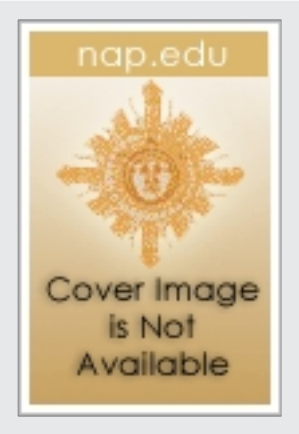

Review of a Research Protocol Prepared by the University of Utah: Letter Report

DETAILS

20 pages | 8.5 x 11 | null
ISBN null | DOI 10.17226/10315

AUTHORS

Committee to Review a Research Protocol Prepared by the University of Utah, Board on Radiation Effects Research, National Research Council

THE NATIONAL ACADEMIES
Advisers to the Nation on Science, Engineering, and Medicine
National Academy of Sciences
National Academy of Engineering
Institute of Medicine
National Research Council

DIVISION ON EARTH AND LIFE STUDIES
Board on Radiation Effects Research

Dr. James M. Smith
Chief, Radiation Studies Branch
Centers for Disease Control and Prevention
4770 Buford Highway, NE
Mailstop F35
Atlanta, Georgia 30341-3742

February 22, 2002

Dear Dr. Smith:

On October 25, 2001, the National Research Council's Committee to Review a Research Protocol Prepared by the University of Utah met at the Beckman Center of the National Academies in Irvine, California. The task before the committee was to review critically the proposed methods and analyses and to assess whether they were appropriate and complete. This task differs somewhat from the usual ones the committee addresses which are to review draft reports issued at the end of a study. However, in this instance, the very limited information provided the committee precluded a detailed evaluation of the study and its likelihood of success. Accordingly, the committee focused its attention on identifying improvements rather than commenting on the basic proposal.

The committee notes, however, that a study of thyroid disease in the Utah-Nevada-Arizona population has the promise to produce important information. The Chernobyl studies have clearly demonstrated an excess of thyroid cancer following much higher ^{131}I exposures, while the Hanford fallout study, with a lower average ^{131}I dose, did not show any thyroid effects. The Utah study has a higher dose distribution than the Hanford study, so that it could potentially have greater statistical power and be more informative in the low dose range. It may also have less uncertainty in the dosimetry than Hanford because of the substantial number of external exposure measurements made in the study areas at the time of the Nevada Test Site nuclear detonations. The past performance of these investigators in locating and enlisting the participation of the study subjects in the previous cycle of this study was very good, which is important for minimizing selection bias. The potential value of this study thus provides a context in which to evaluate the proposed study design, the dosimetric, epidemiologic and clinical procedures, the analytic methods and the statistical power. However, an appropriate statistical analysis is needed to demonstrate whether this potential could be realized. The committee has attempted to be critical in a constructive manner, even to the extent of providing a technical framework for the estimation of statistical power, given the sources of uncertainty in the study data (see Appendix B).

2001 Wisconsin Ave, NW, Washington, DC 20007 USA
202-334-2232 (telephone) 202-334-1639 (fax) E-mail: national-academies.org

The Radiation Studies Branch of the Centers for Disease Control and Prevention (CDC) charged the committee to address the following:

1. Are the study objectives attainable?

2. Are the proposed study design and the methods for data collection and statistical analyses unambiguous, and do they adequately address the study objectives? In particular, we would like the Committee to comment on the following:

a. The proposed expansion of the study cohort beyond the previous follow-up study of 1985-1986.

b. The adequacy of the proposed design in dealing with possible detection, selection, and information biases.

c. The appropriateness of the proposed outcome measures.

d. The approach for accounting for uncertainty and variability in individualized dose estimates and the dose-response analysis.

3. Has the statistical power of the study been appropriately addressed?

Present at the October 25, 2001 meeting, in addition to the members of the committee and the BRER staff, were representatives of the investigative team at the University of Utah, Joseph L. Lyon, the principal investigator; Wayne Meikle, the coinvestigator; Mary Bishop Stone, the program manager; Stephen Alder, the project's statistician, and Owen Hoffman, who was recently hired as a consultant to the project, and Judy Qualters and Felix Rogers of CDC. Dr. Lyon and his colleagues described the proposed study and responded to the committee's questions. The committee's comments and recommendations set out in the paragraphs that follow stem from the written, albeit incomplete, documentation of the proposed research, the presentations, and responses to the committee's queries. Our comments and recommendations are organized around the questions previously listed.

Question 1. Are the study objectives attainable?

The protocol is not sufficiently developed nor rigorous enough to permit a judgment of whether all the study objectives are attainable. Four objectives are stated:

"a. to test whether a healthy group of children, inadvertently exposed to radioactive iodine will have an increased number of thyroid neoplasms related to this exposure while controlling for other causes of thyroid cancer including medical x-rays and occupational radiation."

Insufficient and incomplete information is presented in the protocol, and too few analyses have apparently been carried out, for the committee to judge the attainability of this objective in this cohort. In particular two elements necessary to ensure that the objective will be met were not adequately described in the protocol: safeguards in the study's design, data collection and analysis to minimize the likelihood and magnitude of biases, and a correctly conceptualized method for determining the statistical power of the protocol and its application in the study design. Some examples of such issues are discussed in the committee's answer to question number 2 and in Appendix A.

"b. to determine if exposure to low doses of radiation was associated with non-neoplastic forms of thyroid disease detected and/or occurring up to 44 years after the exposure."

Again, the evidence presented in the protocol pertinent to this objective is too limited for the committee to judge whether it is attainable; further specification of safeguards against bias and evaluation of the statistical power attainable in this cohort are required (Some examples of such issues are discussed in the committee's answer to question number 2 and in Appendix A). In addition, inadequate attention was paid to ways of detecting and controlling for geographic variation in the prevalence of these thyroid conditions.

"c. to identify, locate, and enroll individuals who were exposed as children to the fallout in the 1950s in Washington and Lincoln counties, but who moved from these counties before the study group was assembled in 1965."

The protocol lacks detailed plans for locating the people in question (especially women), having them screened for thyroid disease wherever they reside, and ensuring their screening rates will be high enough to impart confidence in the validity of this part of the study. One would have expected to see a detailed statement of methods and some preliminary data to predict how successfully this objective can be attained.

"d. to update the Phase II dose assignment model with information that has become available during the past 15 years, and to improve the treatment of uncertainty in the dose assignment model."

