Design of Drug Development Programs with Biomarkers: A stratified medicine approach

Paul Frewer
Abstract

In oncology is it becoming increasing common to have targeted treatments based on a patients biomarker status.

The talk will focus on types of designs which could be used in investigating a treatment with this intention.

For example designs allowing assessment of both biomarker positive and negative populations and designs incorporating a number of different investigational medicines targeted at specific patients.

There are advantages and challenges to the designs and we will focus on the statistical considerations to be aware of when developing the clinical plan...
Contents

• Need for patient selection
• Types of biomarker
• Enriched/stratified study designs
• Basket/umbrella studies
Need for patient selection

Targeted oncology drugs are designed to interfere with specific molecules that are involved in tumour growth and spread.

There is a biological rationale to restrict development to patients for whom the specific target molecule appears to be playing a key role in their cancer’s growth and survival.
Types of biomarkers

Biomarker
“a biological molecule found in blood, other body fluids, or tissues that is a sign of a normal or abnormal process, or of a condition or disease,” (NCI)

Categorical

- Is a specific genetic alteration present: yes or no?*
  - EGFR mutation in NSCLC.
  - BRAF in melanoma.

Continuous

- A quantitative assay (e.g. IHC) is collapsed into a binary result.
  - Estrogen receptor expression in breast cancer.
  - PD-L1 expression in NSCLC.

- PSA levels in prostate cancer.

*although there will still be sources of variability: heterogeneity within tumour over time, between labs, interpretation, etc.
From continuous to categorical: how do you define a threshold?

PD-L1 expression in NSCLC data from a phase I study of pembrolizumab (Dolled-Filhart et al., 2016)

<table>
<thead>
<tr>
<th>Percent of cells staining positive</th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>≥50%</td>
<td>19</td>
<td>25</td>
</tr>
<tr>
<td>&lt;50%</td>
<td>8</td>
<td>94</td>
</tr>
</tbody>
</table>

If a ≥50% threshold is used to predict RECIST response then the proportion of “true” responses called correctly is 19/27.

The proportion of “true” non-responses called correctly is 94/119.

50% was chosen because all other thresholds would produce a smaller score when these two proportions are added together.

- No control arm used at this stage: we can’t tell from this experiment whether this biomarker is predictive or prognostic.

- Should we be confident moving into phase III with thresholds defined in early stages of development?

- Jiang et al. (2007) suggest defining the threshold adaptively in phase III.
Drug development with a companion diagnostic

Practical perspectives of personalized healthcare in oncology

Darren R. Hodgson\textsuperscript{1}, Robert Wellings\textsuperscript{1} and Christopher Harbron\textsuperscript{2}

\textsuperscript{1}Clinical Biomarker Group, Oncology Therapy Area, Global Medicines Development, Merck\textregistered, Kenilworth, NJ 07030, USA
\textsuperscript{2}Discovery Statistics, Novartis, Basle, Switzerland

\begin{figure}
\centering
\includegraphics[width=\textwidth]{figure1}
\caption{Estimating the clinical utility of drug and biomarker. The individual contributions of candidate drug and diagnostic to risk, benefit and value cannot be determined without comparators for standard clinical practice in treatment and diagnosis, that is, without treatment of putative diagnostic 'negative' patients. Numbers for illustration only.}
\end{figure}
When is it more efficient to proceed in subset only?

**Biomarker positive population:**

\[ \log \hat{HR}_+ \sim N\left( \theta_+ , \frac{4}{e_+} \right) \quad \text{i.e.,} \quad \log \hat{HR}_+ \frac{\sqrt{e_+}}{2} \sim N\left( \frac{\theta_+}{\sqrt{2}e_+} , 1 \right) \]

**Biomarker negative population:**

\[ \log \hat{HR}_- \sim N\left( 0 , \frac{4}{e_-} \right) \]

**Full population, where \( p \) is the proportion biomarker positive and \( e \) is the total number of events:**

\[ p \log \hat{HR}_+ + (1 - p) \log \hat{HR}_- \sim N\left( p \theta_+ , \frac{4}{e} \right) \quad \text{i.e.,} \quad \left( p \log \hat{HR}_+ + (1 - p) \log \hat{HR}_- \right) \frac{\sqrt{e}}{2} \sim N\left( \frac{p \theta_+}{\sqrt{2}e} , 1 \right) \]

Therefore the number of events a non-enriched study needs, in order to have the same power as an enriched study with \( e_+ \) events, is

\[ e = e_+ / p^2 \]
When is it more efficient to proceed in subset only?

