Clinical trials

How to find an effective biomarker?

Richard M. Simon, Genomic Clinical Trial and Predictive Medicine, Cambridge 2013

4.12.2013

Vendula Švendová
Institute for Medical Informatics, Meduni Graz
What is a biomarker?

• Biological measurement (e.g. gene mutation, HER2 protein positivity, expression level...)
  – *prognostic biomarker*: predicts the progress of the disease (under no or standard treatment)
  – *predictive biomarker*: predicts the response to a specific treatment
Why do we need biomarkers?

• Most cancer patients do not benefit from the systematic treatments they receive

• Being able to better predict which patients are likely to benefit would
  – Benefit patients
  – Control medical costs
  – Improve the success rate of clinical drug development
Prognostic and Predictive Biomarkers in Oncology

• Single gene or protein measurement
  – HER2 amplification (breast cancer)
  – KRAS mutation-negative biomarker (colorectal cancer)

• Index or classifier that summarizes expression levels of multiple genes
  – OncotypeDx recurrence score (breast cancer) = weighted average of expression levels of 21 genes
    “Chemotherapy yes or no?”
Biomarker validity

• Analytical validation
  – Accuracy compared to gold-standard
  – Robust and reproducible if there is no gold-standard

• Clinical validation
  – Does the biomarker predict what it is supposed to predict for independent data?

• Clinical/Medical utility
  – Does use of the biomarker result in patient benefit?
    • Is it actionable?
    • Generally by improving treatment decisions
Problems with many current biomarkers

• Developed in unfocused studies and not designed to address an intended medical use
  – Studies based on convenience samples of patients for whom tissue is available

• Although they correlate with a clinical endpoint, they have no demonstrated medical utility (they are not “actionable”)
Clinical trials
phases not uniquely defined, may overlap

• **Preclinical** – in vitro, in vivo testing on animals

• **Phase I** – Does the drug hit the target? is it safe? toxic side effects? testing on patients (15-30)

• **Phase II** – does it have any therapeutic effect (e.g. shrinkage of a tumor)? testing on patients (100-300)

• **Phase III** – is it more effective or less toxic than standard treatment? What is the patient benefit? Testing on patients (1000-2000)

• **Phase IV** – postmarketing surveillance, watching drug use in public, long term effect
Clinical trials
phases not uniquely defined, may overlap

• Preclinical – in vitro, in vivo testing on animals
• Phase I – Does the drug hit the target? is it safe? toxic side effects? testing on patients (15-30)
• **Phase II** – does it have any therapeutic effect (longer survival, shrinkage of a tumor)? testing on patients (100-300)
• Phase III – is it more effective or less toxic than standard treatment? What is the patient benefit? Testing on patients (1000-2000)
Phase II - goals

• Build a biomarker classifier - define its cut-point
e.g. HER2 – amplified YES/NO – binary classifier
cut-point: when is it YES and when is it NO?

• Validate the classifier – find the best for further testing!
does the treatment decision based on ‘HER2amplified YES/NO’ bring some results in patient benefit?
Is it promising and can be largely tested in Phase III?
General rule: Never build and evaluate model on the same data!

• Results might be very biased!
• Instead:
  – split data into training set and validation set
  – Build on training, evaluate on validation
  – If not enough data available – resample and evaluate on resampled

Prediction is very difficult; especially about the future Neils Bohr
Build on training, evaluate on validation in detail

**Biomarker classifier** = function $C$ that maps the biomarker values $x$ into categories

• If enough of data $D$ - split $D$ into training set $T$ and validation set $V$
  – Classifier $C$ built using $T$ set: $C(x; T) =$ YES or NO
  – Misclassification error rate is estimated by trying the model on $V$ set (evaluation)

• If not enough data ($p \gg n$)
  – Classifier built using full data set: $C(x; D) =$YES or NO
  – Misclassification error rate estimated using resampling methods (cross-validation)
Phase II - goals

• **Build a biomarker classifier - define its cut-point**
  e.g. HER2 – amplified YES/NO – binary classifier
  cut-point: when is it YES and when is it NO?