Too little information is given in the protocol as to the dose-assignment model that has been or will be used, the information that has become available, and the methods proposed for treatment of uncertainty for the committee to determine attainability of this objective. Substantial information on the thyroid-dose method used in Phase II has been published (see, for example, Stevens et al., 1992, and Simon et al., 1990). It would have been highly desirable to summarize that information in the protocol, to indicate the similarities and differences with the proposed dose assignment model, and to indicate what new information and improvements were to be incorporated.

Question 2. Are the proposed study design and methods for data collection and statistical analyses unambiguous, and do they adequately address the study objectives? In particular, we would like the Committee to comment on the following:

a. The proposed expansion of the study cohort beyond the previous follow-up study of 1985-1986.

b. The adequacy of the proposed design in dealing with possible detection, selection, and information biases.

c. The appropriateness of the proposed outcome measures.

d. The approach for accounting for uncertainty and variability in individualized dose estimates and the dose-response analysis.

The statements of the objective of this research—as in the task statement, the study

statement, the questionnaires, and the consent form—are inconsistent and inaccurate. For example, the sole statement of purpose on the exposure questionnaire "this survey is part of a study of the effects of lifestyle and the environment on health" is overly vague, so that it is questionable whether or not this meets the usual standards of informed consent; and a further example of an incorrect statement of purpose is given in Appendix A. Provision of misleading or inaccurate statements of purpose to study subjects is unacceptable. The task and study statements, questionnaires, consent forms, and any other study subject contact documents or scripts should be carefully reviewed and revised to make them consistent and accurate. Similarly, the definitions of the statistical outcomes need to be sharpened and made more specific. For instance, our meeting with the investigators showed that little, if any, thought had been given to which persons should be excluded (because of prior disease) in the piecewise temporal analyses. In addition, the protocol should contain information on the expected number of cases of (rather than just rates of) thyroid cancer in the study in the absence of exposure to fallout radiation from the nuclear detonations at the Nevada Test Site and on the expected number of excess cases under the several scenarios used in the statistical power calculations. This would aid in interpreting the statistical power results. The same applies to thyroid nodules and other disease outcomes.

2a: ***The proposed expansion of the study cohort beyond the previous follow-up study of 1985-1986.*** The estimates of the probable attained study size seem optimistic and perhaps unrealistic. First, they do not take into account the study losses that will occur if parents are not available for dosimetry interviews (for example, with respect to milk consumption). At the meeting, these losses were estimated to be about 20%. That estimate seems to be low in light of the current age distribution of the parents; many will be deceased. There is no explanation or protocol to cover the case in which no parents would be available for interviews. A carefully reasoned estimate or the result of a pilot study is needed rather than an extemporaneous number.

Second, the investigators assumed that they would be able to locate and enlist the participation of 90% of the new augmented-population members. Tracing such people who reside outside of the three-state area, especially women, after about 50 years will be difficult. Furthermore, their participation in going to their local thyroid physicians in diverse locations might lead to greater attrition than that in the Utah-Nevada-Arizona in-state screening program and may thereby be subject to greater self-selection bias, as well as to lack of uniformity in the thyroid screening procedures by the local physicians. More attention to the rates of study losses is warranted. There was no pilot-program information on and virtually no plan for obtaining screening participation from those who reside outside the three-state area. That weakness applies to the subgroups that have moved out of state both before and after 1965.

We recommend that before deciding upon whether to augment the cohort with out-of-staters, a pilot study on a random sample is needed to determine the success rate in locating these subjects, in obtaining thyroid screening participation, and in obtaining milk consumption information from mothers. A separate reliability study should be conducted to determine the reliability between current and (preferably) 1965 or 1986 reports of milk consumption/source information; this is needed to determine the degree of uncertainty in current milk-exposure estimates and its effect on the amount of statistical power gain afforded by augmenting the study cohort.

Third, milk-consumption data will have to be obtained now for virtually all the new additions to the cohort. The milk-consumption data will be obtained about 50 years after the exposure time, and a substantial fraction will have to be obtained from surrogates of the mothers (because of the death or infirmity of the mothers). Studies show that the reliability of food-consumption data obtained several decades after the fact is quite poor (Dwyer et al., 1989). There is likely to be less gain in statistical power than one might expect from adding numbers if the added exposure data are unreliable. That raises a question about the value of study-size augmentation. At a minimum, we recommend that simulations be performed to estimate the effect of unreliable milk-consumption data on the exposure estimates and on the consequent degree of gain in study power afforded by adding subjects with relatively unreliable exposure information to the study.

2b: ***The adequacy of the proposed design in dealing with possible detection, selection, and information biases.*** Several elements in the epidemiologic and statistical methods require further thought and specification to reduce bias and achieve validity. A number of safeguards to minimize information bias were missing from the protocol. Safeguards should be fully developed to maximize study-staff and subject blinding. For instance, the parent dosimetry interviews should be conducted before subjects' thyroid screening. At the screening, a subject should complete the medical-history and other questionnaires first. The examiners (palpation and sonography) should be blinded as to each subject's medical-history report. While it can be argued that the person performing the palpation should be blinded to the sonographic results and vice versa, here it would seem that a consensus of the examiners on the interpretation of their findings would improve the study design. The recommendation for fine-needle aspiration (FNA) should be made after blinded review of sonograms by the three radiologists on the review panel. The cytologic reading should also be performed in a blinded fashion, and surgical specimens should have a blinded pathologic review. A control should be built into the protocol to ensure that all examiners conduct the same proportions of examinations in low- and high-exposure geographic areas; otherwise, examiner differences could produce a bias.

Several additional specifications of criteria must be instituted to ensure objectivity. For example, the biopsy criterion for a "prominent nodule" in a multinodular gland was not specified (for instance, must it be greater than 1 cm in diameter?). Criteria should be given for performance of FNA (for instance, under what circumstances will one, two or more aspirates be obtained?).

The protocol did not indicate which "other potential confounders" will be evaluated for possible inclusion in the analyses. When asked, the investigators did not indicate one of the more important potential confounders: that a subject had participated in the 1965 and 1985 screenings (screening will detect thyroid disease in addition to or earlier than that found in routine medical care).

