

A Technical Review of the Final Report of the Hanford Thyroid Disease Study

March 30, 2004

A. James Rутtenber, Ph.D., M.D.
F. Owen Hoffman, Ph.D.
Raymond J. Carroll, Ph.D.
Duncan C. Thomas, Ph.D.
Sander Greenland, M.A., M.S., Dr.P.H., C. Stat

Summary of Findings

We conclude that the results and conclusions of the Final Report of the Hanford Thyroid Disease Study (HTDS) (Davis et al., 2002) cannot be used to rule out important risks for thyroid cancer, neoplasms, or hypothyroidism from exposures to iodine-131 (I-131) from the Hanford nuclear facility. Specifically, we find that:

- (1) The interval estimates given by HTDS are much too narrow because they ignore large sources of uncertainty; in particular:
 - (a) They ignore major sources of uncertainty in dose assignment, and
 - (b) They do not account for the large losses from the cohort (the 1/3 for whom clinical outcome data were unavailable); and
- (2) The HTDS has misinterpreted its own statistics by relying on statistical significance, and then, on top of that, not calculating or interpreting significance test results correctly:
 - (a) They miscalculate significance levels by making inappropriate Bonferroni adjustments,
 - (b) They misinterpret nonsignificance as lack of evidence of a dose response, when in fact there is a trend in dose-response in some analyses, and
 - (c) They do not take adequate account of their own confidence intervals in formulating their closing statements.

Proper accounting for these problems would reveal that the study does not provide evidence capable of discriminating between no effect and relatively strong effects.

Introduction

In this report, we address the extent to which the HTDS results and interpretations can be used to assess disease causation for individuals exposed to Iodine-131 (I-131) from the Hanford nuclear facility. We focus on three important questions: 1) Were radiation doses to the thyroid and their uncertainties modeled appropriately for the dose-response analyses that were conducted?; 2) Did the study, as conducted and analyzed, have adequate power to detect any likely effects at the chosen (0.05) significance level, as the HTDS authors assert?; and 3) Given the problems with estimates of doses and their uncertainties, and with the consequent low statistical power, were the analytic approaches and interpretations of results adequate and appropriate?

As detailed below, the answers to these questions are: "No," "No," and "No." For (1), we identify a number of problems with the dose estimates made for individual subjects, and substantial unmodeled sources of uncertainty in the radiation dose estimates. For (2), we point out that the statistical power for the study has been mischaracterized and is likely to be far lower than claimed by the HTDS. For (3), we identify problems with

the analytic choices that only further reduced the power, and show that the authors misinterpret lack of statistical significance as lack of evidence.

Hanford Dosimetry Uncertainty

Ideally, dose reconstruction is based on measurements of the exposures received by individual members of the cohort or by measurements of the amount of radioactivity in air, deposition, and foodstuffs. Because of the absence of detailed environmental measurements and individual-specific exposure data, the HTDS relied almost exclusively on mathematical models to simulate the release, environmental transport, exposure and individual thyroid dose to each individual in their cohort. Such a practice is known to be associated with large and complex uncertainty (Hoffman 1991, Hoffman et al, 1993, Hoffman et al, 1996, Hoffman 1999).

When modeled doses are used as surrogates for measured values, it is important to characterize these uncertainties and account for them in the modeling process. Moreover, the full nature and extent of the uncertainty in individual dose estimation should be taken into account explicitly prior to determination of the statistical power of an epidemiological study. Upon the ascertainment of the incidence of disease among members of the cohort, the effect of dosimetry uncertainty in determining the slope of the dose response and the limits of its confidence intervals must also be considered.

The Hanford Environmental Dose Reconstruction Integrated Codes (HEDRIC) attempted to account for errors in dosimetry through a Monte Carlo approach. Probability distributions were assigned to uncertain model parameters, and 100 random alternative realizations of true dose were simulated for the entire cohort of 3191 “in area” study participants. Nevertheless, for individual dose assignment in HTDS some important sources of uncertainty were ignored. Some sources of uncertainty were inappropriately assigned default point estimates (single values that were held constant), some led to a systematic tendency to over and/or underestimate true dose for specific subgroups of the HTDS cohort, and some resulted in unrecognized complex mixtures of classical and shared and unshared Berkson errors. The incomplete handling of uncertainty in the HTDS has resulted in overestimates of statistical power and estimates of confidence intervals that are narrower than they should be.

The statistical theory of exposure measurement error distinguishes two basic ways such errors can come about. The “classical error” model assumes one is measuring exposure with some instrument (e.g., a dosimeter or questionnaire) which returns a value that is distributed in some fashion around the true (unknown) exposure with some error. The “Berkson error” model arises in experimental studies where an investigator aims to apply particular exposures to members of various groups but, due to inter-individual variability, the actual exposures received by individuals within each group is distributed in some fashion around the applied dose for the group.

In the standard theory, both kinds of errors are assumed to have zero mean and to follow some known distribution (e.g., normal). The standard theory also assumes that measurement errors are uncorrelated between individuals. A consequence of classical measurement error is that the slope of an exposure-response relationship is attenuated — biased towards the null — by a factor that depends upon the relative variance of true exposures and measurement errors; in contrast, for certain types of models, Berkson error may produce no bias towards the null, but the variance of the slope estimate will be increased (and power reduced) by a factor similarly depending upon the relative variance of applied exposures and within-group variability.

In epidemiologic studies such as the HTDS, the measurement system is complex and contains elements of both kinds of error. Calculated doses (conditional on such variables as date of birth, place of residence, and dietary habits) are similar to the “applied exposures” in the experimental setting and are interpreted as expressing the mean (or median) of a distribution of unknown true doses for all individuals with similar characteristics; thus, they can be expected to have a Berkson error structure. On the other hand, the questionnaire information that forms one of the inputs to the dosimetry calculations is more likely to have a classical error structure.

To further complicate the situation, measurement errors are likely to be highly correlated between individuals to the extent that they share certain characteristics — e.g., the errors for two people living in the same town will be similar if the assumed deposition in the town was too high or too low. The quantitative implications of mixed Berkson and classical errors and of shared errors will be discussed below.

