ELSEVIER

Inclusion of Women and Minorities in Clinical Trials and the NIH Revitalization Act of 1993—The Perspective of NIH Clinical Trialists

Laurence S. Freedman, MA, Richard Simon, DSc, Mary A. Foulkes, PhD, Lawrence Friedman, MD, Nancy L. Geller, PhD, David J. Gordon, MD, PhD, and Richard Mowery, PhD

Division of Cancer Prevention and Control, National Cancer Institute (L.S.F.), Division of Cancer Treatment, National Cancer Institute (R.S.), Division of AIDS, National Institute of Allergy and Infectious Diseases (M.A.F.), Division of Epidemiology and Clinical Applications, National Heart, Lung, and Blood Institute (L.F., N.L.G.), Division of Heart and Vascular Diseases, National Heart, Lung, and Blood Institute (D.J.G), and Division of Collaborative Research, Extramural Research, National Eye Institute (R.M.), all addresses in Bethesda, Maryland

1. BACKGROUND

The NIH Revitalization Act of 1993 [1] was signed by President Bill Clinton on June 10, 1993. Aside from authorizing NIH [the National Institutes of Health] to carry on its mission, a section of the Act directed the NIH to establish guidelines for inclusion of women and minorities in clinical research. This section will have wide-ranging implications on the conduct of NIH-sponsored clinical research. The statute defines "clinical research" to include "clinical trials" and states that:

"In the case of any clinical trial in which women or members of minority groups will be included as subjects, the Director of NIH shall ensure that the trial is designed and carried out in a manner sufficient to provide for valid analysis of whether the variables being studied in the trial affect women or members of minority groups, as the case may be, differently than other subjects in the trial" (492 B(c)).

The statute further allows exclusions to the requirement for entering women and minorities in clinical trials, as follows:

"In the case of a clinical trial, the guidelines may provide that such inclusion in the trial is not required if there is substantial scientific data demonstrating that there is no significant difference between

Address reprint requests to: Laurence Freedman, Biometry Branch, DCPC, NCI, Executive Plaza North, Suite 344, Bethesda, Maryland 20892, Tel. No.: 301-496-7748.

Received August 17, 1994; accepted February 21, 1995.

Controlled Clinical Trials 16:277–285 (1995)

655 Avenue of the Americas, New York, NY 10010

0197-2456/95/$9.50
SSDI 0197-2456(95)00048-L

(i) the effects that the variables to be studied in the trial have on women or members of minority groups, respectively; and
(ii) the effects that the variables have on the individuals who would serve in the trial in the event that such inclusion were not required" (492B(d)(2)(B)).

NIH convened a Planning Group for writing guidelines on the implementation of this section of the NIH Revitalization Act. The Group was chaired by Dr. Wendy Baldwin, Deputy Director for Extramural Research at NIH. The Planning Group asked the NIH clinical trials community for advice on the interpretation of the Act and its implementation, particularly with regard to the above-quoted passages relating to clinical trials.

In response, a general meeting of NIH clinical trialists was convened at which the issues were discussed, and a subgroup was formed comprising the authors of this paper. The subgroup prepared a document setting out their recommendations to the Planning Group, broadly following the views expressed at the general meeting. The Planning Group enthusiastically accepted the Recommendations and incorporated concepts described in that document into the writing of the NIH Guidelines. These Guidelines were published in the *Federal Register* [2] on March 28, 1994.

Although the Guidelines include the conclusions of the Recommendations, restriction on publication space prevented including the full Recommendations document. Because we feel that it is important to convey the conceptual background to the Guidelines we provide below a version of our Recommendations to the Planning Group.

2. RECOMMENDATIONS FROM THE NIH CLINICAL TRIALS COMMUNITY

A subgroup of biostatisticians, epidemiologists, and clinicians, representing all of the NIH institutes (the authors of this paper), was organized to address the definition of key phrases used in the NIH Revitalization Act provisions concerning the inclusion of women and minorities in clinical research. The subgroup viewed this as being of the greatest importance because the interpretation of these phrases, especially "a valid analysis . . . differently than other subjects in the trial" (492B (c)), would, in their opinion, determine whether this Act would achieve its laudable goal of ensuring that clinical trials address the health needs of women and minorities or whether it would seriously impair the ability of NIH to carry out clinical trials at all. It was therefore crucial that ambiguities in the language of the Act be implemented in such a way as to be consistent with the practical and ethical conduct of clinical trials.