The investigators propose to augment the more highly exposed portion of the cohort from the targeted Utah and Nevada counties by searching state birth records to identify persons who were probably in the targeted Utah and Nevada counties during the peak fallout period but who emigrated from those states before the initial cohort definition in 1965. However, they do not propose the same augmentation for the "low-exposure" Arizona group. The Utah-Nevada group

would therefore contain early out-migrants, but the Arizona group would not, so the subject-selection procedures would not be comparable. Whether that would bias thyroid-disease rates is not known, but the fact that it is not known means that it should be appropriately guarded against by proper study design.

There has been no evaluation of the reproducibility or accuracy of the milk-consumption questionnaire or of the effect of obtaining the answers at substantially different times after the potential exposure, and no evaluation was built into the study. At a minimum, the investigators could reinterview a subsample of perhaps 200 parents or parent surrogates who were interviewed in 1986 and conduct a detailed comparison of responses on the two occasions. Better yet would be comparing the 1986 and current questionnaire data related to milk consumption with the 1965 milk-consumption estimates if they are still available. In fact, if there are enough 1965 data, it would be valuable to analyze thyroid disease in relation to the dose estimates based on the 1965 data as an alternative check on results. Thought could also be given to whether statistical and other methods could be developed to permit using the 1965 data for the corresponding 1986-2002 milk-consumption variables when the former are available.

2c: ***The appropriateness of the proposed outcome measures.*** A more complete description of the clinical procedures is necessary. It should include

- Methods used to minimize examiner bias based on history or medical findings.
- Quality-assurance procedures for laboratory, ultrasound, and examination procedures.
- Pertinent specification of the ultrasonic equipment and how it will be used. The choice of ultrasound equipment needs to consider the method of recording data for subsequent image analysis and display.
- FNA procedure—discussion of the rationale for single vs multiple biopsies of nodules to minimize sampling errors.

We recommend that the investigators lay out a protocol in detail that describes how they will handle persons residing outside the three-state area and that they explicitly consider the maintenance of consistency between those inside and those outside the three-state area, provide information on how the screening will be handled for those residing outside the three-state area, and consider the financial aspects of this protocol.

2d: ***The approach for accounting for uncertainty and variability in individualized dose estimates and the dose-response analysis.*** The approach to uncertainty and variability in individual dose estimates and the dose-response analysis is poorly described and appears contradictory. For example, the choice of the regression method (linear or logistic) for the primary analysis of incidence data differs throughout the protocol. Appendix D of the proposal describes a method for quantifying uncertainty in the risk estimates while allowing for uncertainty in dose, but the committee is not convinced that the approach is appropriate (and we were confused by the material presented to us). In any event, the description of the approach is not transparent; it is important to revise the description by using mathematical notations that define more precisely the concepts presented in this section.

The calculations for the primary source term, ^{131}I in milk, are not obvious. The questionnaires have a number of items pertaining to the milk source, but it is not clear how any of the data, other than decay time, will be used. There is discussion of cow-to-cow variation among breeds of cows in home milk vs dairy milk, cows vs goats, and so on, but the equations presented seem to have only a single term for milk and to use a single value for fallout-to-feed and feed-to-milk factors. Where did those factors come from? Are they national averages, or are they adjusted for local conditions? Are home conditions the same as dairy conditions? This shows again that the extensive information on Phase II dosimetry that is included in Stevens et al. (1992) should have been summarized in the protocol, if, indeed, the methodology of Stevens et al. (1992) is to be used in the proposed study. The protocol should state how the questionnaire data are to be used, or if they are not to be used, why they are being collected. It should also evaluate whether the uncertainties discussed would significantly affect the power of the study, or could introduce significant biases.

To complete the computation of estimated dose for people in the study, it is necessary to estimate the deposition of ^{131}I on the ground from the airborne plume generated by the weapons tests. That was apparently done in Phase II (the proposal gives no information on this point) by assuming that the deposition of ^{131}I was spatially constant in each of the counties. However, as has been shown in a number of studies (Beck and Anspaugh, 1991, Thompson, 1990), deposition concentrations are not spatially constant, and therefore location effects should be examined for their potential effect on the power of the study or in introducing bias, and perhaps taken into account. Weather is at least one of the factors in this spatial differentiation. The proposal claims (page 18) that areal deposition will be updated, taking weather into account, but there is no information on how this will be done or on the sources and nature of the weather information that will be used.

Question 3. Has the statistical power of the study been appropriately addressed?

The protocol developed some expected rates of thyroid cancer for various dose groups based on the National Council on Radiation Protection and Measurements (NCRP) (1985) risk estimates and cancer-registry data but did not use them in the calculations. The protocol spoke of tripling the rates of thyroid cancer observed in cancer registries, but it was unclear if this was done. Similarly, it spoke of using a relative biological effectiveness (RBE) of three (that is, an effectiveness factor of 1/3) but apparently did not do so. The statistical-power projection ended up using the central estimate of the risk coefficient from the previous Utah study to project risk. That estimate is uncertain (being based on only eight cases). The investigators might better have used consensus estimates by the committee on NCRP (1985), Biological Effects of Ionizing Radiation (NRC, 1990), or others. Furthermore, **it is strongly recommended that the investigators not use a single estimate for statistical-power calculations, but vary the input parameters (screening-effect magnitude, RBE, risk coefficient, and input parameters for the dosimetry uncertainties) and produce a range of statistical-power estimates. We also recommend that power calculations be performed for selected nonneoplastic thyroid diseases for which the background prevalence rates are much higher, e.g., autoimmune diseases. This would give a more realistic idea of the likely statistical power of the study**.

Some power calculations are presented in the protocol, but they are based on a simulation

with several puzzling aspects. For example, in the simulation study to estimate power, the sample size was taken to be random and to range in value from 1000-5000 subjects. In power calculations it is not customary to have the sample size be a random quantity (although often a range of power calculations corresponding to plausible sample sizes are presented). Here, however, the range of sample size was very extreme, and did not correspond to the estimates of sample size given in the protocol or in the presentations by other investigators. Formula 4 on page 40 is confused in that σ^2 is described there as the variance of true dose given the estimated dose, whereas it should be the variance of estimated dose or, more precisely, the variance of the expected value of true dose given the input data available for an individual (see NRC, 2000). Even if that is corrected, the formula is valid only when errors are not shared from subject to subject. Because of those logical difficulties, we conclude that the investigators have not made the statistical power of the study clear. The power calculations as described in the protocol do not take into account the fact that the primary analysis will include the previously reported cases of cancer and neoplasia, which have already shown a marginally significant dose-response. The main issue, then, to be addressed is the amount of new information that will be obtained in the update of the study. Formal power calculations should be done separately for the updated study, neglecting all the previously obtained data (so that follow-up of each subject only starting from the time of the previous examination up to the present exam, is considered). Then in a combined analysis, the investigators could consider the likely reduction of the length of the confidence interval, for the dose response already obtained, when the new study information is added to the old. In other words the importance of the new information in adding to the old is of paramount importance.

