DOE Review Release 03/26/2018

Applications of Regression in External Dose Reconstruction

ORAUT-RPRT-0087 Rev. 00
Effective Date: 03/13/2018
Supersedes: None

Subject Expert(s): Thomas R. LaBone, Keith A. McCartney, and Matthew H. Smith

Document Owner Approval:	Signature on File Thomas R. LaBone, Document Owner	Approval Date:	03/08/2018
Concurrence:	Signature on File John M. Byrne, Objective 1 Manager	Concurrence Date:	03/08/2018
Concurrence:	Signature on File Scott R. Siebert, Objective 3 Manager	Concurrence Date:	03/08/2018
Concurrence:	Signature on File Kate Kimpan, Project Director	Concurrence Date:	03/08/2018
Approval:	Signature on File James W. Neton, Associate Director for Science	Approval Date:	03/13/2018

FOR DOCUMENTS MARKED AS A TOTAL REWRITE, REVISION, OR PAGE CHANGE, REPLACE THE PRIOR REVISION AND DISCARD / DESTROY ALL COPIES OF THE PRIOR REVISION.

☒ New ☐ Total Rewrite ☐ Revision ☐ Page Change

PUBLICATION RECORD

EFFECTIVE DATE	REVISION NUMBER	DESCRIPTION
03/13/2018	00	New report initiated to introduce ordinary least squares regression and quantile regression as alternative methods for estimating beta dose based on gamma dose for workers whose dosimeters were relatively insensitive to beta radiation. Initiated by Thomas R. LaBone. Incorporates formal internal and NIOSH review comments. Training required: As determined by the Objective Manager. Initiated by Thomas R. LaBone.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACRONYMS AND ABBREVIATIONS

CDF	cumulative density function
DOE	U.S. Department of Energy
GM	geometric mean
GOF	goodness of fit
GSD	geometric standard deviation
IREP	Interactive RadioEpidemiological Program
mrem	millirem
NIOSH	National Institute for Occupational Safety and Health
OLS	ordinary least squares
ORAU	Oak Ridge Associated Universities
PDF	probability density function
QQ	quantile-quantile
ROS	regression on order statistics
SRDB Ref ID	Site Research Database Reference Identification (number)

1.0 INTRODUCTION

NOTE: This report is intended as reference by statisticians who use the methods described in the report. The details of its application and any necessary justifications are provided in the site-specific reports where the methods are used. The development of methods and software that are used by dose reconstructors to assign dose to claimants is not within the scope of this report.

Either by design or nature, external dosimeters can have different sensitivities to different types and energies of radiation. For example, some dosimeters worn by workers are sensitive to both beta and gamma radiation (i.e., beta/gamma dosimeters) and others are sensitive to gamma radiation but relatively insensitive to beta radiation (i.e., gamma dosimeters). To estimate beta dose for workers who wore gamma dosimeters that are insensitive to beta radiation, health physicists can use the beta/gamma dosimeters readings to develop a model for estimating the beta dose as a function of the gamma dose. Problems of this type are not uncommon in external dosimetry[1] and have traditionally been solved using a regression of the natural logarithm of the ratio of beta dose to gamma dose on standard normal quantiles, specifically lognormal regression on order statistics (ROS) as described in Helsel (2012) and ORAUT (2014).

This report introduces ordinary least squares (OLS) regression (Weisberg 2005) and quantile regression (Koenker 2005; Cade and Noon 2003) as alternative methods for estimating beta dose given the gamma dose. Regardless of method, the goal is to calculate either:

- A point estimate of the beta dose given a point estimate of the gamma dose (where the point estimates of the beta and gamma dose are usually the 50th or 95th percentiles), or

- The parameters of a beta dose distribution (usually lognormal) given a gamma dose distribution (usually normal or lognormal).

The gamma doses can come from personal dosimeter measurements or from coworker models. This discussion will be made more concrete by using the three regression methods to analyze actual beta/gamma doses from a U.S. Department of Energy (DOE) facility. The datasets and R code used in the preparation of this report are available in ORAUT (2017).

2.0 DOSIMETRY DATA

The dataset used to illustrate the different analysis methods consists of reported beta/gamma doses from a DOE site over a 7-year period. The censoring level was 30 mrem for gamma dose and 50 mrem for beta dose (i.e., a gamma dose of less than 30 mrem was reported as <30 mrem). Therefore, the beta/gamma dose pairs can consist of:

1. Uncensored beta dose and uncensored gamma dose,
2. Uncensored beta dose and censored gamma dose,
3. Censored beta dose and uncensored gamma dose, and
4. Censored beta dose and censored gamma dose.

For OLS and quantile regression, beta dose is regressed on the gamma dose, which means that the gamma dose is the independent variable. These regression methods require an uncensored independent variable, so only beta/gamma dose pairs 1 and 3 are applicable. Likewise, for ROS the

[1] Neutron vs. gamma doses, extremity vs. whole-body doses, etc., can also be addressed in the same way described here for beta/gamma doses.

ratio of a beta dose (censored or uncensored) to a censored gamma dose is not well defined,[2] so again only beta/gamma dose pairs 1 and 3 can be used. Excluding the censored gamma doses from the model does not pose significant technical difficulties because the models developed with all three methods can be extrapolated to gamma doses <30 mrem if needed.

There are 5,457 beta/gamma pairs in categories 1 and 3 in the dataset, 1,626 of which are in category 1. The 5,457 beta/gamma pairs are shown in the log-log[3] scatter plot in Figure 2-1. The beta/gamma ratios are displayed on a quantile-quantile (QQ) plot in Figure 2-2, where the censored ratios are in red and the uncensored in black. Even though the beta dose has a single censoring level, the beta/gamma ratios have many different censoring levels which, as discussed in the next section, complicates the analysis of the beta/gamma ratios.

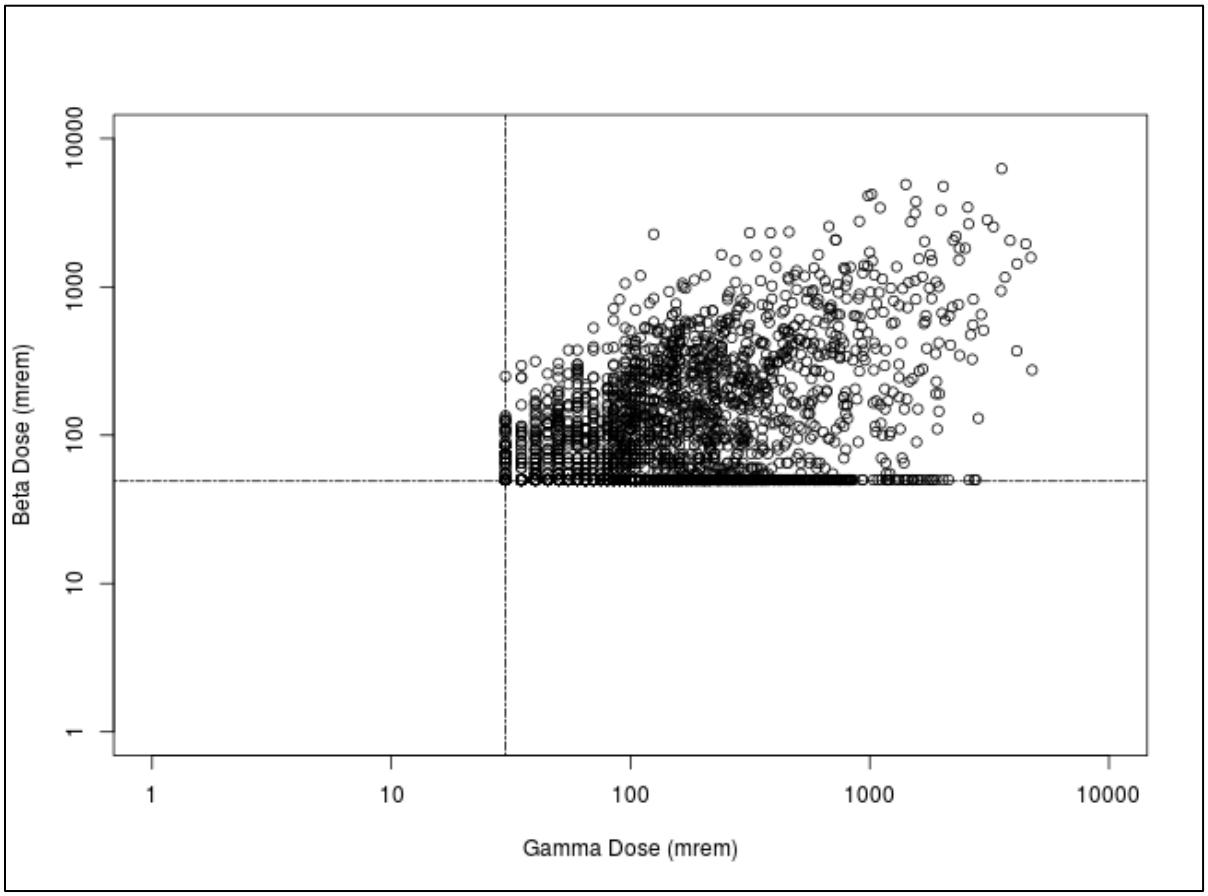

Figure 2-1. Scatter plot of beta versus gamma dose; the beta dose censoring level is 50 mrem (horizontal line) and the gamma dose censoring level is 30 mrem (vertical line).

3.0 IMPUTING CENSORED DOSES

Most external dose data sets are censored, and this dataset is no exception with approximately 70% of the beta doses being censored (reported as <50 mrem). The relationship between the beta and gamma doses in the presence of censored beta doses can be modeled by:

[2] For example, 100 divided by <30 could be anywhere between 100/30 and infinity and <50 divided by <30 could be undefined.

[3] In this report a logarithmic scale on a plot uses a base 10 (common) log. Elsewhere, a log refers to a base e (natural) log.

- Using regression methods that specifically account for censoring (i.e., survival analysis methods (Helsel 2012), or

- Imputing uncensored results for the censored results and then using standard regression on the resulting complete dataset (ORAUT 2015).

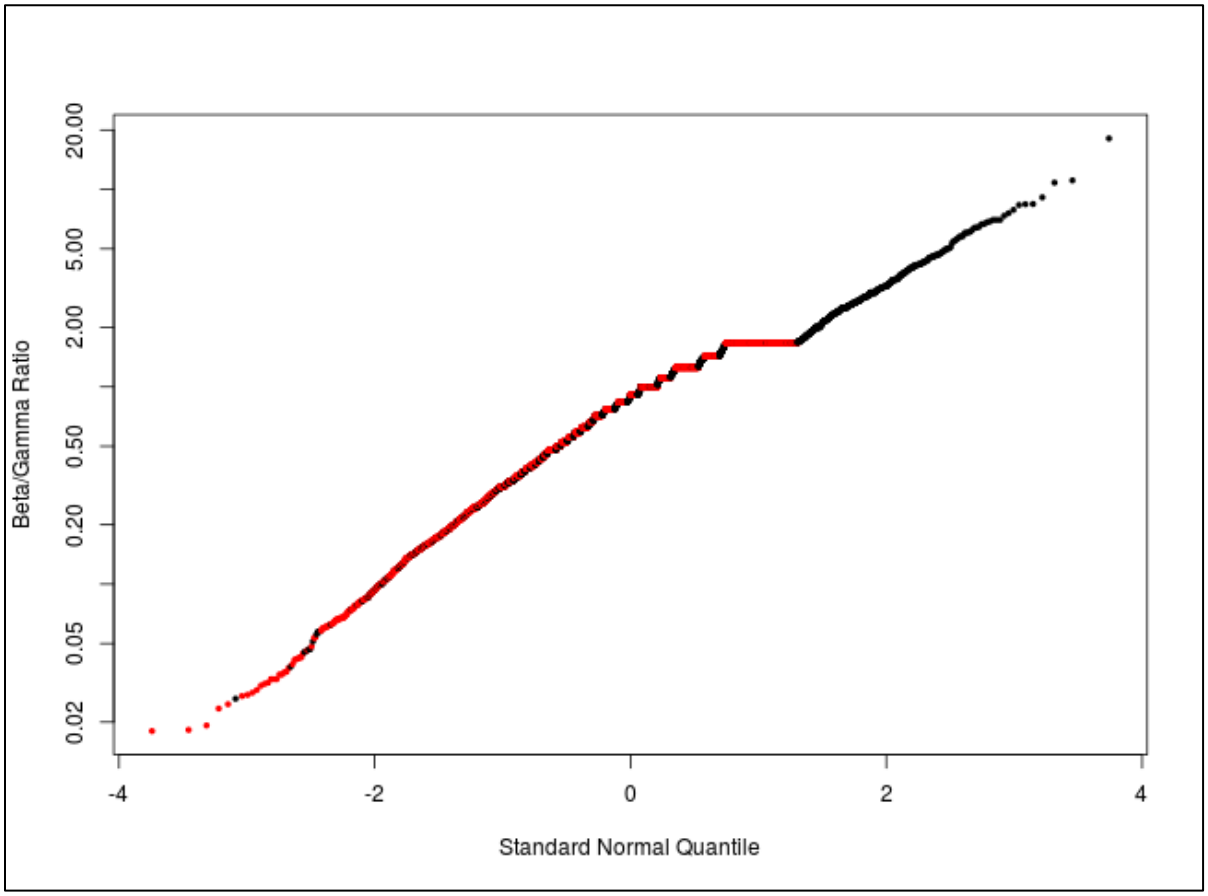

Figure 2-2. Lognormal QQ plot of beta/gamma ratios from censored (red) and uncensored (black) beta doses.

