Global Sensitivity Analysis of Randomized Trials with Informative Drop-Out

ASA Webinar

Daniel Scharfstein
Aidan McDermott
Johns Hopkins University
dscharf@jhu.edu

May 24, 2016

Funding Acknowledgments

- FDA
- PCORI

Missing Data Matters

- While unbiased estimates of treatment effects can be obtained from randomized trials with no missing data, this is no longer true when data are missing on some patients.
- The essential problem is that inference about treatment effects relies on *unverifiable* assumptions about the nature of the mechanism that generates the missing data.
- While we usually know the reasons for missing data, we do not know the distribution of outcomes for patients with missing data, how it compares to that of patients with observed data and whether differences in these distributions can be explained by the observed data.

Robert Temple and Bob O'Neil (FDA)

- "*During almost 30 years of review experience, the issue of missing data in ... clinical trials has been a major concern because of the potential impact on the inferences that can be drawn .... when data are missing .... the analysis and interpretation of the study pose a challenge and the conclusions become more tenuous as the extent of 'missingness' increases.*"

NRC Report and Sensitivity Analysis

- In 2010, the National Research Council (NRC) issued a reported entitled "The Prevention and Treatment of Missing Data in Clinical Trials."
- This report, commissioned by the FDA, provides 18 recommendations targeted at (1) trial design and conduct, (2) analysis and (3) directions for future research.
- Recommendation 15 states
    - *Sensitivity analyses should be part of the primary reporting of findings from clinical trials. Examining sensitivity to the assumptions about the missing data mechanism should be a mandatory component of reporting.*

ICH, EMEA and Sensitivity Analysis

- 1998 International Conference of Harmonization (ICH) Guidance document (E9) entitled "Statistical Principles in Clinical Trials" states: "*it is important to evaluate the robustness of the results to various limitations of the data, assumptions, and analytic approaches to data analysis*"
- European Medicines Agency 2009 draft "Guideline on Missing Data in Confirmatory Clinical Trials" states "*[i]n all submissions with non-negligible amounts of missing data sensitivity analyses should be presented as support to the main analysis.*"

PCORI and Sensitivity Analysis

- In 2012, Li *et al.* issued the report "Minimal Standards in the Prevention and Handling of Missing Data in Observational and Experimental Patient Centered Outcomes Research"
- This report, commissioned by PCORI, provides 10 standards targeted at (1) design, (2) conduct, (3) analysis and (4) reporting.
- Standard 8 echoes the NRC report, stating
  - *Examining sensitivity to the assumptions about the missing data mechanism (i.e., sensitivity analysis) should be a mandatory component of the study protocol, analysis, and reporting.*

Sensitivity Analysis

The set of possible assumptions about the missing data mechanism is very large and cannot be fully explored. There are different approaches to sensitivity analysis:

- Ad-hoc
- Local
- Global

Ad-hoc Sensitivity Analysis

- Analyzing data using a few different analytic methods, such as last or baseline observation carried forward, complete or available-case analysis, mixed models or multiple imputation, and evaluate whether the resulting inferences are consistent.
- The problem with this approach is that the assumptions that underlie these methods are very strong and for many of these methods unreasonable.
- More importantly, just because the inferences are consistent does not mean that there are no other reasonable assumptions under which the inference about the treatment effect is different.

Local Sensitivity Analysis

- Specify a reasonable benchmark assumption (e.g., missing at random) and evaluate the robustness of the results within a small neighborhood of this assumption.
- What if there are assumptions outside the local neighborhood which are plausible?

Global Sensitivity Analysis

- Evaluate robustness of results across a much broader range of assumptions that include a reasonable benchmark assumption and a collection of additional assumptions that trend toward best and worst case assumptions.
- Emphasized in Chapter 5 of the NRC report.
- This approach is substantially more informative because it operates like "stress testing" in reliability engineering, where a product is systematically subjected to increasingly exaggerated forces/conditions in order to determine its breaking point.

Global Sensitivity Analysis

- In the missing data setting, global sensitivity analysis allows one to see how far one needs to deviate from the benchmark assumption in order for inferences to change.
- "Tipping point" analysis (Yan, Lee and Li, 2009; Campbell, Pennello and Yue, 2011)
- If the assumptions under which the inferences change are judged to be sufficiently far from the benchmark assumption, then greater credibility is lent to the benchmark analysis; if not, the benchmark analysis can be considered to be fragile.

Global Sensitivity Analysis

- Restrict consideration to follow-up randomized study designs that prescribe that measurements of an outcome of interest are to be taken on each study participant at fixed time-points.
- Focus on monotone missing data pattern
- Consider the case where interest is focused on a comparison of treatment arm means at the last scheduled visit.

