical situations. The other term for this is precision. Validity is the ability to correctly measure the phenomenon that is intended to be measured.

The difference between validity and reliability is illustrated in Figure 1. When the blood pressure (BP) of an individual is measured as 132/72 mmHg, how confident are you that this really is the level? Variations can occur from a variety of sources mentioned earlier. In the case of BP, external sources such as nonstandardized instruments, cuff size, patient not being fully relaxed, and the white-coat effect, can cause uncertainties. In case of the sphygmomanometer, not being careful in the gradual deflation of the cuff, in making the reading at the right moment, and in missing Korotkoff sounds are among the aspects that would vary from observer to observer.

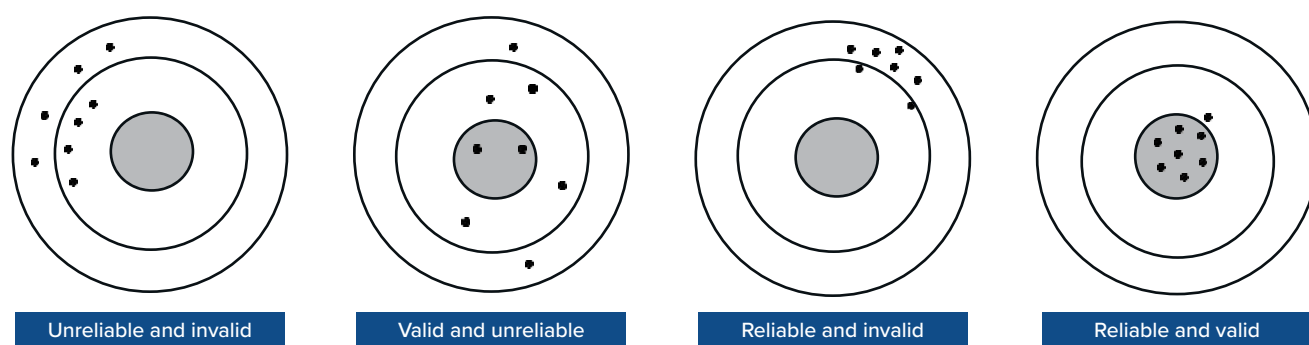


Figure 1: Dart game illustrating validity and reliability

Reliability of research findings

Other things being equal, a study based on 400 patients gives more reliable results than the one based on 60 patients. As will be explained in a later article, the CI becomes narrow, and the statistical test of a hypothesis procedure attains more power to detect a given difference when the sample size is large. Thus, a large sample is insurance for the reliability of the results.

A measurement is considered reliable if it can be reproduced under identical conditions. The variation would always be there, but it must be minimal. Body temperature is considered reliable because it depends only on the proper application of a clinical thermometer. The reading is obtained almost without human intervention. It is univariate in this sense. As opposed to this, the measurement of the T4 level is multivariate as it depends on several interactive factors such as proper collection of the sample, appropriate analytical methods, and correct reporting. Thus, it is less reliable.

The other question is regarding the reliability of the instruments used in the research. The use of the same instrument for measuring the same level should give the same answer. For example, no BP instrument so far available is fully reliable as it can give different readings in identical conditions.

For instruments such as a questionnaire, the reliability is assessed by tools such as test-retest reliability, split half consistency, and Cronbach alpha.⁵

Validity of research

Epistemic uncertainties relating to the research tools such as a questionnaire, laboratory test, and diagnostic criterion are assessed in terms of their validity. Proper choice of research material is important to achieve valid results. For example, pain is measured in several ways, yet they fail to capture it fully. Would you prefer a visual analogue scale or a verbal rating scale to assess the intensity of pain? Two such devices may not give the same results—thus choice is important. One is more 'appropriate' than the other in a given situation. Thus, no measure of pain is fully valid. Such uncertainties remain a concern in many research situations.

The validity of measurements assumes slightly different meanings in different contexts. In the context of a single measurement, validity is its ability to be close to reality. This depends on the care adopted in measurement. For example, the validity of liver function test results depends on the quality of reagents, analysis method, care in recording, etc. With adequate care, the value recorded can be believed to be true.

Similar concerns can be expressed for the validity of a response. When a patient is asked about the duration of complaints, how correctly is s/he able to report? A person of age 73 years may not report vision problems considering that it is natural at his age. A smoker may not report about cough, and a VDRL-positive man may intentionally suppress his sexual history. Such instances adversely affect the validity and increase the uncertainty level in research.