We interpreted the Act as demanding appropriate representation of subjects of different gender and race/ethnicity in clinical trials so as to provide the opportunity for detecting major qualitative differences (if they exist) among gender and racial/ethnic subgroups and to identify more subtle differences that might, if warranted, be explored in further specifically targeted studies. This is a policy that we strongly support. Other interpretations may serve less well the health needs of women, minorities, and all other constituencies.

We begin with giving our definitions for key phrases in the Act. These are "clinical trial," "valid analysis," and "significant difference." (*Readers should return*

to Section 1 to review the use of these items in the Act.) Following these definitions, we will explain in detail our interpretation of the sections of the Act that deal with clinical trials, and our underlying rationale.

3. DEFINITIONS

3.1 Definition of "Clinical Trial"

The term "clinical trial" is used several times in the NIH Revitalization Act. Within the context of the Act, we understand "clinical trial" to refer to broadly based Phase III clinical investigations.

By this we mean prospectively designed studies involving human subjects that evaluate the effectiveness of an experimental intervention in comparison with a control or standard intervention or that compare two or more existing interventions, that are usually designed to enroll at least several hundred subjects, and that are designed to provide evidence that is potentially sufficiently definitive to lead to a broad change in public health policy or change in standard of care. These studies are performed after evidence of efficacy and safety have been obtained in preliminary investigations. The definition includes both pharmacologic and nonpharmacologic interventions, given for disease prevention, prophylaxis, diagnosis, or therapy.

3.2 Definition of "Valid Analysis"

". . . the trial is designed and carried out in a manner sufficient to provide for a *valid analysis* of whether the variables being studied in the trial affect women or members of minority groups, as the case may be, differently than other subjects in the trial."

The term "*valid analysis*" is scientifically understood to mean an unbiased assessment. Such an assessment will, on average, yield the correct estimate of the difference in outcomes between two groups of subjects. The main requirements to ensure a valid analysis of the question of interest are (1) allocation of study participants of both genders and from different racial/ethnic subgroups to the intervention and control groups by an unbiased procedure such as randomization, (2) unbiased assessment of the outcome of study participants, and (3) use of unbiased statistical analyses and proper methods of inference to estimate and compare the intervention effects among the gender and racial/ethnic subgroups.

3.3 Definition of "Significant Difference"

"Such inclusion (of women or minorities) in the trial is not required if there is substantial scientific data demonstrating that there is no *significant difference* between

(i) the effects . . . on women or members of minority groups, and
(ii) the effects . . . on the individuals who would serve as subjects in the trial in the event that such inclusion were not required."

A "significant difference" in the context of this clause is understood scientifically to mean a difference that is of clinical or public health importance. For example,

an intervention having a clearly harmful effect in a minority subgroup compared to a clearly beneficial effect in other subjects would constitute a significant difference between the effect in the two groups (i.e., groups (i) and (ii) above).

This definition differs from the commonly used "statistically significant difference" which refers to the event that, for a given set of data, the statistical test for a difference between the effects in two groups achieves statistical significance. Statistical significance depends on the amount of information in the data set. With a very large amount of information, one could find a statistically significant, but small, difference that is of very little importance. Conversely, with less information one could find a large difference of potential importance that is not statistically significant.

The definitions presented in subsections 3.1, 3.2, and 3.3 fit into an underlying rationale that guides our interpretation of the Act. This rationale is presented in the next section.

4. RATIONALE

4.1 Phase III Trials and Subgroup Analysis

Clinical trials are commonly categorized into three phases. Phase I trials are small studies investigating the feasibility of giving a new intervention, including an evaluation of toxicity and an assessment of the subjects' compliance with the intervention. Phase II trials are preliminary studies of the efficacy of a new intervention, typically having some or all of the following three characteristics: no comparison group, limited numbers of subjects (usually less than 100), and an outcome that is a surrogate for the disease endpoint of real interest. They are often employed as preliminary screens for eliminating noneffective interventions from further study. Phase III trials are much larger studies aiming at a clear evaluation of the effectiveness of a new intervention, usually involving comparison with a control or standard intervention. According to our definition in 3.1, the Act focuses specifically on Phase III trials when it refers to "clinical trials."

In the above definitions, the term efficacy is used to denote the ability of an intervention to affect the target disease, whereas the term effectiveness denotes the overall impact on the disease in the presence of real constraints such as subject noncompliance, comorbidity, and so on.