In summary, it is important that the investigators attempt to include in their statistical power simulation model all the sources of uncertainty, and that the simulation model treat the "shared" and "unshared" sources of uncertainty appropriately. A technical report prepared by one of our members, given as Appendix B, outlines a statistical approach to do so.

General Recommendations

The committee's recommendation is that the decision on whether the study is conducted should be based on a scientifically defensible protocol in which an improved design is advanced and the proposed methods and analyses are justified in detail. Specifically:

- **The thyroid dose assignment should be compared with the methodology used in 1986, and the improvements should be clearly identified**. A preliminary reliability/validity study should be conducted to compare milk consumption/source data obtained in 2002 with that obtained in 1965, if possible, or at least in 1986. This would help determine the value of the proposed augmentation of subjects and of the study in general.

- **The power of the study should be more carefully investigated and summarized under a variety of conditions reflecting the uncertainties in the size of the study population, completeness of data, and uncertainties in exposure assessments.**

- The proposed augmentation of study size using out-of-staters merits consideration if it meets several criteria: the study design for the augmentation guards against bias; current recall of milk consumption/sources is sufficiently reliable and accurate to assure that the added subjects would significantly improve study power; and pilot data demonstrate the ability to locate, screen and obtain parental milk consumption reports from a high proportion of these subjects.

- **The letters, questionnaires, and consent forms should not be used until they are appropriately revised to be consistent with the study objectives and to reduce potential bias.**

- The information provided on the number of people to be interviewed, the length of interview time, and the number of interview teams appears inconsistent with the project schedule. The study schedule should be reviewed carefully.

- **Adequate safeguards need to be built into the protocol to minimize biases in the study design, data collection and analysis. Thorough safeguards should be developed to ensure appropriate blinding of both study-staff and subjects.** The study details should be reviewed and revised as appropriate to ensure to the maximum extent possible that "blind" is indeed blind. For instance, coded identification numbers on documents for purposes of person identification should be replaced with random numbers. This needs to be corrected before interviews begin.

If you desire elaboration on the comments above or the accompanying appendixes, please do not hesitate to call or write to Dr. Isaf Al-Nabulsi or me.

Sincerely yours,

William J. Schull
Chairman

COMMITTEE TO REVIEW
A RESEARCH PROTOCOL PREPARED BY THE UNIVERSITY OF UTAH

WILLIAM J. SCHULL (Chair), Professor Emeritus, Houston, TX
BRUCE B. BOECKER, Scientist Emeritus, Albuquerque, NM
ANDRÉ BOUVILLE, National Cancer Institute, Bethesda, MD
A. BERTRAND BRILL, Vanderbilt University Medical School, Nashville, TN
MELVIN W. CARTER, Neely Professor Emeritus, Dunwoody, GA
EDMUND A. C. CROUCH, Cambridge Environmental Inc., Cambridge, MA
SHARON M. FRIEDMAN, Lehigh University, Bethlehem, PA
SUSAN E. LEDERER, Yale University School of Medicine, New Haven, CT
MILTON LEVENSON, Menlo Park, CA
DONALD E. MYERS, University of Arizona, Tucson, AZ
ROY E. SHORE, New York University School of Medicine, New York, NY
DANIEL O. STRAM, University of Southern California, Los Angeles, CA

NATIONAL RESEARCH COUNCIL STAFF
ISAF AL-NABULSI, Study Director
DIANNE STARE, Project Assistant
DORIS E. TAYLOR, Staff Assistant
NORMAN GROSSBLATT, Editor

SPONSOR'S PROJECT OFFICER
JAMES SMITH, Centers for Disease Control and Prevention

References

Beck, H. L. and Anspaugh, L. R. Development of the County Data Base: Estimates of Exposure Rates and Times of Arrival of Fallout in the ORERP Phase-II Area. Comparison of Cumulative Deposition Density Estimates Based on Analyses of Retrospective and Historical Soil Samples. United States Department of Energy Report DOE/NV-32, 1991.

Dwyer, J. T., Gardner, J., Halvorsen, K., Krall, E. A., Cohen, A., Valadian, I. Memory of Food Intake in the Distant Past. Am J Epidemiol. 130: 1033-1046, 1989.

NCRP (National Council on Radiation Protection and Measurements). Induction of Thyroid Cancer by Ionizing Radiation. Report No. 80. Recommendations of the National Council on Radiation Protection and Measurements. Bethesda, MD. 1985.

NRC (National Research Council). Review of the Hanford Thyroid Disease Study Draft Final Report. Washington, DC: National Academy Press, 2000.

NRC (National Research Council). Health Effects of Exposure to Low Levels of Ionizing Radiation. The Committee on the Biological Effects of Ionizing Radiation (BEIR V). Washington, DC: National Academy Press, 1990.

Simon, S. L., Lloyd, R. D., Till, J. E., Hawthorne, H. A., Gren, D. C., Rallison, M., and Stevens, W. Development of a Method to Estimate Dose from Fallout Radioiodine to Persons in a Thyroid Cohort Study. Health Physics 59(5): 669-691, 1990.

Stevens, W., Till, J. E., Thomas, D. C., Lyon, J. L., Kerber, R. A., Preston-Martin, S., Simon, S. L., Rallison, M. L., and Lloyd, R. D. Assessment of Leukemia and Thyroid Disease in Relation to Fallout in Utah: Report of a Cohort Study of Thyroid Disease and Radioactive Fallout from the Nevada Test Site. Supported by the National Cancer Institute (Contract #N01-CO-23917), U. S. Public Health Service, Department of Health and Human Services. Salt Lake City, UT: University of Utah, 1992.

Thompson C. B. Estimates of Exposure Rates and Fallout Arrival Times Near the Nevada Test Site. Health Physics 59:555-563, 1990.

Appendix A
Comments on the Questionnaires and the Protocol

The questionnaires and other information provided to the study subjects or their parents (or parental surrogates) contain obvious inconsistencies between each other and with the study protocol presented. Examples follow (the references in parentheses, such as Tab 1 or Box 2, refer to the location in the documents the committee received from the investigators).