With the lognormal ROS it is fairly easy to account for the censored beta doses using methods specifically designed for the task. This is not the case with the more complicated regressions discussed later, so it is advantageous to replace the censored doses with estimates of uncensored doses and modeling the dataset as if it were uncensored. This simplifies computations by replacing censored beta doses with random values drawn from the distribution of beta doses that are below the censoring level. The distribution of beta doses below the censoring level is referred to as the imputation model. One way to develop this model is to fit a lognormal distribution to all beta doses as shown in Figure 3-1 and extrapolate the fit into the censored region. The imputation model is used to generate lognormally distributed random beta doses below 50 mrem that are used in the regression in place of the censored beta doses (Krishnamoorthy, Mallick, and Mathew 2009, p. 4).

The result of the imputation is shown in Figure 3-2, where the yellow dots are the imputed beta doses. The dataset can now be considered to be complete and uncensored when doing calculations. This method is discussed in ORAUT-RPRT-0071, *External Dose Coworker Methodology* (ORAUT 2015), where it is used for developing coworker models.

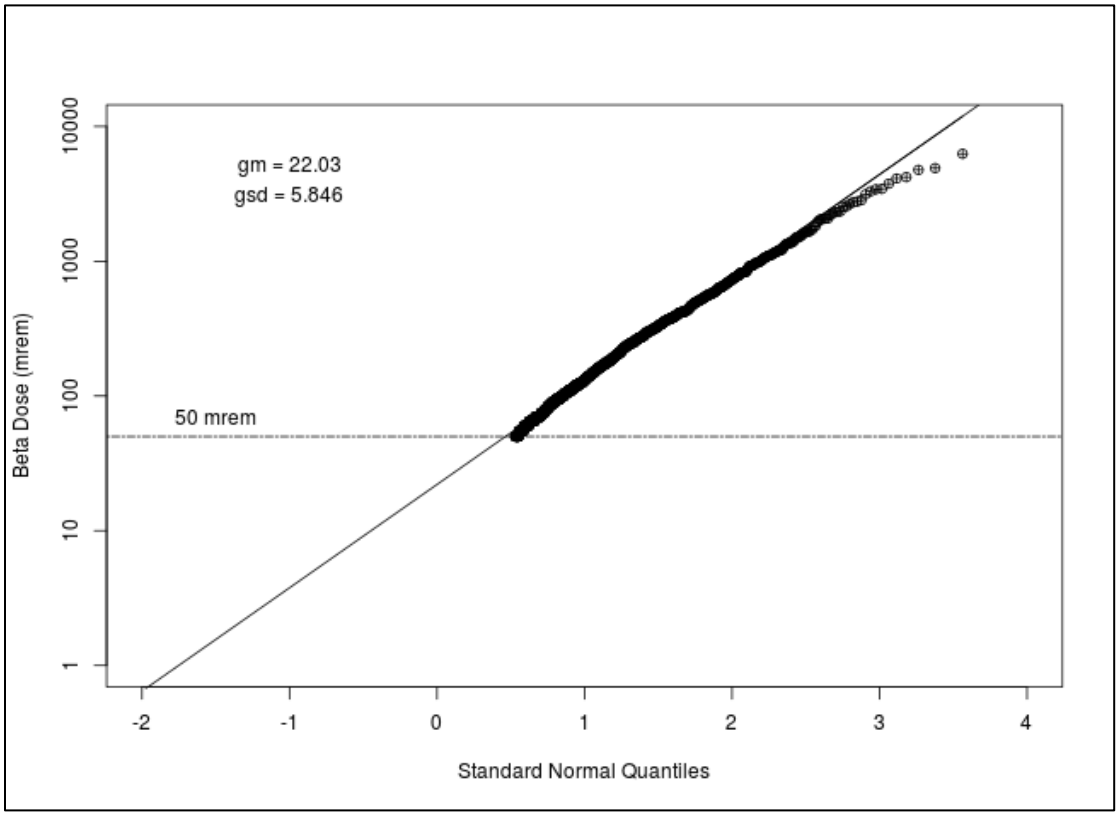

Figure 3-1. Lognormal QQ plot of all beta doses. The lognormal ROS fit gives the parameters for the model used to impute censored beta doses.

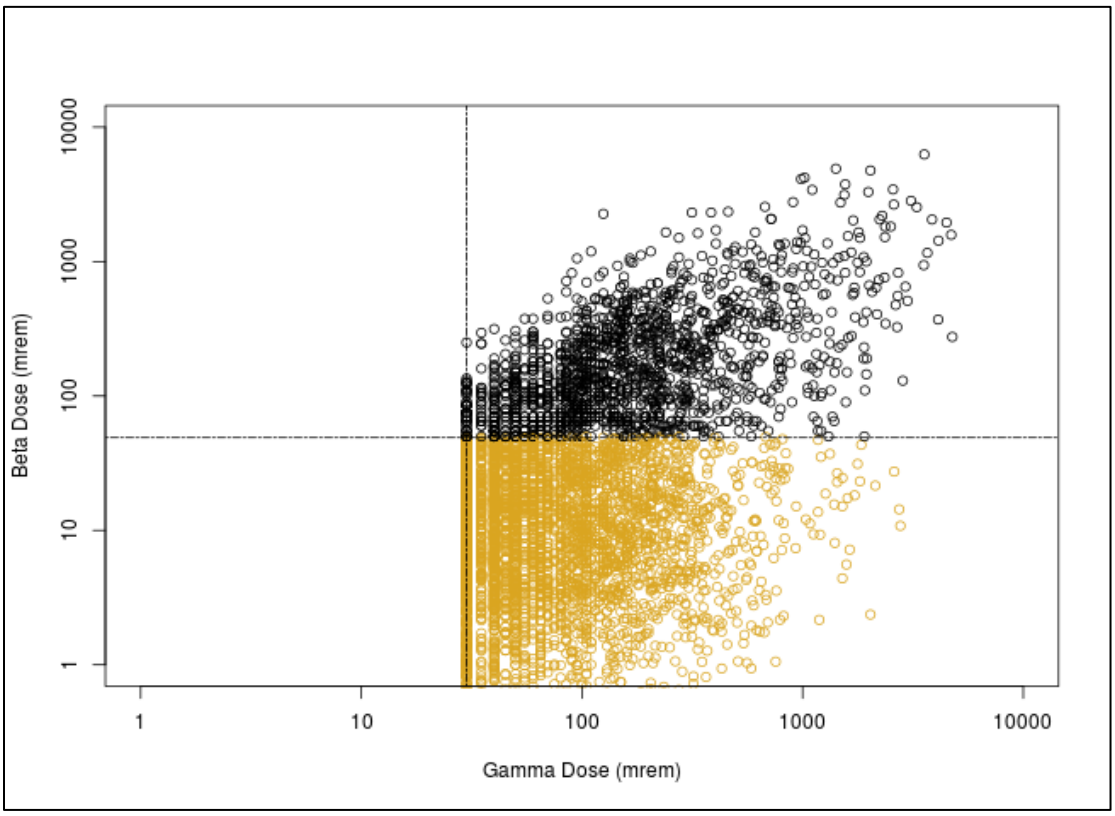

Figure 3-2. Scatter plot of uncensored (black) and imputed (yellow) beta doses versus gamma doses.

4.0 REGRESSION ON ORDER STATISTICS

ROS is used extensively in the dose reconstruction program and is discussed in other Project documents such as ORAUT-RPRT-0053, *Analysis of Stratified Coworker Datasets* (ORAUT 2014). The ROS used here consists of the OLS regression of the log of the beta/gamma ratio on the standard normal quantiles of the beta/gamma ratios:

$$E\left[\log\left(H_\beta / H_\gamma\right)\right] = \beta_0 + \beta_1 Q \quad (4\text{-}1)$$

where

$E[\log(H_\beta/H_\gamma)]$ = expectation (mean) of the log of the ratio of beta to gamma dose
H_β = observed beta doses (mrem)
H_γ = observed gamma doses (mrem)
β_0 = intercept in the formal model
β_1 = slope in the formal model
Q = quantiles of the observed $\log(H_\beta/H_\gamma)$ calculated assuming a standard normal distribution $N(0,1)$

This model shows that the mean of the log beta/gamma ratio is normally distributed $N(\beta_0,\beta_1)$, meaning the beta/gamma ratio is lognormally distributed where β_0 is the log mean and β_1 the log standard deviation of the lognormal distribution. The fitted model is:

$$\widehat{\log\left(H_\beta / H_\gamma\right)} = b_0 + b_1 Q \quad (4\text{-}2)$$

where

$\widehat{\log(H_\beta/H_\gamma)}$ = predicted value of the log beta/gamma ratio for a given quantile
b_0 = –1.4114737 and
b_1 = 1.386666

are the fitted parameters estimated from the regression.

The estimated geometric mean (GM) is:

$$GM = \exp\left(b_0\right) = 0.244 \quad (4\text{-}3)$$

and the estimated geometric standard deviation (GSD) is:

$$GSD = \exp\left(b_1\right) = 4.001 \quad (4\text{-}4)$$

The quantiles Q can be readily calculated from the complete dataset[4] (the one with the imputed values) or the censored dataset, but the censored dataset is used in the example given here. The predicted values of the log beta/gamma ratio are the black line shown in Figure 4-1. The result of the analysis is the GM and GSD of the lognormal model. The poor fit of the lognormal model to the log

[4] If the dataset consists of uncensored data, the mean and standard deviation of the log beta/gamma ratio can be used to calculate the GM and GSD. The ROS approach is only needed when some of the data are censored.

beta/gamma can be seen in the plot in Figure 4-1. Such unattractive fits are not uncommon in practice, but the lognormal model has typically been used anyway because:

- Of its great familiarity and utility,
- Fitting other distributions to the data requires more complicated methods like maximum likelihood (which has come into use in the Project only recently), and
- Only distributions that are available with the Interactive Radio-Epidemiological Program (IREP) can be used, none of which are guaranteed to adequately fit a given dataset. In these cases it is common for the lognormal distribution to be used to give the "best available fit".

For the estimated 95th percentile (red dot in Figure 4-1) from the fitted lognormal distribution, the lack-of-fit does not matter much because the ROS line of best fit coincides with the observed ratios in the area of the 95th percentile. However, for the 50th percentile (blue dot in Figure 4-1) the lack-of-fit does matter because the predicted and observed ratios do not agree.

