Estimands, Missing Data, and Sensitivity Analysis

Geert Molenberghs

geert.molenberghs@uhasselt.be & geert.molenberghs@kuleuven.be

Interuniversity Institute for Biostatistics and statistical Bioinformatics (I-BioStat)
Universiteit Hasselt & KU Leuven, Belgium

www.ibiostat.be

U Penn Symposium

April 18, 2018
Survey population: The collection of units (individuals) about which the researcher wants to make quantitative statements.

Sample frame: The set of units (individuals) that has non-zero probability of being selected.

Sample: The subset of units that have been selected.

Probability sampling: The family of probabilistic (stochastic) methods by which a subset of the units from the sample frame is selected.

Design properties: The entire collection of methodological aspects that leads to the selection of a sample.
**Sample size:** The number of units in the sample.

**Analysis and inference:** The collection of statistical techniques by which population estimands are estimated.

Examples: estimation of means, averages, totals, linear regression, ANOVA, logistic regression, loglinear models.

**Estimand:** The true population quantity (e.g., the average body mass index of the US population).

**Estimator:** A (stochastic) function of the sample data, with the aim to “come close” to the estimand.

**Estimate:** A particular realization of the estimator, for the particular sample taken (e.g., 22.37).
Your M.o.t.R. Clinical Trial

• Setting:

<table>
<thead>
<tr>
<th>Potential outcomes</th>
<th>$(T_{0j}, T_{1j})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Individual treatment effect</td>
<td>$\Delta_{Tj} = T_{1j} - T_{0j}$</td>
</tr>
<tr>
<td>Expected treatment effect</td>
<td>$\beta = E(T_{1j} - T_{0j})$</td>
</tr>
</tbody>
</table>

• No missing data $\implies$ 50% of missing data

• Fair to say: **Estimand** is $\beta = E(T_1 - T_0)$ in population

• **Randomization:** Treatment effect estimable from observed data:

• **Estimator:** $\overline{T}_1 - \overline{T}_0$
- Information coming from:
  - data
  - design
  - assumptions

- Would be different in an epidemiological study
Surrogate Endpoints Evaluation: Potential Outcomes

Alonso, Van der Elst, Molenberghs (Statistical Modeling 2016)

- Setting:

<table>
<thead>
<tr>
<th>Potential outcomes</th>
<th>$(T_{0j}, T_{1j})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Individual causal effect</td>
<td>$\Delta T_j = T_{1j} - T_{0j}$</td>
</tr>
<tr>
<td>Expected causal effect</td>
<td>$\beta = E(T_{1j} - T_{0j})$</td>
</tr>
<tr>
<td>Surrogate</td>
<td>$S_j$</td>
</tr>
</tbody>
</table>
• **Predictive causal association:**

\[ \rho_\psi = \text{corr}(\Delta T_j, S_j) \]

• **(Un)identifiability:**

\[ \rho_{T_0T_1} \text{ not identifiable} \]

• **Information coming from:**

  ▶ data
  ▶ design
  ▶ assumptions \[\rightarrow\] sensitivity
Sensitivity analysis for age-related macular degeneration trial:
Surrogate Endpoints Evaluation: Full Causal Paradigm

Alonso et al. (Biometrics 2015)

- Setting:

<table>
<thead>
<tr>
<th>Treatment potential outcomes</th>
<th>$(T_{0j}, T_{1j})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment individual causal effect</td>
<td>$\Delta_{Tj} = T_{1j} - T_{0j}$</td>
</tr>
<tr>
<td>Surrogate potential outcomes</td>
<td>$(S_{0j}, S_{1j})$</td>
</tr>
<tr>
<td>Surrogate individual causal effect</td>
<td>$\Delta_{Sj} = S_{1j} - S_{0j}$</td>
</tr>
</tbody>
</table>

- Individual causal association (ICA):

$$\rho_\Delta = \text{corr}(\Delta_{Tj}, \Delta_{Sj})$$
• Joint distribution unidentifiable

• Capture assumptions in **causal diagrams** $\rightarrow$ reduced forms of $\rho_\Delta$

• **Information coming from:**
  - data
  - design
  - assumptions $\rightarrow$ sensitivity

• Meta-analytic version in multiple trials
(a) Trial-level surrogacy

Treatment effect on the surrogate endpoint vs. Treatment effect on the true endpoint.

(b) Individual-level surrogacy

Results for the surrogate endpoint vs. Results for the true endpoint.

(c) MICA

Percentage distribution of $\rho_M$.
### Terms of Enrichment

**Enriched data**

<table>
<thead>
<tr>
<th>Coarse data</th>
<th>Augmented data</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Incomplete data</strong></td>
<td><strong>Randomized studies</strong></td>
</tr>
<tr>
<td><strong>Censored data</strong></td>
<td>Random effects</td>
</tr>
<tr>
<td>Joint models</td>
<td>Latent classes</td>
</tr>
<tr>
<td>Grouped data</td>
<td>Latent variables</td>
</tr>
<tr>
<td>Non-compliance</td>
<td>Mixtures</td>
</tr>
</tbody>
</table>

---

Estimands, Missing Data, and Sensitivity Analysis, U Penn
Increasing Complexity

- **Standard clinical trial:** design compensates for what is unobserved

- **Surrogacy:** augmentation: sensitivity is **design**-based

- **Incomplete data/non-compliance:** coarsening: sensitivity is **(non-)observation**-based
  - (Subjective) choices unavoidable
  - Interference of intercurrent events
  - Scenarios needed about $f(y_i^m | y_i^o, x_i, \theta)$ (Devan, p. 9)
  - **Such scenarios should preserve estimand**
  - Easy and elegant with MI
Concluding Reflections

• Devan starts with the right point question: WHY?

• Both: taxonomy is a GOOD thing

▷ Devan: Proper definitions needed: objective/question — endpoint — estimand
▷ Tom: principal stratification can be of help

• Sensitivity analysis ☝️ Estimands