Appendix S1.Supporting Information.Dosimetric error model and other statistical details

In this Appendix we outline the two regression calibration methods for dose error correction used in the paper. The first is an adaptation of the method presented in the paper of Kukush*et al.* [13]for dealing with dose measurement error in the Ukrainian-US Thyroid Screening Study. The method outlined differs from that of Kukush*et al.* [13] in using a sum of log-normal distributions to model the underlying quasi-true dose distribution rather than the single log-normal (or piecewise-constant) distribution assumed by Kukush*et al.* [13]. We contrast this with a very similar method that makes slightly more stringent distributional assumptions. In the main paper we present results of fitting the latest (2010) Ukrainian-US data using these two types of regression-calibration adjustment. In the equations that follow we distinguish those equations corresponding exclusively to the second method (SS1), (SS2), ..., to distinguish them from (S1), (S2), ..., corresponding to the adapted method of Kukush*et al.* [13] or the second method where this does not differ.

According to the dosimetric model described in Likhtarev*et al.*[16], the calculated thyroid dose of person i is expressed as:

$$D_i^{ins} = f_i \frac{Q_i^{mes}}{M_i^{est}} \qquad \text{(S1)}$$

where Q_i^{mes} is the measured content of $^{131}\mathrm{I}$ in the thyroid gland of person i at time t_{mes}, M_i^{est} is the estimate of the thyroid mass, and f_i is a multiplier which takes into account the parameter values of an ecology-metabolism model for person i. While f_i is also measured with error, empirical evaluations of its variation suggest that its error is for most individuals much smaller than the errors in Q_i^{mes} and in M_i^{est} (see Likhtarev*et al.* [16]) and we shall ignore this component in the analysis. In principle, the uncertainty in f_i can be included in the evaluation in many

ways, using, for example, the methods of Stram and Kopecky[12]. It is known that most of the uncertainty in f_i is associated with unshared error, the result of individual behavior of the cohort members. We outline a method similar in principle to the likelihood integration method of Fearn*et al.*[14], which was developed with application to the residential radon case-control studies. It has elements in common with the method of Stram and Kopecky[12].

The uncertainties in Q_i^{mes} and M_i^{est} are as follows:

• The measured activity is associated with a multiplicative error, which is determined by the characteristics of the measuring instrument [55,56], so that:

$$Q_i^{mes} = Q_i^{tr} V_i^Q \tag{S2}$$

where Q_i^{tr} is the true ^{131}I content in the thyroid gland, and V_i^Q is the independent multiplicative measurement error. If we assume that V_i^Q is log-normally distributed, so that $\ln(V_i^Q) \sim N(0, {\sigma_{Q,i}}^2)$, then the conditional pdf $Q_i^{mes} \mid Q_i^{tr}$ is:

$$P[Q_i^{mes} \mid Q_i^{tr}] = \frac{1}{\sigma_{Q,i}\sqrt{2\pi} Q_i^{mes}} \exp\left[-\frac{[\ln(Q_i^{mes}) - \ln(Q_i^{tr})]^2}{2{\sigma_{Q,i}}^2} \right] \tag{S3}$$

Expressions (S2)-(S3) of course define a classical multiplicative error model.

• The true values of the thyroid mass M_i^{tr} are determined according to a Berkson measurement error model as:

$$M_i^{tr} = M_i^{mes} V_i^M \tag{S4}$$

where M_i^{mes} is the median of the thyroid mass for a given gender-age group, and V_i^M an independent multiplicative error. The measured median thyroid mass values, M_i^{mes}, by age and oblast, are those used by Likhtarev*et al.*[15]. If we assume that V_i^M is log-normally distributed, so

that $\ln(V_i^M) \sim N(0, {\sigma_{M,i}}^2)$, then the conditional pdf $M_i^{tr} \mid M_i^{mes}$ is:

$$P[M_i^{tr} \mid M_i^{mes}] = \frac{1}{\sigma_{M,i}\sqrt{2\pi} M_i^{tr}} \exp\left[-\frac{[\ln(M_i^{tr}) - \ln(M_i^{mes})]^2}{2{\sigma_{M,i}}^2} \right] \quad \text{(S5)}$$

