Finding the (biomarker-defined) subgroup of patients who benefit from a novel therapy: no time for a game of hide and seek

Lisa M McShane, PhD
Biometric Research Program,
Division of Cancer Treatment and Diagnosis
National Cancer Institute

April 12, 2022
DISCLOSURES

I have no financial relationships to disclose.

- and -

I will not discuss off label use and/or investigational use in my presentation.

- and -

The views expressed represent my own and do not necessarily represent the views or policies of the National Cancer Institute, National Institutes of Health, or the U.S. Department of Health and Human Services.

Having worked in oncology for the last 25+ years, the examples in my talk will all relate to cancer, but hopefully it will be clear how the statistical principles apply more generally.
Drug development in the era of biomarker-driven precision medicine

• Q1: Does the drug benefit any patients?
• Q2: If the drug does not benefit entire patient population, is there a subset that it does benefit?
• Q3: If the drug benefits only a subset, is there a biomarker (or “signature”) that defines the subset?
  • “Predictive” (therapy selection / treatment effect modifier) biomarker
• Q4: If a biomarker is needed, how do we measure it?
An optimal treatment-selection biomarker test might not be ready when a new therapeutic is ready for evaluation in a clinical trial

- Insufficient understanding of biology or mechanism of action of drug to confidently identify a biomarker or signature
- Not sure how to best measure the biomarker
- For quantitative biomarkers, unsure of best clinical cut-off
- Difficulties developing a robust, reproducible assay
Clinical trial design considerations when faced with uncertainty about a biomarker-defined subgroup most likely to benefit

• Extracting a candidate biomarker from the literature or preliminary data

• Biomarker assay reproducibility

• Statistical design and analysis considerations for prospective biomarker-based subgroup testing in a definitive trial of a new therapeutic
Clinical trial design considerations when faced with uncertainty about a biomarker-defined subgroup most likely to benefit

- Extracting a candidate biomarker from the literature or preliminary data
- Biomarker assay reproducibility
- Statistical design and analysis considerations for prospective biomarker-based subgroup testing in a definitive trial of a new therapeutic
Extracting a candidate biomarker from the literature or preliminary data

• **Preclinical studies** – how to extrapolate biomarker results from preclinical models to humans, lack of preclinical models (e.g., for immunotherapies)

• **Phase I/II clinical trials** - small samples sizes, research grade biomarker assays, earlier endpoints (e.g., tumor response), heterogeneous patients (e.g., different tumor types), often non-randomized

• **Retrospective biomarker evaluation on “convenience” specimens** – generally heterogeneous patients & treatments, poor study design and analysis, unreliable/incomplete data, biomarker assays not sufficiently described

• **Retrospective biomarker evaluation on specimens from similar completed trials** – possibly different cancer types or disease stages, or different therapies in same class, different version of the biomarker assay
Tumor Mutational Burden (TMB) cut-offs, association with overall survival, by tumor type

Samstein et al. *Nat Genet* 2019;51:202-206 (Figure 2); MSKCC institutional case series

- **TMB =** nonsynonymous mutational load/burden (mutations per megabase) by MSK-IMPACT assay using both tumor and germline DNA
- Cut-off defined as top 20%-tile of TMB for each tumor type, which varied widely
- Observed positive prognostic effect of TMB-High vs. TMB-Low on overall survival after ICI treatment for most tumor subtypes and all drug classes
- Two-sided log-rank P value for the comparison of TMB-High and TMB-Low survival curves < 0.05 for 4/11 tumor types (influenced by sample size)

---

**2 subgroup factors:** tumor type and TMB (High vs. Low)
Tumor Mutational Burden (TMB) association with tumor response, by FoundationOne CDx


- TMB = DNA base substitution mutations (including synonymous) per megabase, by FoundationOne CDx assay
- Prespecified definition of TMB-High was TMB ≥ 10 across all tumor types
- Objective response rate (ORR, pooled across tumor types) is 6% (95% CI 5-8%) for TMB-Low vs. 29% (95% CI 21-39%) for TMB-High
- In TMB-H, ORRs by tumor type range from 0% (no TMB-High) to ≈47% (discounting 2/2 = 100% in thyroid)
- In TMB-L, ORRs by tumor type range from ≈3% to ≈12%
- Supported tumor agnostic FDA approval for pembrolizumab with selection by FoundationOne CDx

Multicohort, open-label, phase 2 KEYNOTE-158 study of pembrolizumab
Tumor Mutational Burden (TMB) association with tumor response, by MSK-IMPACT