1. Consent form, page 1 (Tab 1): Specifies two purposes of the study—"if being exposed to fallout relates to changes in being able to have children or thyroid disease in families"— that are not among the stated objectives given in the protocol.

2. Consent form, page 2 (Tab 1) Item A6: Notes that “less than 5% of subjects will be chosen", but the proposal says 10% (on page 29, under "data management").

3. Consent form for fine-needle biopsy, page 2 (Tab 1): Says "treated, if needed, free of charge", but the original consent form says that costs will be borne by insurance if the subject is insured. Such treatment would not be free of charge even if only copayments are required.

4. Refusal Script, page 1 (Tab 6): States that the person would receive a visit from "two medical doctors who specialize in thyroid disease", but the protocol (page 20) states that the team would consist of a nurse practitioner or physician's assistant and a certified sonographer. Neither of the latter is necessarily a medical doctor, and neither necessarily specializes in thyroid disease.

5. New Subject Location Script, page I (Tab 7): States as eligibility criteria “born 1946 through 1958 Lived 1 year in Washington or Lincoln counties in period 1951 through 1958", whereas the eligibility criteria given in the protocol (page 10) states “born 1947 through 1953 Lived in Washington county May1953 through June 1953, or Lincoln county May 1952 through May 1953.” The two sets of criteria are substantially different.

6. Questionnaire Preparation Booklet (Tab 13, page 5 of 14): Asks the subject about medical history after age 17, but the Interview Booklet (for the parent) (Tab 14, page 21 of 31) asks only about the child's medical history up to age 15, and the Exposure Questionnaire (Tab 15, page 79) asks about the child's medical history up to age 18.

7. Questionnaire Preparation Booklet, page 11 (Tab 13): Requests information about "you/your spouse" pregnancies. But more is needed to obtain accurate reproductive histories such as information on pregnancies in all sex partners-and the response could be misleading if the spouse has been pregnant with parties other than the respondent.

8. Exposure Questionnaire, page 4 (Tab 15): Gives incorrect instructions: they cover a much longer period and wider geographic area than contemplated in the protocol.

9. Radiation Dose Determination: Protocol gives no real information on the model to be used for radiation-dose determination. It states that the model is shown in Figure 5 on page 18, but that diagram is meaningless without further description. Figure 5 should be replaced with a

figure that shows clearly the steps involved in the dose-estimation process to be used in Phase III. The current Figure 5 is extremely confusing: What does "environmental transport" (Box 8) contain, given that it has inputs of "source term", "animals", "cows/goats", and "commercial milk producers pooled"? How does "radiation released and transported" (Box 2) get to "animals ate contaminated vegetation" (Box 4) without going through "environmental transport" (Box 8)? No reference is provided for either the proposed model or any predecessors.

10. The effort required, or the time required, appears to have been substantially underestimated. Field work is said to require 2 years (page 20). Having two teams examine 5000 subjects means 1250 examinations per team per year, or about six examinations per team per working day. That is impossible, given that examinations are expected to take 60-90 minutes each.

The study protocol does not appear to have been adequately assessed with respect to potential sources of bias. In view of the likely marginal statistical power of the study, a thorough discussion, and preferably a quantitative evaluation, of potential sources of bias is essential, even for small potential biases. Examples follow:

1. Page 20: To minimize the possibility of bias in the review of ultrasound images, it is important that survey locations not be identifiable to reviewers on the basis of characteristics of the images. Will random study identification designation be used, and how will care be taken to avoid minor instrumentation or operator variation that could identify survey subjects in high- vs low-dose regions?

2. Page 22: The contact diagram shown on page 22 has the examiners interpreting the questionnaire. If a nodule is found, the subject is told and schedules fine-needle biopsy before completing the questionnaire. Subjects with immediately detected nodules thus have a potentially traumatic event thrust on them before they complete the questionnaire. That could bias their answers. Moreover, the examiners themselves will know the nodule status of the subject before they perform the review of the questionnaire with the subject—another potential source of bias.

3. Page 43: It seems that the "control" group will be substantially different from the "dosed" groups because of the inclusion of all the late movers. What will be done to evaluate whether that introduces a bias?

4. Page 48, Table 13: The table gives insufficient information to make any judgment about recall bias. In particular, there is no indication of the standard deviations. The table contains information on only 1528 subjects. Why so few, and was the method of selection determined to minimize the introducing of bias?

5. Page 49: The examiners are supposed to be unaware of exposure status, but they are going to start administering the questionnaire before they do the physical examination, and they are presumably going to talk to the subject and so might be made aware of the exposure status from the information so gained. There are instructions to the examiners to avoid such small talk, but it appears that it would be difficult or impossible to control or prevent it.

Appendix B

Power and Uncertainty Analysis of Epidemiologic Studies of Radiation-Related Disease Risk in which Dose Estimates are Based on a Complex Dosimetry System: Some Observations.

Daniel O. Stram, Department of Preventive Medicine, School of Medicine, University of Southern California, Los Angeles, California.

ABSTRACT

This paper discusses practical effects of dosimetry error relevant to the design and analysis of an epidemiologic study of disease risk and exposure. It focuses on shared error in radiation-dose estimates for such studies as the Hanford Thyroid Disease Study or the Utah Thyroid Cohort Study, which use a complex dosimetry system that produces multiple replications of possible dose for the cohort. I argue that a simple estimation of shared multiplicative error components via direct examination of the replications of dose for each person provides information useful for estimating the power of a study to detect a radiation effect. Uncertainty analysis (construction of confidence intervals) can be approached in the same way in simple cases. I also offer some suggestions for Monte Carlo-based confidence intervals.

1. INTRODUCTION

Several recent epidemiologic studies have used a complex dosimetry to estimate radiation-related health effects of exposures to fallout or nuclear-plant releases. Two examples are the Utah Thyroid Cohort Study [1] [2] and the Hanford Thyroid Disease Study (Draft Report). In both, individual doses of radioactive iodine (^{131}I) to the thyroid gland were estimated decades after exposure. In this report, I make a number of observations concerning analysis of study power to detect a simple dose-response relationship and analysis of the uncertainty in estimated dose-response relationships. Those issues involve the uncertainty of dose itself in the study. The setting differs from the traditional measurement-error problem in that errors in dosimetry are not independent from subject to subject. For example, in the ^{131}I setting, the basic approach (cf [3]) is to estimate total deposition of ^{131}I onto grass, uptake by cows on pasture, transfer of iodine into cow's milk, milk consumption by children, and uptake to the thyroid gland. For each subject, i, in the study, a set of personal data, W_i, regarding age and location of residence during the exposure period, source (backyard cow vs local dairy) and amount of milk consumed, and so on, is collected. The dosimetry system uses those data to impute a dose estimate. Shared uncertainties result, obviously, if such quantities as the total deposition onto grass or the average fraction of ^{131}I that is excreted into cow's milk are misidentified. Other uncertainties can be regarded as independent from person to person (for example, differences between true and reported milk consumption).