Case Study: Quetiapine Bipolar Trial

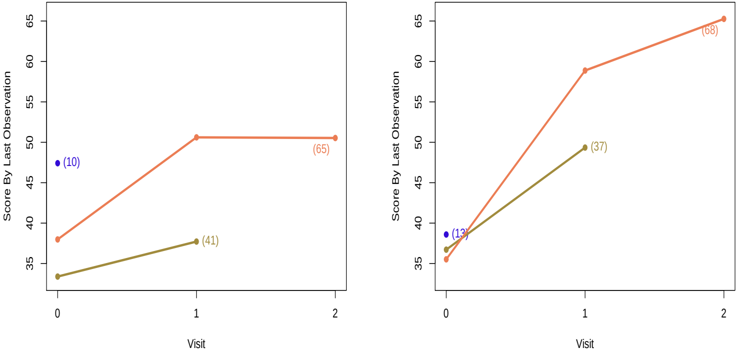
- Patients with bipolar disorder randomized equally to one of three treatment arms: placebo, Quetiapine 300 mg/day or Quetiapine 600 mg/day (Calabrese *et al.*, 2005).
- Randomization was stratified by type of bipolar disorder.
- Short-form version of the Quality of Life Enjoyment Satisfaction Questionnaire (QLESSF, Endicott *et al.*, 1993), was scheduled to be measured at baseline, week 4 and week 8.

Quetiapine Bipolar Trial

- Focus on the subset of 234 patients with bipolar 1 disorder who were randomized to either the placebo (n=116) or 600 mg/day (n=118) arms.
- Only 65 patients (56%) in placebo arm and 68 patients (58%) in the 600mg/day arm had a complete set of QLESSF scores.
- Patients with complete data tend to have higher average QLESSF scores, suggesting that a complete-case analysis could be biased.

Observed Data

Figure: Treatment-specific (left: placebo; right: 600 mg/day Quetiapine) trajectories of mean QLESSF scores, stratified by last available measurement.



Central Question

What is the difference in the mean QLESSF score at week 8 between Quetiapine 600 mg/day and placebo in the counterfactual world in which all patients were followed to that week?

Global Sensitivity Analysis

- Inference about the treatment arm means requires two types of assumptions:
  (i) *unverifiable* assumptions about the distribution of outcomes among those with missing data and
  (ii) additional testable assumptions that serve to increase the efficiency of estimation.

Global Sensitivity Analysis

- Type (i) assumptions are necessary to identify the treatment-specific means.
- By *identification*, we mean that we can write it as a function that depends only on the distribution of the observed data.
- When a parameter is identified we can hope to estimate it as precisely as we desire with a sufficiently large sample size,
- In the absence of identification, statistical inference is fruitless as we would be unable to learn about the true parameter value even if the sample size were infinite.

Global Sensitivity Analysis

- To address the identifiability issue, it is essential to conduct a sensitivity analysis, whereby the data analysis is repeated under different type (i) assumptions, so as to investigate the extent to which the conclusions of the trial are dependent on these subjective, unverifiable assumptions.
- The usefulness of a sensitivity analysis ultimately depends on the plausibility of the unverifiable assumptions.
- It is key that any sensitivity analysis methodology allow the formulation of these assumptions in a transparent and easy to communicate manner.

Global Sensitivity Analysis

- There are an infinite number of ways of positing type (i) assumptions.
- Ultimately, however, these assumptions prescribe how missing outcomes should be "imputed."
- A reasonable way to posit these assumptions is to
  - stratify individuals with missing outcomes according to the data that we were able to collect on them and the occasions at which the data were collected
  - separately for each stratum, hypothesize a connection (or link) between the distribution of the missing outcome with the distribution of the outcome among those with the observed outcome and who share the same recorded data.

Global Sensitivity Analysis

- Type (i) assumptions will not suffice when the repeated outcomes are continuous or categorical with many levels. This is because of *data sparsity*.
- For example, the stratum of people who share the same recorded data will typically be small. As a result, it is necessary to draw strength across strata by "smoothing."
- Without smoothing, the data analysis will rarely be informative because the uncertainty concerning the treatment arm means will often be too large to be of substantive use.
- As a result, it is necessary to impose type (ii) smoothing assumptions.
- Type (ii) assumptions should be scrutinized with standard model checking techniques.

Global Sensitivity Analysis

- The global sensitivity framework proceeds by parameterizing (i.e., indexing) the connections (i.e., type (i) assumptions) via sensitivity analysis parameters.
- The parameterization is configured so that a specific value of the sensitivity analysis parameters (typically set to zero) corresponds to a benchmark connection that is considered reasonably plausible and sensitivity analysis parameters further from the benchmark value represent more extreme departures from the benchmark connection.

Global Sensitivity Analysis

- The global sensitivity analysis strategy that we propose is focused on separate inferences for each treatment arm, which are then combined to evaluate treatment effects.
- Until later, our focus will be on estimation of the mean outcome at week 8 (in a world without missing outcomes) for one of the treatment groups and we will suppress reference to treatment assignment.

Notation: Quetiapine Bipolar Trial

- Y_0, Y_1, Y_2: QLESSF scores scheduled to be collected at baseline, week 4 and week 8.
- Let R_k be the indicator that Y_k is observed.
- We assume $R_0 = 1$ and that $R_k = 0$ implies $R_{k+1} = 0$ (i.e., missingness is monotone).
- Patient is on-study at visit k if $R_k = 1$
- Patient discontinued prior to visit k if $R_k = 0$
- Patient last seen at visit $k-1$ if $R_{k-1} = 1$ and $R_k = 0$.
- Y_k^{obs} equals to Y_k if $R_k = 1$ and equals to *nil* if $R_k = 0$.