From this it follows that:

$$E\left[\frac{1}{M_i^{tr}} \middle| M_i^{mes} \right] = E\left[\frac{1}{M_i^{mes} V_i^M} \middle| M_i^{mes} \right] = \frac{1}{M_i^{mes}} E\left[\frac{1}{V_i^M} \middle| M_i^{mes} \right] = \frac{1}{M_i^{mes}} \exp\left[\frac{{\sigma_{M,i}}^2}{2} \right] \quad \text{(S6)}$$

As in Kukush*et al.* [13] we assume that the random variables Q_i^{tr} / M_i^{mes} and V_i^Q and V_i^M are jointly independent, and we define $M_i^{est} = M_i^{mes}$. This is strictly weaker than the assumption made by the US Working Group (USWG), that the random variables $(Q_i^{tr}, V_i^Q, M_i^{mes}, V_i^M)$ are jointly independent. In the Ukrainian-US study, the parameters $\sigma_{Q,i}^2$ and $\sigma_{M,i}^2$ have been reasonably reliably estimated from various sources [16] (although this is not the case for thyroid mass measurements outside the age range 5-15, for which more or less subjective assignments were made), and therefore we assume that these variables are known.

The (unobservable) error-free dose is denoted by D_i^{tr} and is given by:

$$D_i^{tr} = f_i Q_i^{tr} / M_i^{tr} \quad \text{(S7)}$$

Based on the analysis of real samples taken from an epidemiological study [16], we provisionally assume that the logarithm of the quasi-true dose $Y_i = f_i Q_i^{tr} / M_i^{mes}$ (it is not the true dose because of the presence of M_i^{mes} rather than M_i^{tr}) as a distribution which can be represented by a mixture of K normal distributions, and hence can be written as:

$$\ln(f_i Q_i^{tr} / M_i^{mes}) \sim \sum_{k=1}^{K} p_k N(\ln(\mu_k), \sigma_k^2) \quad \text{(S8)}$$

where $\sum_{k=1}^{K} p_k = 1$. It should be noted that in general all the parameters, in particular K, $(p_k)_{k=1}^{K}$, $(\mu_k)_{k=1}^{K}$. $(\sigma_k^2)_{k=1}^{K}$ are unknown, and must be estimated. We outline a bit later the likelihood-based method by which this was done. Equivalently, the pdf of the quasi-true dose $Y_i = f_i Q_i^{tr} / M_i^{mes}$ is given by:

$$P[Y_i] = \sum_{k=1}^{K} \frac{p_k}{\sigma_k \sqrt{2\pi} Y_i} \exp\left[-\frac{[\ln(Y_i) - \ln(\mu_k)]^2}{2\sigma_k^{\,2}} \right] \tag{S9}$$

Notice that:

$$D_i^{mes} = f_i Q_i^{mes} / M_i^{mes} = (f_i Q_i^{tr} / M_i^{mes}) V_i^Q = Y_i V_i^Q \tag{S10}$$

Based on (S9) and (S10) the joint pdf of (D_i^{mes}, Y_i) is given by:

$$P[D_i^{mes}, Y_i] = \frac{1}{\sigma_{Q,i} 2\pi D_i^{mes} Y_i} \exp\left[-\frac{[\ln(D_i^{mes}) - \ln(Y_i)]^2}{2\sigma_{Q,i}^{\,2}} \right] \mathrm{x} \sum_{k=1}^{K} \frac{p_k}{\sigma_k} \exp\left[-\frac{[\ln(Y_i) - \ln(\mu_k)]^2}{2\sigma_k^{\,2}} \right] \tag{S11}$$

