SAMPLE SIZE AND THE NUREMBERG CODE

L. Jane Goldsmith and Paul Simmons
University of Louisville Medical School, Louisville, KY 40292

Keywords: Ethics, Power, Sample size, Human Subjects

In post World War II Germany in August 1947, the Nuremberg Code was formulated by judges and physicians during the trial of 23 Nazi physicians and scientists accused of murder and torture in the conduct of medical experiments in European concentration camps during World War II (Shuster, 1998). Characterized as the most authoritative set of rules for the protection of human subjects in medical research, the Nuremberg Code has never been adopted by any country or institution as a legal guideline for research. The primary reason is principle 5:

5. "No experiment should be conducted where there is an *a priori* reason to believe that death or disabling injury will occur; except perhaps, in those experiments where the experimental physicians also serve as subjects."

Although never adopted *in toto*, the Code has had a profound impact on research (Shuster, 1997). Principle 1, inspired by the horrific research conducted on unwilling concentration camp prisoners, laid the foundation for "informed consent":

1. "The voluntary consent of the human subject is absolutely essential . . . "

In modern industrialized countries of the west, prospective research subjects must be informed of possible harm and sign a consent form.

Beneficence
Beneficence has been defined as "the obligation to protect persons from harm by maximizing benefits and minimizing potential harms" (Miller, 2001). Although not as well known as principle 1, Nuremberg principle 2 can be construed to have relevance to beneficence through accurate sample size determination:

2. "The experiment should be such as to yield fruitful results for the good of society, . . . "

Other pertinent items from Nuremberg are:

4. "The experiment should be so conducted as to avoid all unnecessary physical and mental suffering and injury."

8. "The experiment should be conducted only by scientifically qualified persons. The highest degree of skill and care should be required through all stages of the experiment of those who conduct or engage in the experiment."

Recent Codes
More recent codes of ethics for research on human subjects are the Declarations of Helsinki (1964, 1975, 1983, 1989, 1996, 2000) and the Belmont Report (1979), which governs research in the U.S. These codes, continuing the impetus of the Nuremberg Code, address in greater detail three main issues in research on human subjects:

- Autonomy—Defines a prospective human subject as a self-governing human being with protections and rights
- Beneficence—Admonishes researchers to do good and minimize harm, as described before
- Justice—Requires the proper distribution of burdens and benefits.

Ethics for Statisticians
Statisticians have a special appreciation for the importance of appropriate, careful sample size determination. Sometimes statisticians allude to stewardship: "Selecting an insufficient sample size yields a study with inadequate sensitivity, whereas selecting an excessive sample size wastes resources." (Muller, *et al*, 1992)

The American Statistical Association in 1999 approved "**Ethical Guidelines for Statistical Practice**." Concerned about research subjects, the following admonishments appear under heading **D. "Responsibilities to Research Subjects"**:

2. "Avoid the use of excessive or inadequate numbers of research subjects by making informed recommendations for study size. These recommendations may be based on prospective power analysis, the planned precision of the study endpoints(s), or other methods . . . "

3. "Avoid excessive risk to research subjects and excessive imposition on their time and privacy."

Statisticians are also aware of the amount of work and time needed to accomplish adequate sample size determinations, which can take a month or more (Muller, *et al*, 1992). The pay-off in terms of resources is also well-known. For example, in the chapter "Maximizing Power in Randomized Designs When N is Small" (Venter and Maxwell, 1999), 4 research designs are examined for the same effect size through 5 hypothesized values for the within-subject correlation. The number of needed subjects ranges from 13 to 128 for different designs that yield the same statistical power. Sample size was calculated 20 times for this comparison.

The inefficiencies associated with the common practice of collecting uninformative data have also been documented (Cohen, 1983, and Goldsmith, 1995).