Notation: Quetiapine Bipolar Trial

- The observed data for an individual are

  $$O = (Y_0, R_1, Y_1^{obs}, R_2, Y_2^{obs}),$$

  which has some distribution P^* contained within a set of distributions $\mathcal{M}$ (type (ii) assumptions discussed later).
- The superscript $*$ will be used to denote the true value of the quantity to which it is appended.
- Any distribution $P \in \mathcal{M}$ can be represented in terms of the following distributions:
  - $f(Y_0)$
  - $P[R_1 = 1 | Y_0]$
  - $f(Y_1 | R_1 = 1, Y_0)$
  - $P[R_2 = 1 | R_1 = 1, Y_1, Y_0]$
  - $f(Y_2 | R_2 = 1, Y_1, Y_0)$.

Notation: Quetiapine Bipolar Trial

- We assume that n independent and identically distributed copies of O are observed.
- The goal is to use these data to draw inference about $\mu^* = E^*[Y_2]$.
- When necessary, we will use the subscript i to denote data for individual i.

Benchmark Assumption (Missing at Random)

- $A(y_0)$: patients with $Y_0 = y_0$.
- $B(y_0, y_1)$: patients on-study at visit 1 with $Y_0 = y_0$ and $Y_1 = y_1$.

Missing at random posits the following type (i) “linking” assumptions:

- In each stratum $A(y_0)$, the distribution of Y_1 and Y_2 is the same for those last seen at visit 0 as those on-study at visit 1.
- In each stratum $B(y_0, y_1)$, the distribution of Y_2 is the same for those last seen at visit 1 as those on-study at visit 2.

Benchmark Assumption (Missing at Random)

Mathematically, we can express these assumptions as follows:

$$f^*(Y_1, Y_2|R_1 = 0, A(y_0)) = f^*(Y_1, Y_2|R_1 = 1, A(y_0)) \text{ for all } y_0$$

and

$$f^*(Y_2|R_2 = 0, B(y_0, y_1)) = f^*(Y_2|R_2 = 1, B(y_0, y_1)) \text{ for all } y_1, y_0$$

Benchmark Assumption (Missing at Random)

Using Bayes' rule, we can re-write these expressions as:

$$P^*[R_1 = 0 | Y_2 = y_2, Y_1 = y_1, A(y_0)] = P^*[R_1 | A(y_0)]$$

and

$$P^*[R_2 = 0 | Y_2 = y_2, B(y_1, y_0)] = P^*[R_2 = 0 | R_1 = 1, B(y_1, y_0)]$$

Missing at random implies:

- The decision to discontinue the study before visit 1 is like the flip of a coin with probability depending on the value of the outcome at visit 0.
- For those on-study at visit 1, the decision to discontinue the study before visit 2 is like the flip of a coin with probability depending on the value of the outcomes at visits 1 and 0.

Benchmark Assumption (Missing at Random)

- MAR is a type (i) assumption. It is "unverifiable."
- For patients last seen at visit k, we cannot learn from the observed data about the conditional (on observed history) distribution of outcomes after visit k.
- For patients last seen at visit k, any assumption that we would make about the conditional (on observed history) distribution of the outcomes after visit k will be unverifiable from the data available to us.
- For patients last seen at visit k, the assumption that the conditional (on observed history) distribution of outcomes after visit k is the same as those who remain on-study after visit k and have the same observed history is unverifiable.

Benchmark Assumption (Missing at Random)

Under MAR, μ^* is identified. That is, it can be expressed as a function of the distribution of the observed data. Specifically,

$$\mu^* = \mu(P^*) = \int_{y_0} \int_{y_1} \int_{y_2} y_2 dF_2^*(y_2|y_1, y_0) dF_1^*(y_1|y_0) dF_0^*(y_0)$$

where

- $F_2^*(y_2|y_1, y_0) = P^*[Y_2 \leq y_2 | R_2 = 1, B(y_1, y_0)]$
- $F_1^*(y_1|y_0) = P^*[Y_1 \leq y_1 | R_1 = 1, A(y_0)]$
- $F_0^*(y_0) = P^*[Y_0 \leq y_0]$.

Missing Not at Random (MNAR)

The MAR assumption is not the only one that is (1) unverifiable and (2) allows identification of μ^*.