• **Validate the classifier – find the best for further testing!**
  
  does the treatment decision based on ‘HER2 amplified YES/NO’ bring some results in **patient benefit**?
  Is it promising and can be largely tested in Phase III?
Cut-point estimation

• When do we say patient is biomarker positive and when negative (= what is the biomarker cut-point)?

K candidate cut-points \(b_1, b_2, ..., b_K\)

\(p_i\) – probability of response (e.g. tumor shrinkage) for patients with biomarker level \(\leq b_i\) and assume \(p_1 \leq p_2 \leq \cdots \leq p_K\) (regression models)

\(p^*\) - response rate of interest

• 1\textsuperscript{st} stage: \(n_1\) patients, \(H_{0i}: p_i \geq p^*\)
• If not rejected for any \(p_i\), then 2\textsuperscript{nd} stage: more patients, same test
• If rejected for some \(p_i\), then 2\textsuperscript{nd} stage: testing on more patients with biomarker level above the largest \(i\) for which \(H_{0i}\) was rejected
• The smallest value of \(b_i\) for which \(H_{0i}\) was rejected is considered the cut-point
Phase II - goals

• Build a biomarker classifier - define its cut-point
e.g. HER2 – amplified YES/NO – binary classifier
cut-point: when is it YES and when is it NO?

• Validate the classifier – find the best for further testing!
does the treatment decision based on ‘HER2amplified YES/NO’ bring some results in patient benefit?
Is it promising and can be largely tested in Phase III?
Marker Strategy validation designs

Randomize

Use classifier test
- Biomarker +
  - New treatment
- Biomarker -
  - Standard treatment

Don’t use test
- Standard treatment

- inefficient:
  - requires large sample size
  - Many patients will receive the same treatment, regardless the randomization
Modified Marker Strategy
validation designs

Greater power for evaluating medical utility of the classifier

Perform classifier test

Biomarker +

Randomize

New treatment

Standard treatment

Biomarker -

Patient off study
Biology is not black and white

• Cancer biology is complex
• It is not always possible to have the right single predictive classifier identified with an appropriate cut-point by the time the phase III trial is ready to start
• Phases do often overlap
clinical trials
phases not uniquely defined, may overlap

- **Preclinical** – in vitro, in vivo testing on animals
- **Phase I** – Does the drug hit the target? is it safe? toxic side effects? testing on patients (15-30)
- **Phase II** – does it have any therapeutic effect (longer survival, shrinkage of a tumor)? testing on patients (100-300)
- **Phase III** – is it more effective or less toxic than standard treatment? What is the patient benefit? Testing on patients (1000-2000)
Phase III - goals

• Who is un/likely to benefit from the new treatment (medical utility)?

• What is the reduction in hazard compared to standard treatment?

• Ideal plan: detect reduction in hazard by 25-40%
Hazard ratio

- Hazard of death for patients
  - on the new treatment
  - on the control (standard treatment)
- Hazard ratio = \( \frac{\text{on the new treatment}}{\text{on the control}} \)

- Test+ patients: Hazard ratio \( \frac{0.42}{0.7} = 0.60 \)
  40% reduction in hazard for test+
- Test- patients: Hazard ratio \( \frac{0.7}{0.7} = 1.0 \)
  0% reduction in hazard for test-
- If 33% of patients test positive, then hazard ratio for unselected population is
  - \( 0.33 \times 0.60 + 0.67 \times 1 = 0.87 \)
  - 13% reduction in hazard
Randomize-all design
traditional

Might be optimal to show overall superiority of a drug
Enrichment (targeted) design was used for HER2 biomarker

- Does not demonstrate lack of benefit in marker negative patients (loss of information)
- But requires dramatically fewer randomized patients than the standard design
Stratification design
“gold standard”

When there is no good enough proof that the tests-negative patients do not benefit from the treatment – better to include all patients

