Assessing New Quantitative Imaging Biomarkers

Chaya Moskowitz, Ph.D.
Department of Epidemiology and Biostatistics
Memorial Sloan Kettering Cancer Center

With thanks to Todd Alonzo and Alicia Toledano
O’Connor, J. P. B. et al. (2016) Imaging biomarker roadmap for cancer studies
O'Connor, J. P. B. et al. (2016) Imaging biomarker roadmap for cancer studies
Some Goals of Early Phase Studies

- **Analytic validity**
  - Does the imaging biomarker measure what it is supposed to measure?

- **Clinical validity**
  - Is the imaging biomarker associated with the clinical (patient) outcome?
Analytic Validity

• Early phase studies
  – Preclinical, laboratory studies
  – Early clinical development

• Study endpoints and metrics
  – Analytic accuracy
    • Analytic sensitivity, specificity, predictive values, ROC curves
  – Reliability
    • Repeatability, reproducibility
Reliability (Precision) vs. Accuracy

A
One accurate hit, Not reliable

B
Not accurate, Reliable

C
Not accurate, Not reliable

D
Accurate, Reliable

• Potential utility of an imaging biomarker can be greatly impacted by lack of reliability

• Poor reliability can make measured change in parameter difficult to interpret

• Developing reliable biomarkers can be difficult

• Acceptable magnitude depends on use
  – Strong agreement is a necessary component of any subjective procedure intended for diagnostic use
Sources of Variability

• Patient-related
  – Disease or treatment-related
  – Other biophysiological sources

• Imaging system-related
  – Scanner-related
  – Reader-related
Repeatability and Reproducibility

• **Repeatability**: consistency of results when same biomarker is assessed at short intervals on same subjects using **same** equipment, **same** reader, in **same** center

• **Reproducibility**: consistency of results when same biomarker is assessed at short intervals on same subjects using **different** equipment, **different** reader, or in **different** centers

Barnhart and Barboriak, *Translational Oncology* (2009)
Study Designs

• **Repeatability:**
  \[ K \] repeated measurements \((K \geq 2)\) on \(n\) subjects
  – Identical conditions
  – Test-retest, “coffee-break studies”, intra-observer variability

• **Reproducibility:**
  \[ K \] methods/readers measure \((K \geq 2)\) \(n\) subjects
  – Vary component(s) systematically
  – Method comparison, inter-observer variability
• **Descriptive statistics**  
  – Means, variances, correlations

• **Plots**  
  – Pairwise scatter plots  
    • Plot of difference vs average  
    • Mean difference  
    • 95% Limits of Agreement (mean difference ± 2 x standard deviation)

• **Primary metrics usually rely on:**  
  – Absolute differences between measurements  
  – Components of variance

Assessing Repeatability

- Frequently based on within-subject standard deviation, $\sigma_w$
  - Repeatability coefficient: $RC = 2.77 \sigma_w$
  - Repeatability limit: \((-RC, RC)\)
  - Interpretation: interval within which any two readings by same reader would fall for 95% of subjects
Primary Aim: Determine the test-retest performance, assessed by the RC of $K^{\text{trans}}$ and IAUGC90$^{\text{bn}}$ and measured by median pixel values of the whole prostate.
Examples of Repeatability Studies

Evaluating Variability in Tumor Measurements from Same-day Repeat CT Scans of Patients with Non–Small Cell Lung Cancer

• On same scale as measurements
• Relative difference vs. simple difference

Zhao et al., Radiology (2009)
Assessing Reproducibility

• Frequently based on between-subject standard deviation, $\sigma_B$
  – Intraclass correlation coefficient
    • ICC = $\sigma^2_B / (\sigma^2_B + \sigma^2_W)$
    • Interpretation: Proportion of total variance due to the different readers/methods
  – Concordance correlation coefficient (Lin, Biometrics (1989))
    • $\rho_c = (2\sigma_{X_1 X_2}) / (\sigma^2_{X_1} + \sigma^2_{X_2} + (\mu_{X_1} - \mu_{X_2})^2)$
    • Interpretation: Quantifies agreement between two measurements
Examples of Reproducibility Studies

Reproducibility of Measurement of Apparent Diffusion Coefficients of Malignant Hepatic Tumors: Effect of DWI Techniques and Calculation Methods