Missing Not at Random (MNAR)

The first part of the MAR assumption is equivalent to

$$\begin{aligned} & f^*(Y_2|R_1 = 0, Y_1 = y_1, A(y_0)) \\ & = f^*(Y_2|R_1 = 1, Y_1 = y_1, A(y_0)) \text{ for all } y_1, y_0 \end{aligned} \tag{1}$$

and

$$f^*(Y_1|R_1 = 0, A(y_0)) = f^*(Y_1|R_1 = 1, A(y_0)) \text{ for all } y_0$$

In building a class of MNAR models, we will retain (1):

- Among patients in $A(y_0)$ with $Y_1 = y_1$, the distribution of Y_2 is the same for those last seen at visit 0 as those on-study at visit 1.
- The decision to discontinue the study before visit 1 is independent of Y_2 (i.e., the future outcome) after conditioning on the Y_0 (i.e., the past outcome) and Y_1 (i.e., the most recent outcome).
- Non-future dependence

Missing Not at Random (MNAR)

Exponential Tilting

$$\begin{aligned} &f^*(Y_1|R_1 = 0, A_0(y_0)) \\ &\quad \propto f^*(Y_1|R_1 = 1, A(y_0)) \exp\{\alpha r(Y_1)\} \text{ for all } y_0 \end{aligned}$$

$$\begin{aligned} &f^*(Y_2|R_2 = 0, B(y_1, y_0)) \\ &\quad \propto f^*(Y_2|R_2 = 1, B(y_1, y_0)) \exp\{\alpha r(Y_2)\} \text{ for all } y_0, y_1 \end{aligned}$$

- $r(y)$ is a specified increasing function; α is a sensitivity analysis parameter.
- $\alpha = 0$ is MAR.

Missing Not at Random (MNAR)

When $\alpha > 0$ (< 0)

- In the stratum $A(y_0)$, the distribution of Y_1 for patients last seen at visit 0 is weighted more heavily to higher (lower) values than the distribution of Y_1 for patients on study at visit 1.
- In the stratum $B(y_1, y_0)$, the distribution of Y_2 for patients last seen at visit 1 is weighted more heavily to higher (lower) values than the distribution of Y_2 for patients on study at visit 2.

The amount of "tilting" increases with the magnitude of α.

Missing Not at Random (MNAR)

Using Bayes' rule, we can re-write these expressions as:

$$\text{logit } P^*[R_1 = 0 | Y_2 = y_2, Y_1 = y_1, A(y_0)] = l_1^*(y_0) + \alpha r(y_1)$$

and

$$\text{logit } P^*[R_2 = 0 | Y_2 = y_2, B(y_1, y_0)] = l_2^*(y_1, y_0) + \alpha r(y_2)$$

where

$$\begin{aligned} l_1^*(y_0; \alpha) &= \text{logit } P^*[R_1 = 0 | A(y_0)] - \\ &\quad \log E^*[\exp\{\alpha r(Y_1)\} | R_1 = 1, A(y_0)] \end{aligned}$$

and

$$\begin{aligned} l_2^*(y_1, y_0; \alpha) &= \text{logit } P^*[R_2 = 0 | B(y_1, y_0)] - \\ &\quad \log E^*[\exp\{\alpha r(Y_2)\} | R_2 = 1, B(y_1, y_0)] \end{aligned}$$

Missing Not at Random (MNAR)

Written in this way:

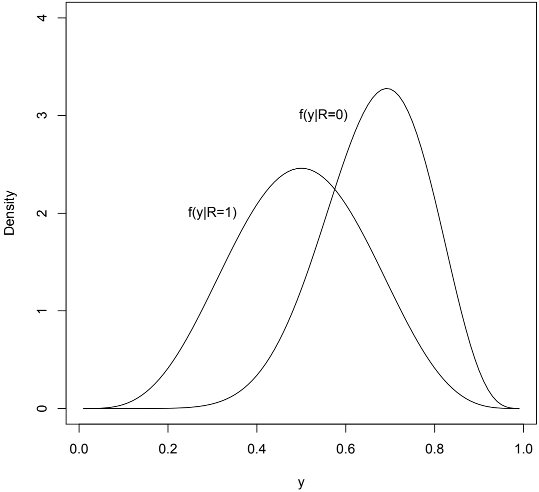
- The decision to discontinue the study before visit 1 is like the flip of a coin with probability depending on the value of the outcome at visit 0 *and* the value of the outcome at visit 1.
- For those on-study at visit 1, the decision to discontinue the study before visit 2 is like the flip of a coin with probability depending on the value of the outcomes at visits 1 and 0 *and* the value of the outcome at visit 2.

Exponential Tilting Explained

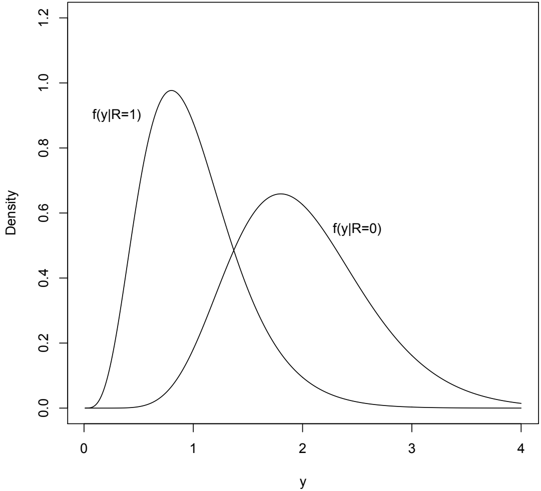
$$f(Y|R=0) \propto f(Y|R=1)\exp\{\alpha r(Y)\}$$