Valero et al. JAMA Oncology 2021;7(5):739-743

**Table. Distribution of Patients According to Tumor Mutational Burden by Cancer Type**

<table>
<thead>
<tr>
<th>Cancer type</th>
<th>Overall No. of patients</th>
<th>TMB-L (&lt;10 mutations per Mb)</th>
<th>TMB-H (≥10 mutations per Mb)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Patients, No. (%)</td>
<td>Response rate, %</td>
<td>Patients, No. (%)</td>
</tr>
<tr>
<td>NSCLC</td>
<td>663</td>
<td>459 (69)</td>
<td>204 (31)</td>
</tr>
<tr>
<td>Melanoma</td>
<td>214</td>
<td>101 (47)</td>
<td>113 (53)</td>
</tr>
<tr>
<td>Kidney</td>
<td>92</td>
<td>92 (100)</td>
<td>0</td>
</tr>
<tr>
<td>Sarcoma</td>
<td>84</td>
<td>80 (95)</td>
<td>4 (5)</td>
</tr>
<tr>
<td>Bladder</td>
<td>82</td>
<td>56 (68)</td>
<td>26 (32)</td>
</tr>
<tr>
<td>Head and neck</td>
<td>74</td>
<td>61 (82)</td>
<td>13 (18)</td>
</tr>
<tr>
<td>Gastric</td>
<td>67</td>
<td>62 (93)</td>
<td>5 (7)</td>
</tr>
<tr>
<td>SCLC</td>
<td>54</td>
<td>32 (59)</td>
<td>22 (41)</td>
</tr>
<tr>
<td>Hepatobiliary</td>
<td>53</td>
<td>49 (92)</td>
<td>4 (8)</td>
</tr>
<tr>
<td>Colorectal</td>
<td>50</td>
<td>43 (86)</td>
<td>7 (14)</td>
</tr>
<tr>
<td>Endometrial</td>
<td>47</td>
<td>44 (94)</td>
<td>3 (6)</td>
</tr>
<tr>
<td>Esophageal</td>
<td>45</td>
<td>41 (91)</td>
<td>4 (9)</td>
</tr>
<tr>
<td>Pancreatic</td>
<td>36</td>
<td>34 (94)</td>
<td>2 (6)</td>
</tr>
<tr>
<td>Mesothelioma</td>
<td>35</td>
<td>34 (97)</td>
<td>1 (3)</td>
</tr>
<tr>
<td>Ovarian</td>
<td>29</td>
<td>28 (97)</td>
<td>1 (3)</td>
</tr>
<tr>
<td>Unknown primary</td>
<td>28</td>
<td>23 (82)</td>
<td>5 (18)</td>
</tr>
<tr>
<td>Breast</td>
<td>25</td>
<td>23 (92)</td>
<td>2 (8)</td>
</tr>
<tr>
<td>Total</td>
<td>1678</td>
<td>1262 (75)</td>
<td>416 (25)</td>
</tr>
</tbody>
</table>

- MSKCC institutional case series of 1678 patients included 16 tumor types treated with anti-PD-1 or anti-PD-L1 therapy
- TMB = nonsynonymous mutational burden by MSK-IMPACT assay (mutations per megabase)
- TMB-High was pre-specified as TMB ≥ 10 across all tumor types
- Response rates generally higher in TMB-High subgroup (11/16 tumor types, excluding unknown). However, high-TMB proportion and magnitude of association between TMB and response rates varied widely across tumor types.
Clinical trial design considerations when faced with uncertainty about a biomarker-defined subgroup most likely to benefit

- Extracting a candidate biomarker from the literature or preliminary data
- Biomarker assay reproducibility
- Statistical design and analysis considerations for prospective biomarker-based subgroup testing in a definitive trial of a new therapeutic
Different bioinformatic pipelines and algorithms produce variable results (and may change over time)

Example: Tumor Mutational Burden (TMB) by 17 different tests
Variability due to different assay methods and bioinformatic pipelines

Ki67 reproducibility across laboratories using different staining and scoring methods

Ki67 (% positive invasive tumor cells), 8 labs assessing different TMA sections, same set of 100 breast tumors

Centrally stained, locally scored
Median range: 10% to 28%
ICC: 0.71, 95% CI=(0.47,0.78)

Locally stained, locally scored
Median range: 5% to 33%
ICC: 0.59, 95% CI=(0.37,0.68)


FDA approval summary for abemaciclib with endocrine therapy for high-risk early breast cancer

• Indication for abemaciclib + endocrine therapy limited to patients with ≥ 4 pathologic positive axillary lymph nodes (pALN) or 1-3 pALN and tumor histologic grade 3 and/or tumor size ≥ 50 mm whose tumors had Ki-67 ≥ 20%

• FDA simultaneously approved Ki-67 companion diagnostic (CDx) Ki67 MIB-1 IHC pharmDx (Dako Omnis, Carpinteria) that was used in trial.