Increasingly, the approach to uncertainty in dose in such studies seems to be to develop a dosimetry system that gives not just one estimate of dose, but rather many replications (100 in the case of the Hanford study) of possible dose for each subject. Moreover, the estimates are not generated independently for each subject; instead, each run of the dosimetry system provides a new realization of possible dose for the entire study population, and the uncertainties in shared

characteristics result in a complex correlation between the dose estimates from subject to subject.

2. AN IDEALIZED DOSIMETRY SYSTEM

Let us pretend that this type of dosimetry system has evolved to the state where we can regard, in a Bayesian framework, each replication of dose to be a sample from the distribution, $f(X_1, X_2, ..., X_N \mid \mathbf{W})$, for true dose, given the full set of input data, **W**, for all N subjects in the study. Because many replications, r, are available, we can calculate for each subject, i, the expected value of the unknown true dose, X_i, given the input data, W_i, available for the subject simply as the average of that subjects simulated X_i. We will call the expected value, $Z_i = E(X_i \mid W_i)$, the estimated dose for each subject. We assume that the data, W_i, are available for each subject (and the dosimetry system is in place) before the collection of outcome data, D_i, for each subject (for example, as in the Hanford study). Finally we assume that D_i is independent of W_i, given X_i (so that W_i consists of only "surrogate" variables).

3. FURTHER SIMPLIFICATIONS

I will now make some observations regarding the use of estimated dose, Z_i, rather than true dose in the analysis of the dose-response relationship $D_i \mid X_i$. We assume that $E(D_i \mid X_i)$ is of simple linear form, that is,

$$E(D_i \mid X_i) = a + b\, X_i \tag{3.1}$$

For now, to simplify discussion further, we assume that D_i is a continuous outcome distributed symmetrically around its expected value, given true X_i. In addition, we can adopt a simple model for the joint distribution of the true doses, X_i, around the expected true doses, Z_i, that incorporates both shared and unshared dosimetry errors. Suppose that we have

$$X_i = \varepsilon_{SM}\, \varepsilon_{Mi} Z_i + \varepsilon_{SA} + \varepsilon_{Ai} \tag{3.2}$$

where ε_{SM} is shared multiplicative error, ε_{Mi} is unshared multiplicative error, and ε_{SA} and ε_{Ai} are shared and unshared additive error, respectively. Assume that the ε are all independent with $E(\varepsilon_{SM}) = E(\varepsilon_{Mi}) = 1$ and $E(\varepsilon_{SA}) = E(\varepsilon_{Ai}) = 0$. Let us first consider the situation when the shared multiplicative and additive error components are both fixed but unknown quantities. We can use standard regression techniques to fit

$$E(D_i \mid Z_i) = a^* + b^*\, Z_i \tag{3.3}$$

There will be a bias in estimating b using b^* of size equal to ε_{SM}. Specifically, the estimate of b^* obtained by fitting Equation 3.3 will be consistent for $b \times \varepsilon_{SM}$. Now consider $Var(\hat{b}^*)$, still treating ε_{SM} and ε_{SA} as fixed. For continuous D_i and small b^*, this will be approximately equal to

$$\frac{Var(D)}{NVar(Z)} \tag{3.4}$$

and we term this quantity the naïve estimate of variance of $\hat{b}$ because it neglects the effects of shared error in the dose estimates. Finally, consider the total variance of $\hat{b}^*$ over the distribution of ε_{SM} and ε_{SA}. We will have

$$Var(\hat{b}^*) = Var\{E(\hat{b}^* \mid \varepsilon_{Sm}, \varepsilon_{SA})\} + E\{Var(\hat{b}^* \mid \varepsilon_{Sm}, \varepsilon_{SA})\}$$

$$\cong b^2\sigma^2_{SM} + \frac{Var(D)}{NVar(Z)} \qquad (3.5)$$

This expression implies that under the null hypothesis that $b = 0$, the expectation (over the shared error components) of the naïve estimate of the variance of $\hat{b}^*$ is equal to the true variance. However, if $|b| > 0$, the naïve variance estimate (3.4) is in fact biased downward by the amount $b^2\sigma^2_{SM}$ compared with the true variance. From that, we may conclude the following:

1. Ignoring shared error in the dosimetry system does not affect the asymptotic size of the test of the null hypothesis that $b = 0$.
2. However, sample sizes or the power of a test calculated under a specific alternative hypothesis, $|b| > 0$, will, if they ignore shared dosimetry error, be incorrect. Power will be overstated or, equivalently, the necessary sample size will be understated.
3. Confidence intervals will also be affected. Ignoring shared dosimetry error while constructing confidence intervals based on either the Wald or Score test will result in confidence intervals that are too narrow.
4. However, it is the upper bounds of $|b|$, and not the lower bounds, that are most affected; in particular, a confidence interval ignoring dosimetry error that does not overlap 0 will not overlap 0 once the shared errors in the dosimetry are properly handled. That follows because the validity of a test of the value $b = 0$ does not depend on shared dosimetry error, because the variance estimate of $\hat{b}$ under the null hypothesis is correct.

Estimation of all the variance parameters in the model of equation (3.2) can be achieved by consideration of the relationship between the variances and covariances between subjects across the replications of the dosimetry. In this model, the covariance of X_{ir} with X_{jr} over the simulations is equal to

$$Z_i Z_j \sigma^2_{SM} + \sigma^2_{SA} \qquad (3.6)$$

Let C_{ij} for each subject i and j denote the sample covariances between X_{ir} and X_{jr} over the replications, r. If we fit by OLS regression the model

$$E(C_{ij} \mid Z_j) = \alpha + \beta Z_i Z_j \qquad (3.7)$$

for all $i \neq j$, the intercept and slope estimates will provide estimates of σ^2_{SA} and σ^2_{SM}, respectively. Similarly, the variance of X_{ir} over the replications is equal to

$$Z_i^2[(\sigma^2_{SM} + 1)(\sigma^2_M + 1) - 1] + \sigma^2_{SA} + \sigma^2_A \qquad (3.8)$$

Linear regression of the sample variance estimates S_i^2 on Z_i^2 will allow the estimation of $\sigma_{SA}^2 + \sigma_A^2$ (intercept term) and $[(\sigma_{SM}^2 + 1)(\sigma_M^2 + 1) - 1]$ (slope term in the regression). We now have two simple equations to solve for the two remaining unknowns σ_M^2 and σ_A^2.