Phase III clinical trials are designed to investigate specific clinical inquiries that are framed as primary questions. These questions typically develop from numerous preliminary studies and have a strong scientific basis. Usually there will be just one primary question that is to be tested in a trial, such as the question whether "intervention A reduces the mortality from disease B compared to the standard or control intervention"; occasionally a more complex trial may be designed to address two or more primary questions. The primary question plays a central role in the design, conduct, and analysis of the trial. In particular, the sample size, that is, the number of subjects required in the trial, is determined with reference to the primary question. A sample size is chosen to provide high statistical power at a stated level (often 90%) for detecting a given overall intervention effect as statistically significant. In other words, we enter enough subjects in the study to make it very likely that should an important overall intervention effect exist, the

trial will detect it. The procedures for monitoring the interim data and the statistical analysis of the final results are also determined on the basis of the primary question.

Extensive experience with clinical trials analyses has led to a philosophy in which the analysis of Phase III clinical trials is divided into two components. The first component is the test of the primary question posed by the trial, in which one examines the effect of intervention over the total group of subjects. The other component is the conduct of secondary analyses to identify questions with sufficient scientific basis to be tested as primary questions in future trials. One set of secondary analyses commonly conducted involves the examination of intervention effects within defined subgroups of subjects. Subgroups may be defined by demographic variables, such as age, gender, race, or ethnicity, as well as biologic variables, such as type or stage of disease. Subgroup analyses should generally be interpreted conservatively, bearing in mind the potentially large number of ways of subdividing the subjects and the consequent opportunities for random variation to cause apparent differences between the intervention effects in various subgroups. If major qualitative differences are found between the intervention effects within separate subgroups, then further studies to examine such differences may be warranted.

The determination of how an intervention should be used almost always depends on results obtained from the totality of related clinical trials and clinical studies; one trial is rarely sufficient to be interpreted in isolation. Analyses of the relation of gender and minority status to outcome and subject characteristics should also take advantage of the statistical methods of summarizing results over many studies, known as meta-analysis. The opportunity to provide definitive answers to questions about subgroup differences in intervention effect is greatest in the context of meta-analysis of multiple clinical trials.

We generally interpret the Act as requiring that, wherever possible, there is an appropriate representation according to gender and race/ethnicity, thus allowing subgroups defined by gender and race/ethnicity to be investigated in subgroup analyses. The general philosophy relating to subgroup analysis in clinical trials, outlined above, should then carry over to these specific gender and racial/ethnic subgroup analyses. Our recommendations for the definition of key phrases in the Act are guided by this general view.

4.2 The Role of Preliminary Evidence in Subgroup Analysis

We considered the role of preliminary evidence regarding differences among intervention effects in different gender or racial/ethnic subgroups. Preparatory to any Phase III clinical trial, certain data are typically obtained. Such data are necessary for the design of an appropriate Phase III trial and include observational clinical study data, basic laboratory (i.e., in vitro and animal) data, and clinical, physiologic, or biochemical data from Phase I and Phase II studies.

It is important that, whenever possible, such preliminary human data be obtained on a diverse population, that is, in subjects of both genders and from different racial/ethnic groups. When designing a Phase III trial, these data must be examined to determine if there are substantial differences observed between the subgroups.

If substantial and significant differences between intervention effects in subgroups are found, then the Phase III trial must be designed to take account of

them. Essentially, the primary question addressed by the trial must take cognizance of the real possibility that the intervention effect differs substantially in certain subgroups. For example, if men and women are thought to respond quite differently to an intervention, then the Phase III trial should be designed to answer two separate primary questions, one for men and the other for women, with adequate sample size for each question.

Another possibility is that the preliminary data strongly support there being no substantial difference between the intervention effects within gender and racial/ethnic subgroups. In this case, there is not a strong scientific rationale for requiring inclusion in the Phase III trial of appropriate representation of members from different gender and racial/ethnic groups. Nevertheless, we consider it *prudent* to include such representation as part of the trial, so that one can gather further evidence in the Phase III trial itself of the comparability of the intervention effects within subgroups. This interpretation is consistent with the 1990 NIH Policy, on which the 1993 Act builds.