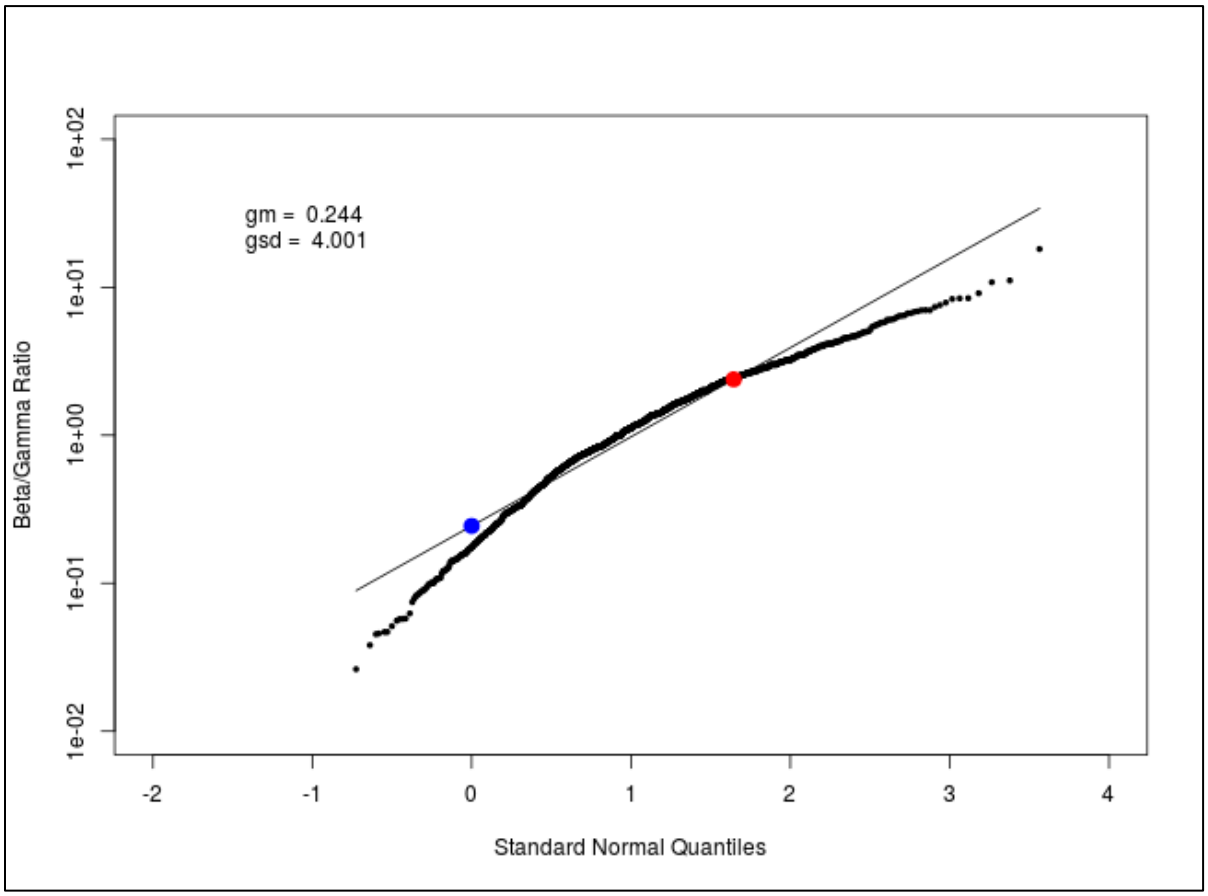

Figure 4-1. Lognormal QQ plot of beta/gamma ratios where the line is the fit from the lognormal ROS. The predicted 50th-percentile beta/gamma ratio is in blue and the 95th-percentile beta/gamma ratio in red.

The scope of this report is limited to the calculation and reporting of the GM and GSD of the beta/gamma ratio and does not specifically address how this information is used in practice. However, standard applications are given in the next two sections.

4.1 POINT ESTIMATES

If the gamma dose is assumed to have a single discrete value, the lognormal beta/gamma ratio and resulting beta dose will also have a single discrete value. In this simple case, a specific percentile of the beta/gamma ratio is calculated from the *GM* = 0.244 and *GSD* = 4.001 of the lognormal distribution along with the standard normal quantile associated with the desired percentile. For example, the 95th percentile R_{95} of the beta/gamma distribution has an associated standard normal quantile of *qnorm*(0.95) = 1.645, giving:

$$R_{95} = GM \times GSD^{1.645} \tag{4-5}$$

$$\widehat{R_{95}} = 0.244 \times 4.001^{1.645} = 2.385 \tag{4-6}$$

The predicted 95th-percentile beta dose in mrem for a 100-mrem gamma dose is therefore:

$$\widehat{H_{\beta 95}} = H_{\gamma} \times \widehat{R_{95}} = 100 \times 2.385 = 239 \tag{4-7}$$

These fixed dose estimates can then be used in IREP when calculating probability of causation.

4.2 DISTRIBUTION ESTIMATES

If the gamma dose is assumed to have a distribution, that distribution is propagated through the lognormal beta/gamma distribution using Monte Carlo methods to give a distribution of beta doses. A suitable distribution like the lognormal is then fit to these beta doses to give parameters for IREP entry. For example, assume a normally distributed gamma dose with a mean of 100 mrem and a standard deviation of 30 mrem. The Monte Carlo simulation sketched in Figure 4-2 propagates the *N*(100,30) gamma dose distribution through the *LN*(μ_{log},σ_{log}) = *LN(log(GM),log(GSD))* beta/gamma ratio distribution to give 10,000 beta doses.

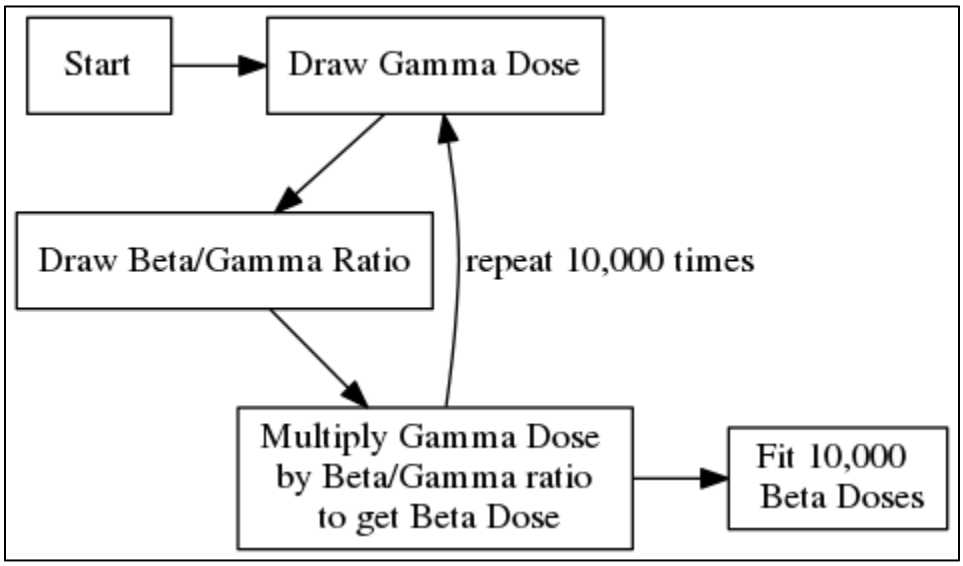

Figure 4-2. Monte Carlo simulation for estimating distribution of beta doses given the distributions of beta/gamma ratios and gamma doses.

When these beta doses are put on a lognormal probability plot as in Figure 4-3, it is clear that these doses are lognormally distributed and subsequent ROS fits yield *GM* = 23.446 and *GSD* = 4.203.

These lognormal parameter estimates are then used in IREP when calculating probability of causation.

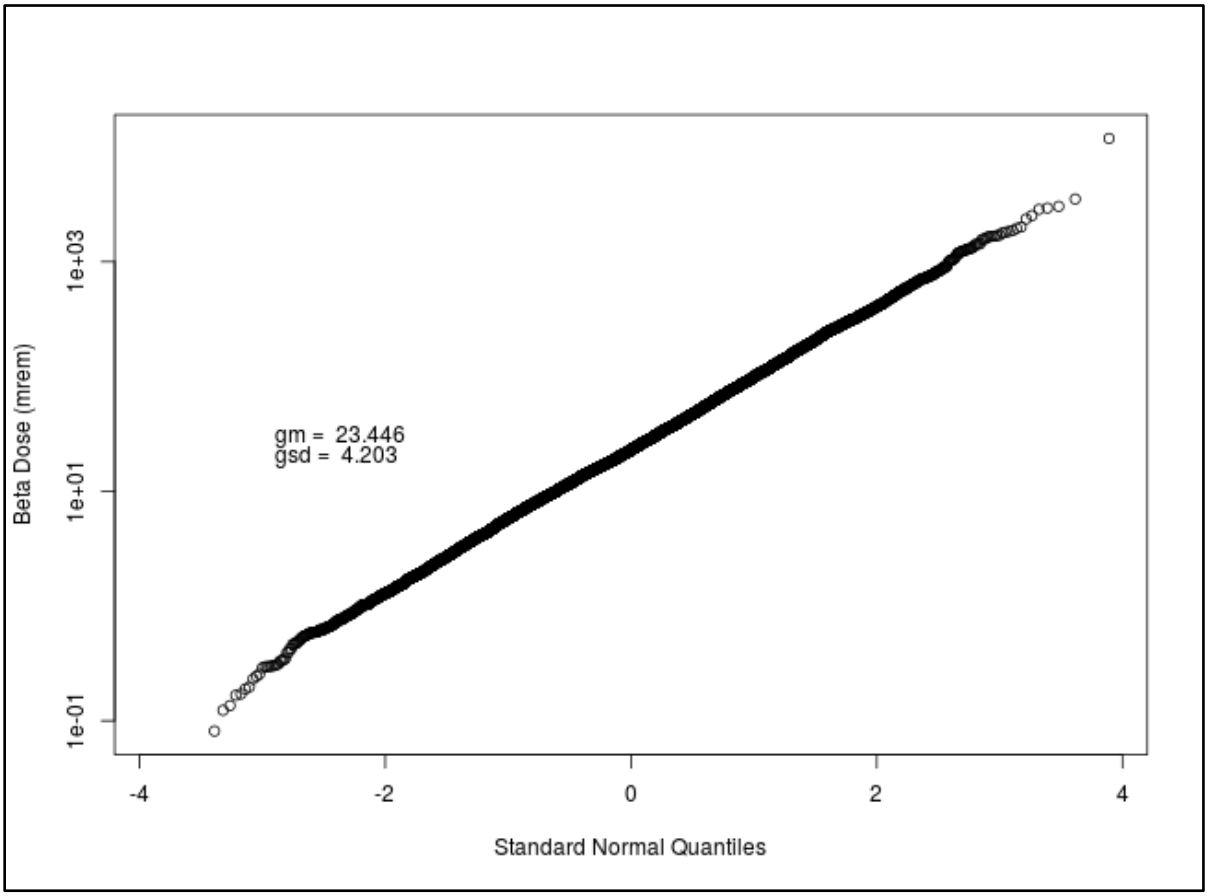

Figure 4-3. Lognormal QQ plot of Monte Carlo simulated beta doses for a given gamma dose of 100 mrem. The lognormal ROS fit to these data give the parameters for the estimated beta dose distribution.