Such an analysis will give some useful information about the additional variation in $\hat{b}$ expected because of shared error, and this can be directly incorporated into the power calculations if the model of equation (3.1) is correct. However, the most direct approach to calculating power under specific alternative hypotheses is undoubtedly by simulation. The logical approach is as follows. If we run the dosimetry system many times, the Z_i values are computed as $E(X_i \mid W_i)$ for each subject. Next, for each of the random replications, $\mathbf{X}_r$, new random values of outcome variable D_{ir} are generated under the model of interest, assuming that each X_{ir} is true dose (so that D_{ir} is taken as independent given X_{ir}). For each set of outcome data, D_{ir} is regressed upon the values Z_i (which remain fixed over the simulations) to obtain a new estimate, $\hat{b}_r$, of b. The usual test statistic, $\hat{b}^2 / Var(\hat{b})$, for testing the null hypothesis $b = 0$, here approximated as $N\hat{b}_r^2 Var(Z)/Var(D)$, is computed and compared with the naïve critical value from the χ_1^2 distribution. The naïve critical value may be used here because of observation 1 above. The number of times that the test statistic falls into the critical region is tabulated and used to compute an approximate power of the test.

4. UNCERTAINTY ANALYSIS

We first restrict our attention to simple models, such as that of equation (3.1), where unbiased estimates, $\hat{b}$, of b can be obtained by regression of D_i on the Z_i. Consider the construction of confidence intervals as the set of b for which the test statistic

$$\frac{(\hat{b} - b)^2}{Var(\hat{b} \mid b)} \tag{4.1}$$

takes values less than the chi-square critical value $\chi_{1,1-\alpha}^2$. By the use of the notation $Var(\hat{b} \mid b)$, we are explicitly expressing, as in Equation 3.5, the dependence of the variance of the estimator on the test value of b. If we think that the model of equation (3.2) holds and can estimate $Var(\varepsilon_{SM}) = \sigma_{SM}^2$, we may consider approximating the test statistic as

$$\frac{(b - \hat{b})^2}{b^2 \sigma^2{}_{SM} + \dfrac{Var(D)}{NVar(Z)}} \tag{4.2}$$

However, this approximation works only for small values of b. Note, for example, that this value approaches $1/\sigma_{SM}^2$ as $b \to \infty$, implying infinite upper bounds for small enough values of type I error α, which, of course, is unrealistic. A better approximation replaces $Var(D)$ with

$E\{Var(D|Z)\}$, which is actually a function of b and the variances of the random-error terms. The simple constraint that $Var(D|Z) \geq Var(D|X) \geq 0$ will in combination with the variance parameters, impose an upper bound on possible values of b less than infinity. Of course this dependence greatly complicates the calculation of the criteria even under the simple model for dosimetry error given in equation (3.2).

Let us consider estimating the denominator of Expression 4.1 directly by simulation. This will require a series of simulations, one for each test value of b considered. Throughout the simulations, the observed value, $\hat{b}$, of the dose-response relation obtained by using Z_i is held fixed. For each complete set, $\mathbf{X}_r$, of replications of the dosimetry, a new set of responses, D_{ri}, are simulated according to the dose-response model, using the test value, b. The outcomes are analyzed by using Z_i to obtain a new dose-response estimate $\hat{b}_r$. The variance, $Var(\hat{b}_r \mid b)$, over the simulations is computed and used in Equation 4.1, to determine whether the test value of b is inside or outside the critical region.

This brute-force simulation approach will work only for relatively simple models. In particular, we have assumed that the sampling variance of $\hat{b}$ does not depend on the intercept parameter, a, and this will generally not hold for binary or counted data. If the intercept parameter markedly affects the variance of $\hat{b}_r$ we will need a two-dimensional search that requires simulations for each possible pair of values of (a,b) to determine whether they jointly satisfy the equation

$$\begin{bmatrix} \hat{a}-a \\ \hat{b}-b \end{bmatrix}^T Var\left(\begin{bmatrix} \hat{a} \\ \hat{b} \end{bmatrix} \middle| \begin{bmatrix} a \\ b \end{bmatrix} \right)^{-1} \begin{bmatrix} \hat{a}-a \\ \hat{b}-b \end{bmatrix} < \chi^2_{2,1-\alpha}$$

Here, the variance term is the variance-covariance matrix of $\hat{a}_r$ and $\hat{b}_r$ estimated by simulation at the test values (a,b).

5. MONTE CARLO MAXIMAL LIKELIHOOD

For models that are nonlinear in X (so that directly regressing D_i on Z_i produces biased estimates) or for models with many background risk parameters or many interactions between dose-response relation and other factors, it might be infeasible to consider construction of confidence intervals in the manner described above. We outline briefly here a simulation-based approach for maximal likelihood that in principle can be used both to approximate maximal likelihood estimates and to construct approximations to full likelihood-based confidence limits (based on the change in the log-likelihood). This discussion is based on Geyer 1996 [4]. Let $l(a,b)$, be the log-likelihood ratio for testing the null hypothesis that $(a,b) = (a_0,b_0)$. Because the full likelihood, $f(\mathbf{D},\mathbf{W})$ is equal to the integral $\int f(\mathbf{D},\mathbf{X} \mid \mathbf{W}) d\mathbf{X}$, we have

