EDRN: Some Statistical Considerations from a Regulatory Point of View

Gregory Campbell, Ph.D.

Director, Division of Biostatistics

CDRH/FDA

Statistics in Planning

- Selection of subjects (do not avoid the difficult cases; guarantee a realistic case mix)
- Mask the test interpreters from diagnostic truth
- Use multiple clinics and lab testing sites

Prospective versus Retrospective

- Virtues of prospective approaches
- Dangers of retrospective data-dredging for confirmatory analysis
- Retrospective rescue and multiplicity

What is (Statistical) Bias (and Why Should I Care)?

- Bias -- Systematic (non-random) error in the estimate of a treatment effect
- Bias can obscure the true diagnostic effect (misestimate it), making an ineffective diagnostic test look effective or an effective one appear ineffective
- Object: eliminate or identify and either reduce or estimate the bias

Bias in Design

- Spectrum bias-- (do not avoid the difficult cases; guarantee a realistic case mix)
- Observer bias--
  - Mask the test interpreters from diagnostic truth
  - Recall bias; fatigue bias; learning curve bias
- Misclassification bias
  - (verification, work-up)
- Inconclusive bias (do not drop out cases)
- Begg (1987)

Spectrum Bias

- The spectrum (or case mix) of the study does not conform with the mix in the population for which the test is intended.
- The types of cases and controls in the study do not represent the population of interest.
- Typically, the cases are more extreme and the controls healthier, both of which contribute to over-estimates of test performance.

Verification Bias

- The bias that occurs when how people are verified according to disease status differs depending on the outcome of other tests.
- Related to this bias is the workup bias; namely that patients are worked-up in a different manner depending on the results of tests; that is, the results of the diagnostic test affect the clinical work-up that establishes the diagnosis.

Sensitivity and Specificity

- Sensitivity (SENS) – fraction of responders who test positive
- Specificity (SPEC) – fraction of non-responders who test negative
- A test is useful (informative) if

$$\text{SENS} + \text{SPEC} > 1$$

Common Reporting Practices that are Statistically Inappropriate

- Using the words “sensitivity” and “specificity” in the comparison of a new test to an imperfect standard.
- Using an algorithm to define the standard (combining several comparative methods) that includes the outcome of the new test.
- Using results from <u>discrepant resolution</u> alone to estimate sensitivity or specificity between a new test and a comparative method.
- FDA Draft Guidance (2003)

Predictive Values

- Positive Predictive Value (PPV) – fraction of test positives who respond

- Negative Predictive Value (NPV) – fraction of test negatives who do not respond

- A test is useful if

  $$PPV + NPV > 1$$

The ROC Plot

- Receiver Operating Characteristic (ROC) Plot—a plot of all (1-Sp, Se) values
- It is a visual representation of the global performance of the diagnostic test.
- ROC plot shows the trade-off of sensitivity and specificity
- It is really a 2-D representation of a 3-D plot. What is the hidden variable (dimension)?
- CLSI GP-10

Variability

- Point estimates of sensitivity, specificity, predictive values, and agreement are not sufficient.
- Confidence intervals reflect the uncertainty of the estimates.
- Focus is often on confidence intervals to characterize the performance of the test and not on hypothesis testing.

Repeatability and Reproducibility of the Test

- Variability of all sorts:
  - Between and Within Day
  - Between and Within Site
  - Between and Within Lot
  - Between and Within Instrument or Chip
  - Between and Within Operator
- Sometimes we use Components of Variance (or components of CV (coefficient of variation))
- Is the variability such that it is valid to combine over sites? Or days, lots, instruments, etc.
- Variability of commercial test, of the non-commercial test

When Does a Diagnostic Test Work?

- Does the diagnostic test add anything to what is already known?
- Example: A diagnostic test for bone mineral density would need to show that it is better than just using a person’s age.

Covariate Modeling: Age in Bone Densitometry

- Problem: If you sample from subjects with major fractures and those without, there is likely to be an age difference that could confound the assessment of bone density. This is a bias.
- Example: Age in bone densitometry
- Is densitometry just a surrogate for age?
- Solution is to build a logistic model using age and bone density (BD) to predict fracture risk and see if BD adds anything.

Intention-to-Diagnose (ITD)

- In any case, everyone in the study should be kept track of and reported
- ITD is the analog of Intention-to-Treat (ITT)
- In diagnostic tests, there are indeterminates, samples without enough, the test fails to work, “no call”, etc. Sometimes this proportion is significant. For ITD analysis, these samples without diagnosis should be reported as failed to correctly diagnose.

Trials in Which Only the Test Positives Are Randomized

- Unable to estimate SENS, SPEC, or NPV directly from the data of the trial; only PPV is unbiased.
- No overall performance of the diagnostic can be reported, much less the treatment-by-test interaction.
- The diagnostic would need to be evaluated in a separate study. If the diagnostic is directly predicting response this is problematic. If a non-commercial diagnostic is used in the drug trial but a commercial product is required, this is an additional complication.

References

- Begg CB. (1987). Biases in the assessment of diagnostic tests. Stat in Med. 6, 411-423.
- Bossuyt, P.M., Reitsma, J.B., Bruns, D.E., Gatsonis, C.A., Glasziou, P.R., Irwig, L.M., Lijmer, J.G., Moher, D., Rennie, D. & de Vet, H.C.W. (2003). Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. *Clinical Chemistry* **49**, 1-6 (also in *Ann. Intern. Med., BMJ* and *Radiology in 2003).*
- CDRH, FDA. (2003). Statistical Guidance on Reporting Results from Studies Evaluating Diagnostic Tests: Draft Guidance (March, 2003).
- Pepe MS. (2003). The Statistical Evaluation of Medical Tests for Classification and Prediction. Oxford Press.
- Zhou X-H, Obuchowski NA, McClish DK. (2002). Statistical Methods in Diagnostic Medicine. Wiley.