# False Indications of Dose-Response Nonlinearity in Large Epidemiologic Cancer Radiation Cohort Studies; A Simulation Exercise

Authors: Beyea, Jan, and Hoffmann, George R.

Source: Radiation Research, 199(4) : 354-372

Published By: Radiation Research Society

URL: https://doi.org/10.1667/RADE-21-00217.1

RADIATION RESEARCH **199**, 354–372 (2023)
0033-7587/23 $15.00
©2023 by Radiation Research Society.
All rights of reproduction in any form reserved.
DOI: 10.1667/RADE-21-00217.1

# False Indications of Dose-Response Nonlinearity in Large Epidemiologic Cancer Radiation Cohort Studies; A Simulation Exercise

Jan Beyea,[a,1] George R. Hoffmann

[a] Senior Scientist Emeritus, Consulting in the Public Interest, Lambertville, New Jersey 08530; [b] Distinguished Professor of Science Emeritus, Department of Biology, College of the Holy Cross, Worcester, Massachusetts 01610

**Beyea J, Hoffmann GR. False Indications of Dose-Response Nonlinearity in Large Epidemiologic Cancer Radiation Cohort Studies; A Simulation Exercise. Radiat Res. 199, 354–372 (2023).**

**This study explores the likely prevalence of false indications of dose-response nonlinearity in large epidemiologic cancer radiation cohort studies (A-bomb survivors, INWORKS, Techa River). Reasons: Increasing numbers of tests of nonlinearity are being made in studies. Hypothesized nonlinear dose-response models have been justified to policy makers by analyses that rely in part on isolated findings that could be statistical fluctuations. After removing dose nonlinearity (linearization) by adjusting person-years of observation at each dose category, indications of nonlinearity, necessarily false, were counted in 5,000 randomized replications of six datasets. The average frequency of any false positive for five indicators of nonlinearity tested against a linear null was roughly 25% in Monte Carlo simulations per study, consistent with binomial calculations, increasing to ~50% within 6 studies assessed. Comparable frequencies were found using Akaike's information criterion (AIC) for model selection or multi-model averaging. False above-zero threshold doses were found more than 50% of the time, averaging to 0.05 Gy, consistent with findings in the 6 studies. Such bias, uncorrected, could distort meta-analyses of multiple studies, because meta-analyses can incorporate high P value findings. AIC-based correction for the extra threshold parameter lowered these false occurrences to 8 to 19%. Given the simulation rates, the possibility of false positives might be noted when isolated findings of nonlinearity are discussed in a regulatory context. When reporting a threshold dose with a P value > 0.05, it would be informative to note the expected high false prevalence rate due to bias.** © 2023 by Radiation Research Society

## INTRODUCTION

Regulation of exposure to ionizing radiation is based on the idea that cancer outcomes are roughly linear with radiation dose and have no dose threshold (*1–5*). This has become the regulatory default or null model. The default in radiation epidemiology is usually tested by fitting cancer counts to mathematical models of response that differ as a function of the dose variable, while relying on null hypothesis testing to assess any need for nonlinear components to be incorporated, such as quadraticity, curvature, or dose threshold. Commonly, a graph is also presented showing a smoothed "loess" fit, along with confidence bands (*6–9*). Thus, a reader can make visual judgments about dose response, supralinear or sublinear, even if authors make no explicit inferences about nonlinearity based on the image. Modest changes to linearity at low doses are part of the regulatory default, as quantified by a scaling factor, the low dose extrapolation factor (LDEF) (*10*).

The regulatory default is not set in stone and is subject to change by policy makers. For instance, in 2018 critics of the linear non-threshold theory (LNT) were able to write sections of a proposed dose-response rule at the USEPA (*11–14*), although subsequently vacated by a Federal Court (*15*). A stated underlying argument by some critics is, "If the LNT is correct, then it should be able to explain the findings in all studies..." (*16, 17*). This reasoning may explain the emphasis placed on isolated findings by some LNT critics as part of their argumentation (*16–20*). However, such thinking does not account for multiple tests for nonlinearity within and across studies.

Isolated indications of nonlinearities in dose response, including those labeled as statistically significant, can occur due to statistical fluctuations in cancer counts at different dose categories. As an example, consider replications where counts at higher doses fluctuate upwards. Positive curvature will be introduced. Positive curvature will also be introduced if counts at lower doses fluctuate downward. If the fluctuations are large enough, fitting software will find a false positive for nonlinearity.

Isolated findings need to be interpreted based on the statistics of multiple comparisons, which is not a concept familiar to everyone with influence on regulatory policy. Moreover, for novel assessment methods, especially visual assessment, the necessary statistical properties needed to

[1] Corresponding author: Consulting in the Public Interest, 53 Clinton Street, Lambertville, NJ, 08530; email: jbeyea@cipi.com.

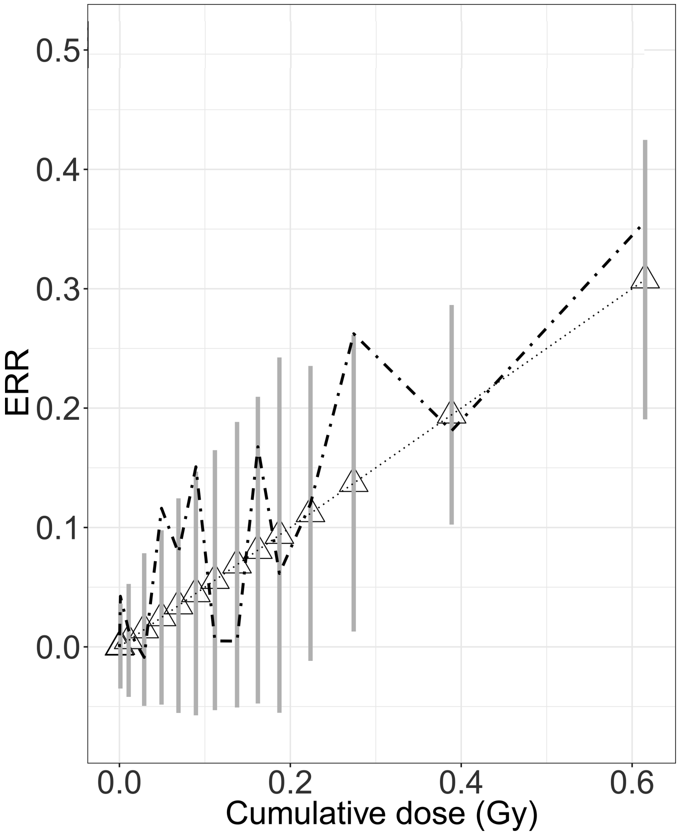

**FIG. 1.** Comparison of excess relative risk (ERR), indicated by cancers/person-time, before and after linearization of the 2017 A-bomb LSS data. The ''after'' data (solid triangles) were obtained by scaling person-time for each data point so that the revised risk lay exactly on a linearized line. No change was made in cancer count data. For plotting purposes, the risks were normalized to the lowest fitted risk value.

interpret multiple results within and across studies may not have been well established.

The estimation of the prevalence of false findings of five types of nonlinearity in multiple comparisons within and across 6 epidemiologic cohort studies is the subject of this paper. The 6 publicly available cohort datasets analyzed were 2 Techa River cohorts (incidence and mortality) (*9, 21*), the INWORKS worker study (*22*), and three Atomic bomb survivor studies (*7, 8, 23*). Using simulation, four sets of questions were explored about multiple comparisons in radiation epidemiology studies.

1. Are there analysis methods in radiation epidemiology that could overstate evidence for nonlinearity, for instance, because the method underestimates the chances that a finding is a false positive?
2. In general, what is the expected aggregated false positive rate for a study, given the many tests for nonlinearity carried out in studies? Specifically, what is the likelihood that at least one false positive for nonlinearity would show up in a single study given, *n*, possibly correlated tests?
3. What is the likelihood that at least one false positive for nonlinearity would show up in multiple studies? How do the simulation results compare with results in the published studies?
4. Does the reporting of three dose-threshold values averaging to 0.043 Gy in four epidemiology studies provide evidence of a true dose threshold, or does the 75%-occurrence rate reflect statistical noise or above-zero bias? Specifically, what is the expected frequency of above-zero dose thresholds irrespective of P values?

Based on the answers to these four sets of questions, implications for regulatory analysis and radiation research are discussed. To look for false findings in analysis methods that do not rely on null-hypothesis testing, the Akaike information criterion (AIC) was also considered for selecting models, as was multi-model averaging.

To quantify expected prevalence of indications of nonlinearity, we developed a straightforward way to transform risk values in publicly available cohort datasets to remove all nonlinearity from dose responses (herein referred to as ''linearization''). Randomness was added to dose category counts using simulation to generate 5,000 replications to analyze for nonlinearity. The linearization procedure ensured that the regulatory null hypothesis of linearity without threshold was known to be true for the modified datasets (Fig. 1). Thus, any finding of a nonlinear response in the replicates would have to be false, caused by statistical fluctuation, no matter what the assessment method might be: standard, novel, parametric, non-parametric or visual.

Although the main assessments made in this article were carried out on cohort data with grouped counts, it was also of interest to see if a more detailed analysis based on individual dose and cancer outcome would produce any improvement in results. For this purpose, an additional simulation step was required to disaggregate individual cases from grouped data. While we worked with cohort data sets, the methodology is also applicable to case-control studies, as we briefly describe in the Discussion section.

Our article is not about finding the true dose response or supporting the LNT scientifically. Our concern here is that whatever method of inference is chosen, null hypothesis testing or multi-model inference, results should be presented in ways that minimize misinterpretation about isolated or statistically biased findings.

This paper has two distinct audiences. The first consists of those who do statistical modeling and/or quantitative analysis to inform regulatory decision-making. The second audience consists of those who do not fall in either of these two categories but who are consumers of epidemiologic data or have an interest in dose-response relationships. Those in the second group may want to skip some of the detail in the Materials and Methods section. In addition, note that the final section, Summary and Conclusions, can serve as an executive summary.

## MATERIALS AND METHODS

This section is divided into five subsections, including ''Datasets and treatments,'' followed by four subsections corresponding to the four sets of questions asked in the Introduction:

- Datasets and treatment, which includes: *Dataset sources; Linearization of risks; Relative risk data; Consideration of study covariates and standardization*.
- Methods for question 1 (Are there high false positive rates for single tests?): *Indicators of nonlinearity; Evaluation of nonlinear shapes; Fitting cancer count data*.
- Methods for question 2 (What is the expected rate of at least one false positive in a single study?): *Binomial calculations for a single study*.
- Methods for question 3 (What is the expected rate of at least one false positive in multiple studies?): *Binomial calculations for multiple studies; AIC selection of best model; Multi-model averaging*.
- Methods for question 4 (What is the expected frequency of above-zero dose thresholds?): *Toy threshold-dose-response model; Fitting simulated individualized survival data*.
- Process flowcharts can be found in the supplementary material[2] (https://doi.org/10.1667/RADE-21-00217.1.S1), figs. S1 to S4 (https://doi.org/10.1667/RADE-21-00217.1.S2): *Flowchart overview; Flowchart for linearizing the datasets; Flowchart for analyzing the linearized datasets; Flowchart for finding maximum likelihood and confidence intervals*.

### *Dataset Sources and Treatments*

*Datasets.* Six publicly available cohort datasets were analyzed: two Techa River cohorts (incidence and mortality) (*9, 21*), the INWORKS worker study (*22*), and three A-bomb survivor studies (*7, 8, 23*). Analysis began with pairs of counts and person-years of observation obtained for a sequence of dose categories. For the three A-bomb survivor studies, the needed pairs were extracted from their publicly available, stratified datasets (*24–26*). Data for the Life Span Study (LSS) cohorts were chosen for analysis. Data were grouped either directly or by adjusting person-years for covariates (standardization), using mean corrections determined in the respective studies, as discussed later in this section. For the default analyses of the A-bomb datasets, the maximum dose included was ~0.6 Gy to keep the focus on the lower dose range and to better match dose ranges in the other studies analyzed. Even in studies with large regression sample sizes, the dose ranges can be restricted to the lower regions when looking for nonlinear effects at low doses (*8*).

Pairs of counts and person-years for the INWORKS study (*22*) and both the Techa River incidence cohort study (*9*) and the mortality cohort study (*21*) were obtained from data found in the published literature. For ease of reproducibility and to show the differences in reference dose levels, the 6 sets of paired data are collected in the Supplementary Material (tables S1–S3; https://doi.org/10.1667/RADE-21-00217.1.S2). The number of dose categories (the regression sample size) was, 11, 7 and 7, respectively, for the INWORKS and the two Techa River studies. In contrast, the A-bomb datasets have 14–16 dose categories over the same dose range and around 25 over the total dose range.