Interobserver Agreement for ADC Measurement Presenting With ICC

<table>
<thead>
<tr>
<th></th>
<th>Breath-hold DWI</th>
<th>Respiratory-triggered DWI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ICC</td>
<td>LOA(^a)</td>
</tr>
<tr>
<td>First</td>
<td>ADC(_{0/500})</td>
<td>0.979 (0.921-0.993)</td>
</tr>
<tr>
<td></td>
<td>ADC(_{50/500})</td>
<td>0.974 (0.934-0.990)</td>
</tr>
<tr>
<td>Second</td>
<td>ADC(_{0/500})</td>
<td>0.983 (0.942-0.994)</td>
</tr>
<tr>
<td></td>
<td>ADC(_{0/1000})</td>
<td>0.964 (0.911-0.986)</td>
</tr>
<tr>
<td>First</td>
<td>ADC(_{0/500})</td>
<td>0.974 (0.934-0.990)</td>
</tr>
<tr>
<td></td>
<td>ADC(_{0/1000})</td>
<td>0.803 (0.555-0.919)</td>
</tr>
</tbody>
</table>

Numbers in parentheses are 95% confidence interval.

Kim et al., J of Magnetic Resonance Imaging (2012)
Other Considerations

• Many other possible methods

• Continuous vs. categorical data
  – Contributes to choice of metrics
  – Kappa statistics for categorical data

• Estimation rather than testing
  – P-values less interesting
  – Confidence intervals

• Studies designed to evaluate both repeatability and reproducibility
Clinical Validity

• Mid-phase studies
  – Clinical studies
  – Retrospective and prospective

• Study endpoints and metrics
  – Clinical sensitivity, specificity, predictive values, ROC curves
  – Risk of the patient outcome for people with or without the imaging biomarker
Examples of Patient Outcomes

- Presence or absence of disease
- Tumor response rate
- Time to recurrence
- Progression free survival
- Disease free survival
- Overall survival
Survival Analysis

• Statistical methods for analyzing data where the outcome is the time to an event

• Applicable for data from single-arm clinical trials, randomized clinical trials, and cohort studies

• Important for:
  – Studies where not all patients enter at the same time (staggered entry)
  – Data analyzed before all patients have experienced the outcome (censoring)
Censoring

• Exact time event occurs is not known

• Different type of censoring:
  – Right censoring: event has not yet occurred
    • Most common type of censoring
    • Examples: study ends or patients are lost to follow-up
  – Interval censoring: event occurred between two time-points, but we don’t know exactly when
    • Examples: outcome occurs between two scheduled follow-up visits
  – Left censoring: Event occurs before the study starts
    • Not usually found in clinical trials
Survival Data Example
• Primary interest:
  \( T = \) the time until the event

• Instead observe:
  \( C = \) the time at which an observation is censored
  \( Y = \min(T, C) \)
  \( \delta = 1 \) if the observation is censored, 0 otherwise

• **Don’t throw away information on \( C \)!**

• **Record and give to your statistician: \( \delta \) and \( Y \)**
  – Example 1: Recurrence yes/no and time to recurrence or last follow-up
  – Example 2: Recurrence yes/no and date of recurrence or last follow-up
Survival Function

• $S(t) = \text{Prob}(T > t)$

• Interpretation: Probability an individual experiences the event after time $t$; probability of surviving beyond time $t$.

• Starts at 1 and decreases towards 0:
  – $S(t) = 1$ for $t=0$, $S(t) = 0$ for $t=\infty$

• Nonincreasing function
Hazard Function

• \( h(t) = \lim_{\Delta t \to 0} \frac{\text{Prob}(t \leq T \leq \Delta t + t \mid T \geq t)}{\Delta t} \)

• Interpretation: Probability of experiencing the event in the next instant given survival without the event until time \( t \)

• Also called hazard rate, instantaneous failure rate, age-specific failure rate

• Mathematically related to the survival function

• Can have many shapes but can never be negative (\( h(t) \geq 0 \))
Estimating the Survival Function

• To estimate $S(t) = \text{Prob}(T>t)$, why not just take the proportion of people with event times greater than $t$?
  – Ignores censoring

• Two main ways:
  – Parametric estimate
    • Assumes the times-to-event follow a particular probability distribution function
  – Non-parametric estimate
    • Empirical estimate
Kaplan-Meier Estimate

\[ \hat{S}_{KM}(t) = \prod_{t_i \leq t} \left( 1 - \frac{D_i}{N_i} \right) \]

- \( N_i \) = Number of people at risk of having the event at the \( i^{th} \) time
- \( D_i \) = Number of people having the event at the \( i^{th} \) time.

