mHealth Training Institute

UCLA

August 2015

Statistical methodology for mHealth

David Elashoff, PhD
Professor of Medicine,
Biostatistics and Biomathematics
Director, Department of Medicine
Statistics Core
Leader, UCLA CTSA Biostatistics
Program

http://obssr.od.nih.gov

Background

- Research on statistical methods for analysis of high-throughput genomic/proteomic studies
- Studies very often failed due to poor experimental design
- Careful consideration of study design with appropriate analytic methods critical for success in “Big Data” research world

Key Terms

- Sensitivity: Probability of testing postive among those with disease. (True positive rate, TPR)
- Specificity: Probability of testing negative among those with no disease. 1-Specificity is the False positive rate, FPR
- TPR and FPR may differ between populations
- Accuracy: Probability of a correct diagnosis. Depends on population prevalence

Study Design

Study Design

	Phase	Objective	Study Design
1	Preclinical Exploratory	Promising directions identified	Case-control (convenient samples)
2	Clinical Assay and Validation	Determine if a clinical assay detects established disease	Case-control (population based)
3	Retrospective Longitudinal	Determine if the biomarker detects disease before it becomes clinical. Define a "screen positive" rule.	Nested case-control in a population cohort
4	Prospective Screening	Extent and characteristics of disease detected by the test are determined and the false referral rate is identified	Cross-sectional cohort of people
5	Cancer Control	Impact of screening on reducing the burden of disease on the population is quantified	Randomized trial (ideally)

Pepe MS, Etzioni R, Feng Z, Potter JD, Thompson ML, Thornquist M, Winget M, Yasui Y.
Phases of biomarker development for early detection of cancer.
J Natl Cancer Inst. 2001 Jul 18;93(14):1054-61.

Phase I: “Discovery”

- Typically utilizes high throughput molecular, clinical, imaging or sensor technologies
- Often case-control designs of the worst sort
- Covariates often not considered
- Subjects can be matched
- Careful consideration of design of discovery studies is critical!

Phase 2-3: PRoBE Design

- PRoBE design represents set of design standards
- NCI EDRN has following study considerations:

  I. Clinical Context

  II. Performance Criteria

  III. The Biomarker Test

  IV. Study Size

Pepe MS, Feng Z, Janes H, Bossuyt PM, Potter JD. Pivotal evaluation of the accuracy of a biomarker used for classification or prediction: standards for study design.J Natl Cancer Inst. 2008 Oct 15;100(20):1432-8.

Clinical Context

- Central Question: For what population and in what clinical setting is the biomarker intended.
- Study population should represent that intended for clinical application (multi-institution) to avoid extrapolation bias
- Study population should be random selection from prospectively collected cohort with outcome blinding.

Confounding/Matching

- Confounding occurs when cases and controls differ on factors related to the biomarker and those that are predictive of disease.
  - Smoking status
  - Age
  - Ethnicity
- Matching can eliminate this confounding but:
  - Can render study population non-representative
  - Can make eventual biomarker test only useful in a comparative but not objective sense
  - Removes ability to assess combined predictive effects of biomarkers and covariates.

Performance Criteria

- Performance measures and acceptable levels depend on clinical context
- TPR and FPR are typical measures used.
- Minimal acceptable levels of TPR and FPR will vary based on:
  - Marker type (Ex. Diagnostic vs Screening)
  - Existing classification methods (add or replace?)
  - Error consequences
  - Cost

Examples

- Diagnostic Test:
  - Subjects with CT detected indeterminate pulmonary nodules may all be sent to biopsy (FPR=100%). If test with high TPR can be developed with even a FPR of 75% it may be useful.
- Screening Test:
  - Screening tests require very low FPR to avoid large number of unnecessary follow-ups.

Biomarker Test

- Blinded assaying of samples using consistent protocols
- Biomarker test can be combination of multiple biomarkers and clinical/demographic factors
- Utilizes pre-defined algorithm to combine quantitative factors, where algorithm development and assessment should be separate.

Statistical Modeling

Multimarker Methods for Model Development

- Two main philosophical approaches

  1. Create a model with relatively small number of markers that are individually assessed within model. (Ex. Logistic Regression)

  2. Create a “profile” model using a large number of markers without individual marker assessment.

Approach #1

- Discovery set used to identify small number (20-50) of targets
- Target prioritization based on
  - Statistical Significance (ex. T-test)
  - Magnitude of Difference
  - Classification ability (ex. ROC curve)
  - Uniqueness (ex. Correlation with other markers)
  - Biological Relevance (ex. tumor associated protein)

Approach #1 (cont.)

- Follow-up validation (phase 2) will:
  - Validate individual targets (statistical significance)
  - Construct multimarker model (ex: regression or decision rule models)
  - Estimate model parameters
  - Can add clinical/demographic (or existing biomarkers) factors to classification model.

Approach #2

- Using discovery set information:
  - Identify a large number of useful targets
  - Construct profile model using set of targets
  - Validate model on additional samples

Profile Models (Supervised Learning)

- Weighted Voting
- Neural Networks (ANN)
- Support Vector Machines (SVM)
- K-nearest neighbors (KNN)
- Linear Discriminant Analysis (LDA)
- Many, many more…

Test Assessment

- Regardless of type of model we want to compute:
  - Sensitivity (TPR)
  - Specificity (1-FPR)
  - Accuracy
  - Area under ROC curve
- Often require validation techniques to accurately compute these parameters

Modeling considerations

- No model type is the “best”
- Evaluation of model assumptions required
- Understanding of variables (missing data, outliers, distributions, measurement scale)
- Understanding of relationship between variables (collinearity)
- Automated modeling does not think and does not evaluate

Validation

Why Validation?

