2013.04.19

Charles Metz: scholar

James A Hanley

McGill University, Department of Epidemiology, Biostatistics, and Occupational Health,

1020 Pine Avenue West, Montreal, Quebec, H3A 1A2, Canada

James.Hanley@McGill.CA ; http://www.biostat.mcgill.ca/hanley

As sad as it is, Charlie Metz's passing[1] is a chance to honor his work, and also to be inspired by the first hand accounts of the sides of Charlie that cannot be measured by Pubmed or the Science Citation Index.

I became aware of his work when, in 1980, Barbara McNeil asked if I would help her assistant who was using Dorfman and Alf's (D+A) model to fit eight ROC curves. In her study of "when are radiologists are more accurate?" each of four radiologists had read CT images of the head twice, once with, once without patient histories. In one of the two 'conditions,' one radiologist had not used a sufficient number of the rating categories, and so the fitting of a binormal model was problematic. I had never heard of ROC analysis, and so, for orientation, she gave me Charlie's tutorial (2).

In 1981, when we met with John Swets and Charlie, they shared with us a draft of their textbook (3). Although the statistical methods have been considerably refined since then, the broad design and statistical principles continue to apply, and so I still to recommend it to beginning researchers – it has John's elegant exposition, and Charlie's extensive and careful coverage of the first principles of and the first approaches to, the statistical-analysis of paired, multi-reader ROC data.

Although he was more ‘parametric’ than I, Charlie and I got along very well. I can still remember his reactions when, in our meeting, I showed him a figure I had hand-drawn. It had a pair of overlapping (bivariate) frequency contours that I was using (more as a schematic than anything else) to work out the covariances between two non-parametric AUCs derived from the same set of patients. In drawing them, I gave the contours elliptical (i.e., bivariate normal) shapes. I tried to stay non-parametric as much as I could. But I did use a rough table, derived from a pair of bivariate *normal* distributions, to derive Kendall’s tau type correlations for use in sample size calculations. Charlie’s first reaction was: “I have come up with exactly the same pair of bivariate Gaussian distributions for a fully-parametric model for two diagnostic tests. I can’t believe we both hit on this same idea of overlapping bivariate distributions independently of each other!” And then with increasing joy he reasoned that if two people had independently come upon the same idea, it must be worth more than if only one of us had. His approach, a bivariate version of the binormal model fitted by Dorfman and Alf in the late 1960s(4), was a breakthrough, and provided a much better way to deal with paired comparisons; one no longer had to rely on univariate fits, accompanied by awkward covariance corrections, the way Charlie was forced to in his contribution to the Swets and Pickett book. He replayed his joy at our ‘independent bivariate moments’ for me when we met again many years later.

When we met in 1981, he had already overcome the technical challenges of fitting his bivariate-binormal model, but he only published it in 1984(5). This “don’t publish it until you have *fully* tested it” attitude was characteristic of Charlie. He was less concerned with accelerating his curriculum vitae than he was about getting the technique exactly right. It was as if his algorithms were being used to extract information from medical images from an individual patient, and the result mattered to the care of that patient, so it had better be right. I wish all biostatisticians were as careful with their development work, and their claims for their methods, and as academically disinterested, as Charlie. This ‘ultimate scholarship' was again evident in his work with ROC curves from continuously-distributed data, which he didn't publish until 1998(6). It took the cajoling of several biostatisticians, and an invitation to the Joint Statistical Meetings to convince him to finally publish it, and to do so in a biostatistical journal. He took the question of whether his approach deserved the 'Maximum Likelihood' label much more seriously, and tested it much more extensively, than many biostatisticians do. To me, it is a masterful piece of work, 'semi-' and indeed 'almost-fully-' parametric. Charlie was one of the very few who really understood and carefully enunciated the bi-normal model. If you were writing or even just speaking about it, he did not let you get away with any shorthand or hand waving.

One feature of the binormal model that intrigued him is the fact that the unequal variance version usually fit the data better than the equal variance one: but that meant that the likelihood ratio was not monotonic and that the ROC curve was not fully concave. When he and I met with Don Dorfman and his colleagues in Iowa in the 1990s, we had a number of discussions about this non-monotonicity issue, and Don told us that he had some solid empirical evidence that this was not just an artifact: it had been observed in controlled experimental laboratory type conditions where one might have expected a monotone ratio. This shook up the scholar in Charlie, and he mulled over it for the rest of the day. By our meeting the next day, he still had no comeback, and for once I saw him resort to a less than scientific explanation, "Don, those data are on pigeons, not humans, and we all know that pigeons have only half a brain. To convince me, show me a non-monotone ratio in controlled *human* data."

In these small-group settings, I got to see other sides of Charlie, what else he enjoyed in life, and his humor. Once, we discussed access to imaging tests in the Canadian health care system: I told him I had undergone brain imaging as a volunteer research subject, but also -- at a time when there were fewer MRI facilities -- as a patient. The CT scan, undergone within a day of being ordered, was negative, but I waited 10 weeks for an MRI scan, thankfully also negative. A physician colleague told me that the ‘instant CT but delayed MRI’ was good medical practice at the time, even if the wait for reassurance was stressful. Charlie then told me about two U.S. family doctors, both hot-air balloon enthusiasts, who had become lost while aloft early one Sunday morning. They finally passed over a town square and on noticing a person on the ground, shouted down to him, “Can you please tell us where are we?” When the person responded, “you are in a hot air balloon,” the one doctor said to the other, “that person must be a neurologist.” When the colleague asked why, he replied, “Because what he told us was absolutely correct, but of absolutely no use to anyone.” When analyzing data on the first uses of CT back in 1977, I had learned that the high diagnostic yield of neurologic imaging examinations did not necessarily mean clearer treatment options; Charlie told me had also seen this aspect of neurology -- not in a research setting, but up close and personally.