The full nature and extent of the uncertainty in individual dose estimation must be taken into account in order to correctly determine the statistical power of the analyses presented in the HTDS. The effect of dosimetry uncertainty in determining the slopes for the dose-response models and the limits of its confidence intervals must also be considered before these estimates can be appropriately interpreted.

Inter-individual Stochastic Variability

Inter-individual variability of true dose within the HTDS cohort can be substantial due to the heterogeneity of I-131 deposition, interception by vegetation, food chain transport, individual variations in consumption rates, and person-to-person variations in thyroid mass and fractional uptake from blood to the thyroid gland. It appears from the documentation of HTDS and from correspondence with Bruce Napier that the only model variables that explicitly accounted for inter-individual variability of true dose, given commonalities of age, gender, location, and dietary source of the individual, were the uncertainties assigned to the dose conversion factor and the GSD assigned to the individual’s intake of foods.

Within the HEDRIC/STRM/RATCHET/DESCARTES/CIDER suite of codes used for the HTDS dose reconstruction, it appears as if the dose conversion factor (DCF) and the default dietary consumption rates were the only uncertain variables that are allowed to vary independently from individual to individual within a given realization. All other parameter uncertainties in the HEDR suite of codes were strongly correlated and thus shared among individuals within the cohort (per each of the 100 realizations of the cohort dose). Therefore, within any realization of the dose for all members of the HTDS cohort, the true inter-individual variability of dose in the HTDS cohort will be substantially understated and the degree of correlation of uncertainty among individuals will be overstated, i.e., the rank order of dose assignments among individuals will remain nearly the same from realization to realization for individuals consuming the given milk type or source. Thus, the mean cohort dose will vary substantially from realization to realization, but the relative inter-individual variability of dose will not. This effect could lead to an overestimation of the uncertainty in the cohort mean dose (which is not reported in HTDS) and an underestimation of the relative variance of the cohort dose.

This problem could have been remedied by using a two dimensional Monte Carlo procedure to separate uncertainty due to stochastic inter-individual variability from uncertainty due to lack of knowledge about true fixed quantities and true geometric means (GMs) and geometric standard deviations (GSDs) that define stochastic variation of true dose between members of the HTDS cohort [see IAEA (1989) and NCRP Commentary No. 14 (1996) for details].

For each of the 100 alternative realizations of the true distribution of cohort doses, unique GMs and GSDs would be specified from a random sample from probability density functions (PDFs) that represent uncertainty in these quantities. In other words, the GMs and GSDs that define stochastic inter-individual variability of true dose themselves are uncertain. Therefore, in each realization, a unique value of GM and GSD would be sampled from PDFs representing uncertainty in the true values of GM and GSD. From the unique GM and GSD thus obtained, an estimate of true stochastic variability of within-cohort doses would be simulated, so that the cohort mean dose and the relative variability of individual doses would vary from realization to realization. Within a realization, the inter-individual cohort GSD would thus be increased from what is currently calculated within HEDR.

Bias and Uncertainty for Milk Pathway Dosimetry

Commercial Cow-Milk Transfer Coefficient: It is well documented that the distribution assumed in HEDR for commercial cow milk (0.012 d/L) is biased on the high side from what has been reported in the literature and includes an uncertainty that is unrealistically small (standard deviation, 0.002 d/L) (Snyder et al. 1994; Napier et al,

2000; Nuclear Regulatory Commission, 1997). This issue was raised in the 1993 peer review conducted for Pacific Northwest Laboratory (Hoffman et al., 1993) and by the NAS review of HTDS (NRC, 2000). However, the authors of HEDR (Napier et al, 2000) chose not to revise their assumptions and declined to examine the effect of changes in these assumptions on the power of the study. Had the assigned distributions for the commercial milk transfer coefficient been wider and centered on values that were more consistent with central estimates reported in the literature, the true dose would be lower and the uncertainty higher for those consuming commercial sources of fresh milk. This source of bias in the dose estimate has not been accounted for within HTDS. The overall effect of this source of bias is to overstate the statistical power of the study and understate the slope and confidence interval of the dose response.

Note that in the I-131 fallout dose reconstructions of the University of Utah and the National Cancer Institute (Stevens et al., 1992, NCI 1997), the commercial milk transfer coefficient for I-131 was assumed to be about 0.004 d/L with a GSD approaching 2.0. The expert elicitation performed by the Nuclear Regulatory Commission (1997) show that uncertainty on the average milk transfer coefficient for a very large region (subsequent to an accidental release of elemental I-131) could be as low as 0.0005 d/L and as high as 0.05 d/L, with most experts reporting uncertainties of the mean value extending down to 0.001 d/L.

The standard deviation of 0.002 d/L assumed in HEDR for the milk transfer coefficient of commercial cows leads to a coefficient of variation of merely 17%, which translates to a GSD of about 1.18. This is a much smaller uncertainty than what would be indicated by GSDs typically assumed for this parameter that range from about 1.4 to 2.1 (Stevens et al., 1992; NCI, 1997; Apostoaei et al., 1999, 2003).

In responding to the NAS review of the draft report of HTDS, the authors of HEDR (Napier et al. 2000 [Appendix 22 of HTDS]), appear to have confused a frequency distribution that describes stochastic variability of true values of the milk transfer coefficient with a distribution that is intended to represent uncertainty for an average value that would prevail throughout a large region of the USA and Europe. Although they acknowledged that their original distribution did not include the lower values reported in the expert elicitation, they chose not to revise their assumptions on the basis that such a revision would not substantially affect the mean value resulting from the use of their original assumptions for HTDS. This logic results in a bias towards overestimation of mean dose to the HTDS subcohort that consumed commercial sources of fresh milk and an underestimation of uncertainty in the doses assigned to these persons.

A systematic overestimation of true dose and a systematic underestimation of the dose uncertainty for this subgroup could compromise the reported HTDS estimates of

statistical power. These errors in estimation could also reduce the central value and the upper confidence limit of the slope of the HTDS reported dose response. Furthermore, the resulting systematic misclassification of dose for specific subgroups could be differentially distributed, thereby obscuring the relationship between dose and risk in dose-response models.