The Techa River datasets had an unusual pattern of counts per category, where the counts in the lowest category (reference category) were much lower than the counts in the next two higher dose categories. To explore the impact of such a reversal on the frequency of false positives, modifications were made to the datasets for sensitivity analysis. The count in the reference category was scaled upward to match the count in the second dose category, while simultaneously scaling person years to keep risk the same.

For the 2017 A-bomb dataset, we chose the Not-In-City group (NIC) as the reference level to avoid questions about the reference level being above a possible dose region of hormesis (*27, 28*). Only the 2017 dataset separates data for the NIC group. Because such a reference level is not standard, sensitivity analyses were performed. For example, the Not-In-City group was combined with the group of survivors at closer distances, namely those located between 3 and 10 km from the epicenter at the time of the bombings.

It was pointed out to us by a reviewer that, in an earlier set of simulations for the two-stage clonal model, which can produce highly discontinuous dose response functions, reliable parameter estimates for the dose response function simulated were only obtained by individualized regression. Parameter estimates for this model were not reliable when individuals were grouped into Poisson dose categories (*29*). Although the dose response models considered in this article are all continuous in dose, the likelihoods for the threshold and 2-slope spline models can have discontinuous derivatives. We therefore looked to see if individualizing risks would make a difference in the results for these two dose-response models.

To this end, two new datasets were created with 17,500 simulated Techa River individuals to be used in dose threshold analysis and 1,750 to be used in the 2-slope spline analysis. Each simulated individual was assigned a dose and a survival history, which incorporated radiation-induced mortality. The radiation contribution to survival was generated by a Cox proportional hazards model, with hazard function linear in dose (Supplementary text S-1; https://doi.org/10.1667/RADE-21-00219.1.S2).

### *Linearization of Risks*

The data were linearized by adjusting person-years of observation at each dose category, which allows the risk at any dose category to be arbitrarily set, while preserving relative variance of risk. The unadjusted risk at each dose category in a cohort study is determined by the number of cancer cases per person-year at risk (counts/pyr). The relative error (standard deviation over the mean) in these risk values depends only on the count variable, if any experimental error in person-years is neglected, as is standard. To obtain a synthetic set of pure linear data with Poisson errors and therefore relative errors matching the underlying study, the counts in each dose category were left unchanged. Only the number of person years in a dose category was modified, which allowed changes in risk without change in count. Person-years were scaled at each dose category, $k = 1$ to $n$, as in Eq. (1), so that the ratio of count to person-years increased linearly with the category's dose at index $k$, designated as $d_k$.

$$Linearized\ person\ years(k) = count(k) \times \frac{person\ years(k=1)}{count(k=1) \times (1 + s \times d_k)}. \tag{1}$$

The variable $s =$ slope. Transformation stages for each dose category in the 2017 A-bomb dataset are shown in Supplementary tables S-2 and S-3 (https://doi.org/10.1667/RADE-21-00219.1.S2). After linearization, the numbers and positions of the dose categories and the corresponding counts and count variances remain. The process is equivalent to assigning every individual study subject in a dose category the same value of person-years multiplied by a dose dependent factor whose inverse is linear in dose.

Once the risk was set to pure linearity of dose response, Monte Carlo simulation techniques were used (*30*) to quantify prevalence of nonlinearity. Poisson variations were added to the count data using the open source R-function, *rpois* (*31*). Each of 5,000 unique dose-response curves was generated for each of the 6 linearized datasets by

[2] Editor's note. The online version of this article (DOI: https://doi.org/10.1667/RADE-21-00217.1) contains supplementary information that is available to all authorized users.

drawing from the Poisson distribution centered on the average count data in each dose category. Random samples of the modified data were then drawn and used to compute test statistics whose distributions provided estimates of the test statistic under the null (*30*). Sensitivity runs were made with the same starting random number seed, so that the same replications were analyzed for each different set of assumptions.

Person-years were kept unchanged from replication to replication. The choice of linear slope is a free parameter that can be chosen to match published values or varied to see if changes in slope affect the results of analysis. A published linear slope may have been affected by some underlying and uncaptured nonlinearity, and, of course, its value is subject to statistical and other errors so it may deviate from other studies.

### *Relative Risk Data*

In addition to the absolute risks discussed above, we also explored a nonstandard relative risk model, where the data were made relative by dividing all risk values by the reference risk (risk at the lowest dose). This model is not based on a reparameterization that leaves the data alone but makes the parameters relative. Such a reparameterization was not considered here, because it does not produce results that differ from the absolute risk model in our simple case without covariates (results not shown). To distinguish this type of relative risk model from the reparametrized versions, it is labeled here as a ''relative risk data model.'' An analyst might consider such a relative risk model for exploring nonlinearity, which is easy to analyze with well-established and accessible software routines, should proprietary software, such as AMFIT, not be accessible for analyzing the reparametrized version. Alternatively, an analyst might simply be exploring alternatives to standard approaches.

### *Consideration of Study Covariates and Standardization*

Only univariate analysis with dose as the independent variable was carried out in this article. In the case of the A-bomb datasets, it was possible to adjust risks by important sex-averaged covariates, such as age at exposure and age of attainment. This was possible because the necessary sex-averaged adjustment functions for relative risk models (not absolute risk models) and their fitted coefficients were included in the published A-bomb articles (Supplementary table S-4; https://doi.org/10.1667/RADE-21-00219.1.S2). Sex-averaged coefficients for absolute risk models were not available for all three A-bomb datasets, so the coefficients for relative risk models were used.

As a sensitivity test, these adjustment factors were used to produce standardized person-years for each stratum prior to linearization. Use of these covariate-adjusted risks in linearization of the datasets produced a change in absolute risk scale factor only. There was no change in relative risk at any dose category, a result that was initially surprising to us but was confirmed by algebraic analysis (Supplementary text S-2; https://doi.org/10.1667/RADE-21-00219.1.S2), as well as confirmed numerically by running simulations and obtaining the same false finding frequencies.

The fact that standardization does not change the linearization by more than a scale factor is a fortunate result, because no covariate adjustment functions are publicly available for either the Techa River Cohort or the INWORKS worker study. Neither could we adjust background by city and location for the A-bomb data, as has routinely been done in studies carried out at the Radiation Effects Research Foundation (RERF). We also were unable to adjust for smoking in the 2017 dataset as was done in the corresponding LSS study (*8*). Although the relative variation between our crude and standardized risk estimates for the A-bomb datasets amounted to less than 10% in the 0 to 0.6 Gy range (results not shown), such a modest change would not be expected for the INWORKS study, where worker age would be correlated with cumulative dose.

### *Methods for Question 1 (Are there High False Positive Rates for a Single Test?)*

*Indicators of Nonlinearity.* All of the five indicators of nonlinearity used or inferable from smoothed graphs in the six published studies were tested in each replication of linearized data. Specifically, we catalogued dose thresholds, curvature and quadratic terms in the replicates, as well as the dose response shapes, supralinearity and sublinearity (including hormesis). The tests are described further in Table 1. Individual test results were tabulated and aggregated for each replicate. For null-hypothesis testing, a (false) positive was counted when 95%-confidence intervals for the relevant parameter excluded the linear null.

A separate count was made for false-positive curvatures having absolute magnitudes greater than 1 and thus a greater than twofold difference in risk at 1 Gy compared to what would be expected from the slope at low doses. These we deem to be of regulatory interest.

Perhaps, the most important finding of a nonlinearity for regulatory policy would be evidence of a dose threshold. The threshold model used here is a simple, 1-parameter extension of the linear default:

$$Y(d) - Y(0) = \begin{pmatrix} B(d - D),\ for\ d > D \\ 0,\ for\ d \le D \end{pmatrix}. \tag{2}$$

Where $d$ is dose, $D$ is the threshold dose value, B is the slope above threshold, and $Y$ is the risk of cancer.

The case of most interest is when $D$ is found to have a narrow confidence range that excludes the null of 0. However, in addition to counting false positives for $D$, tabulations were made for all $D$ values above zero in the simulations, no matter how weak the statistical evidence. This was done for several reasons. First, results about dose thresholds that do not rise to the level of a false positive are routinely published with P values, forming a collection that could be used for meta-analysis. Second, dose threshold findings are sometimes presented or discussed in the literature as indicators of nonlinearity without considering the strength of the evidence (*32, 33*), making it of interest to quantify the likelihood of such indications being statistical fluctuations. It can be argued that such unsupported claims are so obviously flawed that no quantitative response is needed. However, dismissal out of hand may not be sufficient to carry the day with audiences who influence regulation; quantitative illustrations of the frequency of false indications may be helpful.

Average values for the threshold dose, which might serve as a measure of bias away from the null, were also calculated. In an attempt to gain insight into above-zero threshold values, separate counts were made for those threshold dose values that satisfied the AIC criterion for an improved fit, which penalizes the statistical likelihood for each extra parameter introduced (*34*).

### *Evaluation of Nonlinear Shapes*

Nonlinear shapes beyond the 3-parametric functions, threshold, quadraticity and curvature, were addressed using four methods. The methods were loess smoothing, 2-slope spline, the occurrence of five points lying on one side of the linear line, and multi-model averaging. A 2-slope spline dose response function was of particular interest, because it was used in the Techa River mortality study. For AIC selection and multi-model averaging in our analyses, the 2-slope spline function provided the only shape analysis. The 2-slope spline fits were made to each replicate using the R-functions, 'lSpline' and optimizer,'mle.'

The 5-consecutive-point method was of interest because such dose-response curves tend to look nonlinear. They are an example of what has been called the ''clustering illusion'' (*35*). To qualify as a nonlinearity, the required five points had to start with the second data point and cluster on one side of the linear fit to the data, all by at least 0.1 standard deviation.

**TABLE 1**
**Overview and Description of Indicators of Nonlinearity Assessed in Replicates[a]**

| Fitted parameter | Indication[a] | Method | Comment | Sensitivity analysis |
|---|---|---|---|---|
| Threshold dose (false positive) | Confidence interval (CI) excluding zero | Profile likelihood[b] | Ill-behaved profile likelihood[c] | *t* test critical values |
| Threshold dose (>0) | Central value >0 | Profile likelihood[b] | Ill-behaved profile likelihood[c] | |
| Quadratic term | CI excluding zero | *t* test for weighted regression; profile likelihood for Poisson | Well-behaved profile likelihood | Profile likelihood step search |
| Curvature[d] | CI excluding zero | *t* test; profile likelihood for Poisson | Ill-behaved profile likelihood[d] | Likelihood ratio test |
| Curvature magnitude >1 | CI excluding zero | *t* test; profile likelihood for Poisson | Ill-behaved profile likelihood[d] | |
| Loess shape test[e, f] | CI exclude linear fit at a dose < 0.25 Gy[e] | R-routine, ''loess'' | | |
| 2-slope spline shape test[f] | Lower limit of breakpoint dose >0 | Profile likelihood[b] | Ill-behaved profile likelihood[c] | Likelihood ratio test |
| 5-point shape test | 5-points on one side of linear line[g] | Occurrences counted | Percent not obvious for relativized data | |
| AIC model score | Model with lowest AIC score[h] | Subtract penalty from max loglikelihood | AIC = –2 × penalized loglikelihood | |
| Ratio of ERR/Gy's in multi-model fit | Ratio of ERR/Gy's more than factor of 2 up or down below 0.25 Gy compared to 1 Gy[i] | AIC-weighted average of 5 models, with 2-slope spline for shape | CI of multi-model average had to also exclude linear fit at some dose | |

[a] False indications = either false positives (nominal *alpha* = 0.05) or above-zero dose threshold. The threshold function was constant up to the breakpoint, a linear function of dose thereafter.

[b] For dose threshold and 2-slope spline, a segmented search between dose categories using standard maximum likelihood estimation (mle) routines produced the same results as a brute force step-by-step search over the entire breakpoint range, only faster.

[c] Boundary at zero; profile likelihood for dose breakpoint often had multiple peaks and almost always had discontinuous derivatives at doses corresponding to dose categories (Supplementary figs. S-6 and S-7; https://doi.org/10.1667/RADE-21-00219.1.S2). Likelihood at zero breakpoint for 2-slope spline was usually discontinuous.

