Accuracy requires precision: a comment on understanding and using statistics in nuclear medicine

Article *in* Journal of Nuclear Medicine · April 1980

Source: PubMed

1 author:

Ronald Collé
National Institute of Standards and Technology
97 PUBLICATIONS **1,020** CITATIONS

SEE PROFILE

LETTERS TO THE EDITOR

Early Brain Images, an Important Definition

I enjoyed the recent paper of D. E. Tanasescu, et al. (*1*) and I hope that many of my colleagues in Germany and Europe have read it. The authors compare early and delayed Tc-99m glucoheptonate brain images, performed with an Anger camera, and conclude that early Tc-99m GH brain images are inferior to the delayed ones and cannot give much additional information to the physician. I welcome this paper and the former presentations of this group demonstrating the effectiveness of brain scanning when 3- or 4-hr delayed images are made. But I think that the first problem in this connection is to define what an early brain image is. We can agree that delayed scans are superior compared with 20- to 30-min photos. And I believe, further, that there is no difference between 30- and 60-min images.

Beginning in 1968 our brain studies have been carried out following a multiple step approach: first a dynamic study, followed by early images 1 to 5 min after injection in three planes. Then we perform 1- and 3-hr delayed brain scans (*2*). Some years ago we postulated that delayed images (after 3 or more hours) are the most important in arriving at a diagnosis. But we have also found the early pictures to be necessary in patients having an AV malformation, i.e., angiomas, aneurysm, etc. In the detection and differentiation of meningiomas, the 1- to 5-min scintiphotos are very important, because they permit correct diagnosis of meningiomas (*3,4*). Photos made 1 hr after injection are not very informative and can therefore be omitted. In my opinion 20- to 30-min images will not demonstrate a high level of vascular radioactivity in a lesion.

Our earlier studies were performed using Tc-99m as pertechnetate, and we concluded, in agreement with H. Rösler, that with this tracer the nuclide at 1 to 5 min after injection must be intravascular, therefore demonstrating highly vascularized lesions at such a time. Using Tc-99m complexed (Sn^{2+}) with citrate, DTPA and glucoheptonate, we observed, in the early photos, the same effects as with pertechnetate. We conclude, therefore, that images at 1-5 min also permit the demonstration of high vascularity if Tc-99m chelates are used. For some years all our brain studies have been performed using these tracers.

I think, therefore, that we must make really early photos immediately after tracer injection, but that images at 20, 30, or 60 min can be dispensed with. I would be happy to find other writers agreeing, for without such an agreement, we cannot compare the results of our studies.

Finally, I must discourage the use of a second tracer dose, after the delayed scintigrams are done, in order to get very early images. This doubles the radiation dose to the patient, and it is not at all necessary. I think early images immediately following a dynamic study are not so expensive that they cannot be performed routinely. In our department we have been doing this for many years.

J. SAUER
Zentralkrankenhaus
Bremen, Germany

REFERENCES

1. TANASESCU DE, WOLFSTEIN RS, BRACHMAN MB, et al: Early and delayed Tc-99m glucoheptonate brain scintigraphy: Are routine early images indicated? *J Nucl Med* 20: 287-290, 1979
2. SAUER J: *Die hirndiagnostischen Möglichkeiten der Szintillationskamera nach Anger. Results of Clinical Nuclear Medicine. Diagnostic, Therapy, Research.* Stuttgart, New York, F. K. Schattauer, 1971, pp 1053-1060
3. SAUER J: *Hirntumordiagnostik durch szintigraphische Verfahren heute. Gem. Kongreß der Deutschen und der Österreichischen Röntgengesellschaft 1973.* Thieme, Stuttgart 1974, pp 68-76
4. SAUER J: *Wie ist die szintigraphische Hirndiagnostik für die Routine zu optimieren? Nuklearmedizin. Quality Factors in Nuclear Medicine.* Stuttgart, New York, F. K. Schattauer, 1977, pp 593-597