- If $[Y|R=1] \sim N(\mu, \sigma^2)$ and $r(Y) = Y$,
  $[Y|R=0] \sim N(\mu + \alpha\sigma^2, \sigma^2)$
- If $[Y|R=1] \sim Beta(a, b)$ and $r(Y) = \log(Y)$,
  $[Y|R=0] \sim Beta(a + \alpha, b)$, $\alpha > -a$.
- If $[Y|R=1] \sim Gamma(a, b)$ and $r(Y) = \log(Y)$,
  $[Y|R=0] \sim Gamma(a + \alpha, b)$, $\alpha > -a$.
- If $[Y|R=1] \sim Gamma(a, b)$ and $r(Y) = Y$,
  $[Y|R=0] \sim Gamma(a, b - \alpha)$, $\alpha < b$.
- If $[Y|R=1] \sim Bernoulli(p)$ and $r(Y) = Y$,
  $[Y|R=0] \sim Bernoulli\left(\frac{p\exp(\alpha)}{p\exp(\alpha)+1-p}\right)$.

Beta



Gamma



Missing Not at Random (MNAR)

For given α, μ^* is identified. Specifically, $\mu^* = \mu(P^*; \alpha)$ equals

$$\int_{y_0}\int_{y_1}\int_{y_2} y_2 \left\{ dF_2^*(y_2|y_1,y_0)\{1-H_2^*(y_1,y_0)\} + \frac{dF_2^*(y_2|y_1,y_0)\exp\{\alpha r(y_2)\}}{\int_{y_2'} dF_2^*(y_2'|y_1,y_0)\exp\{\alpha r(y_2')\}} H_2^*(y_1,y_0) \right\} \times$$

$$\left\{ dF_1^*(y_1|y_0)\{1-H_1^*(y_0)\} + \frac{dF_1^*(y_1|y_0)\exp\{\alpha r(y_1)\}}{\int_{y_1'} dF_1^*(y_1'|y_0)\exp\{\alpha r(y_1')\}} H_1^*(y_0) \right\} dF_0^*(y_0)$$

where $H_2^*(y_1, y_0) = P^*[R_2 = 0|B(y_1, y_0)]$ and
$H_1^*(y_0) = P^*[R_1 = 0|A(y_0)]$

- μ^* is written as a function of the distribution of the observed data (depending on α).

Inference

For given α, the above formula shows that μ^* depends on

- $F_2^*(y_2|y_1, y_0) = P^*[Y_2 \leq y_2|R_2 = 1, B(y_1, y_0)]$
- $F_1^*(y_1|y_0) = P^*[Y_1 \leq y_1|R_1 = 1, A(y_0)]$
- $H_2^*(y_1, y_0) = P^*[R_2 = 0|B(y_1, y_0)]$
- $H_1^*(y_0) = P^*[R_1 = 0|A(y_0)]$.

It is natural to consider estimating μ^* by "plugging in" estimators of these quantities.

How can we estimate these latter quantities? With the exception of $F_0^*(y_0)$, it is tempting to think that we can use non-parametric procedures to estimate these quantities.

Inference

A non-parametric estimate of $F_2^*(y_2|y_1, y_0)$ would take the form:

$$\widehat{F}_2(y_2|y_1, y_0) = \frac{\sum_{i=1}^n R_{2,i} I(Y_{2,i} \leq y_2) I(Y_{1,i} = y_1, Y_{0,i} = y_0)}{\sum_{i=1}^n R_{2,i} I(Y_{1,i} = y_1, Y_{0,i} = y_0)}$$

- This estimator will perform very poorly (i.e., have high levels of uncertainty in moderate sample sizes) because the number of subjects who complete the study (i.e., $R_2 = 1$) and are observed to have outcomes at visits 1 and 0 exactly equal to y_1 and y_0 will be very small and can only be expected to grow very slowly as the sample size increases.
- As a result, a a plug-in estimator of μ^* that uses such non-parametric estimators will perform poorly.

Inference - Type (ii) Assumptions

We make the estimation task slightly easier by assuming that

$$F_2^*(y_2|y_1, y_0) = F_2^*(y_2|y_1)$$

and

$$H_2^*(y_1, y_0) = H_2^*(y_1)$$

Inference - Kernel Smoothing

Estimate $F_2^*(y_2|y_1)$, $F_1^*(y_1|y_0)$, $H_2^*(y_1)$ and $H_1^*(y_0)$ using kernel smoothing techniques.

To motivate this idea, consider the following non-parametric estimate of $F_2^*(y_2|y_1)$

$$\widehat{F}_2(y_2|y_1) = \frac{\sum_{i=1}^n R_{2,i} I(Y_{2,i} \leq y_2) I(Y_{1,i} = y_1)}{\sum_{i=1}^n R_{2,i} I(Y_{1,i} = y_1)}$$

- This estimator will still perform poorly, although better than $\widehat{F}_2(y_2|y_1, y_0)$.
- Replace $I(Y_{1,i} = y_1)$ by $\phi\left(\frac{Y_{1,i}-y_1}{\sigma_{F_2}}\right)$, where $\phi(\cdot)$ is standard normal density and σ_{F_2} is a tuning parameter.