[d] Curvature was determined using the formulation of dose response as, *dose* × (*slope* + *curvature* × *dose*). The profile likelihood shape contains both a valley and a peak (Supplementary fig. S-10; https://doi.org/10.1667/RADE-21-00219.1.S2). It implicitly depends on the ratio of two variables, one of which can be zero.

[e] Only one shape per replicate was counted to avoid double counting.

[f] Separate counts for supra- and sublinearity. Hormesis counts are usually a subset of sublinearity, but not always for a reference level set to the risk at the first datapoint.

[g] Exceedances began with the second dose category, all exceeding 0.1 standard deviation (obtained from square root of counts).

[h] AIC = Akaike information criterion. The loglikelihood penalty is the number of parameters in the model.

[i] If highest dose in dataset is <1 Gy (INWORKS, Techa River), then the comparison was the ERR/Gy at the highest dose. Average of models was used, not medians.

Locally estimated scatterplot smoothing (loess) was taken to be the default shape analysis. To qualify as an apparent sublinear or supralinear dose response, the 95%-loess confidence band had to fall below or rise above, respectively, the linear fit to the data, somewhere in the 0–0.25 Gy interval. The loess method was chosen as the default for shape analysis because it is easy to understand and interpret. It also has been used in numerous radiation epidemiologic cohort studies (*6–9*), including three of those analyzed here. The loess technique is of particular interest in an investigation of false findings, because the method is not recommended for datasets with a small number of data points (*36*) and could reasonably be expected to have non-standard error rates. Further details are given in Supplementary text S-3 (https://doi.org/10.1667/RADE-21-00219.1.S2).

Additional details for indications of nonlinearity:

1. The dose range in which nonlinear shapes were tallied was 0 to 0.25 Gy, somewhat above the 0.1 Gy that is typically cited as the boundary above which linearity is not contested, at least for high dose rates (*1, 37*). The reason for the higher choice here is that some LNT critics have been concerned about apparent deviations higher in the dose-response curve for the 2012 A-bomb mortality study, from 0.2 to 0.7 Gy (*28*). The value of 0.25 was chosen as a reasonable limit for illustrative purposes. In counting totals of nonlinear shapes, only one per simulation was tallied (to avoid double counting).
2. There is ambiguity in the designation of the baseline for hormesis, which determines how low the confidence bands around the smoothed curve must fall to qualify as hormesis. Three options were considered.
   a. The default reference level was the y-axis risk value at the zero-dose intercept for a linear fit to the data, with the idea that such a level represents the null hypothesis for linearity.
   b. A sensitivity case was the y-axis intercept of the fitted function, either loess or 2-slope spline, which some might consider more realistic.
   c. A second sensitivity case was the y-value of the first data point, a choice that requires no modeling assumptions. Because this choice is vulnerable to fluctuations in the count at the first datapoint, excessive false results may be produced when the number of counts is small at the first datapoint.

### *Fitting Cancer Count Data*

Each replicate's data were analyzed for the five nonlinearities by software either by T-test, profile likelihood or likelihood ratio test. Software was set to call a result a positive when the relevant test statistic reached 95%-confidence (*alpha* set to 0.05). Deviations from a 5% false-positive rate, which are known to occur (*38*), could then be assessed by counting the number of software-assigned positives in simulations. Only if the rate in simulations exceeded 10% did we call attention to it, for two reasons. First, some radiation studies use 10% to delineate positives (*22, 39, 40*). Second, different tests are available to test essentially the same null hypothesis but can give different results for finite regression sample sizes (*41*). Finally, test values have their own uncertainty (*42*). Furthermore, an incorrect assumption about the true value of *alpha* and the width of the distribution for a test statistic would likely have little practical impact were the deviation modest.

For scoping purposes, brute force step searches involving many thousands of small increments in dose were made to find the maxima and confidence intervals for profile likelihoods predicted by software. This profile approach allowed graphical visualization of the complexities of profile likelihoods as functions of parameter values, which made it clear that in some cases software algorithms were inappropriate for finding maxima and confidence intervals. As in Neriishi et al. (*43*), 95% confidence limits were taken at $3.84/2 = 1.92$ units above and below the parameter value at the log of the maximum profile likelihood (*44*).

Most of the nonlinear models have only one more parameter than the linear default, so the profile likelihood and likelihood ratio tests were generally the same. For the 2-slope spline model, there were two extra parameters, so lower confidence bounds for likelihood ratio tests were found at 5.99/2 units below the maximum loglikelihood.

When it was not necessary to perform step searches, e.g., for a linear-quadratic dose response, the maximum likelihoods were found using standard optimization software, specifically maximum likelihood estimator, ''mle'' in the R-statistical language, which sped up calculations. For the linear-quadratic fits, false-positive designations were based on the standard errors produced by the software, which are based on assumptions of smooth behavior at maximum likelihood. However special treatment was necessary to use standard software optimizers to speed up calculations for threshold breakpoint and 2-slope spline, which can have multiple likelihood peaks and discontinuous derivatives. For these two functions, threshold and 2-slope spline, segmented searches between dose category boundaries were used to obtain maximum likelihoods and critical points for profile likelihood, as well as for use in likelihood ratio tests. Between datapoints, the likelihood met regularity conditions allowing searches with the *mle* function to be used to find local maxima within the segments. These calculations were also checked with brute force step searches.

Additional details of the fitting process are listed below.

1. Confidence intervals are used in this article as benchmarks, not to indicate a bright line boundary. The bright-line approach is also implicit in the language of positives, true or false; we use such language here for convenience and because of its wide use in many fields as part of null hypothesis statistical inference.
2. Dose-response was analyzed with inverse-variance weighted linear regression and Poisson regression. The variance for weighted regression of absolute data was taken proportional to the cohort count in a dose category and thus based on Poisson statistics. When fitting the excess relative risk version of this model, the first risk point was deleted, since it does not vary from 0 by definition, and no intercept parameter was allowed in the fit. With relativized data, the variance is no longer pure Poisson. For weighted linear regression, the new variance of the ratios was calculated and used in weighting. To make the calculation, the count distributions were approximated as normal distributions and the ratio of counts was also approximated as a normal distribution (*45*). To extend analysis to 2 Gy for sensitivity analyses of the A-bomb datasets, only weighted linear regression with Poisson based weights was used, because Poisson regression for the 0–2 Gy range introduces its own nonlinearity due to exponentiation of the fitted results to log data
3. Weighted linear regression and Poisson regression may give different inference results for finite number of dose categories (i.e., regression sample size), due to the different underlying likelihood functions, normal vs. Poisson. To check that the false positive rates calculated by weighted linear regression would, as regression sample size was increased, converge to the results obtained using Poisson regression, sample sizes were artificially increased. To this end, new data points were interpolated and placed halfway between count positions of the databases to keep the dose category intervals reasonably uniform, as would likely be done in a real study. Next, all data points were scaled to maintain total counts the same. This process was iterated so that the number of data points was approximately increased fourfold. Likelihood ratio tests were made and checked for convergence as sample size increased.

### *Methods for Question 2 (What is the Expected Rate of at Least One False Positive in a Single Study?)*

The number of replicates with at least one false positive among the five tests for nonlinearity were simply counted for the 5,000 replicates of study data. This accounted for any correlations between the five tests.

### *Binomial Calculations for a Single Study*

With the standard choice of 95% confidence bands, false positives for a single test of nonlinearity will happen in ~5% of the simulations, assuming required assumptions about the data and modeling are met. The probability, then, that at least one false positive will show up when $k$ independent tests for nonlinearity are made in a single study can be obtained by a binomial calculation and compared to simulation rates. With $k$ set to 5, the number of tests of nonlinearity available in the A-bomb studies that we considered, the result is 23% assuming independence of the tests.

$$1 - 0.95^5 = 0.23. \tag{3}$$

### *Methods for Question 3 (What is the Expected Rate of at Least one False Positive in Multiple Studies?)*

The generalization of Eq. (3) to multiple studies is given in Eq. (4), where $f_j$ is the aggregate false positive rate in the $j$th of $n$ studies.

$$F_m = 1 - (1 - f_1)(1 - f_2) \ldots (1 - f_n). \tag{4}$$

For the case of five independent statistical tests of nonlinearity in 6 studies, where the rate in each study is 23%, Eq. (4) predicts a 79% chance that at least one false positive would show up in 6 independent studies. This triplet of numbers, 5%, 23%, and 79% is to be compared with modified values discussed later, when the assumption of independent, perfect tests is relaxed. There are two cases to consider. In the first case it is assumed that all five tests of nonlinearity are available for all 6 studies, which is the assumption that is made to generate the numbers in most of the tables and graphs to follow. The total number of tests considered across all studies, then, is 30. In the second case, which is used to compare simulation results with actual study results, the number of test results available in each specific study was used to determine the reduced $f$ values to insert in Eq. (4). In both cases, the $f$ values were determined from simulations, which automatically accounts for correlation between the different statistical tests of nonlinearity.

To improve the match with actual results in the second case, the dose range considered was increased for the A-bomb datasets to 0–2

Gy, and weighted linear regression of absolute data was used to preserve linearity in this dose range. Poisson regression of absolute data was used for the other datasets.

A study-by-study listing of the primary test results available in the 6 studies, which total to 20, is given in Supplementary table S-5 (https://doi.org/10.1667/RADE-21-00219.1.S2). A primary test means a test performed on all data, without any subgrouping. To fully compare simulation results to A-bomb findings, which included a positive finding for nonlinearity in dose response for males in the 2017 A-bomb study, it was necessary to go beyond primary tests. To this end, linearization was carried out for the 2017 A-bomb dataset separately by sex.

To put the curvature result for males in perspective, a Bonferroni correction of $40 = 2 \times 20$ tests was used for comparing with simulation results. The factor of 2 comes from the assumption that all of the six study groups would have looked at their test results by sex and would only have reported a result by sex, if it had a very low P value. Thus, we inferred that there might have been tests conditional on sex that did not reach 95% confidence, which led us to double the Bonferroni adjustment to account for this conservative possibility.

### *AIC Selection of Best Model*

One way to avoid choosing a privileged null, as null hypothesis testing does, is to rank models by the highest statistical likelihood found in fits to the data. However, with this method, adding extra parameters to a model will always lead to smaller model residuals. To compensate, a model's fitted likelihood can be penalized by a term related to the number of parameters. A common choice for the likelihood penalty in radiation epidemiology (*23, 46–49*), which we adopt here, is based on the AIC criterion (*34, 50*), which in turn has roots in information theory (*34*). Researchers have other choices, including ''corrected'' AIC, ''consistent'' AIC, and Bayesian Information criterion (*50*), which may change results depending on sample size (*50*).

To apply the standard AIC method, unity was subtracted from the statistical loglikelihood determined for each fitted parameter. The model generating the highest penalized likelihood (lowest AIC score) was then chosen in each replication of the 5,000 datasets. If AIC selection picked any of the nonlinear dose response functions over the linear default model, a false positive was declared and counted. Although a privileged null is avoided with the AIC penalty method, there is an implicit linear null for nested models like threshold, quadraticity, curvilinearity, and 2-slope spline. In such cases, the AIC criteria produces the same results as would selecting a model using a null-hypothesis test, but with a higher cutoff P value than 0.05, namely 0.157 (*50*).

### *Multi-Model Averaging*

Multi-model averaging (*51*) is an example of multi-model inference. The averages were obtained by weighting the five dose-response functions with normalized AIC weights (*52*). The weights were based on the negative exponentiation of half the AIC values (*51, 52*), which is algebraically equivalent to taking weights proportional to the likelihood determined by software, reduced by a factor of $e^{-p}$, where $p$ is the number of fitted model parameters. The standard error of the multi-model fit at each point on the fitted curve was based on the square root of the weighed sum of the squares of the standard errors produced by fitting software for each model. Multiplication by 1.96 produced the confidence bands.

The final dose-response curve is analogous to a smoothed curve and conceivably might be less likely to produce false positives. However, the choice of a method to compare a multi-model curve to the linear curve used in regulation is not obvious. The approach taken here for illustrative purposes is to require at least a doubling or halving of the crude slope (ERR/Gy) at low doses, while at the same time requiring the 95% confidence band for the fitted curve to exclude the linear fit. Without both of those conditions being met, we presume that regulators would have minimal interest in considering any modifications to regulatory policy.

Some authors narrow down the number of nested models before including them in the mix of models analyzed with AIC weights (*52*). In our case, we only have nested models to begin with, so we include them all in the model average. Note that some authors pick replicate risk medians of multi-models rather than averages (*52*). There are, thus, different versions of multi-model inference, which means that our results will not necessarily generalize.