Reply

The points made by Dr. Sauer have been covered in our paper. We mentioned in the text that early brain images are inferior to delayed ones in detecting CNS lesions, but we also stated that the use of both may be of help. In our table it is shown that in 2% of cases the early studies contained diagnostic information not demonstrated in the delayed images. In addition, in 6% the early study showed more radioactivity in the abnormality than did the delayed images. We discussed the superiority of earlier images in demonstrating vascular abnormalities. We also obtained a routine early static image in the projection in which the flow study was performed. Currently we are not performing routine early studies except in patients with scalp or skull lesions and in those suspected of having AV malformation or meningioma. The purpose of our paper was to publish the statistics of a large series, comparing both early and delayed Tc-99m glucoheptonate brain images. Our final conclusion was that the early TcGH brain scintigram is not a substitute for a delayed study. However, early scintigraphy was helpful in our series in 8% of the cases studied. Thus, if the logistics in a given institution permit early studies to be done routinely, we feel this approach to be warranted.

DOINA E. TANASESCU
Cedars-Sinai Medical Center
Los Angeles, California

Accuracy Requires Precision: A Comment on Understanding and Using Statistics in Nuclear Medicine

In a recent article, Levin (*1*) reviewed some of the fundamental principles of statistics as they apply to the estimation of measurement uncertainties. Although efforts to get practitioners of nuclear medicine to understand and use statistics are laudable, Levin, unfortunately, has reinforced a common misunderstanding of the meaning of "accuracy." The distinction between *precision* and *accuracy* is frequently misunderstood (*2*). Accuracy is not independent of precision. Accuracy requires precision. The problem goes beyond mere definitional semantics, but leads directly to a misunderstanding of the measurement process and the requirements necessary for obtaining accurate measurements.

Consider a measurement result, x, of a quantity obtained from some given measurement process and used to estimate the *true value,* τ, of the quantity. The accuracy of the measurement process is, as stated by Levin, a measure of the *closeness to the truth.* In fact, the *absolute error* of a particular measurement result is just the difference between x and τ (*2*). The exact difference is, of course, unknowable because the true value can never be known exactly. Although the absolute error is unknowable, limits to its magnitude can be inferred and estimated from the measurement process itself. This estimate of the limits to the absolute error is referred to as the *uncertainty.* For reasons that will be demonstrated shortly, the uncertainty, and hence the inaccuracy, of the measurement process may be appreciable even if the absolute error of a particular result is fortuitously negligible or even zero. Foremost, the uncertainty should be a statement, based on a complete and credible assessment, of the likely inaccuracy or the likely limits to the absolute error in the measurement result. The overall or total uncertainty is used to estimate the inaccuracy of x, and can be thought to be comprised of two types of uncertainty, namely random variability and systematic bias.

The random uncertainty is a statement of precision and is a measure of the reproducibility or scatter in a set of successive independent measurements. Precision then is a measure of the *closeness together.* Sample statistics such as the standard deviation, s_x, which are computed entirely from the measurement data and used to estimate the population parameters such as σ_x, are commonly used measures of precision—or, more correctly, measures of *im*precision.

In contradistinction, a *bias* is a deviation from τ that is always of the same magnitude and direction. It cannot be estimated or calculated solely from a given set of replicate measurements, since every measurement is affected by the systematic bias in the same way. A bias is the difference μ_x-τ between the limiting mean (μ_x) associated with the measurement of the particular quantity by the given measurement process, and the true value, τ, of the quantity (*2*). The detection of bias in a measurement process may be achieved by comparison with a standard (a defined true value) or by verification with two or more independent and reliable measurement methods (*3*). There can be many contributing sources of bias in a given measurement process. They can be introduced by the measurement process and are characteristic of it. Such systematic biasses are not amenable to statistical treatments. The biasses should be estimated upper limits for each conceivable source of inaccuracy in the measurement process. Their magnitudes would preferably be based on experimental verification with standards or other methods, but may have to be estimated from experience and judgement.

The very familiar bull's-eye example shown in Fig. 1 and referred to by Levin should help to illustrate the distinction between precision and systematic biasses, and their relation to accuracy. The bull's-eye of the target corresponds to the true value, and the six shots represent individual measurement results. The figure illustrates the concept of an inaccurate measurement due to imprecision, a precise but inaccurate measurement, and an accurate measurement. In the upper case, a systematic bias may or may not be present. It is impossible to know for certain because the precision is inadequate. In the middle case, the measurement is precise but is inaccurate since a bias is present. Finally in the bottom case, we have the desired condition (accuracy). There is no such case as an accurate but imprecise measurement. Accuracy is *not* independent of precision. Precision, in fact, is the first requirement for accuracy (*2*).

