Pediatr Radiol (2015) 45:32–34
DOI 10.1007/s00247-014-3081-2

RESEARCH FORUM

Measurements in radiology: the need for high reproducibility

Giovanni Di Leo

Received: 26 March 2014 / Revised: 13 May 2014 / Accepted: 5 June 2014

"Measure what is measurable, and make measurable what is not so." It was the 17th century when Galileo Galilei, father of the scientific method, made this well-known statement. With this sentence, Galilei told us to avoid subjective, qualitative evaluations of the world around us. Rather, he advised to make our observations as objective as possible, independent from the observer. He first recognized that human beings greatly differ from one another and that, in the evaluation of a given entity, a spectrum of opinions is inevitable. In an era in which technical instruments were far to come, such a situation made the science of nature very unscientific.

In medicine, the uncertainty recognized by Galilei was further formalized by Alexander Pope, who in 1732 wondered "Who shall decide when doctors disagree?" This question must have been a very common one in Pope's day, because medical practice at that time was based largely on tradition and opinion, not science. Regarding the evaluation of clinical variables, Pope's question challenged reproducibility, in particular interobserver reproducibility [1]. It relates to the common experience in which two independent observers provide different results, with this disagreement implying a sort of uncertainty about the truth. From the patient's point of view, it may appear that his or her condition is not an objective one but rather a subjective one for which each clinician can have his or her own opinion. This can be very frustrating to the patient and cause the patient to lose trust in medicine.

In addition to the interobserver reproducibility there is intraobserver reproducibility, i.e. the ability of a single observer to provide the same opinion regarding a patient's condition if he or she is questioned again later. Indeed, self-disagreements occur more frequently than might be expected, in particular if the posed question has more than two mutually exclusive answers (categorical variables).

Intra- and interobserver reproducibility not only apply to categorical and ordinal variables but also, and more strictly, to quantitative (continuous) variables. Examples include cardiac ventricular volumes, vessel diameters, arterial blood pressure, body temperature, etc. From the observer's point of view, the numerical values of such variables are obtained by means of instruments, i.e. technical systems, based on a physical phenomenon, that are sensitive to the quantity being measured. In radiology, many of these instruments are now available as software algorithms implemented on computers.

Although the use of a technical instrument might lead an observer to believe the measurement to be an objective process without uncertainty, we must remember that this process does not proceed by itself and that it needs the observer's intervention. This intervention may apply at any level and certainly affects the final observed value. For example, the measurement of a vessel diameter based on an MR image requires the observer to place a ruler between two distinct points (the vessel boundaries). This step usually proceeds manually using computer software, and the repetition of this action, even if on the same image, rarely provides the same value as that previously obtained. Furthermore, an independent observer might perform this measurement by placing the ruler between different vessel boundaries or even at another level of the vessel course. Therefore, as for categorical variables, the measurement of continuous variables is also characterized by intra- and interobserver variability.

Measurement variability is intrinsic and unavoidable. For every given entity (variable) that we measure (estimate) and for every given combination of technique (CT, MRI, US, etc.) and observer (radiologist, technician, physicist, etc.), we are faced with the associated variability. Part of this variability is explained in terms of the observer intervention and interpretation and a minor part in terms of instrument variability. The

G. Di Leo (✉)
Radiology Unit, IRCCS Policlinico San Donato, Piazza E. Malan, 20097 San Donato Milanese, Italy
e-mail: gianni.dileo77@gmail.com

total variability (uncertainty) in turn reflects as a lack of reproducibility (precision) and it may have a relevant effect in clinical practice, although this effect is often ignored or, at best, neglected. In the next paragraph a working example is presented to show the impact of reproducibility on patient follow-up.

A case example

Let's consider a patient who undergoes an MR examination for staging a solid tumor. Using computer software, a radiologist manually segments the tumor boundaries slice by slice, obtaining a tumor volume of 40 ml. Let's now consider that this patient is treated with neoadjuvant chemotherapy and that after 3 months the MR examination is repeated at another institution. Finally, let's consider that the radiologist who performs the follow-up examination uses a different MR unit and imaging protocol, obtaining a tumor volume of 33 ml. At this time, most readers would report a tumor volume reduction by 7 ml in comparison to the first examination (a 17% reduction), allowing the clinician to define the patient as responding to treatment. But the questions to be answered are: Is the tumor really smaller, or is the 7-ml volume difference a result of variability? How can one be sure that such a volume difference is accurate? What if that difference was 1 ml, or 0.1 ml? Could one still trust in such a small difference? If not, how big a difference is necessary to be considered a treatment response?