5.0 ORDINARY LEAST SQUARES

The plot of the log gamma dose versus the log beta dose in Figure 3-2 is fairly linear (also see Figure 5-1), suggesting that this linear regression model can be fit to the data as follows:

$$E\left[\log\left(H_{\beta}\right)\right] = \beta_0 + \beta_1 \log\left(H_{\gamma}\right) \qquad (5\text{-}1)$$

where

$E[\log(H_{\beta})]$ = expectation (mean) of the log of the beta dose for a given gamma dose

The model in Equation 5-1 is bivariate: it has two variables, a response variable $\log(H_{\beta})$ and a predictor variable $\log(H_{\gamma})$. The model in Equation 4-1 is univariate: it only has one variable, $\log(H_{\beta}/H_{\gamma})$. Collapsing the bivariate relationship into a univariate relationship inevitably results in the loss of some information about the relationship. For example, patterns in the regression of beta dose on gamma dose can indicate that different groups of workers, who were exposed to radically different radiation sources, are being clumped together. This information is lost when the dose ratios are taken, leading to reduced accuracy in predicted beta doses. The mean regression line is the red line

in Figure 5-1, and it is what you would get by performing OLS regression of the log beta dose on the log gamma dose, which gives a fitted model of:

$$\widehat{\log\left(H_{\beta}\right)} = b_0 + b_1 \log\left(H_{\gamma}\right) \tag{5-2}$$

where

$\widehat{\log(H_{\beta})}$ = predicted value of the mean log beta dose for a given log gamma dose
b_0 = –0.247 and
b_1 = 0.717 are the fitted parameters estimated from the regression

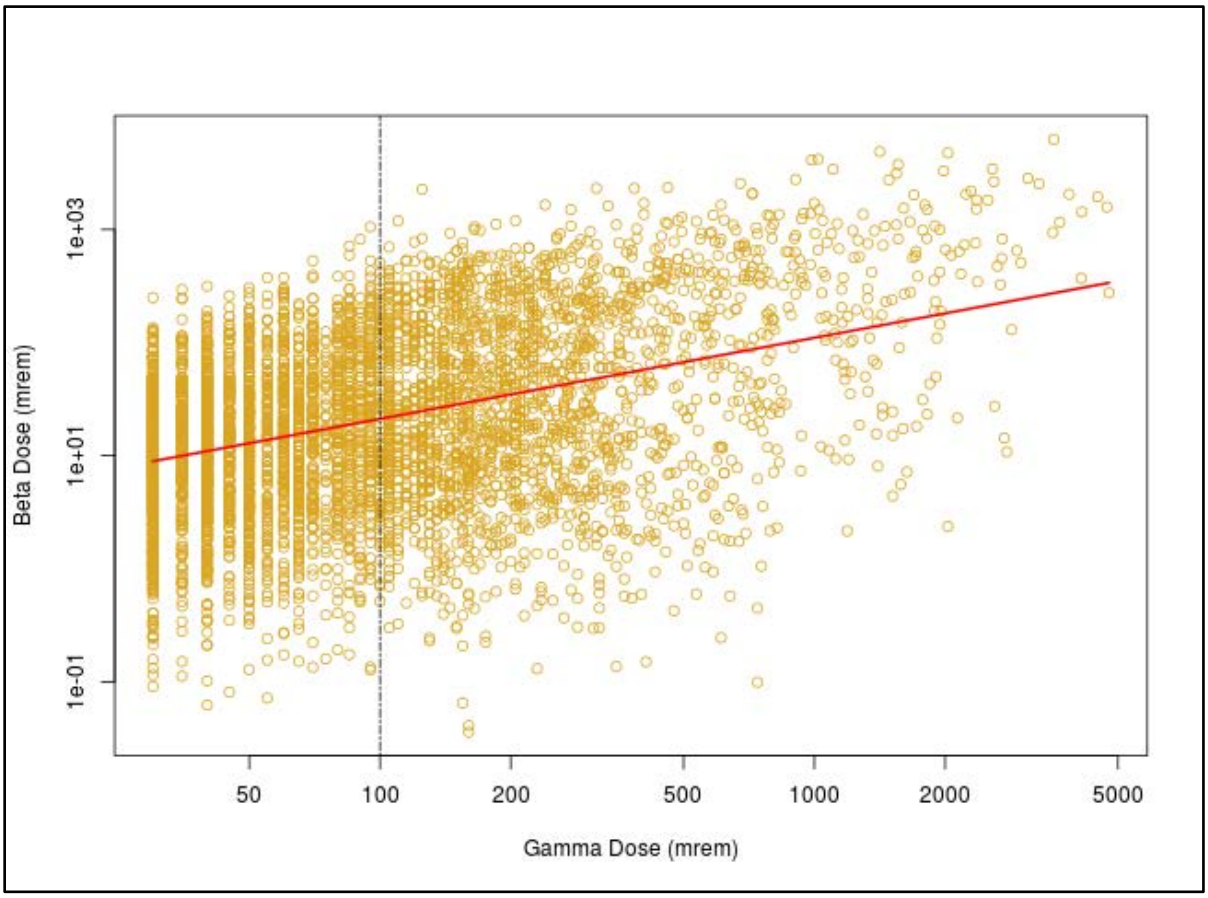

Figure 5-1. Log-log scatter plot of beta doses versus gamma doses. The red line is the OLS regression mean line.

The line is a conditional mean that identifies the mean log beta dose for a given log gamma dose. Taking the exponent of the mean log beta dose gives the median (50th-percentile) beta dose. Therefore, the predicted median beta dose in mrem for a 100-mrem gamma dose (the vertical line in Figure 5-1) is:

$$\widehat{H_{\beta 50}} = \exp\left(-0.247 + 0.717 \log\left(100\right)\right) = 21.2 \tag{5-3}$$

The mean of the log variable being the median of the linear variable is the result of the log transformations.

5.1 QUANTILES

The OLS regression in Figure 5-1 assumes that log beta doses are normally distributed about the mean regression line and that the standard deviation *s* = 1.658328 of the normal distribution is constant for all gamma doses. Therefore, constructing a 95th-percentile line requires simply shifting the mean line upward by 1.645 standard deviations, keeping the same slope. Therefore:

$$\widehat{\log\left(H_{\beta 95}\right)} = b_0 + b_1 \log\left(H_\gamma\right) + sZ_{95} \tag{5-4}$$

(where Z_{95}=1.645 is the number of standard deviations for a N(0,1) distribution associated with the 95th percentile) which in this example is:

$$\widehat{\log\left(H_{\beta 95}\right)} = -0.247 + 0.717 \log\left(H_\gamma\right) + 2.728 \tag{5-5}$$

In Figure 5-2 the blue lines denote the percentiles: the highest blue line is the 99th-percentile beta dose as a function of the gamma dose, the lowest blue line is the 1st-percentile beta dose, and the blue lines in between go from the 5th percentile to the 95th percentile in steps of 5. These percentile lines are symmetrically distributed about the mean line and, while each percentile line has its own unique intercept, they have the same slope as the mean line. The intercepts and slopes for all of the percentile lines in Figure 5-2 are given in Table 5-1.

Table 5-1. Intercepts and slopes for percentile lines in Figure 5-2.

Percentile	Intercept	Slope
1	–4.1053167	0.7167024
5	–2.9751757	0.7167024
10	–2.3727017	0.7167024
15	–1.9662154	0.7167024
20	–1.6431529	0.7167024
25	–1.3659941	0.7167024
30	–1.1170969	0.7167024
35	–0.8864565	0.7167024
40	–0.6676014	0.7167024
45	–0.4558565	0.7167024
50	–0.2474688	0.7167024
55	–0.0390811	0.7167024
60	0.1726638	0.7167024
65	0.3915189	0.7167024
70	0.6221593	0.7167024
75	0.8710564	0.7167024
80	1.1482153	0.7167024
85	1.4712777	0.7167024
90	1.8777641	0.7167024
95	2.4802381	0.7167024
99	3.6103791	0.7167024

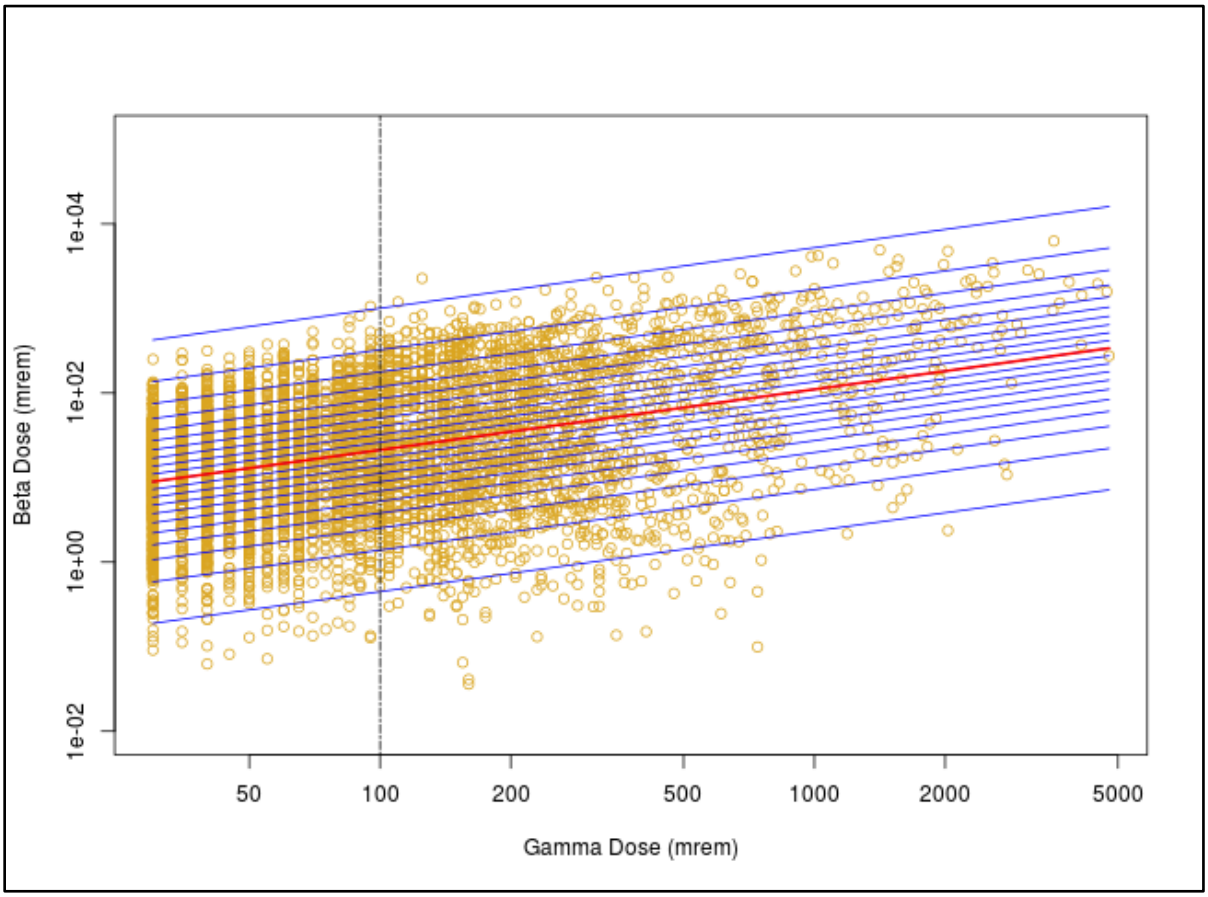

Figure 5-2. Log-log scatter plot of beta doses versus gamma doses. The red line is the OLS regression mean line and the blue lines are the 1st through 99th percentiles based on the OLS model.

As shown below, these 21 equations convey all of the information that is needed to construct point and distributional estimates of the beta dose. The scope of this report is limited to the calculation and reporting of these equations and does not specifically address how this information is used in practice. However, standard applications are given in the next two sections.

5.2 POINT ESTIMATE

To calculate the 95th-percentile beta dose at 100-mrem gamma dose, take the intercept and slope for the 95th percentile from the table above and evaluate it at a gamma dose of 100 mrem:

$$\widehat{\log\left(H_{\beta 95}\right)} = 2.4802381 + 0.7167024\log(100) = 5.780775 \quad (5\text{-}6)$$

Therefore, the predicted 95th-percentile beta dose in mrem for a 100-mrem gamma dose is:

$$\widehat{H_{\beta 95}} = \exp\left(2.4802381 + 0.7167024\log(100)\right) = 324.0101 \quad (5\text{-}7)$$

The point estimates for other percentiles are calculated in the same fashion.

5.3 DISTRIBUTION ESTIMATE USING THE LOGNORMAL DISTRIBUTION

A vertical slice of the plot in Figure 5-2 at 100 mrem rotated so that beta dose is the *x*-axis shows the lognormal probability density function (PDF) in Figure 5-3. This looks like the familiar normal PDF

curve, but it is actually lognormal because the *x*-axis has a log scale. The log mean of this lognormal distribution is μ_{log} = 3.0530679 and the log standard deviation is s_{log} = 1.658328. Perhaps more familiar to most are the corresponding GM [*GM* = exp(μ_{log}) = 21.1802243] and GSD [*GSD* = exp(σ_{log}) = 5.2505248].