### *Methods for Question 4 (What is the Expected Frequency of Above-Zero Dose Thresholds?)*

*Toy Threshold-Dose-Response Model.* To aid in understanding frequency results for threshold doses above zero, a toy threshold-dose-response function, with a threshold dose between the first two datapoints, was introduced for use with absolute risk data. The model can be analyzed graphically, without the need for algebraic fitting of data (Supplementary text S-4 and the figure embedded therein; https://doi.org/10.1667/RADE-21-00219.1.S2).

### *Fitting Simulated Individualized Survival Data*

The Cox proportional hazard model was used to first create individualized linearized datasets using a hazard function linear in dose to predict cancer mortality history. In each replicate, randomization was introduced into the timing of death for each simulated individual. Next, the Cox model was used again, this time to fit the previously created data with nonlinear hazard functions. If a parameter for nonlinearity in a hazard function were declared a positive finding by the software, it would have to be a false positive, because the original dataset was built with a linear function. An embedded flowchart found in Supplementary text S-1 (https://doi.org/10.1667/RADE-21-00219.1.S2) shows the process.

Two nonlinear hazard functions were used, dose threshold and 2-slope-spline. Hypothetical dose breakpoints in the modeling were chosen at each of the 17,500 individually assigned doses, not between them. The ''Surv'' function from the ''survival'' package in the R-statistical language (*31*) was used to provide a response variable for use in fitting with the R-function, ''coxph.'' For loess fitting, survival time of cancer deaths was the outcome fitted to dose. To directly compare individualized results with the standard cohort approach, the individualized results were grouped into dose categories and analyzed with Poisson regression.

Except for loess fitting of survival times, it was not feasible to run 5,000 replications of the Cox hazard analyses, as was done in the main analyses in this paper. We found 200 replications were sufficient to compare the individualized results to those obtained for the grouped data. A similar approach was taken with the 2-slope spline model, but the number of persons analyzed was reduced to 1,750.

## RESULTS

This section includes the following subsections: Results related to questions 1–4; results related to statistical modeling; sensitivity analysis.

### *Results Related to Question 1 (Are there High False Positive Rates for Single Tests?)*

False positives rates over 10% occurred for tests of individual dose response functions such as curvature in some nonstandard situations. Thus, there were analysis methods in radiation epidemiology that could overstate

**TABLE 2**
**False Positive Rates for Nonlinearity, Alone and in Aggregate, for the LSS 2017 A-Bomb Linearized Dataset for Two Dose Ranges, 0 to 0.6 Gy and 0 to 2 Gy (Absolute Risk Data for 5,000 Replications with 95% Confidence Intervals)[a,b]**

| | 0 to 0.6 Gy | 0 to 2 Gy |
|---|---|---|
| Linearized dataset: | Poisson regression[c] | Weighted linear regression[c] |
| False positive for | | |
| Threshold breakpoint | 0.065 (0.058, 0.072) | 0.056 (0.050, 0.062) |
| Quadratic term | 0.052 (0.046, 0.058) | 0.054 (0.048, 0.060) |
| Curvature | 0.053 (0.047, 0.059) | 0.076 (0.069, 0.083) |
| Supralinear (loess shape) | 0.070 (0.063, 0.077) | 0.058 (0.052, 0.064) |
| Sublinear (loess shape) | 0.078 (0.071, 0.085) | 0.063 (0.056, 0.070) |
| At least 1 false positive among the above 5 indicators of nonlinearity in a replication | 0.19 (0.18, 0.20) | 0.19 (0.18, 0.20) |

[a] The numbers shown in the first 5 rows should be compared with the ideal of 0.05. Numbers shown in the last row are aggregates for a study and should be compared to 0.23, which is the expectation for five independent tests each with a false positive rate of 0.05.

[b] 95% CI based on Le, Diop, Al-Emadi, 2013 (https://doi.org/10.29115/SP-2013-0006). The maximum dose considered for threshold and shape determinations was taken to be 0.25 Gy for the dose range, 0 to 0.6 Gy and 0.5 when considering the dose range 0–2 Gy.

[c] Poisson regression does not introduce any nonlinearities of concern in dose response when doses are restricted to <0.6 Gy. Weighted linear regression does not introduce nonlinearities in dose response for any dose range.

evidence for nonlinearity in simulations, but this happened only in two situations. For instance, false positive rates from 13 to 40% occurred across datasets for dose threshold if some standard optimization methods were used to find maximum likelihoods. The excess could be eliminated by using software capable of analyzing data with multiple peaks and discontinuous derivatives in parameter likelihood.

High rates of false positives were also found for some dose-response functions when datasets were analyzed that had a small reference count, even as high as 150. For example, false positives for curvature of relative risk data analyzed with Poisson regression were 66% for Techa River incidence and 28% for Techa River mortality.

Other dose-response functions that gave elevated false positive rates for Techa River were the ‘‘5 point test’’ (5 consecutive points on one side of the linear line). It showed 13% for sublinear and 11% for supralinear when analysis was carried out on relative risk incidence data for Techa River. Even absolute risk data could show elevated rates for Techa River datasets. For example, weighted linear regression of absolute risk data produced false error rates in excess of 10% for the Techa River when it came to threshold (11%), 2-slope spline (16% sublinear, 20% supralinear), and hormesis with the reference being the first data point (38%). The best performing case was Poisson regression of absolute risk data, which had no rates above 10% except for 1 of 3 definitions of hormesis.

A sensitivity analysis demonstrated that it was indeed the low counts, 17 for Techa River incidence data and 150 for the mortality data, that caused the high false positive rates. When the Techa River reference count numbers were artificially increased to match counts in the next highest dose category, 630 and 955 for the two datasets, respectively, all the false positive rates were below 10–11%. Note that in this sensitivity analysis, the reference level person-years were also scaled to keep reference risk values unchanged. Similar problems did not appear in the other four studies analyzed, because the lowest count number in their reference categories was 4,600.

For completeness, results above 10% for all false positives for nonlinearity are listed in Supplementary table S-6 (https://doi.org/10.1667/RADE-21-00219.1.S2). All results, including those below 10%, can be found in Supplementary tables S-7a– S-7d. Count numbers for each dose category in each of the six studies can be found in Supplementary tables S-1 and S-2. Four examples of loess fitting to the linearized 2017 A-bomb data are presented in Supplementary fig. S-5. Additional examples of false dose response shapes found with loess smoothing, 50 in total, are available in Supplementary file 1 (https://doi.org/10.1667/RADE- 21-00219.1.S1).

Although almost all of the results for the A-bomb datasets in this article were calculated for a restricted mid-dose region of 0 to 0.6 Gy, it was of interest to assess a higher dose range for comparison. Table 2 gives such a comparison, showing rates when the dose range is extended from 0.6 Gy to 2 Gy. In the extended dose range case, simple Poisson regression will introduce its own nonlinearities, so weighted linear regression was used in the simulations. The resulting false positive rates for individual tests of nonlinearity as shown in the first five rows of Table 2 were similar for both dose ranges. All values were less than 10%.

Because the true dose response is not expected to be perfectly linear, without any higher order terms (*2*), not all nonlinearities would be of regulatory interest. As stated earlier, a curvature of regulatory interest was defined for this paper to be one with magnitude greater than 1, which means that risks at very low doses will be around half the value predicted by the LNT or less. Although this restriction had modest impacts on the frequency of false positives for curvature in the INWORKS, Techa River, and A-bomb datasets restricted to ~0.6 Gy, there was a large reduction in frequency in simulations when the A-bomb upper dose limit for regression was raised to 2 Gy. The frequencies found for

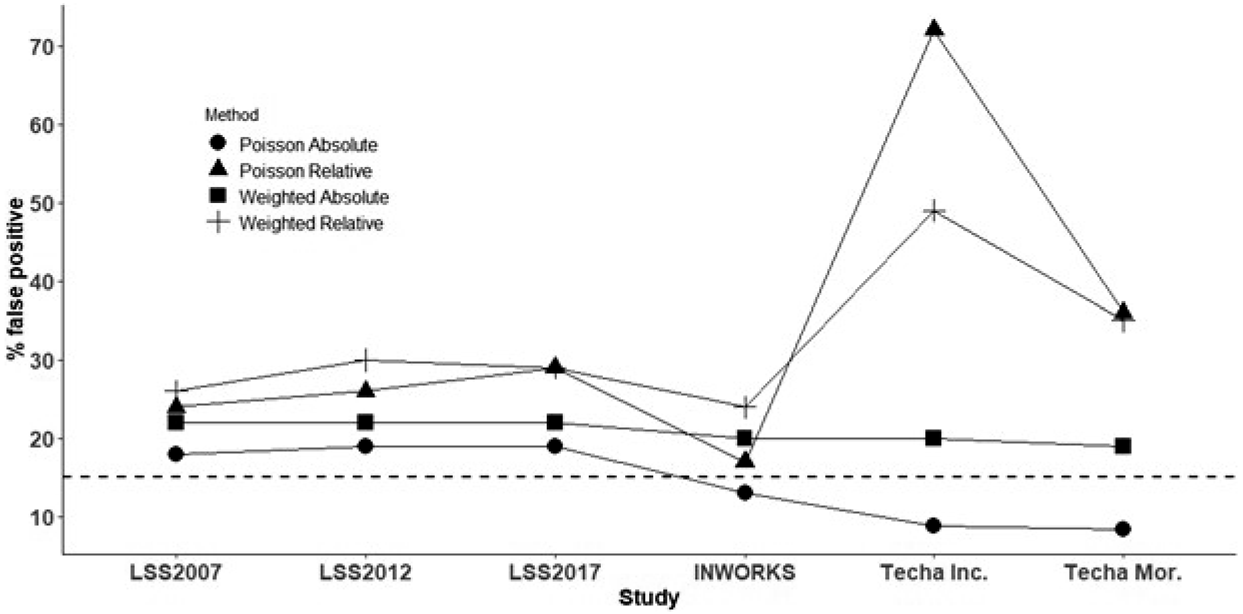

**FIG. 2.** The percentage of times that at least 1 false positive showed up in a single dataset, given five tests for nonlinearity per study. Regression sample sizes are highest at the left, lowest on the right. The dashed line at 15% is the average over the six datasets for Poisson regression of absolute risk data. ''Weighted'' stands for weighted linear regression. The high values for Techa River relative risk data are caused by low count rates in the reference category, which was used as the divisor.

curvature magnitude exceeding unity were 0.16%, 0.3%, and 0.1% for the 2007, 2012, and 2017 datasets, respectively. The rate was also 0.1% for the 2017 dataset, when data for males were analyzed separately. Results for the 0–2 Gy range turn out to be important later, when considering the likelihood of positives for curvature reported in the 2017 A-bomb study.

*Results for Individualizing Risk*

As was largely the case for Poisson regression of absolute risk data, individualizing risks for the Techa River mortality dataset did not produce any false positive findings above 10% for the two indicators of nonlinearity tested, dose threshold and 2-slope spline.

*Linearization Graph*

Figure 1 shows a comparison plot of the (standardized) excess relative risks before and after linearizing the 2017 A-bomb incidence study. It can be seen that there is no nonlinearity left in the linearized curve. The ''before'' data (dashes) in Fig. 1 were standardized to age 70 and age at exposure of 30 years by adjusting person-years to better match published data. Doing so did not change the linearized data beyond a scaling. The ''after'' data (solid triangles) were obtained by scaling person-time for each data point so that the revised risk lay exactly on a linearized line. No change was made in cancer count data. What is left of the original study are the number and positions of the dose categories and the corresponding counts and count variances, which are used in regressions. The slope of the linearized curve was taken to match the published version computed over the entire dose range. It is slightly lower than the slope obtained by fitting the dose range used (0 to 0.6 Gy). For plotting purposes, the risks were normalized to the lowest fitted risk value.

*Results Related to Question 2 (What is the Expected Rate of at Least One False Positive in a Single Study?)*

False positives for nonlinearity from null-hypothesis testing were aggregated to obtain a net rate per study replicate, given five tests per study. Figure 2 shows, for each of the 6 datasets considered, the aggregate false-positive rate per study obtained in simulations for the different methods of regression and data treatment. The graph dramatizes the poor performance of relativized risk data when the counts in the reference level are small (Techa River). When the count rates were artificially increased in the sensitivity exercise mentioned earlier, the high aggregate rates for relativized Techa River data dropped to levels comparable to the results for the other datasets shown in Fig. 2 (results not shown).