One must recognize that as a practical matter it is necessary to achieve precision before anything can be learned about the possible presence of a systematic bias. Furthermore, in reality the location of the bull's-eye (i.e., the true value) is generally unknown. Before one can begin to deduce the location of the bull's eye, one must first

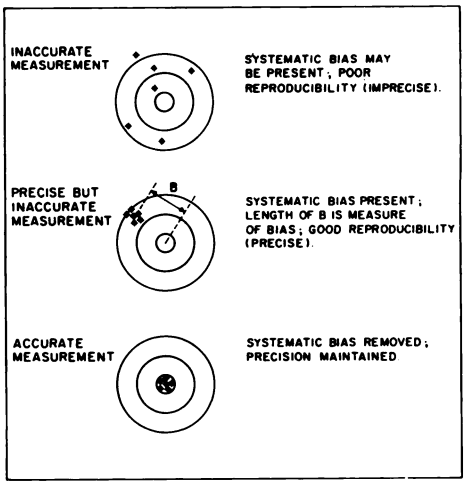

FIG. 1. Illustration of distinction between precision and systematic biasses, and their relation to accuracy.

have precision in order to assess the systematic bias in a reasonable way and eliminate it or make corrections for it. When precision is achieved and maintained, the measurement process is said to be in a state of "statistical control" (*2*). The techniques and details for testing whether or not the measurement process is in statistical control are available (*2-4*) and outside the scope of this letter. Their primary objectives are to test for control, to make predictions in the statistical sense, and to aid in maintaining control over the measurement process.

Precision plays an important role in demonstrating accuracy. Without statistical control over the measurement process it is impossible to make a meaningful or complete assessment of the likely limits to the accuracy in a measurement result. One benefit of making such a complete uncertainty assessment is that the process will require the evaluation and statistical control of many previously unevaluated measurement parameters. This will ultimately aid quality control within laboratories, improve the quality of measurements, and lead to the desired result of accurate measurements.

Levin's extremely restrictive definition of accuracy as equivalent to the absence of bias is inconsistent with the concept of accuracy being "*closeness to the truth,*" and it does not further an understanding of the meaning of accurate measurements or how to achieve them.

R. COLLÉ
National Bureau of Standards
Washington, DC

REFERENCES

1. LEVIN SG: Understanding and using statistics in nuclear medicine. *J Nucl Med* 20: 550-558, 1979
2. EISENHART C: Realistic evaluation of the precision and accuracy of instrument calibration systems. *J Res NBS:* 161-187, 1963 (Reprinted in *Precision Measurement and Calibration—Statistical Concepts and Procedures,* Ku HH, ed, National Bureau of Standards Publ 300, vol 1, Washington, U.S. Govt Printing Office, 1969, pp 21-47)
3. CALI JP, REED WP: The role of the National Bureau of Standards Standard Reference Materials in accurate trace analysis. In *Accuracy in Trace Analysis: Sampling, Sample Handling, Analysis,* Vol 1, LaFleur PD, National Bureau of Standards Spec Pub 422, Washington, U.S. Govt Printing Office, 1976, pp 41-63

4. NATRELLA MG: *Experimental Statistics,* National Bureau of Standards Handbook 91. Washington, U.S. Govt Printing Office, 1963

Reply

Dr. Collé is, of course, correct in his explanation of the measurement theorist's definition of the terms *accuracy* and *precision;* I was taught these many years ago by Dr. Eisenhart at the National Bureau of Standards. But I also realize that all definitions are arbitrary; e.g., most dictionaries use "precise" as a synonym for "accurate".

I used these same words to draw a distinction, often ignored by researchers, that *precision* relates to variability whereas *accuracy* is associated with lack of bias. The problem with using the definitions advocated by NBS and others is that because of the interrelationship between the two words, a technically accurate explanation tends to obscure rather than clarify precisely the distinction that I wanted to draw.

SHELDON G. LEVIN
Armed Forces Radiobiology Research Institute
Bethesda, Maryland

Editor's Comment

(In the minds of many people, including those who edit dictionaries, the words "accurate" and "precise" are nearly alike. Since most of us who work with statistics are not 'purists,' perhaps a simple remedy would be the use of words that are more descriptive. Possibly "repeatability"would be an improvement over "precision," and certainly the difference between "repeatability" and "accuracy" is appreciably more obvious. Editor).