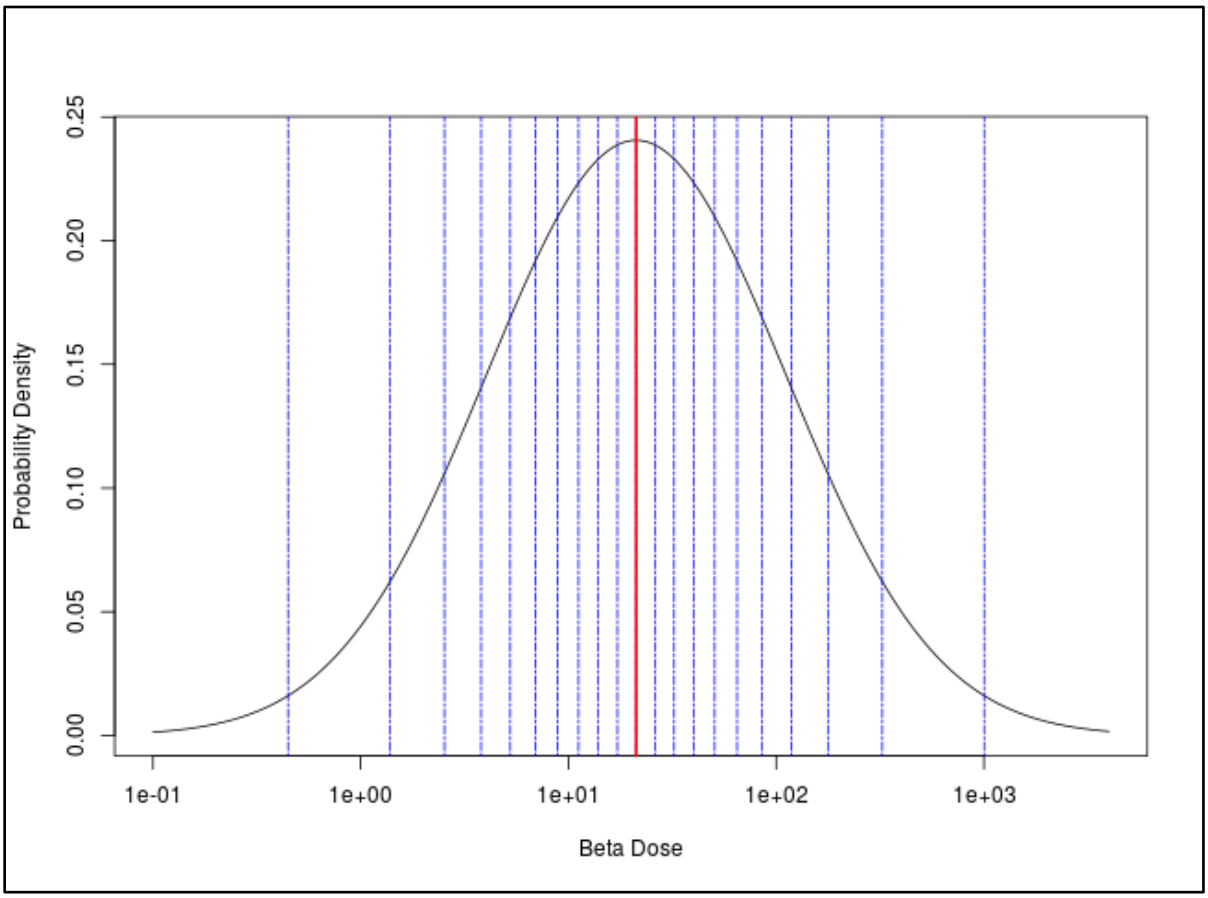

Figure 5-3. Lognormal PDF of beta dose for a gamma dose of 100 mrem.

At another gamma dose, the mean of this normal distribution (the log of *GM*) changes, but the standard deviation (the log of *GSD*) is the same. The blue lines in Figure 5-3 are the same percentiles in Figure 5-2, which are calculated from the lognormal distribution (e.g., the 95th percentile is the log mean plus 1.645 times the log standard deviation). To apply the Monte Carlo uncertainty propagation for the 100 ±30 mrem gamma dose, random gamma doses are drawn from its N(100,30) normal distribution. Each gamma dose has a corresponding lognormal beta dose distribution from which a random draw of a beta dose is made. As illustrated in Figure 5-4, this process is repeated many times to generate a large number of simulated beta doses.

In the final step of the analysis, a suitable model (e.g., lognormal) is fit to the simulated dataset of beta doses to arrive at a beta dose distribution that can be used in IREP. This last step is not explicitly addressed here but rather in site specific reports. Nevertheless, using the lognormal ROS approach, the simulated beta doses in this example can be modeled with a lognormal distribution having *GM* = 21.015 and *GSD* = 5.335. These values are very similar to those that were calculated in Section 4.2, but this is not always the case.

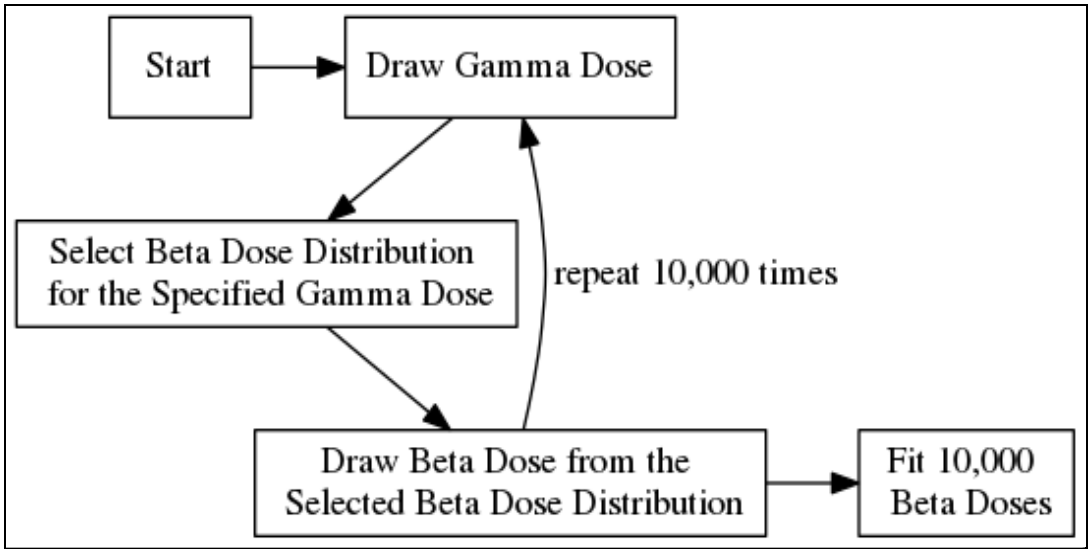

Figure 5-4. Monte Carlo simulation for estimating distribution of beta doses given the OLS regression of beta dose on gamma dose and distribution of gamma doses.

5.4 DISTRIBUTION ESTIMATE USING THE CUMULATIVE DENSITY FUNCTION

Figure 5-3 shows the PDF of the lognormal beta distribution for a gamma dose of 100 mrem. Figure 5-5 shows the corresponding cumulative distribution function (CDF) (the black lines in Figure 5-5 are linear interpolations between the points). By drawing a uniform random number between 0.01 and 0.99 and locating that value on the *y*-axis of Figure 5-5, a random lognormal value can be determined by reading across to the CDF line (usually involving interpolation). This is useful in the discussion of quantile regression below (see Section 6.2) because it provides a method of drawing a random value from a distribution for which there is only an empirical CDF.

The *GM* = 20.478 and *GSD* = 5.008 that were calculated with random draws from a uniform distribution and coupled with the CDF are basically the same as those that were calculated using random draws directly from the lognormal distribution.

5.5 LIMITATIONS OF OLS REGRESSION

In general, OLS regression is more informative and flexible than ROS. For example, beta dose can be regressed on gamma dose and year, which provides information on how the relationship between beta dose and gamma dose changes over time. There are a multitude of tests and diagnostic tools available for OLS regression that allow testing hypotheses and judging how well the model fits the observed data. In addition, evaluation is not limited to a linear function with only an intercept and slope. For example, it allows addition of a quadratic term to the regression. However, OLS regression has a major limitation in that, for any given gamma dose, the percentiles are completely specified by the normal distribution. This means that the percentiles are symmetric around the mean and, for example, there cannot be one relationship (e.g., linear) for the 95th percentile and another relationship (e.g., quadratic) for the 5th percentile. Therefore, the percentiles that are calculated with OLS regression might not accurately model what is actually occurring in the entire dataset. In such cases there is another method called quantile regression that can often provide better results.

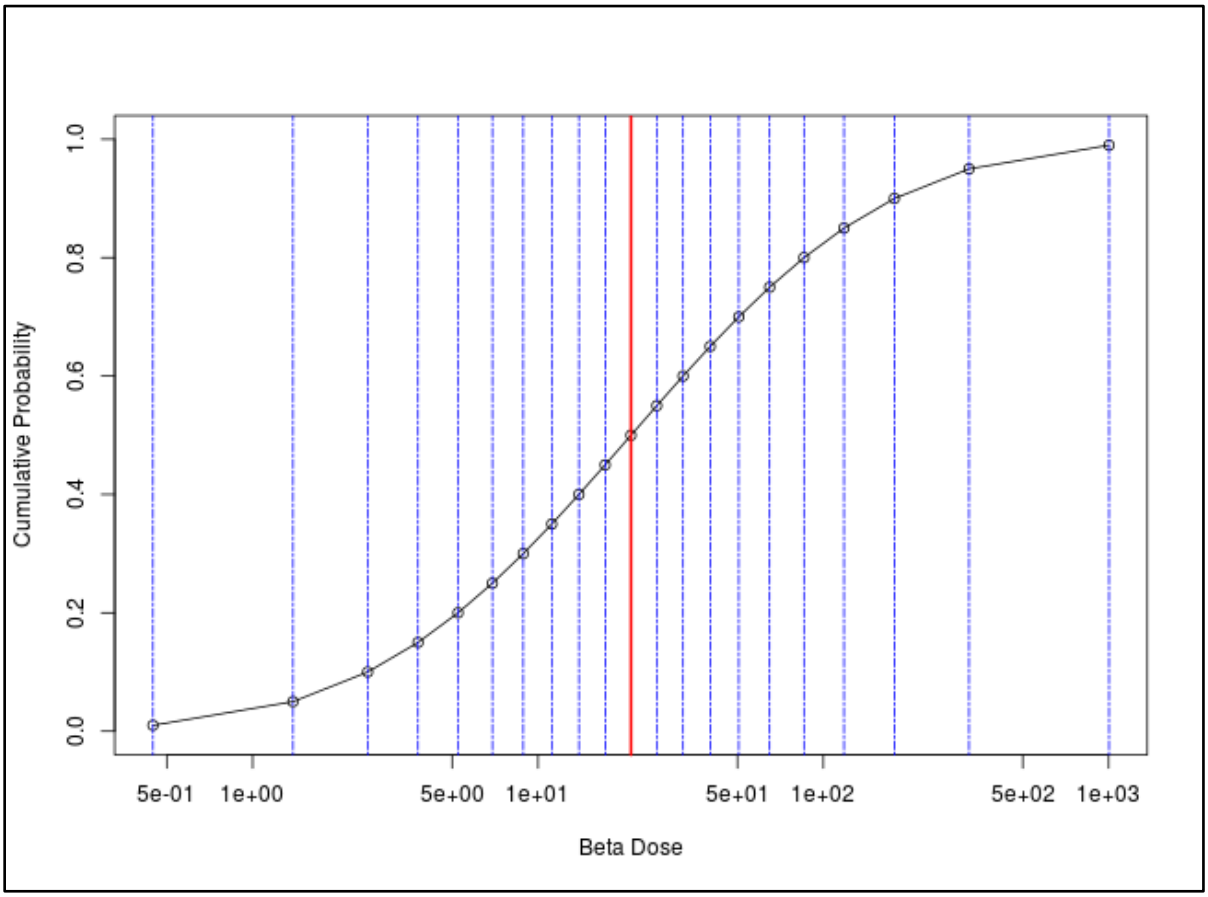

Figure 5-5. Lognormal CDF of beta dose for a gamma dose of 100 mrem.