Mother's Milk Transfer Coefficient: There is evidence that the milk transfer coefficient assumed in HEDR for the transfer of I-131 into mothers' milk has been underestimated. This underestimation would lead to an underestimate of true dose for those who were breast-fed as infants, especially during 1945 (the year with the highest releases from Hanford). This systematic underestimation of true dose would lead to differential misclassification of disease outcomes for a subgroup of the HTDS cohort, which in turn would impact the power of the study and the evaluation of the dose response.

The milk transfer coefficient for mother's milk used in the HEDR suite of models was assumed to be uniform from 0.07 to 0.36 d/L, based on limited data in the literature (Snyder et al., 1994). More recent information from Simon et al., (2002), based on a more extensive review of measurements, gives a lognormal distribution with a GM of 0.37 d/L and a GSD of 1.5. This updated distribution is used in the NCI individual dose calculator for nationwide exposures to NTS fallout I-131. The NTS individual dose calculator from NCI can be found at (<http://ntsi131.nci.nih.gov/>). It is also being applied within the University of Utah revised dose reconstruction model that is being finalized for the refined epidemiological analysis of the special cohort who lived in counties immediately downwind of the NTS.

The median value for the HEDR models for the mother's milk transfer coefficient is 0.22 d/L. This is a factor 1.7 less than the geometric mean value reported by Simon et al. (2002). The maximum value of 0.36 d/L assumed in HEDR is slightly less than the geometric mean value reported by Simon et al., and nearly a factor of 2.3 less than the upper 97% percentile of the lognormal distribution reported by Simon et al. In HTDS, the only individuals explicitly considered to have been on a diet of mother's milk were those for whom such information was provided in the computer-assisted telephone interviews (CATIs). These subjects would have had true doses that were underestimated. The extent of underestimation would depend on the magnitude of dose received in later years when the individual's diet changed to other sources of fresh milk.

Default Assumptions for Subjects without CATIs: There is the potential for a systematic bias towards overestimation of true dose for the 1212 participants in HTDS who were assigned a default diet of cow's milk in the absence of data from a CATI. The degree of overestimation of true dose could be considerable for the HTDS subgroup of subjects who were born in 1945 and who may not have consumed fresh sources of milk, but

were assigned a default diet of fresh cow's milk because of the lack of specific dietary information obtained from a CATI.

It appears, from the information from CATIs, that about 36% of the HTDS study subjects were not on a diet of fresh cow's milk in 1945, the year of the highest I-131 releases from Hanford. Overall, about 69% of the cohort was not on a diet of fresh milk before the age of 6 months, which dropped to about 15% at age 1 year.

As noted above, systematic overestimation of dose for a subgroup in HTDS would be expected to lead to an overestimation of the true statistical power of the study, and to have obscured possible dose-response relationships.

Underestimated Dose Uncertainty

Residence Histories: Uncertainties in HTDS dose assignments have been substantially underestimated by using default dates as surrogates for true dates of changes of residence history, changes in both diet and food sources. This procedure essentially treats the default dates as if they are known without error. No additional analysis of this source of uncertainty is included in HTDS.

Given that the dates of change of residence history and dates of change of dietary sources and amounts are inherently uncertain when interviews are conducted approximately 50 years since the time of initial exposure to Hanford I-131, probability distributions of possibly true dates should have been defined and this source of uncertainty included within the alternative realizations of the true doses

Shared Uncertainty: Uncertainties of the HTDS doses have also been underestimated by allowing discrete time periods and exposure pathways to be treated as uncorrelated when summing uncertain doses for a single person. This procedure is inappropriate as these time periods and exposure pathways have shared sources of uncertainty for estimation of the true dose to a specific person. This is because the estimated dose for an individual for any specific time period is composed of a fixed true component and random error, and the latter errors are likely to be positively correlated across periods within individuals. Therefore, the sum of the error variances from the periods underestimates the total error variance since it ignores the positive covariance terms.

For example, specific alterations to the original CIDER code was made to allow the uncertainty in the I-131 dose conversion factor to be random and independent per individual. The uncertainty in the dose conversion factor is also considered to be independent per time period and exposure pathway (i.e. ingestion versus inhalation) for a given person. Since the uncertainty in the I-131 dose conversion factor is dominated by the uncertainty in the mass of the thyroid gland, the assumption of

independence from one time period to another or for one pathway to the next for a given person is incorrect. A person assigned a dose conversion factor that has an implicitly high mass of the thyroid for one time period should also have a high mass of the thyroid in the subsequent time period. The same is true for the summations of the dose conversion factors for inhalation and ingestion.

In their sensitivity analysis, the authors of HEDR reported that the I-131 dose conversion factor is the most important contributor to the overall uncertainty assigned to an individual's dose. Thus, summation of doses per exposure pathway and time period, assuming statistical independence in the uncertainty of the dose conversion factor, should lead to a suppression of the total relative uncertainty expressed in an individual's dose (whereby the relative uncertainty is expressed as a GSD, a coefficient of variation, or the ratio of the central value to the upper limit of a 95% credibility interval).

The reported distribution of individual GSDs of HTDS dose assignments in Davis et al. (2002) confirms this expectation. The average GSD is merely 2.18, and the lowest is 1.56, with few being above a GSD of 3.0. These GSDs are noticeably low given the value of the GSD assigned to specific model coefficients in the HEDR suite of computer codes. For example, the dose conversion factor has a GSD of 2.0, the milk transfer coefficient for a backyard cow has a GSD of 2.1 (Snyder et al., 1994), and the deposition/interception on pasture grass has an approximate GSD of about 1.8.

Assuming that the total uncertainty in the assigned dose is dominated by the uncertainty in just these three variables, the uncertainty in an individual's dose who consumed milk from a backyard cow should exceed a GSD of 3.0. Note that a GSD of 3.0 exceeds the GSD describing the inter-individual variability of median doses in the entire HTDS cohort (2.69).

A more realistic treatment of dose uncertainty would lead to lower estimates of statistical power and expanded confidence intervals for regression coefficients.