$$\widehat{F}_2(y_2|y_1; \sigma_{F_2}) = \frac{\sum_{i=1}^n R_{2,i} I(Y_{2,i} \leq y_2) \phi\left(\frac{Y_{1,i}-y_1}{\sigma_{F_2}}\right)}{\sum_{i=1}^n R_{2,i} \phi\left(\frac{Y_{1,i}-y_1}{\sigma_{F_2}}\right)}$$

Inference - Kernel Smoothing

- This estimator allows *all* completers to contribute, not just those with Y_1 values equal to y_1
- It assigns weight to completers according to how far their Y_1 values are from y_1, with closer values assigned more weight.
- The larger σ_{F_2}, the larger the influence of values of Y_1 further from y_1 on the estimator.
- As $\sigma_{F_2} \to \infty$, the contribution of each completer to the estimator becomes equal, yielding bias but low variance.
- As $\sigma_{F_2} \to 0$, only completers with Y_1 values equal to y_1 contribute, yielding low bias but high variance.

Inference - Cross-Validation

To address the bias-variance trade-off, cross validation is typically used to select σ_{F_2}.

- Randomly divided dataset into J (typically, 10) approximately equal sized validation sets.
- Let V_j be the indices of the patients in jth validation set.
- Let n_j be the associated number of subjects.
- Let $\widehat{F}_2^{(j)}(y_2|y_1;\sigma_{F_2})$ be the estimator of $F_2^*(y_2|y_1)$ based on the dataset that excludes the jth validation set.
- If σ_{F_2} is a good choice then one would expect

$$CV_{F_2^*(\cdot|\cdot)}(\sigma_{F_2}) = \frac{1}{J}\sum_{j=1}^{J}\left\{\frac{1}{n_j}\sum_{i\in V_j} R_{2,i}\underbrace{\int\left\{I(Y_{2,i}\leq y_2)-\widehat{F}_2^{(j)}(y_2|Y_{1,i};\sigma_{F_2})\right\}^2 d\widehat{F}_2^{\circ}(y_2)}_{\text{Distance for } i\in V_j}\right\}$$

will be small, where $\widehat{F}_2^{\circ}(y_2)$ is the empirical distribution of Y_2 among subjects on-study at visit 2.

Inference - Cross-Validation

- For each individual i in the jth validation set with an observed outcome at visit 2, we measure, by the quantity above the horizontal brace, the distance (or loss) between the collection of indicator variables $\{I(Y_{2,i} \leq y_2) : d\widehat{F}_2^{\circ}(y_2) > 0\}$ and the corresponding collection of predicted values $\{\widehat{F}_2^{(j)}(y_2|Y_{1,i}; \sigma_{F_2}) : d\widehat{F}_2^{\circ}(y_2) > 0\}$.
- The distances for each of these individuals are then summed and divided by the number of subjects in the jth validation set.
- An average across the J validation/training sets is computed.
- We can then estimate $F_2^*(y_2|y_1)$ by $\widehat{F}_2(y_2|y_1; \widehat{\sigma}_{F_2})$, where $\widehat{\sigma}_{F_2} = \text{argmin } CV_{F_2^*(\cdot|\cdot)}(\sigma_{F_2})$.

Inference - Cross-Validation

We use similar ideas to estimate

- $F_1^*(y_1|y_0)$
- $H_2^*(y_1)$
- $H_1^*(y_0)$

In our software, we set $\sigma_{F_2} = \sigma_{F_1} = \sigma_F$ and minimize a single CV function. The software refers to this smoothing parameter as σ_Q.

In our software, we set $\sigma_{H_2} = \sigma_{H_1} = \sigma_H$ and minimize a single CV function. The software refers to this smoothing parameter as σ_P.

Inference - Potential Problem

- The cross-validation procedure for selecting tuning parameters achieves optimal finite-sample bias-variance trade-off for the quantities requiring smoothing.
- This optimal trade-off is usually not optimal for estimating μ^*.
- The plug-in estimator of μ^* could possibly suffer from excessive and asymptotically non-negligible bias due to inadequate tuning.
- This may prevent the plug-in estimator from enjoying regular asymptotic behavior, upon which statistical inference is generally based.
- The resulting estimator may have a slow rate of convergence, and common methods for constructing confidence intervals, such as the Wald and bootstrap intervals, can have poor coverage properties.

Inference - Correct Procedure

- To deal with this, we will "correct" the plug-in estimator.
- We will construct an estimator that is "asymptotically linear" (i.e., can be expressed as the average of i.i.d. random variables plus a remainder term that is asymptotically negligible).
- Our one-step estimator is

$$\text{plug-in} + \text{average of estimated influence functions}$$

- The influence function for a patient by $\psi(O; F, H)$. The estimated influence function is $\psi(O; \widehat{F}, \widehat{H})$.