From (S10) this implies that the unconditional distribution of measured dose, D_i^{mes}, has pdf:

$$P[D_i^{mes}] = \frac{1}{\sqrt{2\pi} D_i^{mes}} \sum_{k=1}^{K} \frac{p_k}{\sqrt{\sigma_{Q,i}^{\,2} + \sigma_k^{\,2}}} \exp\left[-\frac{[\ln(D_i^{mes}) - \ln(\mu_k)]^2}{2[\sigma_{Q,i}^{\,2} + \sigma_k^{\,2}]} \right] \tag{S12}$$

i.e., a weighted sum of log-normal densities. From this expression the various unknown population parameters ($(\mu_k, \sigma_k)_{k=1}^{K}$) can be derived via maximum likelihood techniques, based on the distribution of observed dose in the sample and knowledge of the error distribution log standard deviation, $\sigma_{Q,i}$. By comparison, the second method assumes that the logarithm of Q_i^{tr} has a distribution which is a sum of normal distributions, so that the pdf of Q_i^{tr} is given by:

$$P[Q_i^{tr}] = \sum_{k=1}^{K} \frac{p_k}{\sigma_k \sqrt{2\pi} Q_i^{tr}} \exp\left[-\frac{[\ln(Q_i^{tr}) - \ln(\mu_k)]^2}{2\sigma_k{}^2} \right] \quad \text{(SS9)}$$

Based on (S3) and (SS9) the joint pdf of (Q_i^{tr}, Q_i^{mes}) is given by:

$$P[Q_i^{tr}, Q_i^{mes}] = \frac{1}{\sigma_{Q,i} 2\pi Q_i^{mes} Q_i^{tr}} \exp\left[-\frac{[\ln(Q_i^{mes}) - \ln(Q_i^{tr})]^2}{2\sigma_{Q,i}{}^2} \right] \text{x} \sum_{k=1}^{K} \frac{p_k}{\sigma_k} \exp\left[-\frac{[\ln(Q_i^{tr}) - \ln(\mu_k)]^2}{2\sigma_k{}^2} \right] \quad \text{(SS11)}$$

The second method (expression (SS11)) then implies that the unconditional distribution of measured activity has pdf:

$$P[Q_i^{mes}] = \frac{1}{\sqrt{2\pi} Q_i^{mes}} \sum_{k=1}^{K} \frac{p_k}{\sqrt{\sigma_{Q,i}{}^2 + \sigma_k{}^2}} \exp\left[-\frac{[\ln(Q_i^{mes}) - \ln(\mu_k)]^2}{2[\sigma_{Q,i}{}^2 + \sigma_k{}^2]} \right] \quad \text{(SS12)}$$

i.e., a weighted sum of log-normal densities. Formally, apart from a change of notation (Q_i^{mes} for D_i^{mes}, Q_i^{tr} for Y_i), these are the same equations as (S9), (S11) and (S12). As above, from (SS12) the various unknown population parameters ($(\mu_k, \sigma_k)_{k=1}^{K}$) can be derived via maximum likelihood techniques, based on the distribution of observed activity measurements in the sample. From (S11) and (S12) we derive the conditional distribution of $Y_i = f_i Q_i^{tr} / M_i{}^{mes}$ given D_i^{mes}, which has pdf:

$$P[Y_i \mid D_i^{mes}] = \frac{\sum_{k=1}^{K} \left\{ \begin{array}{l} \dfrac{\sqrt{\sigma_k{}^2 + \sigma_{Q,i}{}^2}}{Y_i \sigma_k \sigma_{Q,i} \sqrt{2\pi}} \exp\left[- \dfrac{\left[\ln(Y_i) - \dfrac{[\ln(\mu_k)\sigma_{Q,i}{}^2 + \ln(D_i^{mes})\sigma_k{}^2]}{[\sigma_k{}^2 + \sigma_{Q,i}{}^2]} \right]^2}{2\left(\sigma_k{}^2 \sigma_{Q,i}{}^2 / [\sigma_k{}^2 + \sigma_{Q,i}{}^2]\right)} \right] \text{x} \\ \dfrac{p_k}{D_i^{mes} \sqrt{\sigma_k{}^2 + \sigma_{Q,i}{}^2} \sqrt{2\pi}} \exp\left[- \dfrac{[\ln(D_i^{mes}) - \ln(\mu_k)]^2}{2[\sigma_k{}^2 + \sigma_{Q,i}{}^2]} \right] \end{array} \right\}}{\sum_{k=1}^{K} \left\{ \dfrac{p_k}{D_i^{mes} \sqrt{\sigma_k{}^2 + \sigma_{Q,i}{}^2} \sqrt{2\pi}} \exp\left[- \dfrac{[\ln(D_i^{mes}) - \ln(\mu_k)]^2}{2[\sigma_k{}^2 + \sigma_{Q,i}{}^2]} \right] \right\}} \qquad \text{(S13)}$$