Incomplete Assessment of Thyroid Doses from the Nevada Test Site

Davis et al. (2002) mentioned that each individual in the cohort received an estimate of their thyroid dose from I-131 deposited throughout the USA from the atmospheric testing of nuclear weapons at the Nevada Test Site. These doses were not described in much detail in the report. Instead, NTS fallout doses for individual subjects were dichotomized based on the median dose for all subjects (5.3 mGy) and then explored as a confounder or effect-modifier in dose-response models. Based on these simple analyses, the HTDS authors concluded that the NTS fallout doses were not confounders or effect modifiers in any model, and could be disregarded. In light of the likelihood

that the statistical power of the analyses is substantially lower than reported by the HTDS, it is likely that the decision to disregard the NTS doses cannot be supported.

We used the NCI web-based calculator to estimate the thyroid doses for a hypothetical individual born in 1943 for every county within the HEDR domain, including the counties of northeastern Oregon and western Idaho (see the expert report of Dr. Owen Hoffman for the detailed results). The NTS individual dose calculator from NCI can be found at <http://ntsi131.nci.nih.gov/>. This web based tool was developed by SENES Oak Ridge, Inc. for the National Cancer Institute. NTS fallout exposure calculations have been made for individuals who were not on a diet of fresh milk, and those consuming average to high amounts of cow's milk (backyard and commercial) and average and high amounts of milk from a dairy goat. As the result of these calculations, we conclude:

- (a) The cut-off dose from NTS I-131 exposure of 5.3 mGy used in HTDS appears to be very low and relevant only to those who did not consume fresh milk; and
- (b) Calculation of county averaged NTS doses for the county of residence included in HTDS shows that typical NTS doses from the consumption of fresh milk products is much larger than 5.3 mGy.

For someone born in 1943 in Benton, County (near Hanford), who consumed fresh store bought milk from 1952 through 1957, the NTS dose would be 15 mGy (90% CI: 5.9, 48 mGy). If this person consumed backyard cow's milk, the NTS dose would be 19 mGy (90% CI: 8.3, 72 mGy). For goat's milk consumption, the NTS dose would be 160 mGy (90% CI 68, 490 mGy).

If this person was born in Benton, Co in 1943, but moved to Nez Perce, County Idaho before 1952, the NTS dose would be 81 mGy (90% CI: 20, 680 mGy) for store-bought milk, and 83 mGy (90% CI: 23, 560 mGy) for milk from a backyard cow. For consumption of goat's milk, the dose would be 660 mGy (90% CI: 180, 5,100 mGy).

If this person lived in Stevens, County (one of the distant counties from Hanford included in the HTDS cohort) the NTS dose would be 25 mGy (90% CI: 11, 110 mGy) for an average consumption of store bought milk, and 36 mGy (90% CI: 16, 150 mGy) for consumption of backyard cow's milk. For goat's milk consumption, the dose would be 290 mGy (90% CI: 120, 1,100 mGy).

The lowest Hanford doses calculated for the HTDS cohort were either for those born in 1946 (regardless of location), or for those who did not consume milk during their first few years of life. However, many participants who reported not consuming fresh milk as infants and young children began to consume fresh milk after the age of about 5 or 7.

For example, the median Hanford dose for the 1946 birth cohort is about 30 mGy. The milk dose from NTS for this birth cohort (assuming residence in Benton County) is nearly half that of Hanford at 17 mGy (90% CI: 7, 50 mGy) for an average consumption of store bought milk and more than 0.66 that of Hanford at 21 mGy (90% CI 8.8, 76 mGy) for milk consumed from a backyard cow. In this example, the upper limit of the 90% CI of the dose received from NTS exceeds the median estimate of dose from HTDS for the 1946 birth cohort.

Thus, with the exception of no milk consumed at all, all other milk consumption scenarios considered by the NCI lead to NTS doses substantially higher than 5.3 mGy from NTS fallout I-131 exposure. These NCI calculated doses would be higher yet for someone born later than 1943 and modestly smaller for someone born before 1943.

Only 8% of those who had CATIs reported zero consumption of raw or processed milk products (Davis et al. 2002). Thus, the fact that HTDS reports that 1,616 subjects had NTS doses less than 5.3 mGy seems rather implausible. Therefore, it appears that the NTS doses calculated for members of the HTDS cohort have been underestimated.

By not exploring thoroughly the possibility of confounding, effect modification, and dose misclassification due to exposure to I-131 in NTS fallout (and fallout from testing in the Pacific and former USSR), the HTDS authors cannot rule out the possibility that true dose-response relationships are present but not observed by the study design.

Analyses of Statistical Power

In the HTDS final report (Davis et al., pp. 47-51), the authors summarized the statistical power calculations that were made based on data from 869 participants in the Pilot Study. They concluded that there would be adequate (> 80%) statistical power for thyroid neoplasia (defined as all thyroid nodules) if the magnitude of the effect was similar to what would be expected from the BEIR V risk model (National Research Council [NRC], 1990) and the risks reported for the Utah study by Kerber et al. (1993). The authors also noted that in the Final Study, the projections of study power were actually exceeded (p. 530).

The authors noted that uncertainties in the individual dose estimates could be expected to reduce study power, and therefore conducted simulation analyses to estimate the impact of these uncertainties on study power. This analysis was expanded in a paper by Stram and Kopecky (2003), with similar results. We believe, however, that the power calculations are far more optimistic than justified, because they incorrectly treated most, if not all, of the uncertainty as of the Berkson type which, as our simulations show, substantially overestimates power.

Below, we explore the assumptions upon which these simulation analyses were based and produce an alternate assessment of the effects of uncertainty on statistical power. Our work takes two steps. First, we describe the Bayesian approach taken by Davis et al. (2002, pp. 206-207) for the analysis of the HTDS data, both to set notation and also to note that in their Bayesian approach to their data analysis, they explicitly allowed for the possibility of classical uncertainty. We then provide an alternative power analysis that explicitly incorporates classical uncertainty, thus showing that the impact of classical uncertainty is to lower power very substantially.