Inference - Uncertainty

- An influence function-based 95% confidence interval takes the form $\widehat{\mu} \pm 1.96\widehat{se}(\widehat{\mu})$, where

$$\widehat{se}(\widehat{\mu}) = \sqrt{E_n[\psi(O; \widehat{F}, \widehat{H})^2]/n}$$

- In equal-tailed studentized bootstrap, the confidence interval takes the form $[\widehat{\mu} - t_{0.975}\widehat{se}(\widehat{\mu}), \widehat{\mu} - t_{0.025}\widehat{se}(\widehat{\mu})]$, where t_q is the qth quantile of $\left\{\frac{\widehat{\mu}^{(b)} - \widehat{\mu}}{\widehat{se}(\widehat{\mu}^{(b)})} : b = 1, \ldots, B\right\}$
- In symmetric studentized bootstrap, the confidence interval takes the form $[\widehat{\mu} - t^*_{0.95}\widehat{se}(\widehat{\mu}), \widehat{\mu} + t^*_{0.95}\widehat{se}(\widehat{\mu})]$, where $t^*_{0.95}$ is selected so that 95% of the distribution of $\left\{\frac{\widehat{\mu}^{(b)} - \widehat{\mu}}{\widehat{se}(\widehat{\mu}^{(b)})} : b = 1, \ldots, B\right\}$ falls between $-t^*_{0.95}$ and $t^*_{0.95}$.
- Useful to replace influence-function based standard error estimator with jackknife standard error estimator.

Quetiapine Bipolar Trial - Fit

- Estimated smoothing parameters for the drop-out model are 11.54 and 9.82 for the placebo and 600 mg arms.
- Estimated smoothing parameters for the outcome model are 6.34 and 8.05 for the placebo and 600 mg arms.
- In the placebo arm, the observed percentages of last being seen at visits 0 and 1 among those at risk at these visits are 8.62% and 38.68%. Model-based estimates are 7.99% and 38.19%.
- For the 600 mg arm, the observed percentages are 11.02% and 35.24% and the model-based estimates are 11.70% and 35.08%.

Quetiapine Bipolar Trial - Fit

- In the placebo arm, the Kolmogorov-Smirnov distances between the empirical distribution of the observed outcomes and the model-based estimates of the distribution of outcomes among those on-study at visits 1 and 2 are 0.013 and 0.033.
- In the 600 mg arm, these distances are 0.013 and 0.022.
- These results suggest that our model for the observed data fits the observed data well.

Quetiapine Bipolar Trial - MAR

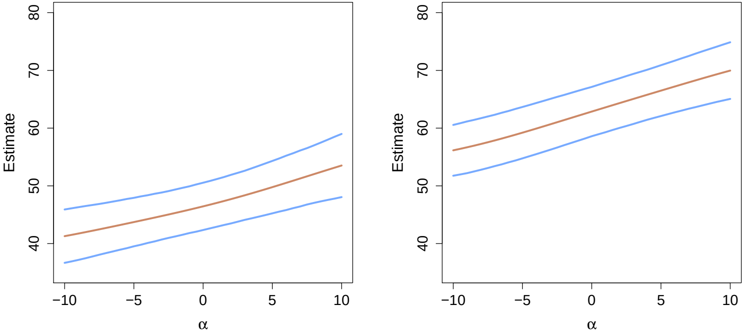
- Under MAR, the estimated values of μ^* are 46.45 (95% CI: 42.35,50.54) and 62.87 (95% CI: 58.60,67.14) for the placebo and 600 mg arms.
- The estimated difference between 600 mg and placebo is 16.42 (95% 10.34, 22.51)
- Statistically and clinically significant improvement in quality of life in favor of Quetiapine.

Quetiapine Bipolar Trial - Sensitivity Analysis

- We set $r(y) = y$ and ranged the sensitivity analysis parameter from -10 and 10 in each treatment arm.
- According to experts, there is no evidence to suggest that there is a differential effect of a unit change in QLESSF on the hazard of drop-out based on its location on the scale.

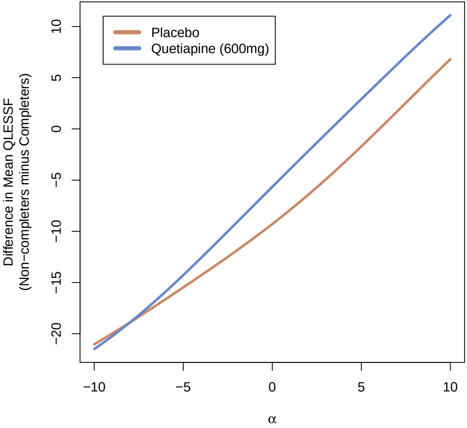
Quetiapine Bipolar Trial - Sensitivity Analysis

Figure: Treatment-specific (left: placebo; right: 600 mg/day Quetiapine) estimates (along with 95% pointwise confidence intervals) of μ^* as a function of α.



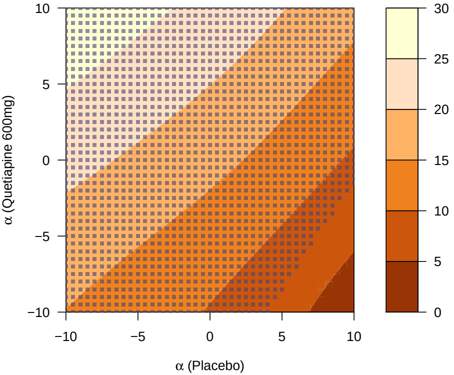
Quetiapine Bipolar Trial - Sensitivity Analysis

Figure: Treatment-specific differences between the estimated mean QLESSF at Visit 2 among non-completers and the estimated mean among completers, as a function of α.