In particular:

$$E[Y_i \mid D_i^{mes}] = \frac{\sum_{k=1}^{K} \left\{ \begin{array}{l} \exp\left[\dfrac{[\ln(\mu_k)\sigma_{Q,i}{}^2 + \ln(D_i^{mes})\sigma_k{}^2] + 0.5\sigma_k{}^2 \sigma_{Q,i}{}^2}{[\sigma_k{}^2 + \sigma_{Q,i}{}^2]} \right] \text{x} \\ \dfrac{p_k}{\sqrt{\sigma_k{}^2 + \sigma_{Q,i}{}^2}} \exp\left[- \dfrac{[\ln(D_i^{mes}) - \ln(\mu_k)]^2}{2[\sigma_k{}^2 + \sigma_{Q,i}{}^2]} \right] \end{array} \right\}}{\sum_{k=1}^{K} \left\{ \dfrac{p_k}{\sqrt{\sigma_k{}^2 + \sigma_{Q,i}{}^2}} \exp\left[- \dfrac{[\ln(D_i^{mes}) - \ln(\mu_k)]^2}{2[\sigma_k{}^2 + \sigma_{Q,i}{}^2]} \right] \right\}} \qquad \text{(S14)}$$

By contrast, the second method ((SS9) and (SS11)) implies that the conditional distribution of the true activity given the measured has pdf:

$$P[Q_i^{tr} \mid Q_i^{mes}] = \frac{\sum_{k=1}^{K} \left\{ \begin{array}{l} \frac{\sqrt{\sigma_k{}^2 + \sigma_{Q,i}{}^2}}{Q_i^{tr}\sigma_k\sigma_{Q,i}\sqrt{2\pi}} \exp\left[-\frac{\left[\ln(Q_i^{tr}) - \frac{[\ln(\mu_k)\sigma_{Q,i}{}^2 + \ln(Q_i^{mes})\sigma_k{}^2]}{[\sigma_k{}^2 + \sigma_{Q,i}{}^2]} \right]^2}{2\left(\sigma_k{}^2\sigma_{Q,i}{}^2 / [\sigma_k{}^2 + \sigma_{Q,i}{}^2]\right)} \right] \text{x} \\ \frac{p_k}{Q_i^{mes}\sqrt{\sigma_k{}^2 + \sigma_{Q,i}{}^2}\sqrt{2\pi}} \exp\left[-\frac{[\ln(Q_i^{mes}) - \ln(\mu_k)]^2}{2[\sigma_k{}^2 + \sigma_{Q,i}{}^2]} \right] \end{array} \right\}}{\sum_{k=1}^{K} \left\{ \frac{p_k}{Q_i^{mes}\sqrt{\sigma_k{}^2 + \sigma_{Q,i}{}^2}\sqrt{2\pi}} \exp\left[-\frac{[\ln(Q_i^{mes}) - \ln(\mu_k)]^2}{2[\sigma_k{}^2 + \sigma_{Q,i}{}^2]} \right] \right\}} \quad \text{(SS13)}$$