Some Bad News

In spite of the ethical guidelines and increased statistical knowledge, it is apparent that research is conducted without proper sample size calculations. Papers such as "Inadequate size of 'negative' clinical trials in dermatology" (Williams and Seed, 1997) document the problem. The scathing editorial entitled "The Scandal of Poor Medical Research" (Altman, 1994) notes:

"When I tell friends outside medicine that many papers published in medical journals are misleading because of methodological weaknesses they are rightly shocked. Huge sums of money are spent annually on research that is seriously flawed through the use of inappropriate designs, unrepresentative samples, small samples, incorrect methods of analysis, and faulty interpretation."

Experience in biostatistics consulting and perusal of medical literature gives evidence that medical researchers do plan clinical studies without an expert biostatistician, obtain approval from human studies committees, and conduct inappropriate studies on humans.

Another Viewpoint

In a viewpoint paper "Why 'underpowered' trials are not necessarily unethical," (Edwards, *et al*, 1997) Edwards and colleagues acknowledge "A systematic review of the literature on the ethics of clinical trials confirmed that 'underpowered' trials are generally regarded as unethical." The authors assert that if a prospective subject is in "equipoise," that is, if the expected harm and benefits of study participation are equal, then a patient should not care whether the sample size is ideal. Figure 1 shows a decision tree (TreeAge Software, 2001) showing patient choices of either the gold standard therapy or randomization to a clinical trial. Arbitrary utilities for Complete recovery (1.0), Recovery with side effects (.8), and Treatment failure (0.0) are shown on the right. In this idealized case, the experimental treatment is expected to perform exactly like the gold standard. The expected utility is the same (.78), whether the subject is randomized or not, and the software notes that the choice is indifferent. The prospective subject is in equipoise regarding this decision. Figure 2, another decision tree, demonstrates the situation in which the experimental therapy is more likely to bring good results. In this case, the software, quite rightly in some peoples' minds, recommends the choice of entering the clinical trial through randomization, as the expected utility, or benefit to the individual patient, is greater in the clinical trial, where the possibility exists of obtaining better therapy. The patient need not care about sample size—he or she makes the decision most advantageous for himself or herself.

Now consider Figure 3. In this hypothetical case, the experimental treatment is not as good as the gold standard. Similar situations arise, in spite of valiant efforts to weed out ineffective therapies in the earlier clinical trial Phases I and II (Lachenbruch, 2001). In this case, the patient can expect to be worse off (have lower expected utility) if he or she chooses to participate in the clinical study. TreeAge software recommends that the patient not be randomized to the clinical trial.

However, many patients, real heroes of medical research, choose to participate in clinical trials when their personal expectations are not excellent. They want to help their fellow man by increasing medical knowledge. Figure 4 shows the tree diagram when the utility functions are adjusted upward to reflect a prospective subject's values regarding medical research. Even though the treatment may be worse, the patient can be advised to choose randomization. The probabilities for advantageous treatment haven't changed from the situation depicted in Figure 3, just the values of the patient.

The patient prefers randomization due to the fact that well-designed research will add to the body of knowledge. This is the "implicit contract with study subjects, who believe that their participation will contribute to the answer of an important pending medical question" (Harrington, 2000).

Meta-analysis

Edwards, *et al*, in their viewpoint article, suggest that even clinical research conducted with inadequate numbers of subjects will eventually be analyzed and

contribute to knowledge. Even ignoring the so-called file-drawer problem, which notes that research that is not statistically significant often remains unpublished and thus not readily available in the medical literature, there are problems with meta-analyses. Two recent articles, "The Promise and Problems of Meta-Analysis" (Bailar, 1997), and "Discrepancies between Meta-Analyses and Subsequent Large Randomized, Controlled Trials" (LeLorier, *et al*, 1997) describe many failings of the current state the art of meta-analysis. In "The Tea Leaves of Small Trials," Harrington describes two meta-analyses that give divergent views of the same treatment (1999).

We are not keeping our covenant with research subjects who care about participating in good research if we expose them to the risks and rigors of undersized clinical trials in hopes that the data may be used in some future halcyon time of perfect meta-analysis.

Data mining in the future

Promising prospective subjects that the data will be stored for future use in research is not adequate insurance that their sacrifice will be put to good use, either. Indeed, the reason for randomized clinical trials instead of comparing new treatments to historical controls is to compare data collected along a time-line that insures comparable adjunct care.

