STATISTICS IN MEDICINE

# The Sense and Sensibility of Sensitivity Analyses

Debbie M. Cheng, Sc.D., and Joseph W. Hogan, Sc.D.

Clinicians and public health officials routinely make evidence-based decisions that affect patient- and population-level health. The need for a principled approach to assessing and interpreting the robustness of findings from scientific investigations is more critical than ever. Randomized clinical trials frequently incorporate sensitivity analyses — for which relevant guidelines are available from the Food and Drug Administration (FDA)[1] and others[2-4] — yet misconceptions persist regarding their purpose and the proper reporting of their findings. Greater clarity is needed, since sensitivity analyses have the potential to influence a trial's conclusions, shape its narrative, and affect both policy and practice.

Sensitivity analyses assess the robustness of a trial's findings[3] by evaluating the degree to which analytic decisions and assumptions affect trial results. They should be preplanned and included in statistical analysis plans (SAPs). They should be designed to evaluate whether analytic choices, or challenges such as missing data, influence inferences regarding treatment effect. Post hoc sensitivity analyses may sometimes be needed; however, their use should be well justified (e.g., by violations of assumptions underlying the primary model) and minimized, because they introduce multiple potential sources of bias — such as data dredging, selective reporting, and inflation of the type 1 error rate — and thus have less credibility than preplanned analyses.

A trial SAP requires prespecification of the analytic approach, including choices about various aspects of the analysis. Examples include using a particular statistical model; making testable assumptions (e.g., linear relationship between two variables); and making untestable assumptions (e.g., about distribution of missing or censored data). Sensitivity analyses examine the effects of making reasonable alternative choices for these various aspects of the analysis. Typically, assumptions about model specification and outcome distribution are testable with observed data, whereas assumptions about missing data and censoring are not. Sensitivity analyses should address these types of assumptions separately.

For example, in a cluster-randomized trial with a binary end point, choices for analysis include a mixed-effects logistic-regression model with a normally distributed random intercept and logistic regression with fixed effects for cluster. Such choices about modeling assumptions can lead to different inferences about treatment effects. Frequently used models for missing data rely on the missing-at-random (MAR) assumption, which presumes that the probability of missingness depends only on observed data, and not on the missing values themselves. Although this assumption cannot be assessed with observed data alone, sensitivity analyses can quantify the degree to which treatment-effect inferences depend on assumptions about missing data. This can be accomplished by imputing missing data across a range of different assumptions and then comparing the resulting treatment-effect inferences across the different scenarios. Multiple imputation with the use of pattern-mixture models provides a structured way to conduct this analysis.[2]

Another option is a tipping-point analysis.[5] Here, data are reanalyzed under a range of missing-data assumptions that differ from those used for the primary analysis to identify the threshold at which qualitative conclusions from the primary analysis would be overturned. A clinically implausible tipping point suggests that the qualitative conclusion of the primary analysis is robust to missing-data assumptions. Generating a range of possible results across different missing-data assumptions provides a representation of uncertainty about inferences attributable to lack of knowledge about the missing data; if the qualitative conclusion about treatment effect doesn't change over a range of plausible assumptions, the primary analysis may be considered robust.

There are common pitfalls related to sensitivity analyses. Some types of analyses appear to assess robustness when in fact they do not. For example, it is common to analyze incomplete longitudinal data with a parametric random-effects model fit using maximum likelihood, which implicitly assumes MAR without having to impute missing data. A second analysis with the same model but using multiple imputation to fill in data under MAR would be expected to yield the same inference about treatment effect and thus would generally not constitute a sensitivity analysis. Similar-

ly, with large samples, a Bayesian analysis with weakly informative prior distributions would not be expected to yield meaningfully different conclusions from a frequentist analysis using the same model. Reporting such comparisons as sensitivity analyses would convey a false sense of robustness.

Comparing analyses that address different research questions or estimate different model parameters also does not constitute a sensitivity analysis.[4] For example, a per-protocol analysis does not assess the robustness of an intention-to-treat analysis because it applies only to participants who adhered to their assigned regimen. Subgroup analyses are not sensitivity analyses because they assess treatment-effect heterogeneity, not sensitivity to analysis choices and assumptions. Sensitivity analyses have the greatest utility and clearest interpretation when the parameter they estimate is the same one estimated in the primary analysis.

The necessity of prespecifying sensitivity analyses is particularly evident when treatment-effect inferences depend on using a model. Without prespecification, there is substantial potential for bias because the decision to pursue further analyses may depend on investigators' perception of the primary findings: if the findings appear to be favorable, investigators may want to avoid the risk of calling results into question; if the findings appear to be unfavorable, they may want to search for "better" results. We caution against both of these approaches and reemphasize that the purpose of sensitivity analyses is to systematically assess the robustness of the primary analysis; results should be reported in those terms, and the prespecified primary analysis must be reported transparently, regardless of sensitivity-analysis results.

Given the principles we've outlined, we have several recommendations for reporting sensitivity-analysis findings.

First, an assessment of sensitivity to key choices and assumptions should be summarized in a table comparing results from the primary analysis and each sensitivity analysis. The table should include treatment-effect estimates, confidence intervals, and the assumptions underlying each analysis.

Second, comparisons of treatment-effect findings should incorporate statistical uncertainty; comparing point estimates alone is neither appropriate nor scientifically meaningful. To assess whether substantive conclusions differ between primary and sensitivity analyses, investigators should consider both point estimates and confidence intervals or present the difference between treatment effects in terms of standard-error units. Although neither method represents a formal test to compare results from different analyses, both are practical approaches incorporating measures of uncertainty. No hard and fast rules exist for determining robustness, but the approaches described above take into consideration measures of uncertainty.