$$
\begin{aligned}
l(a,b) &= \log \frac{f(\mathbf{D} \mid \mathbf{W}; a, b)}{f(\mathbf{D} \mid \mathbf{W}; a_0, b_0)} \\
&= \log \frac{\int f(\mathbf{D}, \mathbf{X} \mid \mathbf{W}; a, b) d\mathbf{X}}{f(\mathbf{D} \mid \mathbf{W}; a_0, b_0)} \\
&= \log \int \frac{f(\mathbf{D}, \mathbf{X} \mid \mathbf{W}; a, b)}{f(\mathbf{D}, \mathbf{X} \mid \mathbf{W}; a_0, b_0)} \frac{f(\mathbf{D}, \mathbf{X} \mid \mathbf{W}; a_0, b_0)}{f(\mathbf{D} \mid \mathbf{W}; a, b)} d\mathbf{X} \\
&= \log \int \frac{f(\mathbf{D} \mid \mathbf{X}, \mathbf{W}; a, b) f(\mathbf{X} \mid \mathbf{W})}{f(\mathbf{D} \mid \mathbf{X}, \mathbf{W}; a_0, b_0) f(\mathbf{X} \mid \mathbf{W})} f(\mathbf{X} \mid \mathbf{D}, \mathbf{W}; a_0, b_0) d\mathbf{X} \\
&= \log \int \frac{f(\mathbf{D} \mid \mathbf{X}; a, b)}{f(\mathbf{D} \mid \mathbf{X}; a_0, b_0)} f(\mathbf{X} \mid \mathbf{D}, \mathbf{W}; a_0, b_0) d\mathbf{X} \\
&= \log E_{a_0, b_0} \left\{ \frac{f(\mathbf{D} \mid \mathbf{X}; a, b)}{f(\mathbf{D} \mid \mathbf{X}; a_0, b_0)} \,\middle|\, \mathbf{D}, \mathbf{W} \right\} \qquad (5.1)
\end{aligned}
$$

If there is a way of generating a total of n samples, $\boldsymbol{X}_r$, from the distribution

$$f(\mathbf{X} \mid \mathbf{D}, \mathbf{W}; a_0, b_0) \qquad (5.2)$$

of true dose, **X**, given both disease, **D**, and input data, **W**, we can approximate Equation 5.1 as

$$l_n(a,b) = \log \left(\frac{1}{n} \sum \frac{f(\mathbf{D} \mid \mathbf{X}_r; a, b)}{f(\mathbf{D} \mid \mathbf{X}_r; a_0, b_0)} \right) \qquad (5.3)$$

Notice that in principle the choice of a_0 and b_0 in 5.2 is arbitrary and that the change in log-likelihood for any two choices of the parameters (a_1, b_1) vs (a_2, b_2) can be written as $l(a_2,b_2)$-$l(a_1,b_1)$. Thus, it appears that we can remove the conditioning on **D** in Equation 5.1 by choosing $b_0 = 0$, so that D_i is independent of X_i. That implies that we can contemplate the calculation of confidence limits for the dose-response parameter b by using the samples, $\mathbf{X}_r$, from $f(\mathbf{X} \mid \mathbf{W})$ provided by the dosimetry system in Equation 5.3.

In general, this simple approach will work well only for b near 0 because as $|b| > 0$ the ratio in the summand of Equation 5.3 becomes extremely variable, requiring a prohibitive amount of computer time to evaluate the expectation numerically. In general, it is best, numerically, to perform the simulation by using the maximal likelihood estimate of a and b as a_0 and b_0 in Expression 5.2. Thus, if the dose-response relation is strongly significant, it will be important to provide a means of sampling from the conditional distribution of **X,** given both **W** and **D**. Rejection techniques, such as the Metropolis Hastings algorithm, can in principle be used to transform samples from the conditional distribution given only **W** to the appropriate distribution, thereby allowing implementation of the relatively complex simulation-based schemes. The basic idea is as follows. An initial $\boldsymbol{X}_r$ is sampled, and then a new $\boldsymbol{X}_{r+1}$ is sampled. For each element *i,* the Hastings criteria are computed

$$p_i = \min\left(1, \frac{f(D_i \mid X_{r+1,i})}{f(D_i \mid X_{r,i})}\right)$$

and $X_{r,i}$ is replaced with X_i with probability p_i; otherwise, $X_{r,i}$ is reused. It is unclear, however, whether the convergence properties of the Monte Carlo EM is based on rejection sampling or of other Monte Carlo methods, such as Gibbs sampling, will be adequate to allow for routine use.

6. DISCUSSION

The techniques described here are based on accepting the notion that the dosimetry system does indeed provide samples from the full joint conditional distribution of true X_i, given the measured input data, W_i, for each subject. In most cases, the uncertainty in dose reflected in the dosimetry system is due to lack of knowledge of parameters (such as true deposition or milk-transfer factors) or data (such as errors in questionnaire data), and the randomness between replications of the dosimetry system represents, at best, a consensus of expert opinion about the likely values of each of the unknown parameters or data. We have simply carried this exercise one step forward; using these approaches is designed to yield a consensus view of the power and uncertainty of a study based on the uncertain dosimetry available.

The simple shared-error model in equation (3.2) clearly may be a vast oversimplification of the error structure of a complicated dosimetry system. It may be useful to expand further on this model to introduce additional terms. For example, subjects whose milk source is the backyard cow will share the uncertainty in the (herd average) values of milk-transfer factors for these animals, not the uncertainty in the (herd average) values of milk-transfer factors for commercial animals. Restricting the estimation of the shared components of variance, using equation (3.7), to similar subjects will allow estimation of separate shared and unshared variance components to reflect this sort of additional complication. Power and uncertainty can be discussed more fully with such an expanded model. The simulation approaches described above, for power and uncertainty, do not depend on the validity of the simple shared-error model given in equation (3.2), however, and are appropriate as long as we can regard each replication from the dosimetry system as representing a sample from the conditional distribution of true dose.

REFERENCES

1. Till, J. E., Simon, S. L., Kerber, R., Lloyd, R. D., Stevens, W., Thomas, D. C., Lyon, J. L., et al. The Utah Thyroid Cohort Study: Analysis of the Dosimetry Results. Health Phys. **68**:472-83, 1995.
2. Kerber, R. A., Till, J. E., Simon, S. L., Lyon, J. L., Thomas, D. C., Preston-Martin, S., Rallison, M. L., et al. A Cohort Study of Thyroid Disease in Relation to Fallout from Nuclear Weapons Testing. JAMA. **270**:2076-82, 1993.
3. NCI (National Cancer Institute). Estimated Exposures and Thyroid Doses Received by the American People from Iodine-131 in Fallout Following Nevada Atmospheric Nuclear Bomb Tests: A Report from the National Cancer Institute. Bethesda, MD: National Cancer Institute. 1997.
4. Geyer, C. Estimation and Optimization of Functions. In Markov Chain Monte Carlo in Practice Chapman and Hall. W.R. Gilks, S.T. Richardson, and D.J. Spiegelhalter, Editors. Chapman and Hall. p. 241-258, 1996.