Figure 2. also shows the better performance obtained with Poisson regression of absolute risk data, even for the Techa River datasets with their low count rate in the reference level. In fact, the Techa River datasets performed better than the A-bomb datasets for Poisson regression of absolute data. Why are they better than the A-bomb datasets in this case? The false positive rates for shapes after loess fitting to data made the difference. The Techa River rates were only 30% of the corresponding rates for the A-bomb studies.

The percentage of studies with at least one false positive averaged over all cases in Fig. 2 was ~25%, which is higher than the 23% predicted by Eq. (3) for independent statistical tests. How can non-independent tests combine to reach and exceed 23%? This happens because in the higher cases, the individual false positive rates are above 5%, compensating for test overlap. Results for single study, aggregate false positive rates that incorporated shape methods other than the default loess fitting, specifically 2-slope spline fits or the five data-point criterion, can be found in Supplementary tables S-8a through S-8d (https://doi.org/10.1667/RADE-21-00219.1.S2).

Even without a low reference count, relativized risk data can be a particularly unreliable choice for Poisson regression, because it is a *ratio* of Poisson counts. The variance in the relative risks is not Poisson and is underestimated in Poisson regression of this type of data, thereby increasing the frequency of false positives.

The average over the six studies for Poisson regression of absolute risk data was 15%. It is shown as the dashed line in Fig. 2. The value of 15% per study is lower than the 23% frequency predicted in Eq. (3) for independent tests with an individual rate of 5%, implying that the tests for nonlinearity are not independent.

AIC selection did not eliminate the multiple comparisons problem associated with null-hypothesis testing of multiple types of dose response function. This was evident with all six datasets, including those whose results are presented in Table 3, specifically, the 2007 and 2017 A-bomb, INWORKS, and Techa River mortality datasets. As shown in Table 3, a nonlinear model was falsely selected as superior to the underlying linear model 24 to 29% of the time, using Poisson regression of absolute data, compared to 15% of the time that null-hypothesis testing falsely selected a positive for nonlinearity averaged over the 6 datasets for the same underlying regression method and data treatment (Fig. 2). Using weighted linear regression of absolute risk data, the corresponding numbers for AIC selection were 35% to 61%, compared to 21% for null-hypothesis testing averaged over the 6 datasets. High frequencies of aggregated false positives were not a surprise given the high effective false-positive cutoff rate ($alpha = 0.157$) implicit in AIC model selection of nested models. These high frequencies could be reduced using versions of AIC with stronger likelihood penalties (*50*), such as ''corrected'' or ''consistent'' AIC (results not shown). AIC selection avoids the need to explicitly pick a privileged null hypothesis, but it is implicitly a multiple comparison and may make false selections.

Multi-model averaging of nested models with AIC weights did not decrease the frequency of false indications of nonlinearity compared with null-hypothesis testing in most cases, using our two-fold requirement for a false positive. When only the first requirement was met, namely that the confidence band for the average shape excluded the linear line at some dose below 0.25 Gy, the frequency of false positives found for multi-model averaging were similar to the results for AIC selection (results not shown). When the requirement of a major deviation at low doses was added, the fraction of time that a multi-model fit qualified in simulations as a false positive under our definition ranged from 10 to 25% for Poisson regression of absolute data (Table 3). The average per-study result of 17.5% was slightly higher than the corresponding average result of 15% for null-hypothesis testing. The two numbers are not directly comparable, because the nonlinear shape functions incorporated were different, specifically loess fitting for null-hypothesis testing and 2-slope spline for multi-model fitting. Loess fitting tends to pick up more false positives. Regardless, given the uncertainties in definitions and calculations, the conclusion is that both techniques, null-hypothesis testing and multi-model averaging can produce isolated false positives for nonlinearity at comparable rates.

Dataset by dataset numbers for both types of shape analysis, loess and 2-slope spline, are archived in Supplementary tables S8a–S8d (https://doi.org/10.1667/RADE-21-00219.1.S2) for null hypothesis testing that can be compared to the results for AIC selection and multi-model fitting given in Table 3.

### *Results Related to Question 3 (What is the Expected Rate of at Least One False Positive in Multiple Studies?)*

The expected rate of at least one false positive occurring in multiple studies was obtained by substituting into Eq. (4) the appropriate single-study, aggregate false positive rates ($f_j$). In the first case considered, it was assumed that all five tests for nonlinearity were available for all six studies, leading to a predicted rate of 61% for Poisson regression of absolute data with dose restricted to a maximum of ~0.6 Gy. For comparison with actual results in published studies, it was necessary to use $f_j$ based on the actual test results, which were less than five in some studies. In this second case, the chances of at least one false positive occurring in the studies was 50%. Based on this 50% figure, there would be a good chance that the single positive for nonlinearity found for primary tests in the actual studies (*8*) could be a false positive, at least if results for subgroup data are ignored. The more general case of subgroup findings is considered in the discussion section.

The simulation-derived, per-study rates ($f$ values) that were inserted into Eq. (4) for the first case equaled 0.18, 0.19, 0.19, 0.13, 0.088, and 0.084. The listed sequence begins with the rates for the three A-bomb studies in chronological order, followed by INWORKS and the two Techa River studies, incidence and mortality. The f-values in the second case were, 0.19, 0.11, 0.19, 0.0, 0.085, and 0.06.

### *Results Related to Question 4 (What is the Expected Frequency of Above-Zero Dose Thresholds?)*

An above-zero threshold falsely occurred more than 50% of the time in the simulations for most of the datasets, due to a bias in the threshold dose, which to our knowledge has not

**TABLE 3**
**Fraction of Time in 5,000 Replications that False Deviations from Linearity were Found in Linearized Datasets using Aikake Information Criterion (AIC)**

| | Criteria | |
|---|---|---|
| | AIC model selection | Multi-model averaging |
| Regression method and dataset (absolute risk data) | At least 1 of 5 nonlinear models outperformed the linear model using AIC ranking criterion[a] | The 95%-confidence band for the AIC-weighted, multi-model fit excluded the linear null at some dose and also deviated substantially from linearity anywhere below 0.25 Gy[b,c] |
| Poisson Regression | | |
| A-bomb LSS 2007 | 0.24 | 0.10 (0.014) |
| A-bomb LSS 2017 | 0.29 | 0.11 (0.017) |
| INWORKS | 0.25 | 0.24 (0.053) |
| Techa River mortality | 0.26 | 0.25 (0.11) |
| Weighted linear regression | | |
| A-bomb LSS 2007 | 0.35 | 0.21 (0.030) |
| A-bomb LSS 2017 | 0.41 | 0.21 (0.056) |
| INWORKS | 0.44 | 0.33 (0.12) |
| Techa River mortality | 0.61 | 0.41 (0.19) |

[a] Rankings based on minimum AIC value, which includes a likelihood penalty for number of parameters. The AIC rankings are equivalent for the nested models considered here to null-hypothesis testing with an *alpha* of 0.157, not 0.05. Nonlinear models included were dose threshold, linear-quadratic, curvature, and 2-slope spline. The largest nonlinear contributor to frequency was usually the 2-slope spline model, sometimes with a value as much as, or more than, the sum of all the others.

[b] Hormesis fraction given in parenthesis. Substantial deviation was defined to require the crude slope (ERR/Gy) to deviate by a factor of 2 up or down from the ERR/Gy at higher doses, either at 1 Gy or at the highest dose in the dataset. A factor of 2 deviation was chosen with the idea that it would be large enough to trigger regulatory interest. Without this criterion, the values were similar to those in the left column for AIC selection. The number of replications for Poisson regression was 1,000 for multi-model averaging.

[c] AIC weights for each model are the relative statistical likelihood obtained from fitting software multiplied by a penalty term, $e^{-p}$, where $p$ is the number of fitted model parameters.

previously been discussed. Figure 3a shows the distribution of dose threshold values found in 5,000 simulations of the linearized 2017-A-bomb dataset determined using Poisson regression methods of absolute risk data below 0.6 Gy.

Experimental error or statistical variation at a true threshold dose of zero must be zero or positive, but the high frequency of skewing was a surprise before the data were plotted. Percentages ranged from 46% to 69%, depending on the regression method and data treatment, when averaged over the 6 datasets (Supplementary table S-9; https://doi.org/10.1667/RADE-21-00219.1.S2). The percentages were also high (73%), when data were individualized.

Because the datasets had been linearized, all of the above-zero threshold dose values were false, caused by statistical noise. No threshold dose peaking occurred, as expected, in the 200 replications of individualized data simulated for the Techa River mortality study, because no dose categorization was involved.

The peaks for dose threshold in Fig. 3a match the dose values assigned to each dose category. About 2/3 of the threshold dose values above zero are contained in the peaks, including the peak very close to zero dose. Data points in Fig. 3a that qualified as false positives did not occur until doses exceeded the 4th peak at 0.029 Gy.

There were four threshold dose values reported among the 6 published studies (*7, 8, 21, 23*), whose average value agrees with the simulation results, 0.043 Gy vs. 0.045 Gy, respectively. These findings are consistent with the published above-zero threshold doses being caused by statistical bias in the threshold dose parameter.

The quadratic parameter whose distribution is shown in Fig. 3b also has a high percentage of false non-zero values, in fact nearly 100%, but the values are equally likely to be negative as positive. Therefore, there is no potential problem for meta-analysis across studies.

Only 1.7% of the above-zero thresholds in Fig. 3a rose to the level of a false positive, so there is no increase in false positives above the expected value of 5%. However, without knowledge of the high expectation rate for dose thresholds with high P values, a reader might see an apparent inconsistency between the reported high P values and the appearance of above-zero dose thresholds in 3 out of 4 cohort studies that tested for threshold nonlinearity (*7, 8, 21, 23*).

Requiring an AIC test to label a threshold model fit as an improvement was another way to discount most of the above-zero dose thresholds found in the simulations. Only 6.3% of dose threshold values in the upper panel were associated with improvements in fit sufficient to satisfy the Aikake information criterion. Across regression methods and data treatment, the percentage of dose threshold values meeting an AIC criterion varied from 6.3 to 23%. There were three above-zero dose thresholds with P values reported in the 6 studies considered, all of which failed to qualify under AIC selection as an improvement in fit. Their

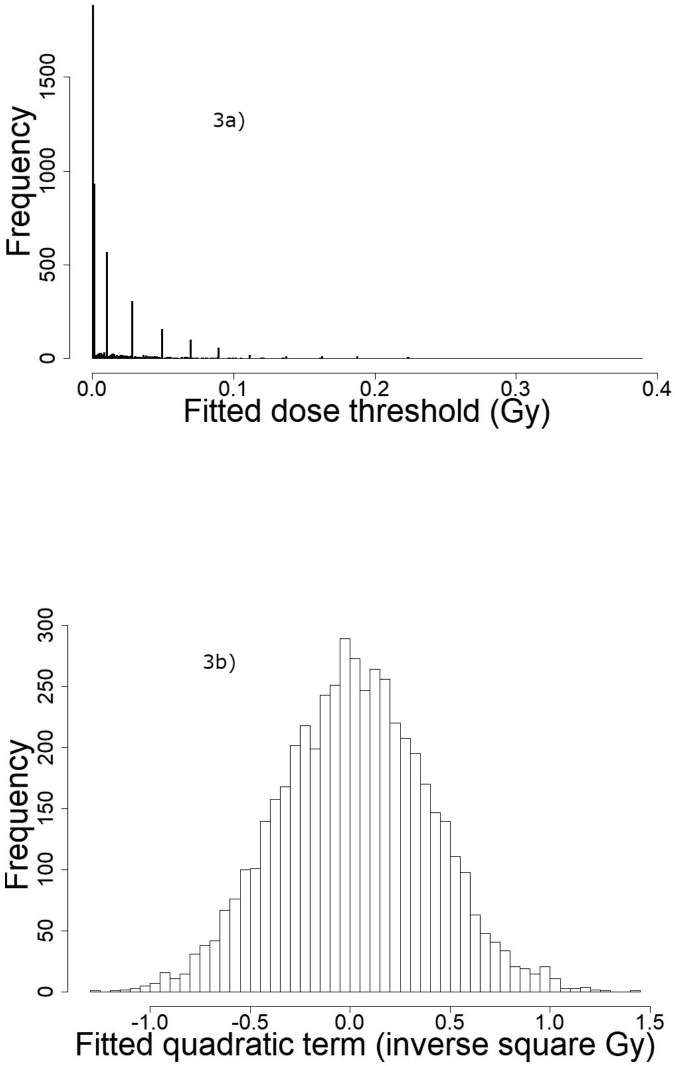

**FIG. 3.** Histograms of parameter values fitted to 5000 dataset replicates. Threshold dose values are shown in upper panel (3a). Values for the quadratic term in the LQ model are shown in lower panel (3b). All non-zero values are false, failing to capture the underlying linearized dose response. Values were obtained from Poisson regression of absolute risk data for the linearized 2017 A-bomb dataset below 0.6 Gy. The peaks in the upper panel for dose threshold match the dose values assigned to each dose category. About 2/3 of the dose values above zero are contained in the peaks, including the peak very close to zero dose.