The severe problems with long-term storage of data, maintenance of confidentiality, and the slow development of the electronic medical record also must be considered. Hoping for eventual success through data mining is not a substitute for well-planned, efficient research that brings results in the most timely fashion.

Back to informed consent

Principle 5 of the Nuremberg Code, proscribing life-threatening research, has been in a sense circumvented by describing risk in an informed consent document. Speaking tongue-in-cheek, perhaps the same approach can be used to facilitate poorly planned studies. If a research institution chooses not to fund an adequate number of expert biostatisticians to plan research (violation of principle 8, "**scientifically qualified persons**") and to approve research designed by tyros which may require too large a sample size (violation of principle 4, "**avoid all unnecessary physical and mental suffering and injury**") or be undersized (violation of principle 2, "**fruitful results for the good of society**"), then perhaps the following paragraph should be included in the informed consent document:

I understand that this research has **not** been carefully planned in accordance with the Nuremberg Code.

In particular, I understand that the experiment design has not been approved by expert biostatisticians, even though they are available in all industrialized countries of the west and probably are available at [insert name of institution]. Careful, state-of-the-art scientific consideration has not been given to sample size calculation, or even to the type of data to be collected. If the experiment is too small, this research may contribute little or nothing to advancing the human condition. If the sample size is too large, too many subjects may be at risk. I choose to participate as a human subject in this research, fully acknowledging that any suffering, loss, or inconvenience on my part is likely to produce little, if any, benefit to mankind. I further acknowledge that the resources of [insert name of funding institution] are likely to be wasted by this poorly designed research.

