7a

Statistics in diagnostic decision making in nuclear medicine

J. Hilden (Copenhagen)

Consider a stream (statisticians say: a population) of cases of a two-way clinical decision problem: either the target disorder is present, or it is not (the 'diseased' vs. the 'non-diseased'). And consider a quantity that holds the result of a diagnostic procedure. By drawing the histograms that describe its distribution in the two subpopulations we can interpret ordinates and areas under the two humps in terms of true and false decisions and get a feel for the trade-off involved, provided that the pre-test probability of disease (percentage diseased) is known.

We use some further classical terminology: "positive" (suggestive of disease) vs. "negative," sensitivity, specificity, "predictive value" and "likelihood ratio."

A rigorous justification of a proposed decision rule clearly requires a quantification of the human and money gains or losses at stake – which is hard even in one-shot diagnostics, harder still in monitoring (dynamic pathology). Classical reference limits are of little use.

Given two quantities whose performance one wants to compare, histograms must be brought on a common scale: the Receiver Operating Characteristic (ROC) diagram offers its service. We shall mention what ROC features to look for and how to interpret them.

Can one combine the two tests? Yes, with proper attention to their simultaneous distributions, one can devise simultaneous procedures or sequential ones (i.e., flow charts).

Pure measurement noise, including reader variation, will flatten the histograms: it always weakens the discrimination. The effect can be mitigated by consensus making and repeated testing (though not always). Imaging equipment comes with 'built-in data reduction,' which is again a coarsening of the true picture. The loss is outweighed by practical advantages and also, importantly, by focussing and standardizing the research efforts that go into procuring the background data needed for all the statistical characterizations I have talked about.

A picture tells you more than a thousand digits. To select the best method of converting an image into a diagnostically informative numerical summary is a demanding research task, especially when the lesion one is looking for can be located anywhere (as opposed to a location known a priori), and when there can be multiple lesions. Typically, region-to-region signal comparisons in terms of 'percentage change' or SUV (standardized uptake value) are called for. They involve a number of fallacies, about which you will hear more today. The statistician's advice can only be: collect good reference data and approach the distribution of the chosen summary quantity in lesion and non-lesion in a purely empirical fashion. Unfortunately, the region-by-region truth does not always become known (when the patient dies from metastases in Region A, a suspicion concerning Region B may remain unresolved). Anyhow, a rational analysis will benefit from distinguishing four types region-to-region interdependence: pathogenetic (a disease focus at A makes a focus at B more likely a priori), diagnostic (a focus at A makes it more difficult to discern a focus at B, or a focus at A prompts a laparotomy, making B directly observable), prognostic (one metastatic focus may suffice to place the patient beyond rescue), and therapeutic (in suspected maxillar sinusitis, a positive finding on the left side, even a false-positive one, may trigger the antibiotic that cures the infection overlooked on the right, thus neutralizing a false-negative finding).

Finally, if you are a producer of evaluations of new tools and want to produce a piece of good and usable research, observe the STARD[1] (The Standards for Reporting of Diagnostic Accuracy) list of matters to attend to and report.

References

1. STARD (The Standards for Reporting of Diagnostic Accuracy), http://www.consort-statement.org/stardstatement.htm