As mentioned above, each measurement has an associated variability. In the present example the observed tumor volume difference between the MR examinations is a result of a complex combination of the following sources of variability:

(1) the intraobserver variability of the radiologist who performed the measurement prior to chemotherapy (R1),
(2) the intraobserver variability of the radiologist who performed the measurement after the chemotherapy (R2),
(3) the interobserver variability between R1 and R2,
(4) the inter-instrumentation variability from the use of two different MR units and imaging protocols,
(5) the biological variability, including the effect of chemotherapy on the individual patient.

All these sources of variability act simultaneously and are unavoidable. As a consequence, the total variability is a weighted sum of all sources. The repeated procedure (the tumor volume measurement), even if applied on the same MR images, rarely provides the same value as that previously obtained. The first three points, for example, are strongly dependent on the radiologists' experience. Technical variables, such as the repetition and echo times and the automated computer algorithm, can also affect the result. The fourth point is easily understood: each manufacturer has its own technical specifications (e.g., magnetic field strength, coil channels, software) and these variables all affect the final tumor volume; imaging protocols greatly affect it as well. Finally, at least a part of the 7-ml observed difference of this working example could be a result of therapy and represent a real reduction of the tumor volume. This last part is what we really want to estimate.

To answer the above questions we need to isolate the effect of chemotherapy by estimating all other sources of variability. To this end, let's imagine that in a preliminary study we obtained, for example, the following results:

(1) intraobserver variability of R1=1 ml,
(2) intraobserver variability of R2=1 ml,
(3) R1–R2 interobserver variability=3 ml,
(4) inter-instrumentation variability=3 ml,

for a total of 8 ml. It is easy to understand that with this total variability one cannot conclude safely whether there was a real reduction in tumor volume because the observed 7-ml difference is likely a result of variability.

Interpretation

In the previous example, any pre- to post-therapy volume differences lower than or equal to 8 ml could be attributed to the total variability associated with that particular measurement. In other words, the observed 7-ml difference of the working example is too small to be considered a real tumor volume reduction. Thus, the patient cannot be classified as a responder.

The total variability of 8 ml of the working example can be considered an intrinsic limit, a threshold under which we cannot go. We can appreciate modifications in the tumor volume only if larger than 8 ml. This threshold is also known as the least significant change, representing the best performance that we can make with this measurement. Although it may appear confounding, the least significant change represents a kind of sensitivity: in other words, in the setting of the example case, we are not sensitive to volume modifications of up to 8 ml. This last statement is true even if the tumor volume is really changed.

The numbers in the working example, of course, are invented. However the real world is not so different. The total interstudy variability of 8 ml is 20% of the tumor volume measured prior the chemotherapy (40 ml). The interstudy reproducibility is the complement to 100% of the total variability, i.e. 80%, a typical value found in the literature. In a very recent paper published on pediatric cardiology, Pediatric Heart Network investigators demonstrated that echocardiography used to quantify left ventricular size and function in

pediatric patients with dilated cardiomyopathy has an interstudy variability as high as 20% [2].

Reproducibility has been inappropriately evaluated for years using correlation coefficients such as the Pearson or the intraclass coefficient [3]. As shown by J. Martin Bland and Douglas G. Altman in 1986, this approach is incorrect for a number of reasons. In their well-known article published on 1986, they presented a valuable method (known as the Bland–Altman method) to estimate reproducibility. Basically, this method consists of obtaining two measurements for each in a series of patients. The difference between individual pairs of values is distributed normally, with the standard deviation being related to the reproducibility of the method used for measurements.

Variability is intrinsic to the measurement processes. It can be reduced but not completely eliminated. A way of reducing variability is to make the measurement as objective as possible, by defining rules and procedures for carrying it out. In the example case, the measurement after chemotherapy could be repeated at the same institution by the same radiologist and using the same scanner.

In clinical practice, radiologists are subconsciously aware of measurements' variability, with the quality of studies being the main determinant for their conclusion. They automatically apply a sort of variability to their measurements, not blindly trusting in numbers. Again, Galilei would advise to scientifically estimate this variability rather than to guess at it. Finally, apart from patient management, a formal estimation of variability is needed to design clinical studies in medical research.

In conclusion, I have shown the importance of a high reproducibility in patient follow-up. Variability is unavoidable and should be estimated in advance for each clinical setting. Physicians should be aware that a lack of reproducibility could result in a wrong interpretation of a patient clinical course and lead to inappropriate decisions.

Conflicts of interest None

References

1. Sardanelli F, Di Leo G (2009) Biostatistics for radiologists. Springer, Milan
2. Lee CK, Margossian R, Sleeper LA et al (2014) Variability of M-mode versus two-dimensional echocardiography measurements in children with dilated cardiomyopathy. Pediatr Cardiol 35:658–667
3. Bland JM, Altman DG (1986) Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 8:307–310