There's no difference—are you sure?

In September 2014, we discussed the importance of an a priori sample-size calculation and other considerations in planning and conducting a study to be sure that it is sufficiently powered to detect the smallest clinically relevant effect size. A reader can then interpret an adequately powered, but negative, study to mean that the difference between treatment groups is not larger than that predetermined effect size and can use the information to support therapeutic decisions. Sample-size calculations are now required by many Institutional Review Boards, funding agencies, and medical journals.

Much less attention has been paid to the issues involved with reporting negative results, particularly when seeming to refute a previous positive report or reports. We will use three examples to illustrate four questions that should be considered: 1) What is the smallest difference that the negative study would have been able to detect given the sample size? 2) If the difference is not statistically significantly different between groups, how informative is a "post hoc" power calculation? 3) Were there limitations in study design, such as measurement error, that may have contributed to a null finding? and 4) Do basic science and other findings support the hypothesis that a difference would be expected?

Growth hormone (GH) cotreatment of poor responders to improve IVF outcomes is our first example. In the March 2016 issue of this journal, Bassiouny et al. randomized 141 participants, reported ongoing delivery rate per cycle start to be 14.7% (10/68) in the GH group and 10.9% (8/73) in the control group (P=.51), and concluded that "there is still no identified impact on pregnancy outcomes." They pointed out that their study was underpowered, yet final enrollment went beyond the original planned enrollment of 120 subjects. How should readers consider the results of this trial, particularly in the context of earlier positive trials summarized in meta-analyses from the Cochrane Group and others which reported substantial (two- to sixfold) increases of pregnancy and delivery rates with GH (1)? Moreover, from translational research, follicular fluid insulin-like growth factor 1 levels are lower in poor responders and in IVF cycles not resulting in success (1), so an impact of GH treatment should not be surprising.

First, how big would the study have needed to be to detect the smallest clinically relevant effect size? Using their 10.9% delivery rate as the base (type I error 5%, power 80%), 554, 229, and 74 subjects would be required per group (assuming equal allocation) to detect absolute increases of 5% (10.9% to 15.9%), 10% (10.9% to 20.9%), and 20% (10.9% to 30.9%) in the outcome, respectively. As the detectable effect size increases, the sample size decreases, which improves study feasibility but at the cost of not being powered to detect a smaller effect size. Their study contained fewer total subjects than would have been required to detect an almost threefold increase. If the delivery rate might be even 1.5-fold higher with addition of GH (certainly clinically relevant) it would be quite unfortunate if a reader decided against use of GH based on the authors' written conclusions without a detailed review of the paper, previous published papers and reviews, and the above considerations.

Second, would it have been helpful to report a "post hoc" power calculation or conclude that there is a "trend"? We think that undertaking a power calculation after the data have been collected and analyzed is less helpful for explaining the observed data than providing the treatment effect size and its associated confidence interval (2). Nonsignificant P values will always correspond to low power, and a "post hoc" power calculation will not change our interpretation that the P value was not statistically significant. Rather, in addition to the treatment effects and associated P values, we advocate for reporting their associated confidence intervals (CIs). The CI will include 0 for absolute risks or 1 for relative risks/odds ratios, as well as the range of values that cannot be rejected. If this range includes a potentially clinically important treatment effect, there would be the suggestion of potential clinical significance for the treatment that might be revealed by larger studies. Alternatively, a narrow CI that includes 0 and excludes effects that are clinically important may help readers to conclude that the intervention is neither statistically nor clinically significant. In the above study, the relative risk can be calculated to be 1.34 (95% CI 0.56–3.20). All values covered by the CI cannot be rejected, including a potential 3.2-fold increase in delivery rate.

Another example of powering a study to detect the smallest clinically relevant effect size is a retrospective cohort study of 517 women in which the ongoing pregnancy rate with transfer of euploid embryos was not associated with the patients' vitamin D levels (3). Despite the study seeming to be large and definitive, the authors provided a post hoc power calculation that reported the detectable effect size to be $\geq$18% absolute difference in ongoing pregnancy rates. Alternatively, the CI could have been highlighted to show readers that an odds ratio of 0.95 (95% CI 0.58–1.58) for ongoing pregnancy in vitamin D replete ($\geq$30 ng/mL) compared with deficient (<20 ng/mL) women does not preclude an increase in odds of almost 1.6-fold, which is still clinically important. The authors were careful to emphasize that further studies are necessary.

