ANALYSIS AND INTERPRETATION OF SIGNALS IN LARGE DATA SETS

Sharon-Lise T. Normand
Harvard Medical School & Harvard School of Public Health

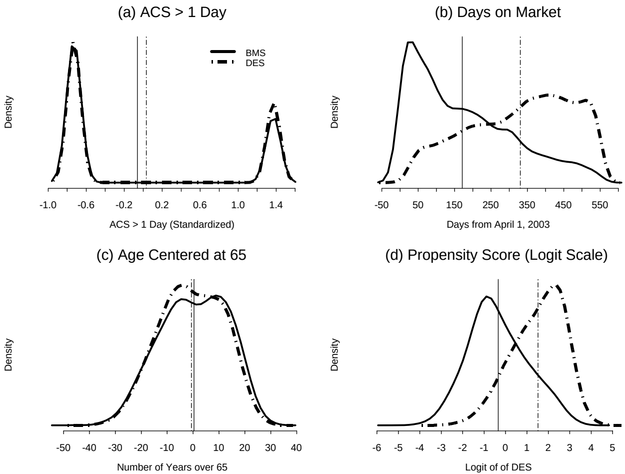
February 16, 2011

WHAT DEFINES LARGE (ACTIVE MEDICAL PRODUCT SURVEILLANCE)?

The amount of **information** depends on:

1. Experimental versus observational unit:
   - Clustered randomized trials with p clusters and n subjects per cluster yields $\leq n \times p$ independent pieces of information
2. Causality:
   - Treatment groups must **overlap** on the basis of pre-treatment variables
   - **Distributions** of pre-treatment variables must be **balanced** across treatment groups
3. Data completeness:
   - How much information in the **observed data** for a specific hypothesis test relative to the full amount of information had the data been **complete**
   - Depends on null and alternative hypotheses

BALANCE Normand SL. Some old and some new statistical tools for outcomes research. Circulation,2008;118:872-884



SIGNAL VERSUS NOISE (ACTIVE MEDICAL PRODUCT SURVEILLANCE)

Suppose data are clustered, e.g., $i = 1, 2, \cdots, n_j$ observations within group j (e.g., a device) for $j = 1, 2, \cdots, J$ (device) groups. Let y_{ji} = mean functional score for patient i treated in group j.

$$y_{ji} \sim N(\alpha_j, \sigma_y^2) \text{ and } \alpha_j \sim N(\mu_\alpha, \sigma_\alpha^2) \tag{1}$$

Want to learn about α_j. For each group j, the estimate of α_j is

$$\hat{\alpha}_j = \omega_j \mu_\alpha + (1 - \omega_j)\bar{y}_j \tag{2}$$

Estimate is a weighted average between the within-group data ($\bar{y}_j$) and the group-level model (μ_α)

SIGNAL VERSUS NOISE (ACTIVE MEDICAL PRODUCT SURVEILLANCE)

$$\left(\frac{\text{No}}{\text{Pooling}}\right)\; 0 \leq \;\omega_j \;=\; \left(\frac{\frac{\sigma_y^2}{n_j}}{\sigma_\alpha^2 + \frac{\sigma_y^2}{n_j}}\right) \;\leq\; 1 \;\left(\frac{\text{Complete}}{\text{Pooling}}\right)$$

- **Noise:** $\frac{\sigma_y}{\sqrt{n_j}}$ (sampling variability)
- **Signal** $= \alpha_j$
- **Strength** $\text{Var}(\alpha_j) = (1-\omega_j)\frac{\sigma_y^2}{n_j} \;\leq\; \text{Var}(\bar{y}_j)$
- **Shrinkage Factor:** $s_j = 1 - \omega_j$

$$\begin{aligned}\hat{\alpha}_j &= \mu_\alpha + s_j(\bar{y}_j - \mu_\alpha) \\ &= \text{overall mean} + \text{Filter} \times \text{(deviation)}\end{aligned}$$

CONCLUDING REMARKS

Large: number of observations is not the basis on which to judge the size of the dataset

Appropriateness of Analytical Approach: is linked directly to the specific question

Variation: is good and can be exploited to bolster inferences