Framework for Bayesian Analysis

In their analysis of the HTDS data, Davis et al. (2002) used a Bayesian approach to correct for the effects of exposure measurement error on the estimated slope of the dose-response relationship. The basic approach can be summarized in terms of three sub-models. First we note the following definitions: Y = disease outcome, X = true dose, W = calculated dose.

- *Disease model*: a logistic model for the probability of disease Y_i given a subjects unknown true log dose $\log(X_i)$, specifically

$$\text{logit Pr}(Y_i=1 | X_i) = \alpha_{j_i} + \beta \log(X_i)$$

where j_i is an indicator variable for the gender of subject i .

- *Exposure model*: true doses are assumed to be distributed according to a mixture of lognormal distributions, conditional on a geographically- and age-defined grouping variable G_i , specifically

$$\log(X_i) \sim N(M_{G_i}, V_{G_i}).$$

- *Measurement model*: the geometric means of the calculated doses W_i are assumed to be lognormally distributed around the true doses X_i with known GSD_i , specifically

$$\log(W_i) \sim N\{\log(X_i), \log^2(GSD_i)\}.$$

A fundamental issue with this formulation of the model is that in the measurement model, all the uncertainties are assumed to be classical. As we show below, classical uncertainty generally leads to a much bigger reduction in power than Berkson error. In addition, their assignment of individuals' true doses was done independently at each iteration, ignoring any shared components of uncertainty (within groups G). This will have the effect of undercorrecting for these components of error. Instead, it appears that a large fraction of dose uncertainty was classical in the implementation of the Bayesian-based analyses of the effects of dose uncertainty on the slope and confidence intervals for a sex-stratified logistic dose-response model (HTDS pp. 206-207). By ignoring the fraction of total uncertainty that is made up of classical errors, the HTDS power calculations have overestimated power. We show this below numerically with much smaller amounts of classical uncertainty than was assumed in the HTDS implementation of the Bayesian analysis of dose uncertainty.

We also note a very puzzling feature of the aforementioned analyses of dose uncertainties. In the case of the application of this method to the model for thyroid cancer, Davis et al. (2002, p. 268) reported that the upper confidence limit of the measurement error corrected dose-response (2.11) is lower than that limit for the median dose estimates (2.61), which is contrary to what one would expect from a classical error model. This discrepancy suggests that Davis, et al. may have compared different models (a logistic model in log dose for the measurement error corrected analysis versus a logistic model in absolute dose for the uncorrected analysis).

Simulation Study of Power

Drs. Carroll and Thomas conducted two parallel simulations with slightly different formulations. We will report the results of one of these studies; the other gave the same qualitative result, namely that the power of the HTDS may well have been greatly overestimated by Davis, et al. In our simulations reported here, we used the model of Reeves, et al. (2001) and Mallick, et al. (2002) to simulate true and calculated doses. Their model consists of four variables: disease status Y , true dose X , calculated dose W , and a latent intermediate variable L between X and W . They specifically then write everything in the log scale as

$$\log(X) = \log(L) + U_b;$$

$$\log(W) = \log(L) + U_c;$$

U_b = Berkson uncertainty;

U_c = Classical uncertainty.

In their work, and in our simulations, the Berkson and Classical uncertainties are normally distributed with mean zero and standard deviations σ_b and σ_c , respectively. The latent intermediate variable L is lognormally distributed.

The parameter values chosen for the simulation are as follows:

- a. The number of study participants was $n = 3,000$.
- b. Mean calculated dose in the log scale = $\log(0.10)$.
- c. Standard deviation of calculated doses in the log scale = $\log(2.7)$.
- d. Standard deviation of uncertainties in the log scale = $\log(2.3)$.
- e. Percentage of error being Berkson = 100%, 90%, 80% and 70%. Note that this is higher than in the Bayesian model formulation of the HTDS.
- f. Correlation of shared Berkson errors = 0.0 and 0.5.
- g. Percentage of men = 50%. Baseline risks for men = $0.0049 - \Delta$ (see below).
- h. Baseline risks for women = $0.0098 - \Delta$.
- i. Excess relative risk = 4.0.

The justification for these values is as follows:

- The median dose in the cohort is 0.097 Gy, suggesting that the mean of the calculated doses in the log scale is $\log(0.097)$.
- The standard deviation of the calculated doses is 0.224Gy. There are however some major outliers, so that for preliminary purposes we set it as = 0.175Gy. This suggests that the standard deviation of calculated dose in the log scale can be calculated in two ways.
 - Assuming log-normality, if σ is the standard deviation of calculated dose in the log scale, then the mean calculated doses divided by the median calculated dose is $\exp(\sigma^2 / 2)$. This suggests a value $\sigma = 1.09$.
 - Alternatively, one can note that the variance of calculated dose divided by the square of the median of calculated dose is $\exp(2\sigma^2) - \exp(\sigma^2)$. This suggests a value $\sigma = 0.93$.
 - We take an intermediate value: $\sigma = 0.99 = \log(2.7)$.
- For each individual, the report computes the median and the upper 95th percentile (page 237), and finds that the ratio of the latter to the former is approximately 4.0. If μ is the mean of the calculated dose in the log scale, and if κ is the standard deviation of the uncertainty in the log scale, then the median calculated dose is approximately $\exp(\mu)$, while the 95th percentile for the individual is approximately $\exp(\mu + 1.645\kappa)$. This suggests that $\exp(1.645\kappa) = 4.0$, or that $\kappa = \log(2.32)$. We took $\kappa = \log(2.3)$.
- The baseline risks for men and women when all error was Berkson were 0.0049 and 0.0098, respectively. When classical error was allowed for, we adjusted the risks slightly by subtracting a quantity Δ chosen to keep the mean number of cases in all simulations at approximately 44. For example, when 80% of the error was Berkson, the baseline risks for men and women were 0.0052 and 0.0103, respectively.
- Because this is a simulation experiment, in which true doses and calculated doses are generated in the log scale, a decision was necessary as to what the calculated doses were in the arithmetic scale of the Gray. We took the mean, assuming that all uncertainties were Berkson. Thus, for an individual, if his or her calculated dose in the log scale was Z_L , and the standard deviation of the uncertainty is κ , then the calculated dose in the Gray scale is $\exp(Z_L + \kappa^2/2)$.
- In the model of Reeves, et al. (2001) and Mallick, et al. (2002), there is a latent variable in the log scale that the latter authors call a “latent intermediate” L . In principle, this is a normal random variable with mean m_L and standard deviation s_L . The difficulty with this is that if one exponentiates L , one can get extremely high true doses in the scale of the Gray. In order to avoid this happening, L was not allowed to be larger than $m_L + 2 s_L$.
- Estimation, hypothesis testing and confidence intervals were computed using likelihood methods. For estimation, we restricted the excess relative risk estimates and the upper end of confidence intervals to be between 0.0 and 40.0.