6.0 QUANTILE REGRESSION

The formal linear model used to fit the example dataset with quantile regression is:

$$Q_x\left(H_\beta\right) = \beta_0 + \beta_1 H_\gamma \tag{6-1}$$

where

$Q_x(H_\beta)$ = xth percentile of the beta dose for a given gamma dose
H_β = observed beta doses (mrem)
H_γ = observed gamma doses (mrem)
β_0 = intercept in the formal model
β_1 = slope in the formal model

All of the data are again presented in Figure 6-1, where the median regression line $Q_{50}(H_\beta)$ is the red line and a vertical black line denotes a gamma dose of 100 mrem. The median regression line is calculated with the quantile regression functions in software such as R and SAS and is not readily calculated with spreadsheets. In this case the mean and median regression lines are not vastly different because the mean of log transformed data is the median of the original data.

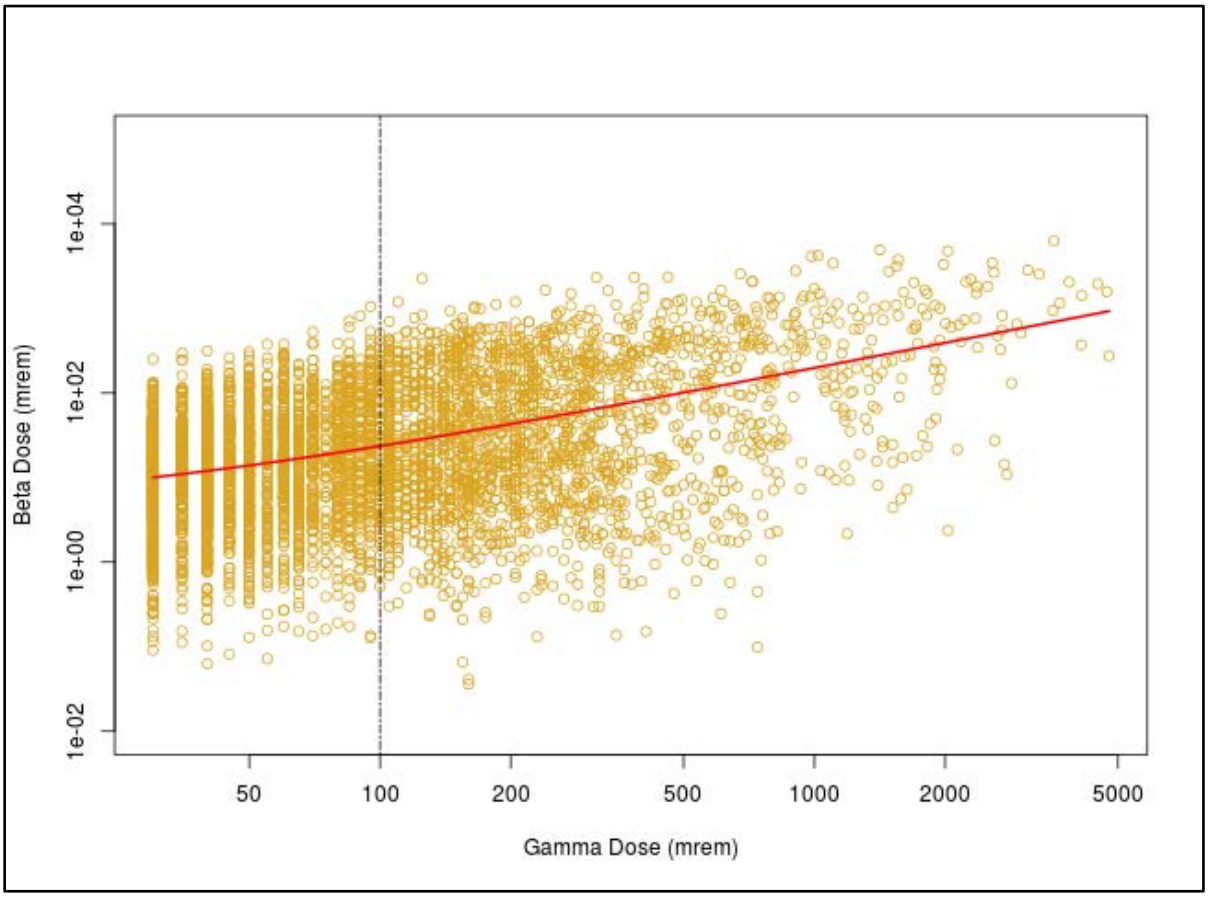

Figure 6-1. Log-log scatter plot of beta doses versus gamma doses. The red line is the quantile regression median line.

Figure 6-2 is the quantile regression version of Figure 5-2. The intercepts and slopes for all of the lines in Figure 6-2 are given in Table 6-1.

The percentiles are not straight lines on the log-log scale because quantile regression ignores the *y*-axis log transform[5] and the *x*-axis was used on a linear scale by choice. These percentile lines are not necessarily symmetrically distributed about the median line, and each percentile line has its own unique intercept and slope. This means that there can, for example, be one relationship (e.g., linear) for the 95th percentile and another (e.g., quadratic or log-linear) for the 5th percentile. This is not possible with OLS.

[5] For example, if 10 rem is the 5th-largest beta dose on a linear scale, it is also the 5th-largest beta dose on a logarithmic scale.

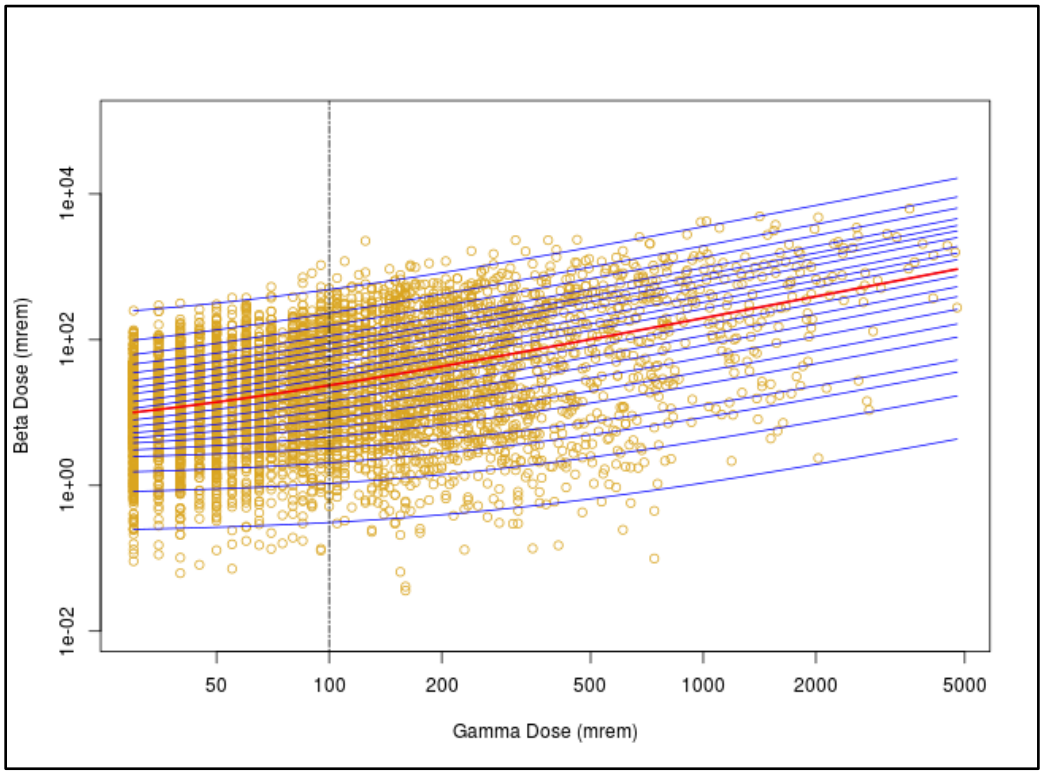

Figure 6-2. Log-log scatter plot of beta doses versus gamma doses. The red line is the quantile regression median line and the blue lines are the 1st through 99th percentiles.

Table 6-1. Intercepts and slopes for percentile lines in Figure 6-2.

Percentile	Intercept	Slope
1	0.223637	0.0008584
5	0.7252814	0.0033916
10	1.3198709	0.0072302
15	2.1775352	0.0105102
20	2.4113781	0.0221591
25	2.8007314	0.0337889
30	2.8612261	0.0539091
35	2.836663	0.0806211
40	3.1505272	0.1104455
45	3.270576	0.1552941
50	4.2145621	0.1935223
55	3.5794242	0.2609813
60	4.4980374	0.3214902
65	6.2546912	0.3959285
70	6.2693667	0.5248273
75	8.0067422	0.6483762
80	11.9819717	0.7683945
85	17.6612903	0.9516129
90	22.5053533	1.3297645
95	41.4285714	1.8928571
99	147.3612116	3.4008423

These 21 equations convey all of the information that is needed to construct point and distributional estimates of the beta dose. As alluded to in Section 5.1, an important distinction between these equations and those for OLS is that all of the equations there are based on the mean regression line whereas all of the equations here represent separate independent regressions. This makes the quantile regression more flexible than OLS and able to better fit the quantiles because the quantile regression lines can take forms different from each other as needed. The scope of this report is limited to the calculation and reporting of the 21 equations and does not specifically address how this information is used in practice. However, standard applications are given in the next two sections.

6.1 POINT ESTIMATE

To calculate the 95th-percentile beta dose in mrem at 100-mrem gamma dose, take the equation for the 95th-percentile line and calculate as in Section 5.2 for the OLS regression:

$$\widehat{H_{\beta 95}} = 41.4285714 + 1.8928571 H_{\gamma} \tag{6-2}$$

Therefore, the predicted 95th-percentile beta dose in mrem for a 100-mrem gamma dose is:

$$\widehat{H_{\beta 95}} = 41.4285714 + 1.8928571(100) = 230.714 \tag{6-3}$$

6.2 DISTRIBUTION ESTIMATE

Figure 6-3 is the empirical CDF from quantile regression and is analogous to the log-normal CDF in Figure 5-5. The CDF is used by drawing a uniform random number between 0.01 and 0.99[6], locating that value on the *y*-axis of Figure 6-3, and reading across to the line to obtain a random beta dose.

[6] Depending on the dataset, it can be difficult to obtain quantile regression fits to extreme quantiles. In this dataset it was feasible to fit from the 1st to the 99th percentile. In a larger dataset it might be possible to extend this to, for example, the 0.1th percentile to the 99.9th percentile.