- Product-Limit estimator

- Most frequently used method for estimating the survival function

- Step function with jumps at the event times
Kaplan-Meier Estimate

To obtain the K-M estimate:

- Order the times from shortest to largest.
- At start of study, \( t_0 \), no one has had event, \( \hat{S}_{KM}(t) = 1 \).
- At each time, calculate

\[
\hat{S}_{KM}(t_j) = \left(1 - \frac{D_j}{N_j}\right) \times \hat{S}_{KM}(t_{j-1}) \tag{1}
\]

- Plot \( \hat{S}_{KM}(t) \) against \( t \)

<table>
<thead>
<tr>
<th>( i )</th>
<th>( t_i )</th>
<th>( D_i )</th>
<th>( N_i )</th>
<th>( (1) )</th>
<th>( \hat{S}_{KM}(t) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>10</td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>5</td>
<td>1</td>
<td>10</td>
<td>(1-1/10) \times 1</td>
<td>0.90</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>1</td>
<td>9</td>
<td>(1-1/9) \times 0.90</td>
<td>0.80</td>
</tr>
<tr>
<td>3</td>
<td>10</td>
<td>1</td>
<td>8</td>
<td>(1-1/8) \times 0.80</td>
<td>0.70</td>
</tr>
<tr>
<td>4</td>
<td>12</td>
<td>0</td>
<td>---</td>
<td></td>
<td>---</td>
</tr>
<tr>
<td>5</td>
<td>15</td>
<td>1</td>
<td>6</td>
<td>(1-1/6) \times 0.70</td>
<td>0.58</td>
</tr>
<tr>
<td>6</td>
<td>16</td>
<td>0</td>
<td>---</td>
<td></td>
<td>---</td>
</tr>
<tr>
<td>7</td>
<td>17</td>
<td>0</td>
<td>---</td>
<td></td>
<td>---</td>
</tr>
<tr>
<td>8</td>
<td>19</td>
<td>1</td>
<td>3</td>
<td>(1-1/3) \times 0.58</td>
<td>0.39</td>
</tr>
<tr>
<td>9</td>
<td>20</td>
<td>1</td>
<td>2</td>
<td>(1-1/2) \times 0.39</td>
<td>0.19</td>
</tr>
<tr>
<td>10</td>
<td>22</td>
<td>0</td>
<td>---</td>
<td></td>
<td>---</td>
</tr>
</tbody>
</table>
Caveats

• Assume probability an observation is censored is unrelated to the probability of having an event
  — Uninformative censoring

• Estimates can be unstable at the tail of the Kaplan-Meier curve when the number of patients remaining at risk gets small

• If the last observation is censored the Kaplan-Meier estimate will not reach 0.
Comparing Two Survival Functions

- $H_0: S_1(t) = S_2(t)$

- Can use:
  - Log-rank test
    - Most frequently used test
    - Takes the whole follow-up period into account
    - Most powerful against consistent differences
  - Modified Wilcoxon test
    - Most powerful against early differences

- State in advance what test you will use

- Sample size/power depends on the number of events
Time-Dependent ROC Curves

- Disease status changes over time
- ROC curves that change as a function of time
- Can define based on the survival function

Heagerty et al. Biometrics 2000
• Treat sensitivity and specificity as time-dependent functions and use Bayes theorem:

\[
\text{Sensitivity}(c, t) = \text{Prob}(X > c \mid D(t) = 1) = \frac{\{1 - S(t \mid X > c)\} P(X > c)}{1 - S(t)}
\]

\[
\text{Specificity}(c, t) = \text{Prob}(X \leq c \mid D(t) = 0) = \frac{\{1 - S(t \mid X \leq c)\} P(X \leq c)}{S(t)}
\]

where \(X\) is the biomarker value

Heagerty et al. Biometrics 2000
• Retrospective study

• Used archived echocardiograph images on 416 patients with chronic systolic heart failure

• Speckle-tracking analysis of left ventricular longitudinal, circumferential, and radial strain and strain rate

• Outcome: prognosis as defined by all-cause mortality, cardiac transplantation, or ventricular assist device placement

• Short- and long-term prognosis
Time-Dependent ROC Curves: Example using Strain in Chronic Heart Failure

Figure 3. Receiver operating characteristic (ROC) curves. In comparison to EF alone, strain and EF demonstrated an improved AUC at both 1 year (0.697 versus 0.633, \(P=0.032\)) (A) and 5 years (0.700 versus 0.638, \(P=0.014\)) (B). In comparison to EF alone, strain rate and EF did not provide incremental value (0.666 versus 0.633, \(P=0.16\)) at 1 year (A) and 5 years (0.668 versus 0.638, \(P=0.13\)) (B). AUC indicates area under the ROC curve; EF, ejection fraction.
Summary

• Critical to assess analytic validity; many studies do not rigorously assess analytic validity

• Consistency of results when biomarker assessed at short intervals on same subjects
  – Using same equipment in same center (Repeatability)
  – Using different equipment in different centers (Reproducibility)

• With time-to-event data, important to properly account for unobserved events