Diagram:
- Measure marker
  - Biomarker +
    - Randomize
      - New drug
      - Control
  - Biomarker -
    - Randomize
      - New drug
      - Control

Interaction effect
test+ vs. test- patients

• If interaction effect is significant, treatment response in test-positive and test-negative patients differ
• But even small treatment effect in test-negative might be of clinical importance
• Also small test-positive sample size might cause non-significance

→ Always consider the size of the estimated treatment effect as well, not only significance
NCI – Biometric Research Branch
http://brb.nci.nih.gov/

- Web based software for planning clinical trials of treatments with a candidate predictive biomarker
  - Sample size planning
  - For binary and survival-disease-free survival end-points
  - Comparison to standard design based on the required number of randomized patients
• **Biomarker Targeted Randomized Design**

Targeted design randomizes only marker positive patients to treatment or control arm. Untargeted design does not measure marker and randomizes all who otherwise are eligible.

• **Biomarker Stratified Randomized Design**

Stratified design randomizes both marker positive and negative patients.

© NIH, 2008
Biomarker Stratified Randomized Design

Stratified design randomizes both marker positive and negative patients.

See references 73-75 in Technical Reports Section

• **Stratified Design with Prospective Analysis Plan and Binary Endpoint**

• **Stratified Design with Prospective Analysis Plan and Time-to-Event Endpoint**

© NIH, 2008
Stratified Design with Prospective Analysis Plan and Binary Endpoint

Randomized trial comparing new treatment (T) to control (C) includes both classifier positive and classifier negative patients. Presumes availability of binary classifier for biomarker predictive of benefit for new treatment.

Choose one analysis plan:

- **Analysis plan A**: Determine sample size for overall test comparing T to C for all randomized patients at reduced two-sided level alpha. If overall test is not significant, then test T vs C in classifier positive subset using (.05-alpha) significance threshold.

  - Overall improvement in response probability for new treatment: 0.15
  - Two-sided significance threshold (alpha): 0.03
  - Power for overall test: 0.90

- **Analysis plan B**: Determine sample size for comparing T to C in classifier positive subset at two-sided .05 level. If that is significant at .05 level, then evaluate classifier negative subset at two-sided .05 level.

  - Improvement in response probability for new treatment in classifier positive patients: 0.30
  - Power: 0.90

- **Analysis plan C**: First test for interaction between size of treatment effect and subset (classifier + or classifier -). If interaction is non-significant, just compare treatments overall at two-sided significance level .05. Otherwise, compare treatments within subsets at two-sided .05 level.

  - Overall improvement in response probability for new treatment: 0.15
  - Significance threshold for interaction test: 0.10
  - Power for overall test: 0.90

[Calculate]
Your Input:

Probability of response for control arm = 0.1
Proportion of patients who are classifier positive = 0.25

Analysis Plan A:
Overall improvement in response probability = 0.15
Two sided significance threshold = 0.03
Power for overall test = 0.9

Result:

Analysis Plan A:
Total number of patients randomized for overall test = 326

Expected number of classifier positive patients = 81

Power for classifier positive subset (significant threshold 0.02):

<table>
<thead>
<tr>
<th>Power</th>
<th>Improvement in response probability for classifier positive patients</th>
</tr>
</thead>
<tbody>
<tr>
<td>.80</td>
<td>0.376</td>
</tr>
<tr>
<td>.70</td>
<td>0.32</td>
</tr>
</tbody>
</table>

Output produced at Wed Nov 13 09:05:21 2013
Summary

• Standard treatment is not beneficial for most patients – we need better biomarkers
• Trials must focus on medical utility – who is likely to benefit and how much? Survival, response rate, hazard ratio
• There are many validation designs for different trial conditions Available sample size, end-point, biomarker quality,…
• NCI software can help you with planning your trial
Clinical trials worldwide (2007)
Density per country inhabitant (in millions)
% - average relative annual growth rate

Nature Review Drug Discovery 2008