Third, investigators should explain the assumptions made with each method and whether testable assumptions are supported by the data. Such contextualization is critical when the conclusions from primary and sensitivity analyses differ substantially. For testable assumptions, if two models provide qualitatively different conclusions, it should be feasible to empirically assess which model is more appropriate. The model that makes fewer assumptions, and is therefore more flexible, may be preferred, although the prespecified primary analysis must still be reported as such. For untestable assumptions, such as those about missing data, when the sensitivity analyses indicate that the treatment effects may differ from those in the primary analysis, authors should report a range of results across a set of sensible assumptions and possibly include a tipping-point analysis. There is no way to determine which of several untestable assumptions is "better"; however, sensitivity analyses provide important information about the robustness of the primary analysis.[4] The FDA has published guidance on implementing a structured approach to sensitivity analyses in the presence of testable and untestable assumptions.[1]

Fourth, investigators should frame the key conclusions in terms of the primary analyses and use the sensitivity analysis to represent the degree of uncertainty attributable to modeling choices and untestable assumptions. If testable assumptions underlying the primary analysis are deemed to be inappropriate, that assessment should be reported. However, authors should avoid using sensitivity analysis to either replace the prespecified primary analyses or draw definitive new conclusions. All analyses should be transparently reported and clearly designated (e.g., primary vs. sensitivity, preplanned vs. post hoc).[3]

Sensitivity analyses are instrumental in evaluating the validity of trial conclusions. Careful thought is required to preplan and report them appropriately because trial data can directly affect patient care, resource distribution, and public health policy. A principled approach to ensuring that sensi-

tivity analyses are conducted, presented, and interpreted responsibly will promote the scientific integrity of trials and advance the quality of both medical knowledge and patient care.

The views expressed in this article are those of the authors and do not represent the views or policy of the *Journal*, where Dr. Hogan is a statistical editor.

Disclosure forms provided by the authors are available at NEJM.org.

From the Department of Biostatistics, Boston University School of Public Health, Boston (D.M.C.); and the Department of Biostatistics, Brown University School of Public Health, Providence, RI (J.W.H.).

This article was published on September 14, 2024, at NEJM.org.

1. Food and Drug Administration. E9(R1) statistical principles for clinical trials: addendum: estimands and sensitivity analysis in clinical trials. Guidance for industry. May 2021 (https://www.fda.gov/media/148473/download).
2. Little RJ, D'Agostino R, Cohen ML, et al. The prevention and treatment of missing data in clinical trials. N Engl J Med 2012;367:1355-60.
3. Thabane L, Mbuagbaw L, Zhang S, et al. A tutorial on sensitivity analyses in clinical trials: the what, why, when and how. BMC Med Res Methodol 2013;13:92.
4. Morris TP, Kahan BC, White IR. Choosing sensitivity analyses for randomised trials: principles. BMC Med Res Methodol 2014;14:11.
5. Goudman L, Molenberghs G, Duarte RV, Moens M. The influence of missing data on disabilities in patients treated with high-dose spinal cord stimulation: a tipping point sensitivity analysis. J Clin Med 2021;10:4897.

DOI: 10.1056/NEJMp2403318


# No One in Charge

Janet R. Gilsdorf, M.D.

I didn't know him, didn't even know his name. He said hello when he took the aisle seat beside me on our 4-hour flight to the West Coast, and then we didn't speak. In silence, I ate my omelet while the stranger beside me ate his French toast. Then I dozed and worked on my knitting project while he watched a movie. Neither of us saw a reason to introduce ourselves.

Then, about 45 minutes before we landed, my seatmate had a discussion with a flight attendant about congested airport lounges. When the flight attendant resumed her duties, I asked him a question about the lounges, and we began to talk. I learned that he was a retired military officer, and he learned that I'm a retired pediatrician. We chatted about our beautiful home state of Michigan and about our children.

I could tell from the way he spoke and what he said that he was a leader. His sentences were carefully constructed, and his words well chosen. His voice was confident and commanding. He explained what he used to tell his troops, the rigors of basic training, the unique, ordered culture of the military.

In a quiet voice, with his eyes focused on the seat ahead of him, he told me that his wife had died about a year earlier, and that his kids were pressuring him to move closer to one of them, to either the Atlantic or the Pacific coast. He praised his late wife, describing how she had made a beautiful home for their family wherever in the world they found themselves, whether in Saudi Arabia, Germany, or Washington, DC. I noted the gold wedding band that still encircled his left ring finger.

He took a deep breath, turned toward me, and said, "She was diagnosed with cancer and was dead 6 weeks later."

"That's pretty tough," I said, looking into his solemn eyes.

He nodded. "The hardest thing about her long hospitalization was that no one was in charge."

No one in charge? I hesitated and then, awkwardly, I explained that every inpatient has a physician of record — the doctor in charge.

What did he mean by "no one in charge?" He was probably referring to "the team," the physicians, nurse practitioners, physician assistants, nurses, residents, and students who provide clinical care at teaching hospitals. His comments made it clear that he saw this group as disorganized and leaderless.

Every medical team has a leader, I thought, the person ultimately responsible for setting the goals, for designing the therapeutic and diagnostic plans, for organizing the rest of the team. Who, I wondered, was that person for my seatmate's wife? It might have changed from day to day or week to week, but the man should have known who was at the top of the chain of command in his wife's care. Like his wife-the-patient, he, too, had been desperately fighting the proverbial war on cancer, and he had felt lost.

"Well," he said in measured tones as he slowly ran his fingers up and down the sides of his coffee cup, "when she was first admitted, a guy introduced himself as her oncologist, but I never saw him again." He paused a moment, and then added, "Never." The word sounded like the thrust of a dagger.