Re: Scintigraphic, Electrocardiographic, and Enzymatic Diagnosis of Perioperative Myocardial Infarction in Patients Undergoing Myocardial Vascularization

I read with interest the study by Burdine and coworkers in the July issue of the *Journal of Nuclear Medicine* (*1*), in which they conclude that Tc-99m pyrophosphate myocardial imaging (TcPPi) is "probably the most valuable means of diagnosing perioperative myocardial infarction."

However, the study design is handicapped because of the lack of external determination of the end point. The authors use combinations of the predictor variables to determine the outcome "myocardial infarction."

Further, the particular combinations of predictor variables to define the outcome event appears to bias the study against the possibility that enzyme elevation is the most valuable variable. By requiring both enzyme elevation and TcPPi to be positive for "definite myocardial infarction," "positive" cases cannot be classified by enzyme elevation alone. To qualify as "probable myocardial infarction," the authors require that enzyme elevation must be accompanied by persistent electrocardiographic changes—usually the least sensitive factor in myocardial infarction.

To illustrate the concerns, I have prepared a hypothetical table of data wherein the "truth" is known. Test A represents the least sensitive test and Test B the most sensitive, with Test C intermediate in sensitivity.

Using the "truth," the sensitivity of Tests A, B, and C are 0.621, 0.947 and 0.800, respectively. The specificity of the three tests are 0.989, 0.994, and 0.994. The predictive value of the three tests when positive (PVP) are 0.855, 0.947, and 0.938. The predictive value of the negative tests (PVN) are 0.961, 0.994, and 0.979, respectively. Thus, B is the most sensitive test; B and C are equally specific. B has the highest predictive values for both positive and negative tests.

TABLE

Group	n	"Truth"	Test results A	B	C	A or (B + C) positive
1	895	−	−	−	−	−
2	5	−	+	−	−	+
3	5	−	+	−	+	+
4	17	+	−	+	−	−
5	3	+	−	−	+	−
6	2	+	+	−	+	+
7	2	+	+	+	−	+
8	16	+	−	+	+	+
9	55	+	+	+	+	+

In contrast, when the criteria A or (B + C) are used, the following figures result. Sensitivity for Tests A, B, and C: 0.812, 0.859, and 0.918. Specificity with these criteria are 1.00, 0.981, and 0.997, respectively. The predictive value of a positive test for the three tests are 1.00, 0.811, and 0.963. Predictive value of negative tests are 0.983, 0.987, and 0.992. Using this analysis, C appears to be preferable to B by each measure of test utility.

The numbers in this example were quickly assembled to illustrate the point that predictor variables should not be used to define the outcome measure. I did attempt to make the incidence of events comparable to those of perioperative myocardial infarction, to make Test A resemble electrocardiographic diagnosis in being the least sensitive of the three methods, and to have Test B with a slight advantage over Test C. The example is not intended to prove that enzyme elevation is the most valuable means of diagnosing perioperative myocardial infarction, although this may be true. Rather, it is to illustrate that, given the approach used by the authors, I cannot conclude that they have demonstrated that TcPPi myocardial imaging is the most valuable means of diagnosing perioperative myocardial infarction.

DENNIS M. DAVIDSON
Stanford University School of Medicine
Stanford, California

REFERENCE

1. BURDINE JA, DEPUEY EG, ORZAN F, et al: Scintigraphic, electrocardiographic, and enzymatic diagnosis of perioperative myocardial infarction in patients undergoing myocardial revascularization. *J Nucl Med* 20: 711-714, 1979

Reply

We agree with Dr. Davidson that the use of predictor variables to determine outcome is less than optimal, but we emphasize that there is no definitive procedure short of necropsy to diagnose perioperative myocardial infarction (POMI). Postoperative assessment of regional wall motion adds valuable information, but is still less than definitive, particularly when damage is confined to the subendocardium. This problem of a lack of a satisfactory "gold standard" hampers all such comparative studies.

While we therefore agree with Dr. Davidson's concerns, we nevertheless believe that he is incorrect in his conclusions. In view