For hypothesis testing and confidence intervals, we used two-sided likelihood ratio statistics. The true excess relative risk was taken to equal 4.0 so that the power when all uncertainty was unshared Berkson was approximately 0.80.

- Scenarios depended on the percentage of uncertainty that is Berkson and the common correlation in the shared Berkson uncertainties. For each of the eight scenarios, 500 simulated data sets were constructed.
- The mean number of disease cases for each scenario was computed by simulating samples of size $n = 3,000$ a total of 5,000 times, and then averaging.

Table 1: First Simulation Experiment

Correlation of Berkson errors	% of uncertainty being Berkson	Mean number of cases	Mean ERR- Gy^{-1}	Median ERR- Gy^{-1}	Mean Upper 95 th for ERR- Gy^{-1}	Median Upper 95 th for ERR- Gy^{-1}	Power
0.00	100%	40.7	5.7	3.9	23.1	40.0	0.75
	90%	40.6	4.8	3.2	20.6	40.0	0.64
	80%	40.7	3.4	2.4	17.0	24.2	0.55
	70%	40.7	3.0	2.0	14.8	20.8	0.43
0.50	100%	40.7	5.9	3.8	22.5	40.0	0.66
	90%	40.6	4.5	3.0	19.2	33.8	0.57
	80%	40.7	3.4	2.4	16.3	23.8	0.49
	70%	40.7	2.7	2.0	13.7	18.6	0.42

The results of the first simulation experiment were as follows.

- Shared (correlated) Berkson uncertainties lower power. In the case that all uncertainty is Berkson, shared uncertainties lowered the power from 75% to 66%. This is consistent with the results of Stram and Kopecky (2003) in their simulation of Berkson uncertainties.
- Classical uncertainties lower power as well, fairly dramatically. Thus for example if all uncertainties are Berkson and unshared, the power is 75%, while if the Berkson uncertainties are shared and 20% of the total uncertainty is classical, then the power drops to 49%.
- The upper end of the 95% confidence interval for excess relative risk is not dramatically affected by shared Berkson uncertainties, but it is lowered a great deal under the existence of classical uncertainties.

We also conducted a second simulation experiment. The only differences were that the baseline risks for men = 0.0024 - Δ , the baseline risks for women = 0.0070 - Δ and the excess relative risk = 6.5. The results, not given here, qualitatively confirm the first simulation experiment, namely major drops in power for shared Berkson uncertainty and even 20% classical uncertainty.

Summary of Problems with Power Estimates for the HTDS

In the estimation of statistical power with data from the pilot study, Davis et al. (2002) did not adequately account for dose uncertainty. Our simulations indicate that the effect of complex mixtures of dosimetry uncertainty on statistical power can lower power substantially.

In the post-hoc power analyses that accounted for dose uncertainty, Davis et al. (2002) and Stram and Kopecky (2003) selected combinations of correlations between Berkson errors that did not adequately characterize the uncertainties in the dose estimates. Our simulations indicate that accounting for the full complications of mixtures of dose uncertainties with more appropriate and complete assumptions will produce power estimates that are substantially lower than those estimated by Davis et al. and Stram and Kopecky.

Inappropriateness of Bonferroni Adjustment

Davis et al. (2002) employed a highly unconventional approach to estimating confidence limits for the set of parameters $\Theta = (\alpha_M, \alpha_F, \beta)$, which is invalid for the problem of testing or estimating dose-response because that response is determined only by the parameter β , not by the intercepts α_M, α_F . Apparently reasoning that because there are three parameters to be estimated, they wish to estimate joint limits such that there is only a 5% probability that *any* one or more of them will be outside these limits by chance, they report 98.33% limits on each parameter separately. When more intercept (α) parameters are present they use even higher percentages: 98.75% for 3 intercepts, 99% for 4.

This approach is unnecessarily conservative and not based on any appropriate statistical theory, since it effectively assumes the parameter estimates are independent and that each parameter is of equal importance in determining the dose-response of interest. The correct method for setting confidence limits on parameters in multi-parameter models is to use the inverse of the Fisher information matrix, which directly allows for the covariances between parameter estimates. Because we are interested in inference on the dose-response (slope) parameter β , there is no need to be concerned about limiting the overall error rate for *all* parameters jointly; instead, what should be reported is the 95% confidence limits on β , allowing for its dependence upon the other parameters. When a 98.33% confidence level is used for the single intervals, the impact of the Bonferroni correction is to increase the confidence interval width by about 22% (2.394/1.96) over what should have been reported; for 98.75% the increase is 27%, and for 99% it is 31%. Wider intervals are less likely to detect dose-response; for several of the reported intervals the adjustment may have changed significant results to non-significant results. Thus the adjustment further reduces the power of the analysis to detect dose-response.

Failure to Account for Uncertainty due to Deaths and Nonparticipation

Of 5,199 persons identified for potential study based on birth parameters (time and location), 3,447 or about 2/3 completed the clinical exam and some withdrew after that. Potential subjects in the 1/3 of the study cohort for whom this clinical outcome data were unavailable had either died before the study or did not agree to participate. It has long been known that such a high rate of cohort loss can seriously bias study results, even if the marginal data available on all cohort members does not suggest any systematic loss (i.e., even if the loss is of equal proportion with respect to exposure categories or disease categories) (Greenland, 1977).