There is some confusion in medical literature about the difference between validity and accuracy. The term accuracy is also used for the closeness of the measurement to the reality with a slight rider. A BP reading 132.3 is more accurate than 132 mmHg and a cholesterol level of 223mg/dl is more accurate than 220–229 mg/dl category when correctly measured. However, such high accuracy can be redundant because this kind of difference does not alter the management of the patients.

The validity of a device is obtained in terms of its predictive value for positive outcome and negative outcome, respectively. A gold standard is necessary against which this kind of validity is evaluated. Procedures to obtain such predictivities will be discussed in a later article. Sensitivity and specificity are also indicators of validity.

In a clinical trial setup, the initial equivalence of two groups of subjects, one of which is put on a drug and the other on a placebo, indicates internal validity. In this case, the difference arising in the outcome can be safely ascribed to the drug, and the treatment effect is correctly portrayed. Imbalances and biases, such as paying more attention to the cases than controls, are threats to validity. Even when no such hiccup is present, the conclusion could still be valid only for the groups of subjects actually studied and generalisation may be difficult.

Generalisability is achieved when the groups under study are adequate representative of the target population. Random sampling, which we will discuss later, as a strategy to achieve a representative sample, affords external validity. But this is just one aspect. The setting of the study, eligibility criteria for inclusion of subjects, the applicability of the test regimen to the patients available in normal clinical practice, suitability of outcome measure, the occurrence of side-effects, and balanced approach – all contribute to external validity. In short, judge that the results would apply to the next eligible patient you encounter.

Validity has various types. Face validity is the apparent correspondence between what is intended and what is actually obtained. For example, cleft lip, deafness, or dental caries lack face validity as a cause of death. Thus, face validity is achieved when the observations look just about right. An instrument or a method is called criterion-valid if it gives nearly the same information as a criterion with established validity. Whenever a new index, a score, or any other method is developed, it is customary that its criterion validity is established by comparing its performance with a standard. It is not uncommon in medical setups that no validated standard is available. For example, no valid measure is available to assess characteristics of balance in people with vestibular dysfunction. The Berg Balance Scale is often used but is not considered a fully valid measure. A relatively new tool is the Dynamic Gait Index. The two can be investigated for agreement. If the agreement is good, the two methods can be called concurrently valid. That is both are equally good (or, equally bad). The fourth type is content validity. This is based on the domain of the content of the measurement or the device. Fever by itself is not content-valid for infection. Would you consider kidney function tests such as urea clearance and diodrast clearance content valid for assessing the health of kidneys? Many such tests are restricted to specific aspects and do not provide a complete picture. Content validity is often established through qualitative expert reviews. The last is construct validity which seeks the agreement of a device with its theoretical concept. Body-mass index is construct-valid for overall obesity but not for central obesity, whereas waist-hip ratio is construct-valid for central obesity and not for overall obesity.

Summary

The quality of medical research can be improved by controlling aleatory and epistemic uncertainties, and by ensuring reliable and valid results. This would require appropriate methodology and careful choice of research tools at the time of planning and execution of the study.

Acknowledgements: None

Financial source: None

References

1. Indrayan A, Vishwakarma G, Malhotra RK, Gupta P, Sachdev HP, Karande S et al. The development of QERM scoring system for comprehensive assessment of the Quality of Empirical Research in Medicine-Part 1. *Journal of Postgraduate Medicine*. 2022 Oct 1;68(4):221-30.doi: 10.4103/jpgm.jpgm_460_22.
2. Kanumba ES, Mabula JB, Rambau P, Chalya PL. Modified Alvarado scoring system as a diagnostic tool for acute appendicitis at Bugando Medical Centre, Mwanza, Tanzania. *BMC surgery*. 2011 Dec;11:1-5.
3. Knaus WA, Wagner DP, Draper EA, Zimmerman JE, Bergner M, Bastos PGet al. The APACHE III prognostic system: risk prediction of hospital mortality for critically ill hospitalized adults. *Chest*. 1991 Dec 1;100(6):1619-36.
4. Merckmanuals. International Prognostic Score in Hodgkin Lymphoma. <https://www.merckmanuals.com/medical-calculators/IPS.htm>- Last accessed 28-03-24.
5. Indrayan A, Malhotra RK. *Medical Biostatistics*, 4th ed. CRC Press, 2018.
6. Tanos P, Yiangou I, Prokopiou G, Kakas A, Tanos V. Gynaecological Artificial Intelligence Diagnostics (GAID) GAID and Its Performance as a Tool for the Specialist Doctor. *InHealthcare* 2024 Jan 16 (Vol. 12, No. 2, p. 223). MDPI.