• Concerns that labs likely to use home brew Ki67 assays rather than the approved CDx!

Royce et al. J Clin Oncol 2022; online
Clinical trial design considerations when faced with uncertainty about a biomarker-defined subgroup most likely to benefit

• Extracting a candidate biomarker from the literature or preliminary data

• Biomarker assay reproducibility

• Statistical design and analysis considerations for prospective biomarker-based subgroup testing in a definitive trial of a new therapeutic
Statistical design and analysis considerations for prospective biomarker-based subgroup testing

• To enrich or not enrich (with or without subsequent subgroup testing)

• Managing multiple subgroups
  • Pre-specified vs. post hoc
  • Type I error control
  • Nested vs. adjacent subgroup testing
Clinical trial enrichment

“Prospective use of any patient characteristic to select a study population in which detection of a drug effect (if one is in fact present) is more likely than it would be in an unselected population”

- Reduce inter-patient and intra-patient heterogeneity
- Prognostic enrichment strategies
  - More events $\Rightarrow$ more statistical power
- Predictive enrichment strategies – choosing patients more likely to respond to the drug treatment (e.g., use of treatment-selection biomarker)

Considerations in use of a biomarker for up-front clinical trial enrichment

**Precision medicine**

- Patients who benefit from new therapy
- Patients who do not benefit from new therapy

**Biomarker-defined subgroup**

**Biomarker very useful for enrichment**

- Patients who benefit from new therapy
- Patients who do not benefit from new therapy

**Biomarker-defined subgroup**

**Biomarker minimally useful for enrichment**

- Patients who benefit from new therapy
- Patients who do not benefit from new therapy

**Biomarker-defined subgroup**

**Trial extremely challenging even with a good enrichment biomarker**

- Patients who do not benefit from new therapy

**Biomarker-defined subgroup**
Principles for statistical management of multiple subgroups in a definitive clinical trial

• Most rigorous analysis pre-specifies subgroups (be judicious in choice and number) and controls overall type I error (e.g., partition $\alpha$)

• Adaptations to subgroup testing or enrichment after trial initiation may be needed (e.g., due to external data), but **must be made blinded** to accruing outcome data


• **Hierarchical testing of treatment effect in “biomarker-positive” subgroup followed by testing in overall (ITT) population is a generally flawed approach.**

Hierarchical testing example 1: Checkmate 649

• Randomized, open-label, phase 3 trial

• First-line, nivolumab plus chemotherapy versus chemotherapy alone for advanced gastric, gastro-oesophageal junction, and oesophageal adenocarcinoma

• Dual primary endpoints, two-sided significance levels (type I error) of 0·03 allocated to OS and 0·02 to PFS. Upon superiority of OS in patients with a PD-L1 CPS* ≥ 5, OS was hierarchically tested in patients with a PD-L1 CPS ≥ 1 with a fraction of α (50% α transmitted=0·015), followed by all randomly assigned patients (ITT, 100% α transmitted=0·015).

*CPS = combined positive score by Dako PD-L1 immunohistochemistry 28-8 pharmDx assay (Dako, Santa Clara, CA, USA)
### Hierarchical testing example: Checkmate 649

**Overall survival (OS)**  

<table>
<thead>
<tr>
<th>Group</th>
<th>Nivo + chemo</th>
<th>Chemo alone</th>
<th>12-mo OS (95% CI)</th>
<th>Med OS (mo) (95% CI)</th>
<th>12-mo OS (95% CI)</th>
<th>Med OS (mo) (95% CI)</th>
<th>HR</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPS ≥ 5</td>
<td>n 473</td>
<td>n 482</td>
<td>57% (53–62)</td>
<td>14.4 (13.1–16.2)</td>
<td>46% (42–51)</td>
<td>11·1 (10.0–12.1)</td>
<td>HR 0.71</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(95% CI)</td>
<td>(95% CI)</td>
<td>(95% CI)</td>
<td>(95% CI)</td>
<td>CI (0.59–0.86)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>12-mo OS</td>
<td>Med OS (mo)</td>
<td>HR 0.77</td>
<td>99.3% CI (0.64–0.92)</td>
<td>p&lt;0.0001</td>
</tr>
<tr>
<td>CPS ≥ 1</td>
<td>641</td>
<td>655</td>
<td>56% (52–59)</td>
<td>14.0 (12.6–15.0)</td>
<td>47% (43–51)</td>
<td>11.3 (10.6–12.3)</td>
<td>HR 0.77</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(95% CI)</td>
<td>(95% CI)</td>
<td>(95% CI)</td>
<td>(95% CI)</td>
<td>CI (0.64–0.92)</td>
</tr>
<tr>
<td>ITT</td>
<td>789</td>
<td>792</td>
<td>55% (51–58)</td>
<td>13.8 (12.6–14.6)</td>
<td>48% (44–51)</td>
<td>11.6 (10.9–12.5)</td>
<td>HR 0.80</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(95% CI)</td>
<td>(95% CI)</td>
<td>(95% CI)</td>
<td>(95% CI)</td>
<td>CI (0.68–0.94)</td>
</tr>
</tbody>
</table>