There is good reason to expect that the loss was nonrandom in ways relevant to the study. It seems reasonable to expect that persons who knew or suspected themselves to have been exposed to Hanford radiation or to have thyroid disease were more likely to have participated. Furthermore, the exposure and disease may have been related to overall mortality. Although seemingly paradoxical, such positive associations of both the exposure and the disease with subject loss can result in a net downward bias in the observed exposure-disease association (Greenland, 2003a), even if the exposure and the disease affect loss independently of one another. An example demonstrating this possibility was first published by Joseph Berkson in 1946 and subsequently came to be known as Berkson's bias or Berkson's paradox (not to be confused with Berkson error, discussed above).

The uncertainty arising from the substantial and potentially nonrandom loss of cohort members could have been addressed using Monte-Carlo techniques akin to those used to address measurement uncertainties (Greenland, 2003b). That this uncertainty was not addressed analytically is another reason why the HTDS report overstates the strength of conclusions with regard to the size of effect that may be present in light of the data (Davis et al., 2003, p. liv).

Interpretation of Results in the HTDS Report

A major concern with the report is the way the authors chose to respond to the following criticism raised by the NRC Subcommittee of the Board on Radiation Effects Research:

"The subcommittee is concerned that the results of the study were reported and interpreted in black and white terms of whether a statistical test was passed or failed. It recommends that confidence limits be provided throughout the report to allow the readers to judge how large a radiation effect might be consistent with the data. It feels that the HTDS investigators probably overstated the strength of their finding that there was no radiation effect." (NRC, 2000, p. 9).

Although the authors did provide confidence limits for their final report, they made little use of them in their interpretations, and the major flaw pointed out in the above NRC quote remains in the Final HTDS report: There is a consistent confusion of lack of statistical significance with lack of evidence, which leads the authors to overstate the strength of their findings. This sort of misinterpretation is common but nonetheless is incorrect, as explained by a number of textbooks (e.g., Royall, 1997; Rothman and Greenland, 1998, Ch. 12). The latter textbook cites numerous articles explaining the fallacy in this confusion.

Correct interpretation of lack of significance in the HTDS report is simply that, when examined using the models and methods of the authors, the data do not overwhelmingly favor any alternative over the null. This correct interpretation leaves open the possibility that the evidence favors the alternative (that there is an effect), albeit not very strongly when using the authors' approach.

For example, on page li of the executive summary, the HTDS reported a slope estimate for prevalence of thyroid ultrasound detected abnormalities (UDAs) of 0.029/Gy, $p=0.14$. While not significant at the 0.05 or even the 0.10 level, this result implies that (using the authors' approach) the evidence supports a slope of 0.029/Gy more than it supports a slope of zero, or no association (albeit not by very much); this support is reflected by the fact that 0.029 is called the maximum likelihood estimate, the value most favored by the data evidence under the chosen model (Royall, 1997). Yet at the end of the Summary and Conclusion on p. liv, in reference to all their results, the authors stated "the results of the HTDS provide no evidence of a statistically significant association."

This statement and others like it in the report conflates absence of statistical significance with absence of evidence. When one separates the two concepts, as one should, one sees that there is lack of significance, but that some of the evidence weakly favors the possibility of small effects. The net result of this clarification shows that, *even under the authors' approach*, the results are more ambiguous than negative, and do not particularly favor the null over possible effects of some concern.

More clarification is attained by examining and correctly interpreting the confidence intervals that the authors present. While we laud the authors for including the confidence intervals, we think that the HTDS report failed to make full use of them in interpreting their results. In the above example for thyroid UDAs, this interval includes the value 0.080/Gy, which if the actual slope, would be of some concern. Under the author's approach this slope value cannot be ruled out, just as the null cannot be ruled out, and which demonstrates the inconclusiveness of the authors' own results. This sort

of examination of the intervals should have been used as a counterpoint to the authors' reliance on statistical significance in formulating their Summary and Conclusions section.

In the preceding criticism of the HTDS statistical interpretation we accepted for the moment the authors' statistical analyses. As we emphasized earlier, however, these analyses are subject to numerous criticisms. The HTDS report relied almost exclusively on the results from a series of regression models that were based on dose modeled as a continuous variable, with few additional explorations of confounding, effect modification, or dose misclassification. In the light of the aforementioned problems with power and dosimetry uncertainty, more thorough analyses would have had serious impact on the interpretation of the study, leading to even more ambiguous results, which in turn should lead to even more cautious and limited interpretation.

In the Summary and Conclusion section of the HTDS final report, the authors concluded that: "...the findings of this study are not inconsistent with the current published literature regarding the effect of exposure to radioactive iodine and the risk of thyroid and parathyroid disease. This is particularly so given the relatively small magnitude of the estimated thyroid radiation doses in members of the HTDS cohort (mean = 174 mGy) and the relatively protracted nature of the exposure over time. There is little evidence in the literature to suggest that people exposed to I-131 at the levels found in this study over a period of months or years would experience higher rates of thyroid or parathyroid disease as a result of their exposure (Davis et al., p. 543)."

The meaning of this statement is unclear. It can be interpreted in at least three different ways: 1) the results, although not showing statistically significant elevations in risk, are not inconsistent with other published studies supporting risks for certain thyroid diseases from I-131 exposures, if the upper bounds of reported confidence intervals are considered (as pointed out for the case of thyroid cancer and I-131 exposure from NTS fallout by the NRC (2000, pp. 111-112); 2) a strict interpretation of the lack of statistical significance for the results is consistent with the findings of no risk for thyroid diseases in some (but clearly not all) studies of I-131 exposures, but not consistent with those that show elevated risks; or 3) a strict interpretation of the results based on statistical significance is not inconsistent because there is little evidence that protracted exposures to I-131 are associated with thyroid disease.