P values were too high, lying above the 0.157 boundary that, for nested models, determines whether a fit is an improvement under the AIC criterion (*50*).

### *Results Related to Statistical Modeling*

*2-Slope Spline and Threshold.* The 2-slope spline function does not have the standard asymptotic properties (*53, 54*) on which standard software may rely to quantify error rates of type I, and convergence of a likelihood test with regression sample size is slow (*55*). The profile likelihood curve was usually discontinuous at 0 breakpoint and always constant thereafter within the first and last regions defined by dose category boundaries. Discontinuous jumps in profile likelihood derivatives, as well as multiple peaks, could appear in other profiled regions.

Although the profile likelihood for the threshold model was continuous at zero breakpoint, and the likelihood was only constant for the last dose region, it too could have discontinuous profile likelihood derivatives and multiple peaks within its boundaries. Note that a two-peak likelihood surface for threshold dose, described as unusual, was reported in the 2017 A-bomb study (*8*). Example profile likelihood curves for 2-slope spline and dose threshold are shown in Supplementary figs. S-6 and S-7 (https://doi.org/10.1667/RADE-21-00219.1.S2).

This nonstandard profile likelihood behavior generally prevented off-the-shelf software from finding the proper parameter values and critical points for threshold dose response and 2-slope splines. The software would not always find the maximum of the maxima within segments, which is why segmented searches were made between dose category values.

Although the 2-slope spline *breakpoint* showed the same peaking structure above zero breakpoint as found in the threshold model, the *lower slope* was not constrained in any way. It split in simulations evenly between positive and negative values (results not shown), showing no preference for either sublinear or supralinear dose response. Thus, false findings across studies would also split between sublinear and supralinear shapes and would not present a problem for meta-analysis.

The frequency of false positives for dose threshold, sublinearity and supralinearity with 2-slope spline were <10% for Poisson regression of absolute risk data, but accurate results required careful optimization of the parameter searches. Calculated rates were high if a naïve search for optimum dose threshold value were made using derivative methods for obtaining confidence intervals (Supplementary table S-7b; https://doi.org/10.1667/RADE-21-00219.1.S2). As expected, differences in false positive rates for 2-slope spline fits between regression methods decreased as regression sample sizes were expanded (Supplementary table S-10 and Supplementary figs. S-8 and S-9; https://doi.org/10.1667/RADE-21-00219.1.S2).

The false positive rates did not converge to 5%. This was initially surprising to us but was consistent with the literature in cases when the null hypothesis (zero breakpoint) lies on the boundary of a constrained parameter space (*53, 54*). Quite a bit is known in such situations, ''... but knowledge is scattered across the literature and considerably less well known among practitioners'' (*56*). A mixture distribution of chi-square statistics is expected that can depend on the configuration of the datapoints (*53*). As a result, the standard numerical values for setting critical points for false positives are not valid in these special cases. This fine point made little practical difference in the simulations: false positive rates for threshold did not increase above 10% for standard Poisson regression and barely increased over 10% for the case of weighted linear regression of absolute data (11% for threshold dose in the Techa River incidence case).

### *Curvature*

Dose response functions with breakpoints were not the only examples of unusual profile likelihood curves. Curvature had a profile likelihood shape that almost always contained both a valley and a peak, an example of which is shown in Supplementary fig. S-10 (https://doi.org/10.1667/RADE-21-00219.1.S2).

### *Individualizing Data*

As for loess shape fitting to individualized survival data, it was possible to carry out the full 5,000 replications. Sublinear and supralinear false positive rates were 10% and 9.3%, respectively. This is 4 to 5 times higher than the results for loess fits to grouped cancer risks at Techa River. However, such a comparison is not between equivalents. One endpoint is survival time, and the other endpoint is accumulated mortality counts.

What was most notable about the loess fitting to individualized survival times for the Techa River mortality dataset was the low frequency of hormesis, with hormesis reference level equal to the average survival time at zero dose. The value of 0.6 per thousand was 300 times smaller than the corresponding hormesis frequency of 18% accumulated using loess fitting to grouped risk data. This finding is not surprising in light of the recommendation to use a loess algorithm only when the number of datapoints is large (*36*), which was not the case with the analysis of grouped counts.

Graphs corresponding to fits to a dose threshold model for a single replication of individualized Techa River mortality data are shown in Supplementary fig. S-11 (https://doi.org/10.1667/RADE-21-00219.1.S2). In the individualized 2-slope spline case (not shown), bumps and oscillations in the profile likelihood for breakpoint were more common than in the corresponding dose threshold plots. Supplementary table S-11 shows that false positive rates found when simulated individuals were analyzed with a Cox model were similar to the false positive rates obtained when groups of these individuals were analyzed as Poisson counts.

### *Sensitivity Analyses*

Sensitivity analyses included varying the dose response slope used in linearization and the smoothing parameter in loess fitting. None of these variations led to any changes in conclusions. Nor did variations in A-bomb parameters, specifically, the choice of reference level, the upper dose range, and the number of persons considered. Details of sensitivity analyses, including those mentioned earlier in the text, can be found in Supplementary text S-5 (https://doi.org/10.1667/RADE-21-00219.1.S2).

## DISCUSSION

This section includes the following subsections: Regulatory default, Threshold dose bias, Hormesis, Multiple comparisons, Simulation frequency results compared to study results, Unknown error rates of type I, Linearization in case/control studies, and Methodological limitations.

### *Regulatory Default*

The choice of a linear dose response function for cancer causation as the null model is a policy choice by regulators, which in the past has coincided with long-standing recommendations by almost all national and international regulatory and scientific organizations (*2–4*). We have assumed in this paper that when considering changes to the status quo, regulators of radiation exposure would continue to give the LNT privileged status as the default dose response model. Were a threshold model to be the regulatory norm for a substance, it would have the privileged status and it would be the linear model that would be expected to meet a ‘‘clearly superior' ranking before adoption (*57*). False positives would still be an issue, but it would be false positives for linearity or other nonconforming responses that would need to be considered.

From the scientific perspective, the linear model is simple and transparent, satisfying Occam's razor in adding only one free parameter to the constant term. It and the pure quadratic function have the least number of parameters of standard models, which makes them experimentally attractive as the models to use first when count numbers are limited.

The pure quadratic model is rarely considered in current radiation cancer epidemiology. Pure linearity, however, is the usual starting point, as indicated by its use as the default in all the epidemiologic studies considered for this paper. Yet, there is potential merit in analyzing dose response in ways that do not privilege the linear LNT model. As we have shown in this paper (Table 3), these methods can also exhibit substantial false indications of nonlinearity within a study, which may be of interest to regulators.

Multi-model fitting, like other data smoothing techniques, has an interesting potential advantage in regulatory debates about dose response. Every multi-model fit is likely to be nonlinear to some degree, which could change the conversation from, ‘‘Is the fit nonlinear?'' to ‘‘How much nonlinearity is there, and does it matter?'' Although not explored here, one complexity of multi-model averaging made evident by the results of this paper is that the threshold dose bias will be averaged into the multi-model shape to some extent, if a threshold dose response is one of the models included in the multi-model average.

### *Threshold Dose Bias*

The simulation threshold results dramatize the fact that a bias exists for threshold dose. Intuitively, after seeing the simulation results in Fig. 3a, one might guess that half a bell curve would fall above zero dose and the other half would be clamped at zero, but an analytic proof is not obvious. Will more data and more dose categories reduce the numbers? The results from analysis of the toy threshold-dose-response function suggest not. Graphical analysis of the toy function indicates that, regardless of the magnitude of the spacing between dose categories, the appearance of above-zero dose thresholds in fits to data will be common. There exists a location for a threshold dose between the first two datapoints that will always reduce the residuals of the fit to the data, if two common conditions are met (Materials and Methods section and Supplementary text S-4; https://doi.org/10.1667/RADE-21-00219.1.S2). This predicted insensitivity to the distance between dose categories was supported by the high frequency of dose thresholds found in the simulation of 17,500 individuals in the Techa River mortality dataset, where the spacing between individual dose categories was very small.

### *Hormesis*

Hormesis is a biphasic dose response in which effects at low doses are opposite to those at high doses. In the present context, this suggests that low-dose radiation exposures are beneficial. Any model that predicts protective effects in a low-dose range would be difficult to fit into a regulatory regime. One reason is that there are concerns about the generalizability of hormetic responses (*58*). A major obstacle in regulatory application would be that the actual breakpoint dose for a hypothetical protective effect would be uncertain, and if one were identified, it could well vary among subgroups owing to genetic variation in susceptibility (*59*).

### *Multiple Comparisons*

The results for aggregated false positives per study (Tables 2 and 3) were restricted to primary tests of nonlinearity and thus underestimate the possible instances of multiple comparisons because they do not account for secondary tests. The 2017 A-bomb study by Grant et al. (*8*) assessed a large number of dose ranges, historical dosimetry schemes, and other sensitivity-test subgroupings. We counted 8 confidence intervals for quadraticity, 10 P values for curvature in the main text, and almost 100 P values for curvature in the appendices. These indicators are not independent. While it is scientifically necessary to analyze all such fits to data to avoid missing any nonlinearities (*60*), the presentation of so many results without any guidance about multiple comparisons might lead some readers astray.

As several thoughtful reviews and commentaries suggest, even the research community is not immune to misinterpretations of statistical significance (*35, 61, 62*). This may be an appropriate moment to pay attention in regulatory analysis to the complications of multiple comparisons and the large numbers of statistical inferences that are made, within and across studies.

Should adjustments be made for multiple comparisons? Adjusting significance to account for multiple comparisons may lead to an increase in the number of false negatives, which could lead to important research findings being overlooked (*60*). From a decision-maker's perspective, however, adjusting for multiple comparisons makes statistical sense, but not necessarily with simple approaches such as a Bonferroni correction. It would be better to consider the costs of the false positives and false negatives (*63, 64*), which might be measured in social, health, and/or monetary units.

A Bonferroni correction may still be useful as a sensitivity calculation, because it gives an extreme, where the cost of a false negative is assumed negligible compared to the cost of a false positive. When a great number of tests are carried out, a false discovery rate calculation will be less conservative than a Bonferroni correction (*65*), but may still be incomplete from a decision-support perspective without some idea of cost functions for false positives and negatives. Overall, given the debate over adjustment and its complexity, there is no simple recommendation that can be made.

### *Simulation Frequency Results Compared to Study Results*

As for false positives, there was only one study among the six that reported a positive for its primary analysis using all of the data without subgrouping. That study result was a curvature test in the 2017 A-bomb study with a P value of 0.03. To compare the result to simulation results is straightforward for this primary analysis, as discussed in the Results section, which predicted a 50% chance of finding a false positive among the 20 tests for nonlinearity in the six studies. Thus, both a true and false positive are compatible with the A-bomb primary finding, making the finding uninformative.

However, there is additional evidence to consider beyond the crossing of the bright-line P value of 5%. In the case of the curvature finding for males in the 2017 A-bomb study, the magnitude of the curvature was 1.3 and the P value was 0.002. Even multiplying 0.002 by a conservative Bonferroni correction of 40 (Materials and Methods) would leave the chances of a false finding to be low at 0.08, which argues against the curvature finding for males being explainable by multiple comparisons, although still possible. In simulations, a false positive for curvature for males equal to or greater than 1.3 occurred only with a frequency of 0.001, consistent with the low P value of 0.002.

There are findings in other A-bomb studies that also might be considered by a review committee. For instance, the 2012 A-bomb study reported a positive for upward curvature (P = 0.02), not as a primary test, but as a test within a 6-category subgroup consisting of a 2-category dose range for three time periods (their table 7). Whether or not quadraticity and dose threshold were also tested in these 6 subgroupings was not specified, and we did not try to simulate this situation. The 2000 A-bomb study (*6*), whose data we did not linearize, presented a smoothed loess fit that trends in the opposite direction for nonlinearity, appearing graphically to be a positive finding for supralinearity, not sublinearity. Uncertainty in the neutron component of dose may also be relevant to interpretations of the A-bomb curvature finding. Increasing the neutron RBE to the higher end of its uncertainty range made the curvature in the A-bomb data negative, as has been reported (*66*).