It then follows that:

$$E[Q_i^{tr} \mid Q_i^{mes}] = \frac{\sum_{k=1}^{K} \left\{ \begin{array}{l} \exp\left[\frac{\ln(\mu_k)\sigma_{Q,i}{}^2 + \ln(Q_i^{mes})\sigma_k{}^2 + 0.5\sigma_k{}^2\sigma_{Q,i}{}^2}{[\sigma_k{}^2 + \sigma_{Q,i}{}^2]} \right] \text{x} \\ \frac{p_k}{\sqrt{\sigma_k{}^2 + \sigma_{Q,i}{}^2}} \exp\left[-\frac{[\ln(Q_i^{mes}) - \ln(\mu_k)]^2}{2[\sigma_k{}^2 + \sigma_{Q,i}{}^2]} \right] \end{array} \right\}}{\sum_{k=1}^{K} \left\{ \frac{p_k}{\sqrt{\sigma_k{}^2 + \sigma_{Q,i}{}^2}} \exp\left[-\frac{[\ln(Q_i^{mes}) - \ln(\mu_k)]^2}{2[\sigma_k{}^2 + \sigma_{Q,i}{}^2]} \right] \right\}} \quad \text{(SS14)}$$

Again, formally, apart from a change of notation (Q_i^{mes} for D_i^{mes}, Q_i^{tr} for Y_i), these are the same equations as (S13) and (S14). Notice that:

$$D_i^{tr} = f_i Q_i^{tr} / M_i^{tr} = (f_i Q_i^{tr} / M_i^{mes}) / V_i^M = Y_i / V_i^M \quad \text{(S15)}$$

By assumption, Y_i, V_i^M are independent, so that by (S6):

$$E[D_i^{tr} \mid D_i^{mes}] = E[Y_i / V_i^M \mid D_i^{mes}] = E[Y_i \mid D_i^{mes}]E[1/V_i^M]$$

$$= \exp\left[\frac{{\sigma_{M,i}}^2}{2}\right] \frac{\sum_{k=1}^{K}\left\{\exp\left[\frac{\ln(\mu_k){\sigma_{Q,i}}^2 + \ln(D_i^{mes}){\sigma_k}^2 + 0.5{\sigma_k}^2{\sigma_{Q,i}}^2}{[{\sigma_k}^2 + {\sigma_{Q,i}}^2]}\right] \mathrm{x} \right.}{\left. \frac{p_k}{D_i^{mes}\sqrt{{\sigma_k}^2 + {\sigma_{Q,i}}^2}\sqrt{2\pi}} \exp\left[-\frac{[\ln(D_i^{mes}) - \ln(\mu_k)]^2}{2[{\sigma_k}^2 + {\sigma_{Q,i}}^2]}\right]\right\}}$$
$$\frac{}{\sum_{k=1}^{K}\left\{\frac{p_k}{D_i^{mes}\sqrt{{\sigma_k}^2 + {\sigma_{Q,i}}^2}\sqrt{2\pi}} \exp\left[-\frac{[\ln(D_i^{mes}) - \ln(\mu_k)]^2}{2[{\sigma_k}^2 + {\sigma_{Q,i}}^2]}\right]\right\}}$$

$$= \exp\left[\frac{{\sigma_{M,i}}^2}{2}\right] \frac{\sum_{k=1}^{K}\left\{\exp\left[\frac{\ln(\mu_k){\sigma_{Q,i}}^2 + \ln(D_i^{mes}){\sigma_k}^2 + 0.5{\sigma_k}^2{\sigma_{Q,i}}^2}{[{\sigma_k}^2 + {\sigma_{Q,i}}^2]}\right] \mathrm{x} \right.}{\left. \frac{p_k}{\sqrt{{\sigma_k}^2 + {\sigma_{Q,i}}^2}} \exp\left[-\frac{[\ln(D_i^{mes}) - \ln(\mu_k)]^2}{2[{\sigma_k}^2 + {\sigma_{Q,i}}^2]}\right]\right\}}$$
$$\frac{}{\sum_{k=1}^{K}\left\{\frac{p_k}{\sqrt{{\sigma_k}^2 + {\sigma_{Q,i}}^2}} \exp\left[-\frac{[\ln(D_i^{mes}) - \ln(\mu_k)]^2}{2[{\sigma_k}^2 + {\sigma_{Q,i}}^2]}\right]\right\}} \qquad \text{(S16)}$$