Signed: _____
(At last) A fully informed prospective human subject

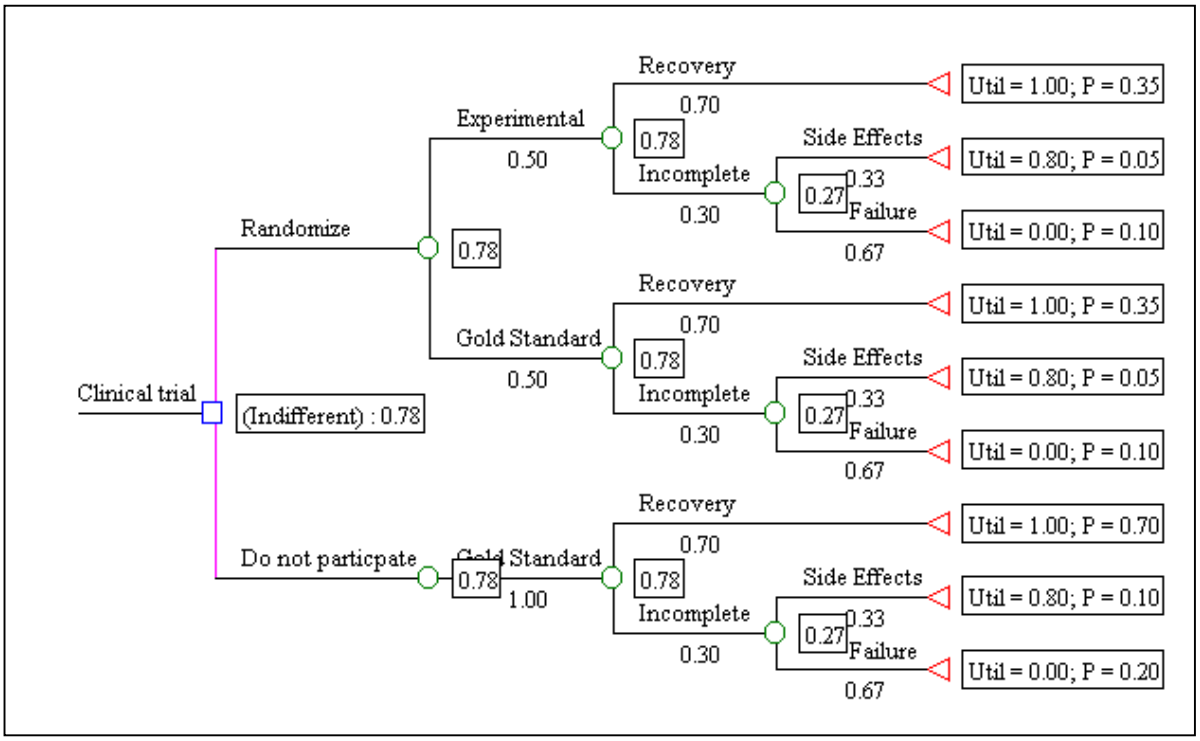

Figure 1. Decision tree showing equipoise for a prospective subject's participation in a clinical trial.

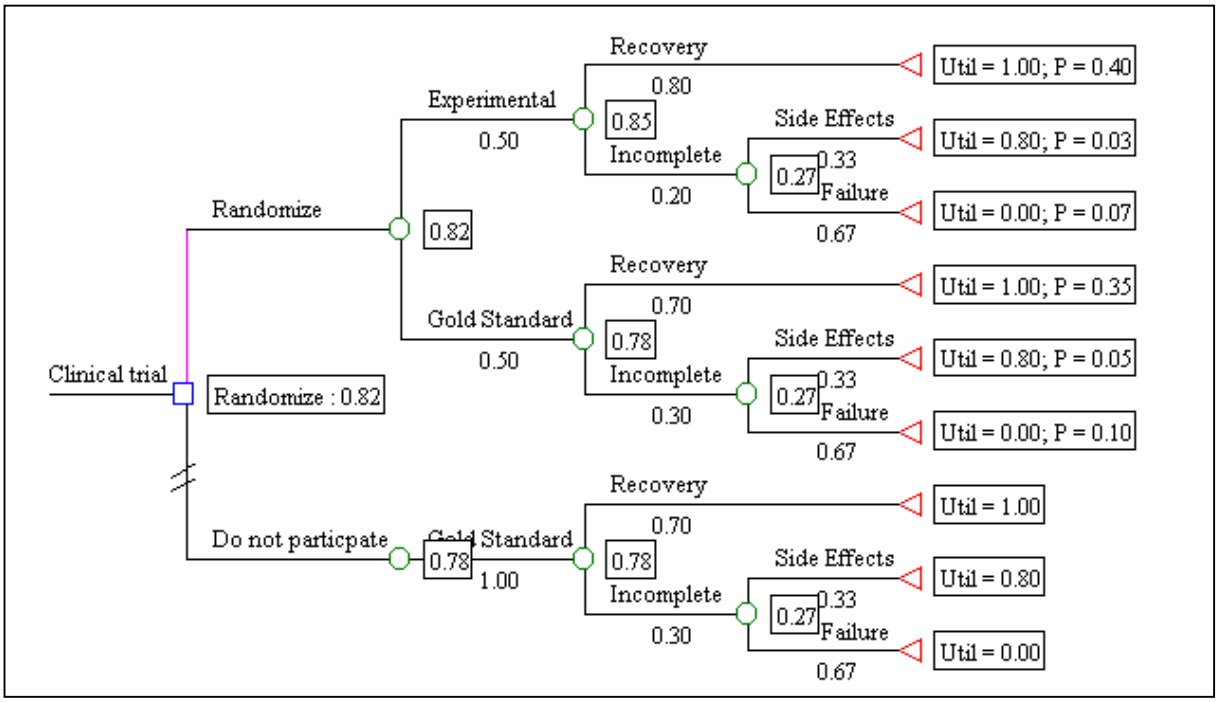

Figure 2. Promising experimental treatment makes participation more desirable.