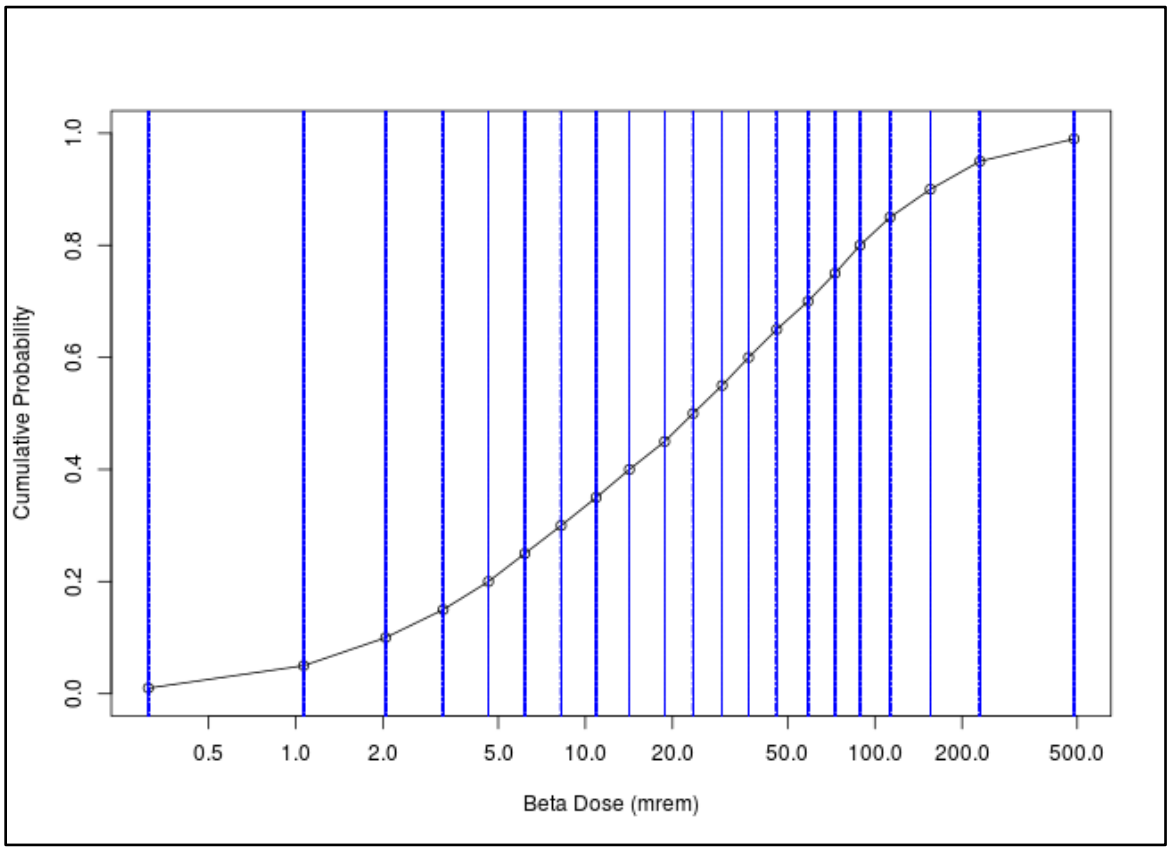

Figure 6-3. Empirical CDF of beta dose for a gamma dose of 100 mrem calculated with quantile regression.

This process is repeated to generate a large number of simulated beta doses, which are then analyzed as discussed in Section 5.4. Using the lognormal ROS approach, the simulated beta doses can be modeled with a lognormal distribution with *GM* = 20.258 and *GSD* = 5.208. The *GM* and *GSD* that are calculated here with quantile regression are very similar to those that are calculated in Section 4.2 using ROS and in Section 5.3 using OLS.

7.0 GOODNESS-OF-FIT TEST

The quality of the fit for quantile plots like the one shown in Figure 6-2 can be difficult to judge by eye, especially when a regression line is on the edge of the data rather than in the middle. A goodness-of-fit (GOF) test is used here to judge how well a given quantile regression model fit the observed data (this test can also be applied to the residuals from OLS regression). This test is based on an analysis of the residuals *r*, which are defined here to be the observed beta doses minus predicted beta doses:

$$r = H_{\beta} - \widehat{H_{\beta}} \tag{7-1}$$

As an example, the residuals for the 95th-percentile line are:

$$r = H_{\beta} - \left(41.4285714 + 1.8928571 H_{\gamma}\right) \tag{7-2}$$

The residual plot in Figure 7-1 is not very revealing as to whether or not the fit is good. Nevertheless, it is encouraging in that about 95% of the residuals are negative (that is, 95% of the observed beta doses are below the 95th-percentile regression line).

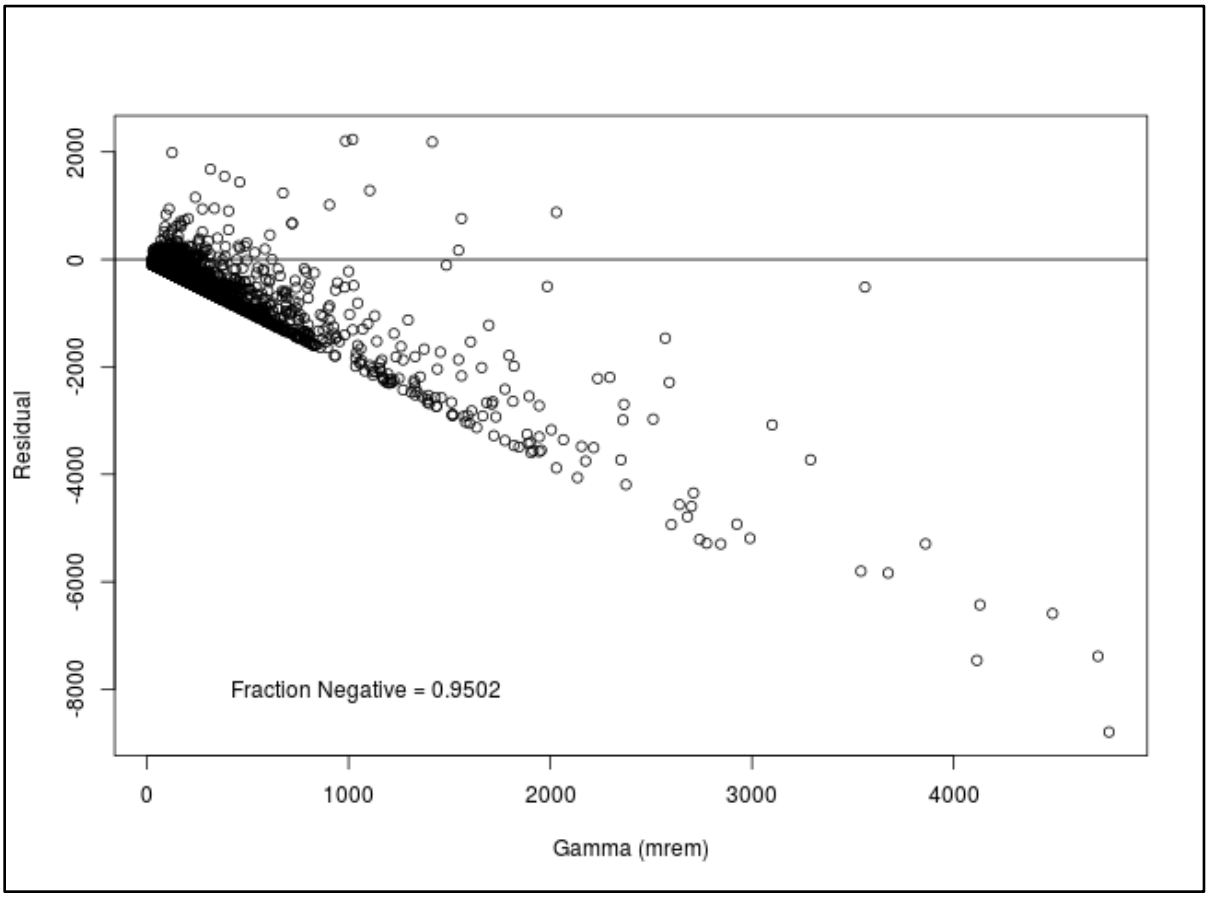

Figure 7-1. Raw residuals (observed beta dose minus the predicted 95th-percentile beta dose) for the 95th-percentile quantile regression.

If the 95th-percentile regression line fits the data, then by definition the probability of observing a negative residual conditioned on the gamma dose should be a constant 0.95. The binary residuals in Figure 7-2 indicate whether the residuals are above the regression line (assigned a value of 0) or below the regression line (assigned a value of 1).

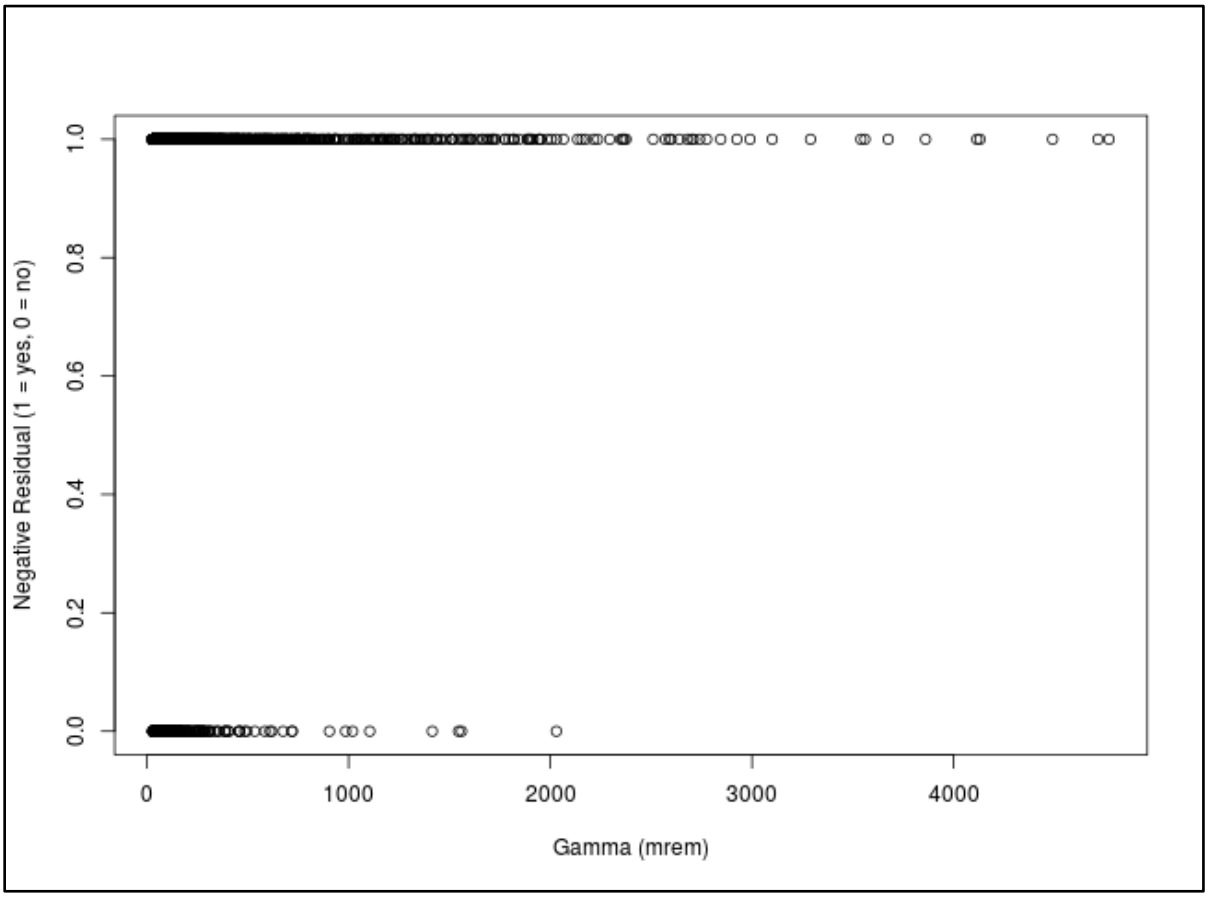

Figure 7-2. Binary residuals for the 95th-percentile quantile regression; 1 indicates a negative residual and 0 indicates a positive residual.

Again, this is not very revealing by itself, but now a logistic regression of the binary residuals on gamma dose can calculate the probability of a negative residual as a function of gamma dose. If the quantile regression fits the data well, this probability should be a constant 0.95 (i.e., the slope of the logistic regression should not be significantly different from zero). The deviance test of the null hypothesis that the slope is equal to zero (Hosmer, Lemeshow and Sturdivant, p. 10; Faraway, p. 32) returns a *p* value of 0.9632, which is taken to mean that the slope is not significantly different from zero at a significance level of 0.05. The GOF plot in Figure 7-3 summarizes everything, including the 95% confidence interval on the predicted probability, and shows that the quantile regression of the 95th percentile beta dose on gamma dose fits well.