By contrast, the second method estimates:

$$E[D_i^{tr} \mid Q_i^{mes}, M_i^{mes}] = E[f_i Q_i^{tr} / M_i^{tr} \mid Q_i^{mes}, M_i^{mes}]$$
$$= f_i E[Q_i^{tr} \mid Q_i^{mes}]E[1/M_i^{tr} \mid M_i^{mes}] \qquad \text{(SS15)}$$

(making use of the independence of $(Q_i^{tr}, V_i^Q, M_i^{mes}, V_i^M)$ in the second equality) and (S6), (SS14) and (SS15) therefore imply that the conditional expectation of the true dose given all the measured quantities (Q_i^{mes}, M_i^{mes}) is:

$$E[D_i^{tr} \mid Q_i^{mes}, M_i^{mes}] = f_i \frac{\exp\left[\frac{{\sigma_{M,i}}^2}{2}\right]}{M_i^{mes}} \frac{\sum_{k=1}^{K}\left\{\exp\left[\frac{\ln(\mu_k){\sigma_{Q,i}}^2 + \ln(Q_i^{mes}){\sigma_k}^2 + 0.5{\sigma_k}^2{\sigma_{Q,i}}^2}{[{\sigma_k}^2 + {\sigma_{Q,i}}^2]}\right] \mathrm{x} \right.}{\left. \frac{p_k}{\sqrt{{\sigma_k}^2 + {\sigma_{Q,i}}^2}} \exp\left[-\frac{[\ln(Q_i^{mes}) - \ln(\mu_k)]^2}{2[{\sigma_k}^2 + {\sigma_{Q,i}}^2]}\right]\right\}}$$
$$\frac{}{\sum_{k=1}^{K}\left\{\frac{p_k}{\sqrt{{\sigma_k}^2 + {\sigma_{Q,i}}^2}} \exp\left[-\frac{[\ln(Q_i^{mes}) - \ln(\mu_k)]^2}{2[{\sigma_k}^2 + {\sigma_{Q,i}}^2]}\right]\right\}}$$

(SS16)

Note: Kukush*et al.* [13] state that "the sample correlation between $\log(Q_i^{tr})$ and $\log(M_i^{mes})$ is equal to 0.26." In that paper for the underlying radio-epidemiological study, the sample correlation between $\ln(Q_i^{mes})$ and $\ln(M_i^{mes})$ was reported, because in simulations made in Kukush*et al.*[13], the actual values of Q_i^{mes} were taken as Q_i^{tr} and classical multiplicative error was imposed on those values. In our data, we found no evidence for correlation between Q_i^{mes} and M_i^{mes}: we estimated the Pearson correlation coefficient between them to be about -0.05, implying that they are largely uncorrelated. As such, there is no evidence to invalidate use of the second regression calibration model, although the first model, which does not make such strong assumptions must be regarded as *a priori* more plausible.

Fitting of thyroid cancer risk model to data

An excess odds ratio (EOR) model was employed, in which the probability of thyroid cancer for an individual with age a at screening, age e at exposure, with true thyroid dose $D_i^{tr} = f_i Q_i^{tr} / M_i^{tr}$ of gender s is given by:

$$P(a,e,s,D \mid \alpha,\beta,\gamma,\kappa,\tau,\psi) = \frac{\exp\left[\beta_0 + \beta_s 1_{s=male} + \sum_{k=1}^{N} \beta_k 1_{a_{k-1}\le a<a_k}\right]\left[1+\alpha D \exp\left(\gamma D + \kappa[e-8] + \tau[a-22] + \psi[a-e-14] + \eta 1_{s=male}\right)\right]}{1+\exp\left[\beta_0 + \beta_s 1_{s=male} + \sum_{k=1}^{N} \beta_k 1_{a_{k-1}\le a<a_k}\right]\left[1+\alpha D \exp\left(\gamma D + \kappa[e-8] + \tau[a-22] + \psi[a-e-14] + \eta 1_{s=male}\right)\right]} \quad \text{(S17)}$$