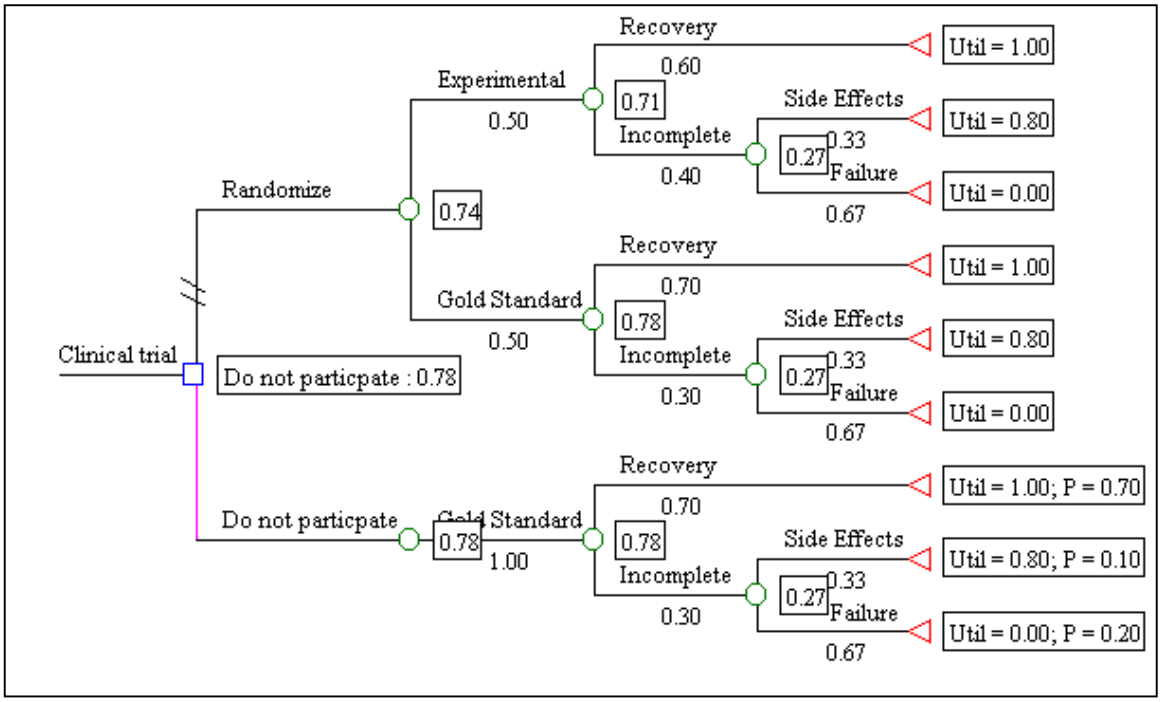

Figure 3. Propect of poor experimental therapy leads to decision of non-participation as human subject.