**ASCO 2021 presentation (abstract #4002) showed HR=0.94 for CPS < 5**
### Hierarchical testing example: Checkmate 649

**Progression free survival (PFS)**


<table>
<thead>
<tr>
<th>Group</th>
<th>Nivo + chemo</th>
<th>12-mo PFS (95% CI)</th>
<th>Med PFS (mo) (95% CI)</th>
<th>Chemo alone</th>
<th>12-mo PFS (95% CI)</th>
<th>Med PFS (mo) (95% CI)</th>
<th>HR CI p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPS ≥ 5</td>
<td>473</td>
<td>57% (53–62)</td>
<td>7.7 (7.0–9.2)</td>
<td>482</td>
<td>46% (42–51)</td>
<td>6.0 (5.6–6.9)</td>
<td>HR 0.68 98% CI (0.56–0.81) p&lt;0.0001</td>
</tr>
<tr>
<td>CPS ≥ 1</td>
<td>641</td>
<td>56% (52–59)</td>
<td>7.5 (7.0–8.4)</td>
<td>655</td>
<td>47% (43–51)</td>
<td>6.9 (6.1–7.0)</td>
<td>HR 0.74 95% CI (0.65–0.85) (not tested)</td>
</tr>
<tr>
<td>ITT</td>
<td>789</td>
<td>55% (51–58)</td>
<td>7.7 (7.1–8.5)</td>
<td>792</td>
<td>48% (44–51)</td>
<td>6.9 (6.6–7.1)</td>
<td>HR 0.77 95% CI (0.68–0.87) (not tested)</td>
</tr>
</tbody>
</table>

ASCO 2021 presentation (abstract #4002) showed HR=0.93 for CPS < 5
Hierarchical testing example 2: MonarchE Trial

Randomized, multi-center, open-label, two-cohort phase 3 trial comparing abemaciclib (CDK4/6 inhibitor) plus endocrine therapy (ET) to ET alone in patients with hormone-receptor-positive (HR+) human epidermal growth factor receptor-negative (HER2-) early breast cancer (EBC) at high risk of disease recurrence on the basis of clinical or pathologic features or Ki-67 score.

- **Cohort 1**: patients with \( \geq 4 \) pathologic positive axillary lymph nodes (pALN) or 1-3 pALN and tumor histologic grade 3 and/or tumor size \( \geq 50 \) mm

- **Cohort 2**: patients with 1-3 pALN and Ki-67 score \( \geq 20\%


Harbeck et al. *Annals of Oncology* 2021;32(12):1571-1581
Hierarchical testing example 2: MonarchE Trial

**Statistical plan**

- Study powered to test the intent-to-treat (ITT) population (Cohort 1 + Cohort 2) for IDFS.
- Gated hierarchical testing strategy included three additional end points:
  - IDFS in patients with Ki-67 score $\geq 20\%$ from cohorts 1 and 2
  - IDFS in patients with Ki-67 score $\geq 20\%$ from cohort 1 alone
  - OS in the ITT population
- Two interim analyses and one final efficacy analysis for IDFS were planned, as well as two interim analyses and one final OS analysis
- Two subgroup factors: cohort 1 vs. 2 and Ki67 ($\geq 20\%$ vs. $< 20\%$)

<p>| TABLE 1. Timing of Prespecified IDFS and OS Analyses in monarchE to Date |</p>
<table>
<thead>
<tr>
<th>Prespecified Analysis</th>
<th>Target/Actual No. of Events</th>
<th>Timing</th>
<th>End Points</th>
<th>Meeting</th>
<th>Statistical Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interim IDFS 1 (IA1)</td>
<td>195/203</td>
<td>September 27, 2019</td>
<td>IDFS in ITT</td>
<td>(p &lt; 0.026)</td>
<td></td>
</tr>
<tr>
<td>Interim IDFS 2 (IA2)</td>
<td>293/323</td>
<td>March 16, 2020</td>
<td>IDFS in ITT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Final IDFS</td>
<td>390/395</td>
<td>July 8, 2020</td>
<td>IDFS in Ki-67 $\geq 20%$ (C1 and C2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>IDFS in Ki-67 $\geq 20%$ (C1)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>OS Interim 1 (OS IA1)$^a$</td>
<td>—/186</td>
<td>April 1, 2021</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Abbreviations: C1, cohort 1; C2, cohort 2; IA, interim analysis; IDFS, invasive disease-free survival; ITT, intent-to-treat; OS, overall survival.