Quetiapine Bipolar Trial - Sensitivity Analysis

Figure: Contour plot of the estimated differences between mean QLESSF at Visit 2 for Quetiapine vs. placebo for various treatment-specific combinations of the sensitivity analysis parameters.



Quetiapine Bipolar Trial - Sensitivity Analysis

- Only when the sensitivity analysis are highly differential (e.g., α(placebo) $= 8$ and α(Quetiapine) $= -8$) are the differences no longer statistically significant.
- Conclusions under MAR are highly robust.

Simulation Study

- Generated 2500 placebo and Quetiapine datasets using the estimated distributions of the observed data from the Quentiapine study as the true data generating mechanisms.
- For given treatment-specific α, these true data generating mechanisms can be mapped to a true value of μ^*.
- For each dataset, the sample size was to set to 116 and 118 in the placebo and Quetiapine arms, respectively.

Simulation Study - Bias/MSE

		Placebo			Quetiapine		
α	Estimator	μ^*	Bias	MSE	μ^*	Bias	MSE
-10	Plug-in	40.85	0.02	4.43	56.07	0.40	4.69
	Corrected		0.43	4.56		0.42	4.72
-5	Plug-in	43.45	0.05	4.29	59.29	0.34	4.55
	Corrected		0.27	4.26		0.24	4.35
-1	Plug-in	46.02	0.28	4.34	62.58	0.50	4.39
	Corrected		0.18	4.22		0.14	4.00
0	Plug-in	46.73	0.36	4.44	63.42	0.55	4.36
	Corrected		0.17	4.27		0.14	3.95
1	Plug-in	47.45	0.43	4.57	64.25	0.59	4.32
	Corrected		0.16	4.36		0.15	3.92
5	Plug-in	50.48	0.66	5.33	67.34	0.59	4.20
	Corrected		0.14	5.11		0.19	4.15
10	Plug-in	54.07	0.51	5.78	70.51	0.07	4.02
	Corrected		0.04	6.30		-0.05	4.66

Simulation Study - Coverage

		Placebo	Quetiapine
α	Procedure	Coverage	Coverage
-10	Wald-IF	91.5%	90.5%
	Wald-JK	95.0%	94.6%
	Bootstrap-IF-ET	94.3%	93.8%
	Bootstap-JK-ET	94.4%	93.4%
	Bootstap-IF-S	95.2%	94.6%
	Bootstap-JK-S	95.0%	94.6%
-5	Wald-IF	93.5%	92.9%
	Wald-JK	95.0%	94.8%
	Bootstrap-IF-ET	95.2%	94.6%
	Bootstap-JK-ET	94.8%	94.6%
	Bootstap-IF-S	95.4%	95.2%
	Bootstap-JK-S	95.1%	95.2%
-1	Wald-IF	93.9%	94.2%
	Wald-JK	94.9%	95.4%
	Bootstrap-IF-ET	95.1%	94.8%
	Bootstap-JK-ET	95.1%	94.6%
	Bootstap-IF-S	95.3%	96.4%
	Bootstap-JK-S	95.1%	96.3%
0	Wald-IF	93.8%	94.0%
	Wald-JK	95.0%	95.4%
	Bootstrap-IF-ET	94.6%	94.5%
	Bootstap-JK-ET	94.6%	94.6%
	Bootstap-IF-S	95.5%	96.6%
	Bootstap-JK-S	95.2%	96.7%

Simulation Study - Coverage

		Placebo	Quetiapine
α	Procedure	Coverage	Coverage
1	Wald-IF	93.3%	93.7%
	Wald-JK	95.1%	95.5%
	Bootstrap-IF-ET	94.6%	94.6%
	Bootstap-JK-ET	94.6%	94.6%
	Bootstap-IF-S	95.5%	96.5%
	Bootstap-JK-S	95.2%	96.5%
5	Wald-IF	90.8%	91.3%
	Wald-JK	95.3%	95.7%
	Bootstrap-IF-ET	93.2%	91.6%
	Bootstap-JK-ET	93.8%	93.0%
	Bootstap-IF-S	95.5%	95.4%
	Bootstap-JK-S	95.8%	96.4%
10	Wald-IF	85.4%	87.8%
	Wald-JK	94.9%	94.5%
	Bootstrap-IF-ET	88.2%	87.0%
	Bootstap-JK-ET	92.2%	89.7%
	Bootstap-IF-S	94.6%	93.9%
	Bootstap-JK-S	95.5%	95.1%

Missing Data Matters

- No substitute for better trial design and procedures to minimize missing data.
- Global sensitivity analysis should be a mandatory component of trial reporting.
- Visit us at www.missingdatamatters.org or email me at dscharf@jhu.edu
- We are happy to collaborate with you on executing global sensitivity analyses using our software.

Upcoming Short Courses

- June 22, 2016 - Johns Hopkins University
- July 26, 2016 - University of Washington
- October 26-28, 2016 - BASS XXIII, Rockville, MD