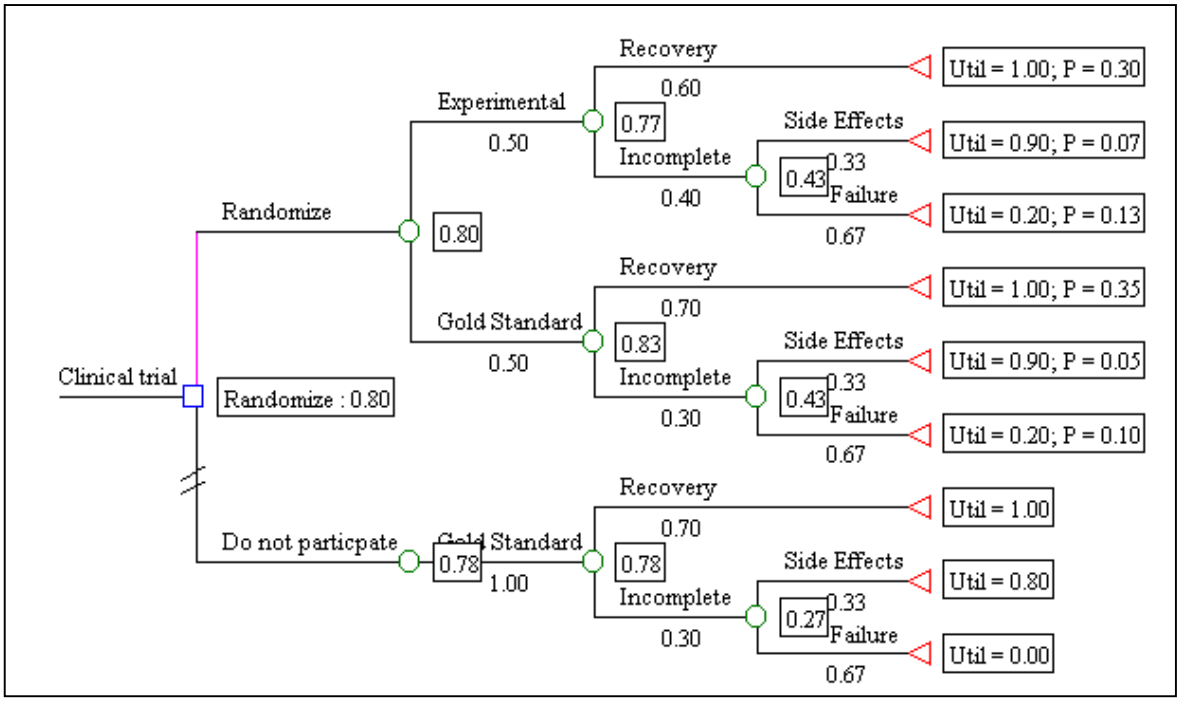

Figure 4. Utilities reflect prospective subject's value system: enter a clinical trial to help fellow man.

REFERENCES

Altman, D.G. (1994) The Scandal of Poor Medical Research. The British Medical Journal. 308(6924): 283-284.

Bailar, J.D. III (1997) The Promise and Problems of Meta-Analysis. 337(8): 559-561.

Belmont Report (1979) http://ohrp.osophs.dhhs.gov/humansubjects/guidance/belmont.htm

Cohen, J. (1983) The Cost of Dichotomization. Applied Psychological Measurement. 7(3):249-253.

Declarations of Helsinki (1964, 1975, 1983, 1989, 1996, 2000) http://www.wits.ac.za/bioethics/helsinki.htm

Edwards, S.J.L., Lilford, R.J., Braunholtz, D., and Jackson, J. (1997) Why "underpowered" trials are not necessarily unethical. The Lancet. 350(9080): 804-807.

Goldsmith, L.J. (1995) Pros and Cons of Cut-points. Proceedings of the Biometrics Section, American Statistical Association, Annual Meeting, Orlando, FL, 272-275.

Harrington, D.P. (1999) The Tea Leaves of Small Trials. Journal of Clinical Oncology. 17(5): 1336-1338.

Harrington, D.P. (2000) The Randomized Clinical Trial. Journal of the American Statistical Association. 95(449): 312-315.

Lachenbruch, P.A. (2001) FDA Statistics. Short course at International Biometrics Society (Eastern North American Region) Spring Meeting. March 25, 2001, Charlotte, SC.

LeLorier, J., Gregoire, G., Benhaddad, A., Lapierre, J., and Derderian, F. (1997) Discrepancies between Meta-Analyses and Subsequent Large Randomized, Controlled Trials. 337(8): 536-542.

Miller, R.L. (2001) Protection of Human Subjects in Research, Instructional Video, University of Louisville.

Muller, K.E., LaVange, L.M., Ramey, S. L., and Ramey, C.T. (1992) Power Calculations for General Linear Multivariate Models Including Repeated Measures Applications. Journal of the American Statistical Association. 87(420): 1209-1226.

Shuster, E. (1997) Fifty Years Later: The Significance of the Nuremberg Code. New England Journal of Medicine. 337(40): 1436-1440.

Shuster, E. (1998) The Nuremberg Code: Hippocratic ethics and human rights. The Lancet. 351(9107): 974-977.

TreeAge Software (2001) TreeAge Software, Inc., 1075 Main St.Williamstown, MA 01267USA

Venter, A., and Maxwell, S.E. (1999) Maximizing Power in Randomized Designs When N Is Small. Chapter 2 in *Statistical Strategies for Small Sample Research*, Hoyle, R. E. (ed.), Sage, Thousand Oaks.

Williams, H.C. and Seed, P. (1997) Inadequate size of 'negative' clinical trials in dermatology 136(1): 151.