No effect in biomarker negative group

Some effect in biomarker negative group

\[ \text{e.g., } \text{HR}_{+ve} = 0.5 \text{ and } \text{HR}_{-ve} = 0.75. \]
Randomize-all design

- may have insufficient patients to assess effect in Biomarker +/-
- imbalance in Biomarker +/-

Targeted design

**Vemurafenib and BRAF V600 mutations. Melanoma – the BRIM-3 study (Chapman, 2011).**

- 2107 patients screened – 675 patients randomized.
- Vemurafenib vs. dacarbazine.
- OS hazard ratio: 0.23 (0.26,0.55)
- PFS hazard ratio: 0.26 (0.2,0.33)

Stratified design

**MARVEL: A Phase III Biomarker Validation Study of Second-Line Therapy in Patients With Advanced Non-Small Cell Lung Cancer (NSCLC)**

- Used EGFR mutation status as a biomarker.
- Initially planned to enroll 957 patients.
- Trial stopped after 23 patients randomized (slow recruitment).
- ClinicalTrials.gov
- NCT00738881

A strata maybe capped, so the All patient population is enriched for the presence or absence of the biomarker.
Adaptive design

- This strategy has been used in phase II oncology studies at AZ.
- Increased flexibility.
- Analysis more challenging.

• In an adaptive design you don’t have to use the primary endpoint to decide which population to go into at the interim analysis.

• For example, PFS is used to select population, OS is used at final analysis.

• However, not straightforward to control the type I error rate.
Possible structure of Phase 3 trial with BM

Key features of phase III biomarker study designs

<table>
<thead>
<tr>
<th>Design</th>
<th>Groups evaluated</th>
<th>Advantages</th>
<th>Disadvantages</th>
<th>Recommended use</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Enrichment</strong></td>
<td>Biomarker-positive subgroup only</td>
<td>Efficient test of treatment benefit in biomarker-positive population</td>
<td>Provides no evidence of treatment benefit in biomarker-negative patients</td>
<td>Biomarkers with very strong credentials; Convincing evidence that the benefit is limited to biomarker-positive patients</td>
</tr>
<tr>
<td><strong>Stratified designs</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Subgroup-specific</strong></td>
<td>Biomarker-positive subgroup and biomarker-negative subgroup (sequentially or in parallel)</td>
<td>Provides clear evidence of treatment benefit in the biomarker-positive population and the biomarker-negative population</td>
<td>Has suboptimal power when treatment effect is heterogeneous across subgroups</td>
<td>Biomarkers with strong credentials; Convincing evidence that the treatment is more effective in biomarker-positive than in biomarker-negative patients</td>
</tr>
<tr>
<td><strong>Biomarker-positive and overall</strong></td>
<td>Biomarker-positive subgroup and overall population (sequentially or in parallel)</td>
<td>Provides clear evidence of treatment benefit in the biomarker-positive population; Good statistical power when the treatment effect is homogeneous across subgroups</td>
<td>Does not provide a clear assessment of benefit in biomarker-negative patients; Can lead to recommendations for use of an ineffective treatment in biomarker-negative patients</td>
<td>Not recommended</td>
</tr>
<tr>
<td><strong>Marker sequential test</strong></td>
<td>Biomarker-positive and biomarker-negative subgroups, and overall population (sequentially)</td>
<td>Provides clear evidence of treatment benefit in the biomarker-positive population and in the biomarker-negative population; Good statistical power when treatment effect is homogeneous across subgroups</td>
<td>Requires a marginal sample size increase compared to subgroup-specific design</td>
<td>Biomarkers with strong credentials; Convincing evidence that the treatment is more effective in biomarker-positive than in biomarker-negative patients</td>
</tr>
<tr>
<td><strong>Fallout</strong></td>
<td>Overall population and biomarker-positive subgroup (sequentially)</td>
<td>Enables statistically rigorous evaluation of treatment benefit in the biomarker-positive patients, if no benefit is detected in the overall population</td>
<td>Does not provide clear assessment of benefit in the biomarker-negative patients</td>
<td>Biomarkers with weak credentials; Treatment assumed to probably benefit all patients and/or biomarker-positive group is only assessed if no benefit is shown in overall population</td>
</tr>
</tbody>
</table>

Enrichment and stratified designs: discussion

Scientific/Statistical

- Enrichment/targetted design is simple to analyse.
- Stratified design makes it possible to assess the predictive ability of the biomarker.