- Why do we need Validation?
  - Marker selection bias in multimarker models.
  - Potential overfitting of model coefficients
  - Overly optimistic estimates of model parameters (TPF, FPF, AUC)

Validation Strategies

- Leave-one-out cross validation (CV)
- K-fold CV (with model reselection and estimation)
- Training/Test
- Independent sample (gold standard)

Leave-one-out CV

- Remove the *i*th observation.
- Fit classification model (with variable selection).
- Use the model to predict the class of the *i*th observation.
- Repeat for all n observations.
- Determine model parameters on basis of accuracy of predictions for left out samples
- Better for small sample sizes

K-Fold Cross-Validation

- Similar to leave-one-out, divide the data equally into k groups and leave out one group at a time.
- Re-run variable selection and model fit and predict status of samples in left out group.
- Can also be run by randomly removing 1/k% of observations and re-running many times.

Training/Test

- Divide samples into training and test groups.
- Typically 50-50 or 67-33.
- Perform marker selection and model.
- Compute model parameters on the basis of test group samples.
- Critical to maintain balance across the groups

Independent Sample

While this is gold standard for validation, it can go wrong in a number of cases:

1. Populations differ in response rate
2. Populations differ in demographic composition
3. Differences in specifics of clinical scenario.

Combining Novel Markers and with Known Factors

- Many clinical scenarios have pre-existing clinical/demographic/biomarkers that need to be accounted for
- The differential AUC (model with know factors vs model with known factors + new markers) can gauge improvement.

Sample Size

Why do we need to consider sample size?

- Avoidance of wasting your time and money!
  - Insufficient power
  - Over power

What do you need to start a sample size computation

- Identify study endpoint(s).
- For each endpoint:
  - What is the effect of intervention or magnitude of the relationship?
  - How much variability?
  - Level of power?
  - One or two sided test?
  - What is the statistical test used to compute power?
  - What is the overall design?

Additional Considerations: Sample Size

- Can compute sample size for power or for estimation
- Account for study dropouts
- Account for multiple comparisons

Ways to Fail: Sample Size

- Not using relevant preliminary data in the calculation if available.
- Sample size calculation does not use methods in the planned statistical analysis
- Prediction modeling with large number of predictors relative to sample size
- Unrealistic assumptions about magnitude of effect

Bad Examples

- “A previous study in this area recruited 150 subjects and found highly significant results (p=0.014), and therefore a similar sample size should be sufficient here.”
- “Our lab usually uses 10 mice per group.”
- “Sample size calculations are not provided because there is no prior information on which to base them.”
- "The throughput of the clinic is around 50 patients a year, of whom 10% may refuse to take part in the study. Therefore over the 2 years of the study, the sample size will be 90 patients.”
- Our statistician computed that the power will be >80% for all study outcomes.

Good Example: Intervention Studies

"A sample size of 38 in each group will be sufficient to detect a difference of 5 points on the Beck scale of suicidal ideation, assuming a standard deviation of 7.7 points, a power of 80%, assuming a two sided significance level of 5% and a two sample t-test. This number has been increased to 60 per group (total of 120), to allow for a predicted drop-out from treatment of around one third. This difference of 5 points is based on our prior study in which….. "

Study Size: Modeling Studies

- Often sample size based on meeting performance criteria rather than inferential testing
- Set either minimal TPR or FPR for continuous markers.
- Power is assessed based on probability of achieving minimal performance given sample size and estimated “true” TPR/FPR values
- Minimal marker performance is set as lower bound for CI

Power/Sample Size for Marker Study Stages

- Variety of criteria for sample size for biomarker studies.
  - Discovery phase: Corrected Statistical significance for discovery studies (typically based on FDR)
  - Model building phase: requires reasonable numbers of cases per marker in model, can be based on confidence limits for TPR, FPR, AUC.
  - Validation phase: Comparison of AUC between biomarker and gold standard or old factors vs old factors plus new factors

Example sample size justification for modeling

- The overall sample size will be 580 (290 cancers and 290 non-cancers). These subjects will be divided up into a cohort of 190 cancers and non-cancers for the model building component and 100 cancers and non-cancers for the model validation component.
- The sample size of 190 of each group for model building will support the following:
- 1) This sample size will be sufficient to include approximately 2-4 clinical, imaging, plasma and nasal markers based on conventional rules of thumb suggesting that we should have 10-15 subjects in the less frequent outcome category per variable in a logistic regression model;
- 2) This will allow us to estimate the AUC for each model with a precision of approximately 0.045 (based on the 95% confidence interval for the AUC and an overall AUC of at least 0.8);

Example sample size justification for validation

- The sample size of 100 of each group for model building will support the following:
- 4) For the model validation component, the sample size of 100 of each group will provide precision of +/- 0.06 for the validation AUC as well as an 80% power to detect differences in AUCs between models of 0.08.
- 5) Finally, with this sample size we will be able to estimate the precision of the sensitivity and specificity (based on the width of the 95% confidence intervals) of between 0.06 to 0.08 depending on the observed values of those performance characteristics.

Reaching out to Experts

- When to contact a biostatistician?
  - Early in process particularly for research design and sample size
  - Study design and planning often requires early involvement with time for iterative improvement
  - Data analytic methods should wait until design and outcomes established