Interpretation 1 is the most supportable of the three, particularly in the light of the HTDS overestimates of statistical power and their incomplete characterization of the effects of dose uncertainty. Interpretation 2 cannot be supported because, as we have shown, that the study power is too low to rule out Type 2 errors, and because there are studies in the scientific literature providing good evidence for associations between I-

131 exposure and thyroid cancer, thyroid neoplasia, and hypothyroidism (see the expert reports of Drs. Hoffman and Ruttenber). Interpretation 3 is one that is not possible to evaluate because the HTDS study did not clarify which subjects received doses over relatively short periods of time (as would have occurred in the spring and summer of 1945) and compare the risks for these subjects with those who received doses over protracted periods, and then compare the risks for both groups with evidence from the scientific literature for risks for I-131 exposure (which is almost exclusively for doses received over short time intervals). [Note: the exposures to I-131 from Hanford during 1945 were extended over many months. Exposures to I-131 from Chernobyl and from fallout were extended over a time period of at least one month per event].

The statement about little published evidence for risks from protracted exposures to thyroid doses of “small magnitude” is also misleading in that it implies that such exposures have actually been studied. We know of no such studies, and none are cited in the literature review of the HTDS final report. The reason that there is little evidence is due to the absence of evidence, not to the existence of studies that actually show no risks from protracted exposures.

In the Summary and Conclusions section of Davis et al. (2002), the authors questioned whether their results were inconclusive because of dose uncertainties in the primary analyses. They then answered their question by citing supporting evidence (i.e., no statistically significant differences between exposure groups or for dose-response models) from three alternative ways of analyzing the data.

The first alternative analysis used qualitative exposures defined in two different ways—by geostrata based on the residence location of the subject’s mother at the time of birth, and by a dichotomous exposure variable based on the reported fresh milk consumption of about 1,257 of 3,440 subjects for whom CATIs were conducted (Davis et al., 2000, pp. 190-192).

As noted above, thyroid exposures from NTS fallout were also received in the geostrata, and doses from these exposures would have been correlated with milk consumption, making analyses with both qualitative exposure alternatives difficult to interpret. Moreover, comparing the cumulative prevalence of thyroid diagnoses between nine geostrata would likely have low statistical power because the exposure classification would be highly uncertain.

Low statistical power is also likely for the analyses with dichotomized exposures, as fewer than half of the total study subjects were included in analyses. We maintain that the similarity in results between the analyses with different exposure metrics is more consistent with low statistical power in each of the three approaches than with the hypothesis that there is truly no relation between exposure and disease.

The second alternative analysis was an assessment of dose uncertainty by two different approaches: 1) a descriptive analysis that plotted the slope parameter for a linear dose-response model and its Bonferroni-adjusted 95% confidence intervals for the 100 realizations of dose; and 2) a Bayesian-based analysis designed to deattenuate the estimates of the regression coefficient in a sex-stratified logistic model (Davis et al., 2000, pp. 205-207). As noted above, the use of Bonferroni-adjusted confidence intervals is overly conservative and it appears the Bayesian-based analyses are inexplicably flawed. The results from these approaches are therefore not informative, and are just as consistent with an interpretation of inadequate statistical power and dose misclassification as with no association between thyroid dose and thyroid disease.

The third analysis was the post-hoc simulation of statistical power. As noted above, the HTDS authors assumed in their Bayesian analysis that dosimetry error was almost entirely classical, but simulated post hoc power with an assumption of almost 100% Berksonian error. Our simulations indicate that this decision resulted in a substantial overestimation of power, and therefore the results do not support the hypothesis that there were no associations between thyroid dose and disease.

In the final two sentences of the final report, the authors stated: "These findings do not definitively rule out the possibility that Hanford radiation exposures are associated with an increase in one or more of the outcomes under investigation. However, it does mean that if such associations exist, they were likely too small to detect using the best epidemiologic methods available."

We agree that the findings do not definitively rule out the possibility that Hanford exposures are associated with an increase in outcomes; in fact the findings are quite compatible with considerable increases in risk of those outcomes, as well as with no increase. With regard to the second sentence, even the best epidemiologic methods are not enough to compensate for a study population that is too small and measurements too uncertain to detect even large risks. Finally, we do not believe that their analytic treatment of uncertain doses, their analyses of the effects of statistical power on the slopes and confidence intervals of the dose-response models, and their use of Bonferroni corrections and statistical significance can be considered "the best epidemiologic methods available," although even the best analytic methods could not compensate for the limitations of the study data; at best they could only show how little can be inferred from the data.

Conclusions

We have identified problems with misclassified doses and possible effect modification by thyroid doses resulting from exposures to NTS fallout. We have also provided evidence that the statistical power for analyses of dose-response relations is much lower than reported and not adequate for accepting hypotheses of no relations between thyroid doses and risks for thyroid cancer, thyroid neoplasia, hypothyroidism, or autoimmune thyroiditis. Finally, we have emphasized that proper interpretation of statistical results would lead to a more cautious view of the evidence than evinced in the HTDS final report.

With regard to the questions we posed in the introduction to this report, to a reasonable degree of scientific probability, we find that:

- 1) The radiation doses to the thyroid and their uncertainties were not modeled appropriately for the dose-response analyses that were conducted. Far more attention should have been paid to dose uncertainties and contributions from NTS and other weapons-test fallout.
- 2) The dose-response models did not have adequate statistical power to conclude that there was no statistically significant evidence of thyroid disease risk from Hanford exposures.
- 3) The statistical analyses and interpretations were strongly influenced by the assumption that the study had adequate statistical power and in general overstated the evidential value of findings that were not statistically significant.

In the light of these findings, and to a reasonable degree of scientific probability, we conclude that, as currently reported, the results from the HTDS cannot be used to rule out important risks for thyroid cancer, neoplasms, or hypothyroidism. Furthermore, because of problems with dosimetry errors, bias, and misclassification, it is not even scientifically defensible to use the reported confidence intervals for regression coefficients to make inferences about the levels of risk that would be consistent with the data. At a minimum, reanalysis of original data would be needed to produce reasonable interval estimates.

We conclude, to a reasonable degree of scientific probability, that the estimates of risk for thyroid diseases that have been reported in the scientific literature for other exposures to I-131 and external penetrating radiation are the most appropriate for determining the contribution of risk for thyroid diseases from Hanford I-131 exposures.