Finally, heterogeneity in study design for measuring risk exposure to the study population can explain part of why there may be both negative and positive studies on the same topic. Several years ago, when researching the effects of overcooking of foods, increased risks of colon, pancreas, prostate and breast cancers were listed, whereas more recently, the risk of breast cancer is no longer mentioned (https://www.cancer.gov/about-cancer/causes-prevention/risk/diet/cooked-meats-fact-sheet). In 1998 (4), in a case-control study, the investigators asked subjects within the Iowa Women's Health Study to fill out detailed questionnaires regarding their doneness preference of beef, hamburger, and bacon and whether their ingestion of well-done meat was a consistent food habit. Participants were given color photographs illustrating the extent of doneness from rare to very well done. A dose-response relationship was found, and women who consistently consumed those meats very well done had an odds ratio of 4.6 (CI 1.36–15.7) compared with women who consumed them rare or medium done. Subsequently, other investigators used Nurses' Health Study participants (5) to prospectively collect detailed information on meats consumed and cooking methods to estimate their levels of exposure to known carcinogenic compounds and associations with subsequent development of breast cancer. In that large prospective cohort, an association between higher exposure and breast cancer was not observed.

There are several epidemiologic principles to consider when evaluating this negative study in the context of basic science, nonhuman animal, and other human studies that have reported a positive association between consumption of meats cooked at high temperature and cancer. One is how exposure is measured, which may contribute to misclassification. People's perceptions of what is meant by well-done can be very different, hence the use of colored photographs in the Iowa Women's Health Study (4). Perhaps more important are the inherent differences among studied populations. Iowa has one of the highest levels of red meat production and intake in the United States. The finding of significance regarding a risk exposure can be easily diluted away by inclusion of a large proportion of individuals without a high level of exposure to the risk. The consistent consumption of very well done red meats could likely be very different in women from Iowa compared with a national group of women in a health care profession. Recall bias can also threaten validity of findings as, for example, women who have had cancer (case subjects) may recall their intake of potential carcinogens differently than women who did not have cancer (control subjects). Both studies did take measures to limit that source of bias.

All those who have written on this subject readily acknowledge that high levels of potent known carcinogens are found in well-cooked red meats and cause mammary cancer in laboratory animals. They also acknowledge that the relationships of well cooked meat to colon, pancreatic, and prostatic cancers are well established. Finally, there has been a disturbing increase in doneness of consumed foods (the number of barbecuing events doubled from 1987 to 1997 [www.barbecuen.com/bbqstats.htm#axzz4amBSh1c9]), and there has been a very concerning recent increase of colon cancer in young adults (https://www.cancer.org/latest-news/study-finds-sharp-rise-in-colon-cancer-and-rectal-cancer-rates-among-young-adults.html). In fertility, there has also been concern regarding the increased production of advanced glycation end-products (AGEs) in foods cooked at high temperature, and one study found an association of follicular fluid levels of AGEs with failure to conceive with the use of IVF (www.lifechoicesandfertility.com).

Unfortunately, overcooking of meats, fish, and vegetables is one of the easiest and least expensive ways for restaurants and individuals to add flavor to foods, and barbecuing has become as American as apple pie. The word "grill" in restaurant names and grill marks on many served foods are rampant. There are more than 100 other flavors that can be added to foods, and recipes are only an internet search away. Every reader of this journal possesses a colon and a pancreas and either a prostate or breasts, and most of our patients have difficulties conceiving. All should be concerned about the impact of overcooking of foods on cancer risk, and we look forward to future studies to clarify the relationship between these food exposures and fertility. We hope that, as well as illustrating some considerations when evaluating negative studies, we have also helped to bring this important topic to the forefront.

David R. Meldrum, M.D.[a,b]
H. Irene Su, M.D., M.S.C.E.[b]
[a] Reproductive Partners San Diego; and [b] Division of Reproductive Endocrinology and Infertility, University of California, San Diego, California

http://dx.doi.org/10.1016/j.fertnstert.2017.06.022

You can discuss this article with its authors and with other ASRM members at https://www.fertstertdialog.com/users/16110-fertility-and-sterility/posts/17987-24446

REFERENCES

1. Bosch E, Labarta E, Kolibianakis E, Rosen M, Meldrum D. Regimen of ovarian stimulation affects oocyte and therefore embryo quality. Fertil Steril 2016; 105:560–70.
2. Hoenig J, Heisey D. The abuse of power: the pervasive fallacy of power calculations for data analysis. Am Stat 2001;55:1–6.
3. Franasiak JM, Molinaro TA, Dubell EK, Scott KL, Ruiz AR, Forman EJ, et al. Vitamin D levels do not affect IVF outcomes following the transfer of euploid blastocysts. Am J Obstet Gynecol 2015;212:315.e1–6.
4. Zheng W, Gustafson DR, Sinha R, Cerhan JR, Moore D, Hong CP, et al. Well-done meat intake and the risk of breast cancer. J Natl Cancer Inst 1998;90: 1724–9.
5. Wu K, Sinha R, Holmes MD, Giovannucci E, Willett W, Cho E. Meat mutagens and breast cancer in postmenopausal women—a cohort analysis. Cancer Epidemiol Biomarkers Prev 2010;19:1301–10.