$^a$Additional OS analyses will be conducted.
Hierarchical testing example 2: MonarchE Trial

<table>
<thead>
<tr>
<th>Prespecified analysis</th>
<th>Endpoints meeting statistical significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interim IDFS 1 (IA1)</td>
<td></td>
</tr>
<tr>
<td>Interim IDFS 2 (IA2)</td>
<td>IDFS in ITT</td>
</tr>
<tr>
<td>Final IDFS</td>
<td>IDFS in Ki67 ≥ 20% (Cohorts 1+2)</td>
</tr>
<tr>
<td></td>
<td>IDFS in Ki67 ≥ 20% (Cohort 1)</td>
</tr>
<tr>
<td>OS Interim 1 (OS IA1)</td>
<td></td>
</tr>
</tbody>
</table>

FDA approved use of abemaciclib + ET in patients at high risk of recurrence meeting the monarchE cohort 1 criteria and whose tumors are Ki-67 ≥ 20% on the basis of the simultaneously approved CDx (Ki67 MIB-1 IHC pharmDx (Dako Omnis, Carpinteria)).

Royce et al. J Clin Oncol 2022; online Jan 27, 2022

- Despite IDFS in ITT statistically significant favoring abemaciclib + ET at all analyses, **ITT OS analysis at all time points (IA2, final IDFS, and OS IA1 data cut-offs) favored ET only arm.**

- In both final IDFS analysis and OS IA1, **remaining subgroups in hierarchy (Ki-67 ≥ 20%),** each had statistically significant IDFS results deepening with time and **OS HRs that numerically favored the abemaciclib plus ET arm** with HR < 1.

- However, cohort 2’s enrollment began approximately 11 months after cohort 1, and cohort 2 alone represented only 9% of total patients (with few events); **statistically significant IDFS improvement in Ki-67 ≥ 20% population (cohorts 1 and 2 combined) was driven by cohort 1.**
Recommended alternatives to hierarchical approach


• For conclusions of favorable efficacy in biomarker-positive subgroup to be extrapolated to include the biomarker-negative subjects
  • Address the multiplicity caused by subgroup testing
  • Sufficient data to obtain reliable treatment effect estimate in biomarker negative subgroup
  • Estimated treatment effect in biomarker negative subgroup should be clinically relevant and at least as large as needed to achieve “statistical significance” in an ITT analysis

• When no expectation of large difference in treatment effects across subgroups, primary analysis should be limited to the entire ITT population
  • Subgroups can be investigated in an exploratory manner

• When genuine uncertainty (equipoise) about whether biomarker associated with magnitude of treatment effect, biomarker stratified design with early futility analysis for biomarker negative subgroup may be warranted
Recommended graphical displays

When subgroups defined by cut-off(s) on a continuous biomarker, examine treatment effect on the biomarker continuum

**Predictiveness curves**


**Subgroup Treatment Effect Pattern Plot (STEPP)**

Lazar et al. *J Clin Oncol* 2010; 28: 4539-4544

Each subpopulation in sliding window contains \( \approx 150 \) patients and \( \approx 50 \) overlapping patients. Solid black lines indicate overall treatment effect, and dotted black lines indicate no effect.
Summary remarks

• Precision medicine brings statistical challenges of increasing number and decreasing size of subgroups; hopefully small subgroup sample sizes are counterbalanced by considerably larger magnitudes of treatment effect.

• Conclusions about treatment effects in subgroups should be based on careful consideration of multiple factors or evidence sources
  • Requires understanding various biomarkers, assays, therapeutic agents, and clinical populations
  • Delicate balance between consideration of multiple factors and pitfalls of post hoc analyses
    • Amatya AK et al. Clinical Cancer Research 2021;27(21):5753-5756

• Need complete and transparent reporting of any studies involving biomarkers
  • Present treatment effects in disjoint/adjacent biomarker-based subgroups rather than nested
  • For continuous biomarkers, present treatment effects along the biomarker continuum
  • Completely describe biomarkers or algorithms for identifying subgroups to allow assessment of comparability across studies and reproducibility in practice
THANK YOU