Practical

- Enrichment studies will have slow enrolment when the prevalence of biomarker +ve population is low.
- Stratified studies can have a similar problem: the biomarker negative population may finish recruitment a lot earlier than the biomarker positive.

These practical issues can be mitigated by testing several targeted agents at the same time in “multi-drug basket” or “umbrella” studies…
Basket studies

**Traditional**

- 1 site of disease
- 1 Mutation
  - Targeted agent

**Basket**

- Many sites
- 1 Mutation
  - Targeted agent

**Multidrug basket**

- Many sites
- Many Mutation
  - Targeted agent

*Vemurafenib and BRAF V600 mutations. Many sites (excluding melanoma).*

*NCI Match*
**Basket studies: design**

Typically single arm studies, with primary endpoint of objective response rate and/or duration of response.

Analysis might (or might not) lump together patients with different sites of disease.

*Vemurafenib and BRAF (non-melanoma).*

- 6 substudies defined by site (NSCLC, breast, etc.)
- In each substudy:
  - Adaptive Simon’s 2-stage design.
  - Between 7 and 19 patients.
  - Aim is to reject a null hypothesis that ORR is 15%.

*NCI Match*

- At least 22 substudies defined by drug target.
- In each substudy:
  - Single-stage design.
  - 31 evaluable subjects.
  - Aim is to reject null hypothesis that ORR is 5%.
**Umbrella studies**

**Enrichment**
- 1 site
- 1 Mutation
  - Targeted agent vs. control

**Umbrella**
- 1 site
- Many Mutation
  - Targeted agent vs. control

*Vemurafenib and BRAF V600 mutations.*
*Melanoma.*

*Lung MAP*
Umbrella studies: design

- Typically, a randomized comparison.
- Primary endpoint PFS or OS.

*Lung MAP*

- 5 substudies based on mutation status.
- In each substudy:
  - Phase II portion has a target PFS hazard ratio of 0.5
  - Go/No-go performed after 55 events.
  - Phase III portion has a target OS hazard ratio of 0.67.
  - Final analysis performed after 256 deaths. (90% power with 2.5% type I error rate).

<table>
<thead>
<tr>
<th>Mutation</th>
<th>Prevalence</th>
</tr>
</thead>
<tbody>
<tr>
<td>PI3K</td>
<td>6-8%</td>
</tr>
<tr>
<td>CDK4/6</td>
<td>12%</td>
</tr>
<tr>
<td>FGFR</td>
<td>9%</td>
</tr>
<tr>
<td>C-MET</td>
<td>16%</td>
</tr>
<tr>
<td>Nonmatch</td>
<td>56%</td>
</tr>
</tbody>
</table>
Lung MAP protocol

- **PI3K** – PIK3CA mutation
- **CDK4/6** – CCND1 amplification, Cdk6 deletion and mutation
- **FGFR** – FGFR amplification, mutation, fusion
- **HGF** – c-Met expression

**Note:**
- CT – chemotherapy (docetaxel or gemcitabine)
- * – Archival formaldehyde fixed-paraffin embedded tumor, fresh core needle biopsy if needed
Basket and umbrella studies: discussion

• Just like enrichment designs, they don’t directly address the predictive ability of a biomarker.

• Many studies so far have used a default design for each sub-study, with limited statistical justification.

• Should analysis lump together patients who have the same mutation but different sites of disease? Or should separate analyses be performed? Ongoing methodology work here: LeBlanc (2009), Simon (2014).

• Mismatches in recruitment rates between substudies is not eliminated and is still difficult in practice.


Acknowledgments

Dominic Magirr for his assistance in developing the slides

Amir Handzel
Questions?
Confidentiality Notice

This file is private and may contain confidential and proprietary information. If you have received this file in error, please notify us and remove it from your system and note that you must not copy, distribute or take any action in reliance on it. Any unauthorized use or disclosure of the contents of this file is not permitted and may be unlawful. AstraZeneca PLC, 1 Francis Crick Avenue, Cambridge Biomedical Campus, Cambridge, CB2 0AA, UK, T: +44(0)203 749 5000, www.astrazeneca.com