I conclude by returning to the 2013 lessons from the history/no history study, and from our very real encounter with the ‘random observer effects’ that Charlie introduced in the 1982 textbook. Three of the radiologists we studied had significantly improved accuracy when reading CT images with fuller knowledge of patient histories, and the other one clearly had not. Radiologists would surely welcome our evidence that they integrate what they see with the history information, but how extensive was our evidence? Charlie’s variance calculation for the average improvement was based partially, as it should be, on the modality x observer interaction sum of squares, i.e., on *how non-uniform* the improvement was across observers, but also, of course, on case variability. His F statistic (or mean/SE t-ratio) was based in large part on the degrees of freedom available to estimate this most critical variance component but also, in part, on case variability: it confirmed that our evidence was not all that extensive. [I learned later that the one radiology resident studied who was not helped by history was considered among the least competent the department had had in quite a few years.]

I was struck by several aspects of that study, which I joined only at the data-analysis stage. A two-percentage points improvement in the AUC does matter; whether one does / does not improve correlates with other assessments of competency; there is a lot of statistical information in a dataset involving images from *just 100 patients*. But most importantly, a study of *4 Harvard radiology residents* does not allow statements about the '*typical'* or '*average'* or '*any other-type*' resident.

Fortunately, in this regard, because of advances in both statistical and information technology, we are far better situated today than in 1982. First, thanks to the statistical innovations by developers such as Dorfman, Berbaum, Obuchowski, Rockette, Hillis, and by Charlie himself, we have better methods for estimating the various components of variance, for amalgamating them, and for computing the appropriate numbers of degrees of freedom. Despite this, I wonder if we sometimes rely too much on simulation studies of the performance of these data-analytic methods, and not enough on broader scientific principles and real experience, to justify whether very small numbers of observers are sufficient. It might be that my one bad experience has unduly influenced my impression that accuracy differences are generally very non-uniform across observers. It would be good to have more data on the sizes of the different variance components, and on which types of settings affect their ordering. Such empirical work would serve as a tribute to Charlie, who first raised the statistical issue and motivated much of the methodological work to deal with it. And it would provide better, more empirically-based, inputs to the planning of the size of studies of improvements in diagnostic performance, and help us distinguish the settings in which the improvements are most relevant.

In the 1980s it was easier to make the excuse that we only have so many residents in any one institution, and that in any case the pairing of subjects (images) will take care of a lot of the variation, and that this the best we could do. But, as is illustrated by our "3 of 4" story, the non-uniformity of effects across readers may matter more than the case variation, and may seriously limit the generality of the inferences we can draw. I have often asked researchers who don't want to face up to this fact whether they would be willing to investigate a new learning aid by relying on the average improvement on a (paired) 100-item test conducted on just 4 human subjects. How often is the number of radiologists the more important '*n*' in the sample size formula, and the number of images a distant second? Unless we actually study enough radiologists, or are privy to the 'worst in many years' type-information – this latter type of information is seldom reported in the Materials and Methods section -- we won't know precisely how big a variance component the non-uniformity (the modality x reader interaction) represents. Fortunately, today, now that images are mainly digital and can be viewed remotely, assembling a sufficiently broad sample of radiologists, from a range of institutions and settings, should no longer be such a limiting factor. The bigger sample size would also allow for performance to be linked to other measurable (and anonymously reportable!) characteristics and competencies of these observers.

My perspective in this tribute has been as a biostatistician in the service of biomedical research. Just as in any other research area, statistical methods are only as useful as the parameters they study, and the information they help to generate and summarize. And, just like other statistical techniques, ROC methods can lead not to just to "type I" and "type II" errors, but also to the more serious type III errors (providing answers to the wrong questions). Charlie Metz would want us to look beyond the mathematics and formulae to the "bigger picture" in radiologic research, and also, always to put scholarship first. Whatever our perspective, let Charlie's scholarship and unwavering principles inspire us to produce work that passes the 3 (journalistic) tests, *"Is it new? Is it true? Does it matter?"*

Acknowledgment: This work was supported by the Natural Sciences and Engineering Research Council of Canada, and the Canadian Institute for Health Research.

REFERENCES

[1] Zou, K. Professor Charles E. Metz Leaves Profound Legacy in ROC Methodology: An Introduction to the Two Metz Memorial Issues http://www.academicradiology.org/article/S1076-6332(12)00481-3/fulltext

[2] Metz, CE. Basic principles of ROC analysis, Semin Nucl Med 1978; 8: 283-298.

[3] Swets JA, Pickett RM. Evaluation of diagnostic systems: Methods from signal detection theory. New York: Academic Press, 1982.

[4] Dorfman DD, Alf E. Maximum likelihood estimation of parameters of signal detection theory and determination of confidence intervals - rating method data. Journal of Mathematical Psychology 1969; 6: 487-496.

[5] Metz CE, Wang PL, Kronman HB. A new approach for testing the significance of differences between ROC curves from correlated data, in Information processing in medical imaging. Deconink F (Ed). The Hague, Nijhoff, pp. 432-445. 1984

[6] Metz CE, Herman BA, Shen JH. Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data. Stat Med. 1998; 17:1033-053.