Therefore the likelihood is given by:

$$L(\alpha,\beta,\gamma,\kappa,\tau,\psi \mid a_i,e_i,s_i,D_i^{tr},\delta_i) = L(\alpha,\beta,\gamma,\kappa,\tau,\psi \mid a_i,e_i,s_i,Q_i^{tr},M_i^{tr},\delta_i)$$
$$= \prod_{i=1}^{N} P(a_i,e_i,s_i,D_i^{tr} \mid \alpha,\beta,\gamma,\kappa,\tau,\psi)^{\delta_i}\left[1-P(a_i,e_i,s_i,D_i^{tr} \mid \alpha,\beta,\gamma,\kappa,\tau,\psi)\right]^{1-\delta_i} \quad \text{(S18)}$$

where δ_i is the indicator of whether person i was a case (=1 if so, =0 otherwise). This of course

is a function of unknown variables f_i, M_i^{tr}, Q_i^{tr}. For the reasons given by Fearn*et al.* [14] and Stram and Kopecky[12], it is sensible to consider instead the integrated likelihood, given by:

$$
\begin{aligned}
&E[L \mid M_i^{mes}, Q_i^{mes}, i = 1,..., N] \\
&= E\left[\prod_{i=1}^{N} P(a_i, e_i, s_i, D_i^{tr} \mid \alpha, \beta, \gamma, \kappa, \tau, \psi)^{\delta_i} \left[1 - P(a_i, e_i, s_i, D_i^{tr} \mid \alpha, \beta, \gamma, \kappa, \tau, \psi) \right]^{1-\delta_i} \mid M_i^{mes}, Q_i^{mes}, i = 1,..., N \right] \\
&= \prod_{i=1}^{N} E\left[\prod_{i=1}^{N} P(a_i, e_i, s_i, D_i^{tr} \mid \alpha, \beta, \gamma, \kappa, \tau, \psi)^{\delta_i} \left[1 - P(a_i, e_i, s_i, D_i^{tr} \mid \alpha, \beta, \gamma, \kappa, \tau, \psi) \right]^{1-\delta_i} \mid M_i^{mes}, Q_i^{mes} \right]
\end{aligned}
\tag{S19}
$$

which can be integrated via Monte Carlo sampling from the distributions (S5) and (S13) (or (SS13)) (and noting (S15)). We use Monte Carlo integration in addition to the two regression calibration methods. For the purposes of this paper, we substitute D_i^{tr} by $E[D_i^{tr} \mid D_i^{mes}]$ for the first regression calibration method (adapted from Kukush*et al.* [13]) as outlined above, or D_i^{tr} by $E[D_i^{tr} \mid Q_i^{mes}, M_i^{mes}]$ as in the second method, in (S18). Whether using these two regression calibration approaches or Monte Carlo integration, the model parameters $\alpha, \beta, \gamma, \kappa, \tau$ are estimated via maximizing the logistic likelihood [21]. In general, because of collinearity in the age and temporal variables (age a at screening, age e at exposure, attained age $a+e$), only one of the age or temporal adjustment parameters, κ, τ or ψ was free to vary. The Monte Carlo sample of doses was generated by various of the study team (VMS, IAL).

Preliminary model fits to dose model

As can be seen from Supporting Information Table S1, there is evidence that a combination of three log-normal models are required to fit the measured dose data, using expression (S12). With the addition of any model up to the third there is a significant improvement in fit (p<0.03), but larger numbers of log-normal distributions yield no significant improvement in fit (p>0.5). The log-likelihood is virtually identical for models with three to seven log-normal distributions.

Supporting InformationFigure S3 demonstrates that the dose is distributed very-nearly log-normally. Therefore for all analyses using the first method, a combination of three log-normal distributions were assumed. Likewise, and as can be seen from Supporting Information Table S2, the second method modeled the true activity distribution using a scaled sum of three log-normal distributions. With the addition of any model up to the third there is a significant improvement in fit (p<0.001), but larger numbers of log-normal distributions yield no significant improvement in fit (p>0.1). [It should be noted that this is therefore slightly different from the approach adopted by Kukush*et al.* [13], who only used a single normal distribution. However, the theoretical simulations of Kukush*et al.* [13] were in effect a proof of principle, not requiring modeling of real data.]