Generally, although data from preliminary studies relating to possible differences among intervention effects in different subgroups must be obtained, evidence of this nature is likely to be less convincing than that deriving from the subgroup analyses that can be performed in usual-sized Phase III trials. This is because the evidence from preliminary studies is likely to be of a more indirect nature (e.g., based on surrogate endpoints), deriving from uncontrolled studies (e.g., nonrandomized Phase II trials), and based on smaller numbers of subjects than in Phase III secondary analyses. For this reason, we consider it likely that data from preliminary studies will, in the majority of cases, neither clearly reveal substantial differences between subgroups of patients, nor strongly negate them. In these cases, Phase III trials should still have appropriate gender and racial/ethnic representation. They would not have the large sample sizes necessary to provide a high statistical power for detecting differences in intervention effects among subgroups, but, with the usual sample sizes and adequate representation, analyses of subgroup effects must be conducted and comparisons between the subgroups made. Depending on the results of these analyses, the results of other relevant clinical research, and the results of meta-analyses of clinical trials, one might initiate subsequent trials to examine more fully these subgroup differences.

In summary, we recommend that the requirement for a valid analysis of subgroup differences in a Phase III clinical trial be made conditional on the preliminary evidence supporting the existence of such differences. The term "valid analysis" itself merely denotes an analysis free from bias, where one is comparing "like with like." Thus, the Act requires that Phase III trials be evaluated in a manner that allows an unbiased analysis of subgroup differences. In addition, where preliminary evidence strongly supports the existence of subgroup differences in intervention effect, enough subjects from each subgroup should be included to allow a statistically powerful assessment of the intervention effect within each subgroup.

This new emphasis on preliminary evidence will shift the focus toward including women and minorities in studies conducted in earlier phases of the development of the intervention. This is indeed consistent with other Sections of the Act that require the recruitment of women and minorities to the broad range of clinical research studies, including early phase trials and developmental studies of new interventions.

4.3 Reasons for Not Requiring High Statistical Power for Subgroup Differences When There Is No Strong Preliminary Evidence of Their Existence

The reasons for not requiring the large sample sizes necessary to provide a high statistical power for detecting differences in intervention effects among subgroups, when the preliminary evidence does not strongly support such subgroup differences, are as follows.

First, as explained above, in the absence of preliminary data that strongly support the existence of subgroup differences, it would be scientifically inappropriate to include consideration of subgroup differences into the primary questions to be posed by the trial. On the other hand, by requiring appropriate representation in the trial of subjects from different gender and racial/ethnic subgroups we provide a mechanism by which further information can be gathered, with the possibility that new data from the trial will raise questions about differences in intervention effects within subgroups to the level of primary questions to be addressed in future Phase III trials.

Second, planning the Phase III trial size to provide definitive answers to these subgroup questions would create difficult ethical problems. It is now becoming standard practice to establish a Data and Safety Monitoring Committee to monitor the data from a Phase III clinical trial as they accumulate. Suppose that interim results indicated an intervention benefit or harm for the group of subjects as a whole. No Data and Safety Monitoring Committee could permit the continuation of the trial to answer questions of intervention effects within gender, racial/ethnic subgroups, or indeed any other defined subgroups, unless there was a strong scientific basis for expecting major differences in intervention effects for such subgroups. And what individual would wish to participate in a trial once the questions were answered for subjects as a whole, unless there *were* a strong scientific basis for believing that the results would differ based on gender or racial/ethnic group? The clinical equipoise situation [3] that must exist at the start of the trial would have changed and the trial could not ethically continue under normal circumstances. Because the effects of the intervention would usually be answered for the group of subjects as a whole long before it would be answered for any subgroup, this problem would arise in most trials that demonstrated benefit or harm. Trials that could continue to term would be mostly those in which no overall difference was demonstrated. In those trials one might be exposing excessive numbers of subjects to potential hazards and discomfort, with no scientifically based hope of benefit. In unusual circumstances, an occasional subgroup based on gender, race, or ethnicity might demonstrate a convincing trend in favor of the intervention, in the absence of an overall intervention effect. In such a case it might be appropriate to continue the trial in that subgroup alone. However, it must be emphasized that such circumstances are uncommon.

Third, determining reliably whether intervention effects differ among subgroups requires huge numbers of subjects. As detailed below, meeting the needs for this would require each Phase III trial to be many times larger than the size required under current standards, depending on the exact number of subgroups of interest.

The sample sizes for clinical trials sponsored by the NIH vary, but a typical clinical trial comparing the effects of two treatments on a mortality or disease incidence endpoint is designed to detect a 25% reduction in the hazard of the

disease event of interest. For standard statistical planning parameters this requires the observation in the trial of 509 such disease events [4]. If the proportion of subjects experiencing an event during the trial is 20%, then 2545 subjects are required. Some NIH-sponsored trials are larger than this and some are smaller. Prevention trials tend to be larger than therapy trials, because the proportion of individuals suffering a disease event tends to be lower in prevention trials.