### *Unknown Error Rates of Type I*

It can be difficult to argue that one fitting method or data treatment is superior to the others. Poisson regression seems natural for radiation cohort studies, but errors other than count variations are assumed to be negligible. If an analyst felt that error types other than count fluctuations were important, such as errors in dosimetry and/or cancer mortality classification (*67*), it might be reasonable to prefer weighted linear regression, which accounts for residual errors using the data directly. On the other hand, it does seem from the results in this paper (Fig. 2) that a relative risk data model can produce very high error rates of Type I when both the number of data categories and the number of counts in the reference level are small. An alternative way of getting relative risks in such cases might be considered, for instance by reparametrizing the regression equations before making maximum likelihood estimates or by relativizing absolute risks after fitting, not before.

### *Linearization in Case/Control Studies*

The linearization method presented here for cohort studies is straightforward. For case-control studies, which do not have person-years to adjust, count data would have to be adjusted. A reasonable approach to linearization in these kinds of studies would be to adjust both case and control counts, so that at each dose category the standard deviation of the log odds ratio is kept constant upon linearization (Supplementary text S-6; https://doi.org/10.1667/RADE-21-00219.1.S2).

### *Methodological Limitations*

Because covariates such as age, sex, and city of bombing were not included in the regressions, the results strictly apply only to univariate regressions and could change somewhat with multivariate analysis. Thus, comparison with univariate study results would be the most direct use of the method described here. It has been assumed that the datasets are independent. Yet, the 2007 and 2017 A-bomb studies have overlapping cancer incidence data, although the methodology, including dosimetry and choice of covariates, has changed between publication dates, introducing some additional effective randomness.

In some studies, uncertainty in dose values has been taken into account in novel ways (*68*). Although variance in dose values could have been introduced when adding Poisson variation to linearized count data, we did not analyze this possibility. The methods considered here included Poisson regression, weighted linear regression, regression of relative risk data, AIC selection, and multi-model fitting. Other analysis methods that were not covered may be of interest to some readers (*32, 46, 47, 68, 69*), including one study that used simulation of data replicates to check confidence limits (*47*).

## SUMMARY AND CONCLUSIONS

### *False Positive Rates for a Single Test of Nonlinearity*

According to simulations of the six linearized cohort studies, there were two situations where false positive rates for null-hypothesis tests of nonlinear cancer dose response exceeded 10%. The first situation was software related. Some common regression software could not correctly analyze models like dose threshold that had discontinuous derivatives and could have multiple peaks for parameter values on the likelihood surface. The excess rates were limited to dose threshold and 2-slope spline models, but the excesses appeared in all datasets.

The second problematic situation was caused by unusually low counts in the reference categories of two datasets, specifically the Techa River incidence and mortality datasets. The excess rates only occurred if nonstandard regression methods or data treatment were used. High false positive rates did not appear for standard Poisson regression of absolute risk data, even with the low counts in the reference category, except in a single instance of the three definitions of hormesis assessed.

Linearization combined with simulation proved to be an informative diagnostic tool for identifying problematic methods for detecting nonlinearities in dose response. The analysis carried out in this paper has largely assumed a regulatory default that is linear without threshold, but the methods introduced here can be adapted to use other types of dose response as the default, should a different one be selected in the future by regulators.

### *Multiple Tests in a Single Study*

Even when individual tests of nonlinearity have low false positive rates, conducting multiple tests for nonlinearity in a study can present a multiple comparisons problem. As many as five tests per study for nonlinear dose-response functions can be identified in epidemiological studies of radiation and cancer (*7–9, 21, 23*). Consequently, the expected frequency of false positives in a study can be substantial, about 19% per study for the three A-bomb datasets according to simulations of linearized data, using Poisson regression of absolute risk data below 0.6 Gy. The value was 15%, when averaged over all six studies (Fig. 2).

The false positive rates quoted above were obtained for null-hypothesis testing. Considering approaches other than null-hypothesis testing, such as AIC model selection or multi-model fitting, did not solve the problem of false nonlinearities turning up in simulations (Table 3).

### *Multiple Tests in Multiple Studies*

When datasets were considered as a group, the chance of at least 1 false positive occurring in the 20 primary test results available in the 6 published studies was estimated to be 50%. Thus, the 1 positive found in the actual six studies using all the data (primary finding) with a P of 0.03 could easily be a false positive. In subgroup analysis by sex, there was strength of evidence to evaluate in the case of the finding of curvature for males in the 2017 A-bomb study. The magnitude of the curvature parameter was greater than 1, and the P value was 0.002. Binomial calculations combined with simulations and a Bonferroni correction suggested that it was unlikely that such a result was explainable by multiple comparisons.

### *Above-Zero Dose Threshold Findings*

In the six studies considered, four of them fitted their data to a dose threshold model. Of these, three reported above-zero dose threshold values averaging to about 0.05 Gy. Although none rose to the level of a positive, their high frequency was unexpected. Simulation provided a simple explanation. Analysis of linearized data showed that a high frequency was to be expected simply on the basis of random fluctuations in counts coupled with a statistical bias in the threshold dose parameter.

Without accounting for threshold dose bias or correcting for the extra model parameter, meta-analyses of above-zero dose thresholds, formal or informal, will be invalid. AIC methods provided a correction method by penalizing threshold fits for the extra parameter introduced. In simulations, this dramatically reduced the frequency of above-zero threshold fits that could be considered an improvement over the linear null. Furthermore, AIC methods could be applied to published P values and were sufficient to declare the published cases poor fits to the data, with no improvement over a linear fit.

### *Implications for Regulatory Analysis*

A single isolated result from a study is sometimes presented to policy makers and the public as ''statistically significant,'' and may therefore be judged by some as convincing evidence of nonlinearity at low dose levels and falsification of the linear regulatory default model. These two words, statistically significant, have developed a magic in the wider culture that can overwhelm nuanced conversation and affect public policy. However, without screening for false positivity, isolated findings should not be considered valid evidence against the regulatory default. Binomial calculations and simulation as part of regulatory analysis may help in judging if a result might be a false positive.

### *Implications for Radiation Research*

When using nonstandard and novel methods of analysis on cohort datasets or subsets of them, or when there is a small number of count numbers or dose categories, it could be helpful for authors to quantify false positive rates. The linearization and simulation techniques presented here may be useful for this purpose.

Regulatory policy is not set by researchers, but their language and the visuals they present can affect it. For instance, epidemiologic studies can be cited and debated by those trying to influence regulatory policy, legislation and public opinion. In discussing their finding of nonlinearity for male cancer dose response, Grant et al. (*8*) were sensitive to the authors' role in shaping regulation, warning against trying at this time to ''confidently guide the development of modified radiation protection policies.'' Their reasoning went beyond study uncertainties to include the contrasting results found in other studies. This is recognition of the multiple comparisons problem and an indirect warning of the possibility of false positives.

The simulation results in our paper strengthen the idea that explicit discussion of the potential for false positives (at least qualitatively), which is common practice in many fields, could be a useful addition to a study in radiation cancer epidemiology if the reporting or discussion of nonlinearity is deemed relevant to regulatory analysis. The recognition in papers of the possibility of false positives could reinforce the need to look at the family of studies before reaching decisions, which is what expert review committees do.

Another option that could be considered is to avoid the shorthand words, ''statistically significant.'' Authors and commenters might instead consider language that is less dependent on the drawing of bright lines, for instance, by discussing degrees of compatibility (*70*).

Attempting to correct for multiple comparisons is not necessarily a better solution for researchers, given the controversy over its desirability in the research context (*60, 64*). Furthermore, adjusting P values or confidence intervals in one study does not account for the comparisons across studies that can raise to high levels the chances of finding a false positive. When results are of regulatory interest, it can be constructive to point out the importance of considering the multiple comparisons problem and the possibility of associated false positives

## SUPPLEMENTARY INFORMATION

Supplementary File S-1 contains graphs of the first 50 dose response curves for 5,000 simulations of the 2017 dataset for a fixed span of 0.7.

Supplementary tables S-1–S-4 provide information on parameters, count data, person years, and regression sample sizes used in modeling. Nonlinearities considered or graphed in the individual 6 studies are listed in Supplementary table S-5. Values for simulated false positive rates for tests of nonlinearity in dose response that exceeded 10% are found in Supplementary table S-6. Supplementary tables S-7a-7d and tables S-8a-8d provide frequencies of false positives for different regression methods and data treatment, with and without aggregation. Aggregated frequencies of false indications of nonlinearity, with and without adding above-zero dose thresholds are found in Supplementary table S-9. Supplementary table S-10 shows how likelihood ratio test statistics tend to converge as regression sample size is artificially increased (for 2-slope spline fits). Supplementary table S-11 compares false positive rates when data were analyzed either individually or grouped. Supplementary figs. S-1–S-4 are flowcharts for the calculations. Supplementary fig. S-5 provides a panel of dose response curves for four replicates of the 2017 dataset smoothed by loess fitting. Supplementary figs. S-6 and S-7 show profile likelihood graphs for 2-slope spline and threshold parameters, respectively. Supplementary figs. S-8 and S-9 show the distribution of the likelihood ratio test for 2-slope spline fits as the regression sample size is artificially increased for the Techa River mortality and LSS2017 datasets. A profile likelihood graph for curvature is shown in Supplementary fig. S-10. Supplementary fig. S-11 shows graphs of threshold dose response and profile likelihood for a single replicate of the fits to simulations of individualized Techa River mortality data. Supplementary fig. S-12 gives a flowchart for the simulation of individualized data (embedded in Supplementary text S-1). Supplementary fig. S-13 shows a toy threshold-dose-response function (embedded in Supplementary text S-4).

Supplementary text S-1 provides more details on simulation of individualized data. Supplementary text S-2 addresses the insensitivity of results to standardization. Supplementary text S-3 gives details of loess smoothing. Supplementary text S-4 describes the toy threshold-dose-response model. Supplementary text S-5 collects the results of sensitivity tests. Supplementary text S-6 describes a way to linearize case-control data.

## ACKNOWLEDGMENTS

We thank Georg Heinze for helpful comments and suggestions on an early version of the manuscript. We thank anonymous reviewers for valuable suggestions, including the suggestion that we explore AIC methods. Both JB and GH contributed to the logic and framing of the manuscript and its writing. JB did the modeling. No outside sources of funding were used. This paper makes use of data obtained from the Radiation Effects Research Foundation (RERF) in Hiroshima and Nagasaki, Japan. RERF is a private, non-profit foundation funded by the Japanese Ministry of Health, Labour and Welfare (MHLW) and the U.S. Department of Energy (DOE), the latter through the National Academy of Sciences. The data include information obtained from the Hiroshima City, Hiroshima Prefecture, Nagasaki City, and Nagasaki Prefecture Tumor Registries and the Hiroshima and Nagasaki Tissue Registries. As requested by RERF, we state that the conclusions in this article are ours and do not necessarily reflect the scientific judgment of RERF or its funding agencies.