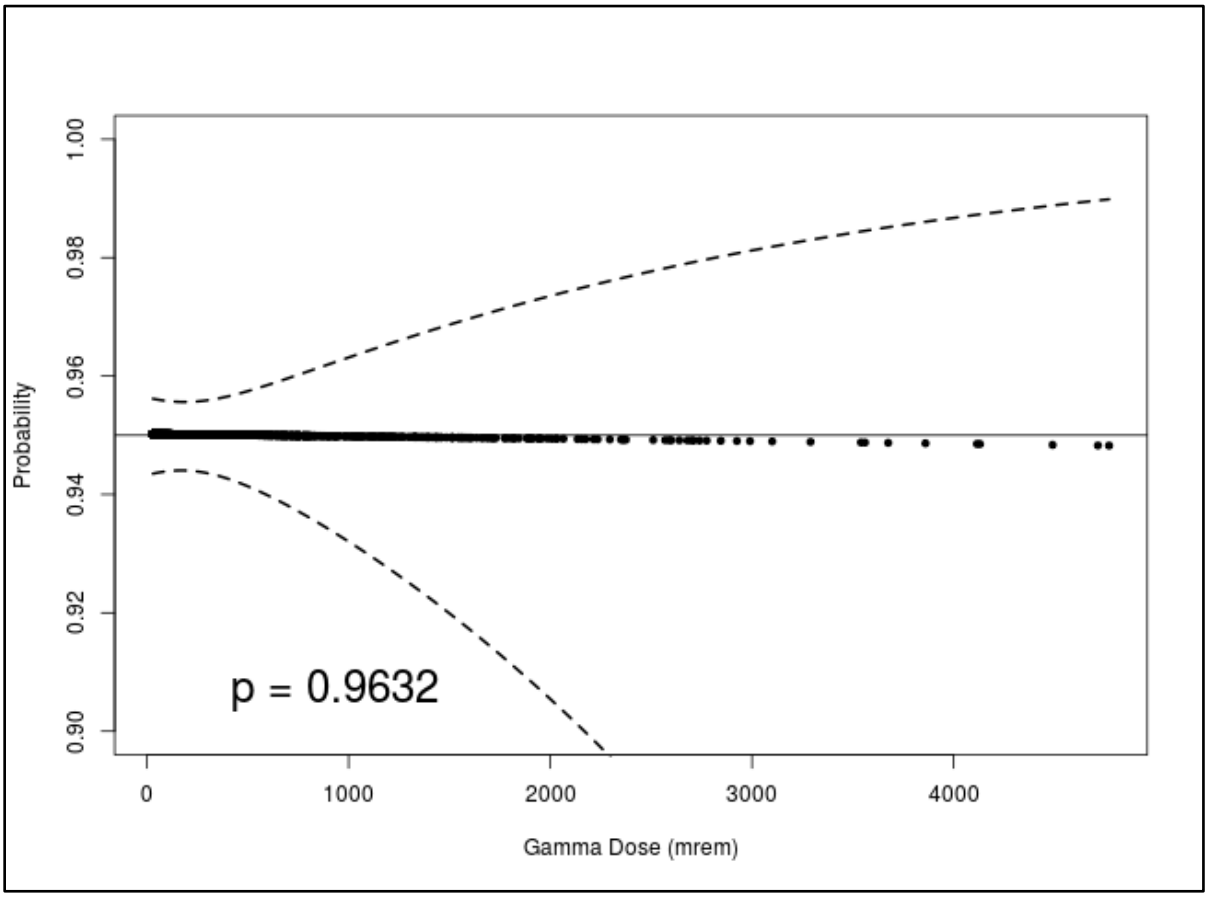

Figure 7-3. Results of logistic regression of binary residuals on gamma dose using the model in Equation 6-1. If the quantile regression adequately fits the 95th percentile of the beta doses, the probability (*y*-axis value) should be 0.95 for all gamma doses (i.e., the line should have a slope of zero). The value of *p* is the result of the deviance test of the null hypothesis that the slope of the line equals zero (the null is not rejected).

The wide 95% confidence bands reflect the fact that there are not many results in the region of the 95th percentile, especially at higher gamma doses. The confidence bands are much tighter for quantiles such as the 50th percentile that lie in the middle of the data rather than on the edge.

7.1 EXAMPLE OF A BAD FIT

To show what a bad fit looks like, fit the model in Equation 7-3 to the complete dataset, which forces the intercept through zero:

$$Q_x\left(H_\beta\right) = \beta_1 H_\gamma \tag{7-3}$$

Here is the quantile regression for the 95th percentile using this model:

$$\widehat{Q_{95}\left(H_\beta\right)} = 2.182 H_\gamma \tag{7-4}$$

The GOF plot as shown in Figure 7-4 shows that the probability of the residual being less than 0.95 is clearly not a constant 0.95, and the null hypothesis that the slope in the logistic regression is equal to zero is rejected at a significance level of $\alpha = 0.05$ with a *p* value of 0.02364. This indicates that the quantile regression of the 95th percentile to the complete dataset using the model in Equation 7-3 has significant lack of fit. Since this test is performed on the residuals, it is independent of the form of the

regression model that generated the residuals. Thus, the same test can be used for the models specified in Equation 6-1 and Equation 7-3.

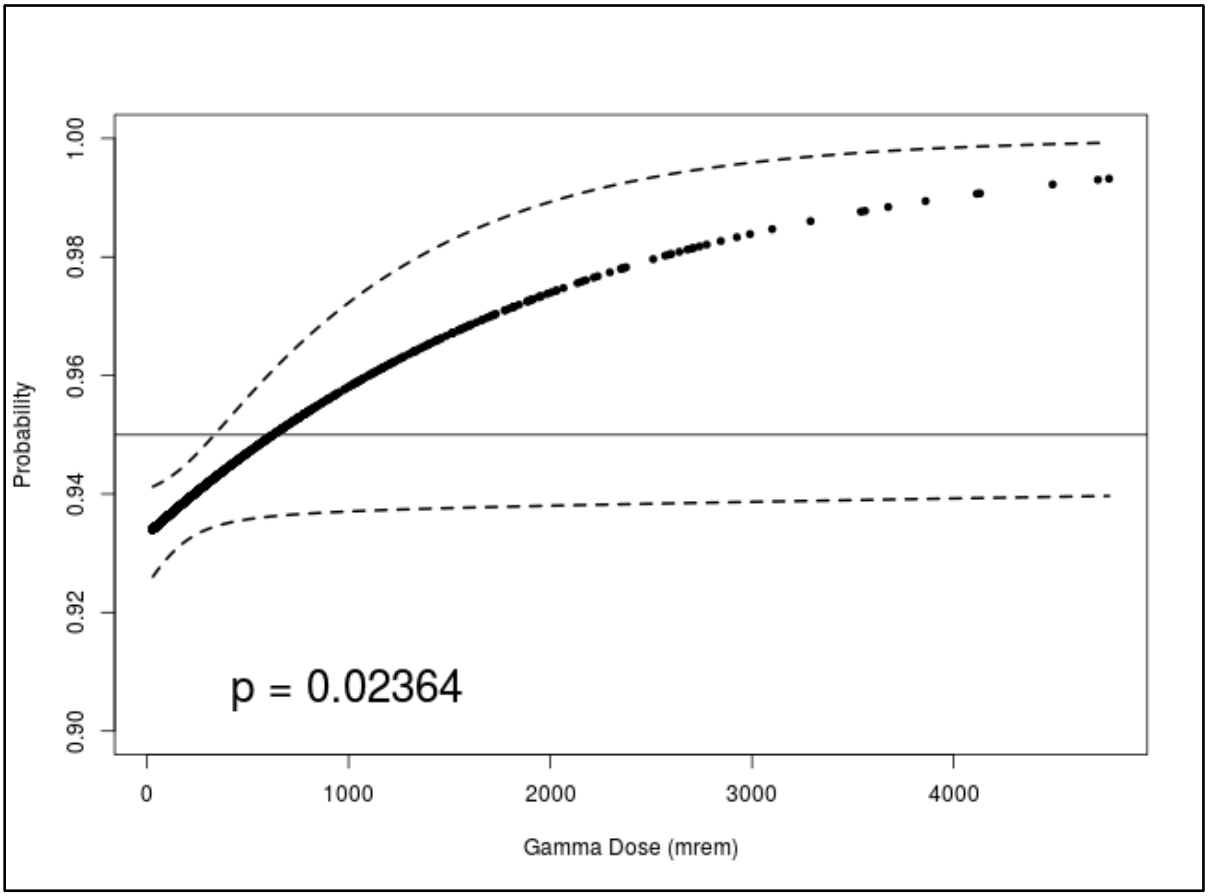

Figure 7-4. Results of logistic regression of binary residuals on gamma dose using the model in Equation 7-3. If the quantile regression adequately fits the 95th percentile of the beta doses, the probability (*y*-axis value) should be 0.95 for all gamma doses (i.e., the line should have a slope of zero). The *p* value is the result of the deviance test of the null hypothesis that the slope of the line equals zero (the null is rejected).

8.0 SUMMARY AND CONCLUSIONS

Historically, the standard Project method has been to take the log of the ratio of beta to gamma dose and regress it on the standard normal quantiles of the data (i.e., lognormal ROS). This procedure gives the GM and GSD of the ratios, which are used to calculate desired quantiles of the beta dose for a given gamma dose along with distributions of beta dose for given distributions of gamma dose. This approach collapses a bivariate distribution (beta versus gamma dose) into a univariate distribution (beta/gamma ratios). There is an inevitable loss of information in this method, but it reduces the complexity of the calculations enough to make them tractable with spreadsheets. The other main problem with lognormal ROS in this application is that the ratios are often not well fit by a lognormal distribution and no formal assessment of GOF test is performed. This practice is probably the result of, historically, no suitable alternative to the lognormal ROS having been identified.

This report has presented OLS regression and quantile regression as alternatives to lognormal ROS. These methods regress beta dose directly on gamma dose, which can reveal relationships that are masked by modeling the ratio of the doses. Another advantage of these types of regression is that other predictors in addition to gamma dose can be included in the analysis. OLS regression assumes

that the beta doses are normally distributed around the mean regression line, which simplifies the calculation of quantiles when this assumption is met. Quantile regression does not assume any particular distribution of the data and is therefore more flexible than OLS regression. The cost of this flexibility is added computational complexity; quantile regression is a computationally intensive method available only in statistics software packages such as SAS, R, and Stata. Quantile regression is not commonly available in spreadsheets, which is perhaps its main limitation and the reason it was not considered earlier in the history of the project.

The choice of which method to use depends on the specifics of the dataset and the desired results for the particular application. For this reason, a statistician and subject matter expert should jointly determine the best method based on those specifics.

REFERENCES

Cade, B. S., and B. R. Noon, 2003, "A Gentle Introduction to Quantile Regression for Ecologists," *Frontiers in Ecology and the Environment*, volume 1, number 8, pp. 412–420. [SRDB Ref ID: 167471]

Faraway, J. J., 2006, *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models*, Chapman & Hall/CRC Taylor & Francis Group, Boca Raton, Florida. [SRDB Ref ID: 169800]

Helsel, D. R., 2012, *Statistics for Censored Environmental Data Using Minitab® and R, Second Edition*, John Wiley & Sons, Hoboken, New Jersey. [SRDB Ref ID: 150353]

Hosmer, D. W. Jr., S. Lemeshow, and R. X. Sturdivant, 2013, *Applied Logistic Regression, Third Edition*, John Wiley & Sons, Hoboken, New Jersey. [SRDB Ref ID: 169439]

Koenker, R., 2005, *Quantile Regression*, number 38, Cambridge University Press, Cambridge, Massachusetts. [SRDB Ref ID: 167513]

Krishnamoorthy, K., A. Mallick, and T. Mathew, 2009, "Model-Based Imputation Approach for Data Analysis in the Presence of Non-Detects," *Annals of Occupational Hygiene*, volume 53, number 3, pp. 249–263. [SRDB Ref ID: 146843]

ORAUT (Oak Ridge Associated Universities Team), 2014, *Analysis of Stratified Coworker Datasets*, ORAUT-RPRT-0053, Rev. 02, Oak Ridge, Tennessee, October 8. [SRDB Ref ID: 136245]

ORAUT (Oak Ridge Associated Universities Team), 2015, *External Dose Coworker Methodology*, ORAUT-RPRT-0071, Rev. 00, Oak Ridge, Tennessee, July 2. [SRDB Ref ID: 145135]

ORAUT (Oak Ridge Associated Universities Team), 2017, "Support Files for ORAUT-RPRT-0087 Rev. 00 - Applications of Quantile Regression in External Dose Reconstruction," Oak Ridge, Tennessee, September 21. [SRDB Ref ID: 167514]

Weisberg, S., 2005, *Applied Linear Regression, Third Edition*, Wiley-Interscience, New York, New York. [SRDB Ref ID: 169440]