The above numbers are based on determining whether the intervention produces a 25% reduction in the hazard of the disease event for the subjects overall, including both genders and different racial/ethnic groups. To achieve high statistical power for comparing the degree of benefit in males with that in females would require an inflation in sample size. The amount of inflation required depends upon what difference in effect is important to detect. If we wish to have high statistical power for detecting a situation where there is a 25% reduction in the hazard of disease in one gender and no reduction in the other gender, then we would require the sample size to be increased by a factor of 4. To detect a situation where there is a 25% reduction in the hazard in one gender and only half of that reduction (12.5%) in the other gender would require the sample size to be increased by a factor of 16. Hence, interpretations of the Act as requiring statistically powerful comparisons of the intervention effects between genders imply an increase in total sample size of trials by a factor ranging from 4 to 16. Interpretations that require statistically powerful comparisons also among racial/ethnic subgroups imply even greater inflation factors depending on the number of subgroups. For example, with five racial/ethnic subgroups, the corresponding inflation factor would range from 10 to 40. Even considering gender alone, an inflation factor of 4 to 16 represents an increase in the sample size of the typical trial described above from 2545 subjects to a number ranging from 10,180 to 40,720. The inflation factors described above are not restricted to trials with outcomes that are disease events, but apply widely to trials regardless of the outcome of interest.

The strategy that we recommend avoids the ethical and scientific problems outlined above but still addresses the underlying concerns regarding representation of gender and racial/ethnic subgroups and is completely consistent with the intentions and the wording of the NIH Revitalization Act.

4.4 Racial/Ethnic Subgroups

With regard to the racial/ethnic subgroups mentioned throughout Section 4, a difficult issue arises over how broad or narrow the division into different subgroups should be. On one hand, division into many racial/ethnic subgroups is tempting in view of the real cultural and biological differences that exist between these groups and the possibility that some of these differences may in fact impact in some way upon the effect of an intervention. On the other hand, from a practical perspective, a limit has to be placed on the number of such subgroups that can realistically be studied in detail for each intervention that is researched. The Act is intended to lead to feasible and real improvements in the representativeness of different racial/ethnic groups in clinical trials. It should particularly emphasize research in those subpopulations that are unusually affected by certain diseases or disorders. With this view in mind, we suggest that, in the above discussion, racial/ethnic subgroups should include any subpopulations in which the disease manifests itself in an extraordinary way. For example, if the incidence of the

disease, disorder or condition is extraordinarily high in a subpopulation compared to the rest of the U.S. population, as with diabetes mellitus in certain Native American groups, this would merit inclusion of the subpopulation as a separate racial/ethnic subgroup, as would unusually early onset of the disease or the presence of disease risk factors specific to the subpopulation. Besides such subpopulations, the requirement to include racial/ethnic subgroups should lead to recruitment across some broad demographic groups so that the requirement for entering a diverse population is met.

5. CONCLUSION

We wholeheartedly support the importance of including appropriate representation of gender and racial/ethnic subgroups in NIH clinical trials whenever possible. We believe the above definitions to be consonant with the general intention of the Act to increase participation of women and minorities in clinical trials and to heighten the awareness of specific disease problems within specific gender and minority groups and subpopulations, while at the same time allowing the clinical research programs to remain productive and to advance the health of the US population.

The following people at NIH contributed substantially to the discussions that led to this paper: Wendy Baldwin, Judy LaRosa, Belinda Seto, Carlos Caban, Bill Blackwelder, Jonas Ellenberg, Mitchell Gail, Sylvan Green, Jack Lee, Roy Milton, Blossom Patterson, George Reed, Seth Steinberg, Clare Weinberg, and Margaret Wu.

REFERENCES

1. NIH Revitalization Act, Subtitle B, Part 1, Sec. 131–133, June 10, 1993
2. NIH Guidelines on the Inclusion of Women and Minorities as Subjects in Clinical Research. Federal Register, Part VIII (59 FR 14508–14513), March 28, 1994 [Also reproduced in NIH Guide to Grants and Contracts (Vol. 23, No 11, March 18, 1994]
3. Freedman D: Equipoise and the ethics of clinical research. N Engl J Med 317:141–145, 1987
4. Schoenfeld D: The asymptomatic properties of comparative tests for comparing survival distributions. Biometrika 68:316–319, 1981