Study design and endpoints in ‘omics-based biomarker studies

Stephen A Williams MD PhD
SomaLogic Inc
Biomarker study design considerations

• Proposed context of use (COU)
• Weight of evidence required for COU
• Nature of evidence for COU – truth standards
• Design: efficient & biased vs. inefficient & pure
• Biomarker selection: candidate vs. hypothesis-free*
• Dimensionality reduction strategies*
• Biological plausibility: important or useless?*
• Performance: how good is good enough?
Context of use: FDA’s Drug development tools qualification program

• "Context of use," or COU, is a comprehensive and clear statement that describes the manner of use, interpretation, and purpose of use of a biomarker in drug development.

• How this helps us:
  – Avoids the concept of “universal validity” where biomarkers are applied in new contexts for which they are not qualified
  – Defines the nature of evidence required to qualify the biomarker in a defined COU
  – Defines the consequences of errors for that COU
  – Defines the intended benefit of success for that COU
A hypothetical example for NASH

• COU#1
  – A noninvasive substitute for liver biopsy, which mimics histopathological scores for liver fat, fibrosis, inflammation and ballooning, with accuracy similar to inter-pathologist variation, for use in phase II studies of NASH, enabling an end-of-phase go/no-go decision acceptable to the sponsor and the regulatory authorities, independent of drug mechanism.
    • Application in phase III or for registration would be a different COU (greater weight of evidence, same nature of evidence)
    • Biopsy substitution at only intermediate time points, retaining biopsy at beginning and end, would require lesser weight of evidence, same nature of evidence
    • Trial enrichment for rapid progressors would be a different nature of evidence (prognosis) but likely similar weight of evidence
Truth standards

• All truth standards are wrong, some are useful

Histopathology?  Time to cirrhosis?
+  Cardiovascular deaths?
+  Transplant rate?

AUC of recognizing Picasso B in this room if A is the truth standard
Weight of evidence for biomarker qualification is modulated by risk:benefit.

Satisficing = “good enough” or “reasonably likely”

Maximizing = “seek the best” or “certainty”
Weight of evidence: What it means

- Single study in a few hundred subjects in a similar population to clinical trials with cross-validation/bootstrap of models
- Above plus independent retrospective validation set in a few hundred subjects
- Above plus exact intended use population in a retrospective study
- Above plus geographic and population variability/robustness in a few thousand subjects
- Above plus prospective application in exact intended use population
- Above plus multiple prospective applications in exact intended use populations
Populations and study efficiency

• What is the “intended use population”? Does your discovery/validation study match it exactly?
  – Easiest if prevalence of cases and controls is near equal, in which case, run a cohort study

• Deliberate mismatches for efficiency:
  – If prevalence of cases and controls is very unequal:
    • Compare the extremes (very efficient for discovery of physiology but does not match intended use)
    • Case:control (efficient but controls who don’t look like cases are missing from the evaluation so performance isn’t representative)
    • Case:cohort (efficient and represents everyone in the intended use population, performance is representative, but physiology is diluted and/or biased e.g. by age, gender)
## Biomarker selection strategies

### Favored list

**PROS**
- Fewer measurements
- Easier control of false discovery rate
- Familiar: psychological acceptance easier
- Uses a-priori evidence – efficient

**CONS**
- Don’t find “black swans” – un-anticipated biology
- Unidimensional selection process misses complex signals
- A-priori evidence might be wrong: failure to replicate is common

### Hypothesis free

**PROS**
- Does not depend on favorites
- Optimal combinations possible – likely better performance
- Black swans (un-anticipated biology) can be found

**CONS**
- Demands large numbers of precise measurements
- Psychological barrier to “Fishing Expeditions”
- Control of false discovery rate requires more samples/skills
Dimensionality reduction strategies

• Purpose: When many measurements are made, to eliminate false positives and to include only the best combination of markers in a model

• Crude strategy: generate unidimensional lists by p-value and chop off the bottom

• Semi-crude strategy: iterate. Use first study to create an enriched list. Use second study to refine enriched list. Repeat....

• Sophisticated mathematical strategies:
  – Stability selection process (penalized popularity contest for marker inclusion in models)
  – Multi-dimensional feature selection of stable analytes
Biological plausibility: Necessary?

**Pros**
- Additional evidence of biomarker relevance (beyond statistical)
- Uncovers new drug targets
- Listed in ICH requirements for surrogate endpoints
- FDA support as criterion for surrogate endpoints
- Causality implications are useful

**Cons**
- Easy to make up a story
- Historically does not discriminate between successes and failures in surrogate endpoints
- Humans like unidimensional or linear stories; biology is more complex than that
- Excludes biomarkers where plausibility is weak/unknown
Biomarker performance: How good is good enough?

- Utilitarian/economic concept:
  - An endpoint is qualified when the value of the true signals is greater than the cost of the false signals
  - May also require superiority to the best available alternative

\[
\text{Value} \times \text{frequency of true positive} + \text{Value} \times \text{frequency of true negative} > \text{Cost} \times \text{frequency of false positive} + \text{Cost} \times \text{frequency of false negative}
\]
Performance metrics (a)

- Binary models: Why ROC curves are not the most important measure

Biomarker distribution

<table>
<thead>
<tr>
<th></th>
<th>Truth</th>
<th>Standard</th>
</tr>
</thead>
<tbody>
<tr>
<td>+</td>
<td>173</td>
<td>20</td>
</tr>
<tr>
<td>-</td>
<td>25</td>
<td>78</td>
</tr>
</tbody>
</table>

Comparison vs. another measure:
- NRI
- IDI

ROC curves
Performance metrics (b)

• Categorical models: Calibration

Biomarker distribution by population deciles or category

Calibration plot: predicted (pink) vs. observed truth (blue)
Summary

• Create a comprehensive evaluation of the context of use (COU) before even thinking of starting a study...
  – This determines the nature of evidence needed, including the truth standard
• For that COU, evaluate the weight of evidence needed (high value low consequence = low weight)
• Determine the appropriate study population and design considerations for efficiency
• Choose a biomarker selection strategy and a dimensionality reduction strategy
• Determine the appropriate performance metrics and how good is good enough in this COU
• Execute the program! Adapt to unexpected complexities.....