Received: November 1, 2021; accepted: January 24, 2023; published online: February 15, 2023

## REFERENCES

1. Health risks from exposure to low levels of ionizing radiation: BEIR VII Phase 2 (Committee to Assess Health Risks from Exposure to Low Level of Ionizing Radiation). Washington, D.C.: National Research Council, National Academies Press; 2006.
2. Shore RE, Beck HL, Boice JD, Jr., Caffrey EA, Davis S, Grogan HA, et al., Recent epidemiologic studies and the linear no-threshold model for radiation protection—considerations regarding NCRP Commentary 27. Health Phys 2019; 116, 235-46.
3. Pawel D, Boyd M, Studies of radiation health effects inform EPA actions. J Radiol Prot 2019; 39, S40-S57.
4. Cléro E, Vaillant L, Hamada N, Zhang W, Preston D, Laurier D, et al., History of radiation detriment and its calculation methodology used in ICRP Publication 103. Journal of Radiological Protection 2019; 39, R19-R35.
5. Implications of recent epidemiologic studies for the linear-nonthreshold model and radiation protection. NCRP Commentary no. 27. Bethesda, MD: National Council on Radiation Protection and Measurements; 2018.
6. Pierce DA, Preston DL, Radiation-related cancer risks at low doses among atomic bomb survivors. Radiat Res 2000; 154, 178-86.
7. Preston DL, Ron E, Tokuoka S, Funamoto S, Nishi N, Soda M, et al., Solid cancer incidence in atomic bomb survivors: 1958–1998. Radiat Res 2007; 168, 1-64.
8. Grant EJ, Brenner A, Sugiyama H, Sakata R, Sakadane A, Utada M, et al., Solid cancer Incidence among the Life Span Study of atomic bomb survivors: 1958–2009. Radiat Res 2017; 187, 513-37.
9. Davis FG, Krestinina LY, Preston D, Epifanova S, Degteva M, Akleyev AV, Solid cancer Incidence in the Techa River Incidence cohort: 1956–2007. Radiat Res 2015; 184, 56-65.
10. Little MP, Evidence for dose and dose rate effects in human and animal radiation studies. Annals of the ICRP 2018; 47, 97-112.
11. Strengthening transparency in regulatory science (proposed rule). Washington: Environmental Protection Agency, Fed. Reg. 83(83): 18768-74, April 30, 2018.
12. Strengthening Transparency in Pivotal Science Underlying Significant Regulatory Actions and Influential Scientific Information. Washington: Environmental Protection Agency, Fed. Reg. 86(3): 469-93, January 6, 2021.
13. Randall D, Comments on EPA's Final Rule, ''Strengthening Transparency'' [cited 2022 Sept. 4] Available from https://www.nas.org/blogs/article/nas-comments-on-epas-final-rule-strengthening-transparency. Nat Assoc Schol; 2021.
14. Rust S, Scientist says some pollution is good for you — a disputed claim Trump's EPA has embraced [cited 2022 April 7] Available from https://www.latimes.com/local/california/la-me-secret-science-20190219-story.html. Los Angeles Times; 2019.
15. Strengthening Transparency in Pivotal Science Underlying Significant Regulatory Actions and Influential Scientific Information; Implementation of Vacatur. Washington: Environmental Protection Agency, Fed. Reg. 86(104): 29515-7, June 2, 2021
16. O'Connor MK, Calabrese EJ, Response to comments on ''Estimating risks of low radiation doses—a critical review of the BEIR VII report and its use of the linear no-threshold (LNT) hypothesis''. Radiat Res 2015; 183, 481-84.
17. O'Connor MK, Risk of low-dose radiation and the BEIR VII report: a critical review of what it does and doesn't say. Phys Med 2017; 43, 153-58.

18. Cuttler JM, Chapter 24 - Treating Neurodegenerative Diseases with Low Doses of Ionizing Radiation. In: Rattan SIS, Kyriazis M editors. The Science of Hormesis in Health and Longevity: Academic Press; 2019. p. 275-83.

19. Doss M, Comment on ‘Implications of recent epidemiologic studies for the linear nonthreshold model and radiation protection’. J Radiol Prot 2019; 39, 650-54.

20. Siegel JA, Sacks B, Welsh JS, Time to Terminate LNT: Radiation Regulators Should Adopt LT. J Radiol 2017; 1, 049-53.

21. Schonfeld SJ, Krestinina LY, Epifanova S, Degteva MO, Akleyev AV, Preston DL, Solid cancer mortality in the Techa River cohort (1950–2007). Radiat Res 2013; 179, 183-89.

22. Richardson DB, Cardis E, Daniels RD, Gillies M, O’Hagan JA, Hamra GB, et al., Risk of cancer from occupational exposure to ionising radiation: retrospective cohort study of workers in France, the United Kingdom, and the United States (INWORKS). BMJ 2015; 351:h5359.

23. Ozasa K, Shimizu Y, Suyama A, Kasagi F, Soda M, Grant EJ, et al., Studies of the mortality of atomic bomb survivors, report 14, 1950-2003: an overview of cancer and noncancer diseases. Radiat Res 2012; 177, 229-43.

24. Radiation Effects Research Foundation [dataset] Life Span Study solid cancer incidence data, 1958-1998, file name, lssinc07.csv. [cited 2020, Dec. 1]. Available from http://www.rerf.or.jp.

25. Radiation Effects Research Foundation [dataset] Life Span Study report 14. Cancer and noncancer disease mortality data, 1950–2003, file name, lss14.csv. [cited 2020 Dec. 1]. Available from https://www.rerf.or.jp.

26. Radiation Effects Research Foundation [dataset] Life Span Study solid cancer incidence data, 1958-2009, file name, sol_col_2017ext_v1 [cited 2019 Aug. 17] Available from https://www.rerf.or.jp.

27. Doss M, Linear no-threshold model vs. radiation hormesis. Dose-Response 2013; 11, 480-97.

28. Ulsh BA, A critical evaluation of the NCRP Commentary 27 endorsement of the linear no-threshold model of radiation effects. Environ Res 2018; 167, 472-87.

29. Kaiser JC, Heidenreich WF, Comparing regression methods for the two-stage clonal expansion model of carcinogenesis. Stat Med 2004; 23, 3333-50.

30. Harrison RL, Introduction to Monte Carlo Simulation. AIP Conf Proc 2010; 1204, 17-21.

31. R_Core_Team, R: a language and environment for statistical computing [cited 2020 Nov. 4] Available from http://www.R-project.org. Vienna: R Foundation for Statistical Computing; 2020.

32. Sasaki MS, Tachibana A, Takeda S, Cancer risk at low doses of ionizing radiation: artificial neural networks inference from atomic bomb survivors. J Radiat Res 2014; 55, 391-406.

33. Calabrese EJ, O’Connor MK, Estimating risk of low radiation doses – a critical review of the BEIR VII report and its use of the linear no-threshold (LNT) hypothesis. Radiat Res 2014; 182, 463-74.

34. Burnham KP, Anderson DR, Multimodel inference: understanding AIC and BIC in model selection. Sociological methods & research 2004; 33, 261-304.

35. Forstmeier W, Wagenmakers E-J, Parker TH, Detecting and avoiding likely false-positive findings – a practical guide. Biol Reviews 2017; 92, 1941-68.

36. Nist, 4.1.4.4. LOESS (aka LOWESS) [cited 2022 April 7] Available from https://www.itl.nist.gov/div898/handbook/pmd/section1/pmd144.htm. NIST/SEMATECH e-Handbook of Statistical Methods; 2022.

37. Aurengo A, Averbeck D, Bonnin A, Guen BL, Masse R, Monier R, et al., Dose-effect relationships and estimation of the carcinogenic effects of low doses of ionizing radiation. Executive Summary. [cited 2021 Oct. 6] Available from: https://www.radiochemistry.org/documents/html/033005_rad.html. French Academy of Sciences - French National Academy of Medicine; 2005.

38. North BV, Curtis D, Sham PC, A note on the calculation of empirical P values from Monte Carlo procedures. Am J Hum Genet 2002; 71, 439-41.

39. Gillies M, Kuznetsova I, Sokolnikov M, Haylock R, O’Hagan J, Tsareva Y, et al. Lung cancer risk from plutonium: a pooled analysis of the Mayak and Sellafield worker cohorts. Radiation Research 2017; 188, 645-60.

40. Grellier J, Atkinson W, Bérard P, Bingham D, Birchall A, Blanchardon E, et al., Risk of lung cancer mortality in nuclear workers from internal exposure to alpha particle-emitting radionuclides. Epidemiology (Cambridge, Mass) 2017; 28, 675-84.

41. Gans DJ, The search for significance: different tests on the same data. J Stat Compu Sim 984; 19, 1-21.

42. Boos DD, Stefanski LA, P-Value Precision and Reproducibility. Am Stat 2011; 65, 213-21.

43. Neriishi K, Nakashima E, Atsushi M, Fujiwara S, Akahoshi M, Hiromu KM, et al., Postoperative cataract cases among atomic bomb survivors: radiation dose response and threshold. Radiat Res 2007; 168, 404-08.

44. Cole SR, Chu H, Greenland S, Maximum likelihood, profile likelihood, and penalized likelihood: a primer. Am J Epidemiol 2014; 179, 252-60.

45. Díaz-Francés E, Rubio FJ, On the existence of a normal approximation to the distribution of the ratio of two independent normal random variables. Stat Papers 2013; 54, 309-23.

46. Nakashima E, Radiation Dose Response Estimation with Emphasis on Low Dose Range Using Restricted Cubic Splines: Application to All Solid Cancer Mortality Data, 1950–2003, in Atomic Bomb Survivors. Health Phys 2015; 109, 15-24.

47. Dropkin G, Low dose radiation risks for women surviving the A-bombs in Japan: generalized additive model. Environ Health 2016; 15, 112.

48. Samartzis D, Nishi N, Cologne J, Funamoto S, Hayashi M, Kodama K, et al., Ionizing radiation exposure and the development of soft-tissue sarcomas in atomic-bomb survivors. J Bone Joint Surg Am 2013; 95, 222-29.

49. Hunter N, Haylock R, Radiation risks of lymphoma and multiple myeloma incidence in the updated NRRW-3 cohort in the UK: 1955–2011. Journal of Radiological Protection 2022; 42, 011517.

50. Dziak JJ, Coffman DL, Lanza ST, Li R, Jermiin LS, Sensitivity and specificity of information criteria. Brief Bioinform 2020; 21, 553-65.

51. Walsh L, Kaiser JC, Multi-model inference of adult and childhood leukaemia excess relative risks based on the Japanese A-bomb survivors mortality data (1950-2000). Radiat Environ Biophys 2011; 50, 21-35.

52. Kaiser J, Jacob P, Meckbach R, Cullings H, Breast cancer risk in atomic bomb survivors from multi-model inference with incidence data 1958–1998. Radiation and environmental biophysics 2012; 51, 1-14.

53. Feder PI, The log likelihood ratio in segmented regression. Annals Stat 1975; 3, 84-97.

54. Susko E, Likelihood ratio tests with boundary constraints using data-dependent degrees of freedom. Biometrika 2013; 100, 1019-23.

55. Song R, Banerjee M, Kosorok MR, Asymptotics for change-point models under varying degrees of mis-specification. Ann Stat 2016; 44, 153-82.

56. Molenberghs G, Verbeke G, Likelihood ratio, Score, and Wald tests in a constrained parameter space. Am Stat 2007; 61, 22-27.

57. Science and Decisions: Advancing Risk Assessment. Washington, DC: National Research Counci, National Academies Press; 2009.

58. Hoffmann GR, Relating hormesis to ethics and policy: conceptual issues and scientific uncertainty. In: Rattan SIS, Le Bourg E

editors. Hormesis in Health and Disease. Boca Raton: CRC Press; 2014. p. 307-37.

59. White RH, Cote I, Zeise L, Fox M, Dominici F, Burke TA, et al., State-of-the-Science Workshop Report: Issues and Approaches in Low-Dose–Response Extrapolation for Environmental Health Risk Assessment. Environ Health Perspect 2009; 117, 283-87.

60. Rothman KJ, No Adjustments Are Needed for Multiple Comparisons. Epidemiology 1990; 1, 43-46.

61. Bishop D, How scientists can stop fooling themselves. Nature 2020; 584, 7819.

62. Coulson M, Healey M, Fidler F, Cumming G, Confidence intervals permit, but don't guarantee, better inference than statistical significance testing. Frontiers in Psych 2010; 1, 1-9.

63. Greenland S, Robins JM, Empirical-Bayes adjustments for multiple comparisons are sometimes useful. Epidemiology 1991, 2244-51.

64. Greenland S, Analysis goals, error-cost sensitivity, and analysis hacking: Essential considerations in hypothesis testing and multiple comparisons. Paediatr Perinatal Epidemiol 2021; 35, 8-23.

65. Benjamini Y, Yekutieli D, False Discovery Rate–Adjusted Multiple Confidence Intervals for Selected Parameters. J Am Stat Assoc 2005; 100, 71-81.

66. Hafner L, Walsh L, Rühm W, Assessing the impact of different neutron RBEs on the all solid cancer radiation risks obtained from the Japanese A-bomb survivors data. Int J Radiat Biol 2022, 1-15.

67. Linet MS, Schubauer-Berigan MK, Berrington de González A, Outcome assessment in epidemiological studies of low-dose radiation exposure and cancer risks: sources, level of ascertainment, and misclassification. JNCI Monographs 2020, 154-75.

68. Little MP, Pawel D, Misumi M, Hamada N, Cullings H, Wakeford R, et al., Lifetime mortality risk from cancer and circulatory disease predicted from the Japanese atomic bomb survivor Life Span Study data taking account of dose measurement error. Radiat Res 2020; 194, 259-76.

69. Furukawa K, Misumi M, Cologne JB, Cullings HM, A Bayesian semiparametric model for radiation dose-response estimation. Risk Anal 2016; 36, 1211-23.

70. Amrhein V, Greenland S, Rewriting results in the language of compatibility. Trends in Ecology & Evolution 2022; 37, 567-68.