Version of 06-11-2015

Original Paper

GM Kendall, R Wakeford, M Athanson, TJ Vincent, EJ Carter, NP McColl, MP Little

Levels of naturally occurring gamma radiation measured in British homes and their prediction in particular residences

GM Kendall (✉)
Cancer Epidemiology Unit, University of Oxford, Richard Doll Building, Old Road Campus, Headington, Oxford, OX3 7LF, UK.
Email: Gerald.Kendall@ceu.ox.ac.uk or Gerald.Kendall@cantab.net
Tel: +44 (0)1865 289600
Fax: +44 (0)1865 289610

R Wakeford
Centre for Occupational and Environmental Health, Institute of Population Health, The University of Manchester, Ellen Wilkinson Building, Oxford Road, Manchester, M13 9PL, UK

M Athanson
Bodleian Library, University of Oxford, Broad Street, Oxford, OX1 3BG, UK

TJ Vincent
Childhood Cancer Research Group, University of Oxford, New Richards Building, Old Road

EJ Carter
Present Address Earth Heritage Trust, Geological Records Centre, University of Worcester, Henwick Grove, Worcester WR2 6AJ.

NP McColl
Centre for Radiation, Chemical and Environmental Hazards, Public Health England, Chilton, Didcot Oxon, OX11 0RQ

MP Little
Radiation Epidemiology Branch, National Cancer Institute, DHHS, NIH, Division of Cancer Epidemiology and Genetics, Bethesda, Maryland 20892-9778, USA.

Abstract Gamma radiation from natural sources (including directly ionising cosmic-rays) is an important component of background radiation. In the present paper indoor measurements of naturally occurring gamma-rays that were undertaken as part of the UK Childhood Cancer Study are summarised, and it is shown that these are broadly compatible with an earlier UK National Survey. The distribution of indoor gamma-ray dose-rates in Great Britain is approximately normal with mean 96 nGy/h and standard deviation 23 nGy/h. Directly ionising cosmic-rays contribute about one-third of the total. The expanded dataset allows a more detailed description than previously of indoor gamma-ray exposures and in particular their geographical variation. Various strategies for predicting indoor natural background gamma-ray dose-rates were explored. In the first of these a geostatistical model was fitted, which assumes an underlying geologically-determined spatial variation, superimposed on which is a Gaussian stochastic process with Matérn correlation structure that models the observed tendency of dose-rates in neighbouring houses to correlate. In the second approach a number of dose-rate interpolation measures were first derived, based on averages over geologically- or administratively-defined areas, or distance-weighted averages of measurements at nearest-neighbour points. Linear regression was then used to derive an optimal linear combination of these interpolation measures. The predictive performances of the two models were compared via cross-validation, using a randomly selected 70% of the data to fit the models, and the remaining 30% to test them. The mean square error (MSE) of the linear-regression model was lower than that of the Gaussian-Matérn model (MSE 378 and 411, respectively). The predictive performance of the two candidate models was also evaluated via simulation; the OLS model performs significantly better than the Gaussian-Matérn model.

Keywords

Gamma radiation, natural background radiation, childhood cancer, leukaemia

Introduction

Exposure to high levels of ionising radiation is a well-established cause of cancer (International Agency for Research on Cancer 2009, 2012) and is one of the few known exogenous risk factors for childhood leukaemia (Belson et al. 2007; United Nations Scientific Committee on the Effects of Atomic Radiation (UNSCEAR) 2008; Wakeford 2013). Evidence comes in particular from data on the survivors of the atomic-bombings of Japan and from other groups receiving moderate or high doses at high dose-rates (United Nations Scientific Committee on the Effects of Atomic Radiation (UNSCEAR) 2008). An important and controversial question is whether the risk estimates obtained from these groups are applicable to the low doses and/or low dose-rates (Muirhead et al. 1993; Wakeford and Tawn 2010) that are the subject of everyday radiological protection.

There have been a number of studies examining potential associations of childhood leukaemia with natural background radiation (Axelson et al. 2002; Richardson et al. 1995; United Kingdom Childhood Cancer Study Investigators 2002a; United Kingdom Childhood Cancer Study Investigators 2002b). However, most have not had adequate statistical power to detect the small increased risk predicted by standard models (Little et al. 2010). Therefore "false negative" findings were to be expected, and even when nominally statistically significant positive associations are found, they are likely to be biased upwards (Land 1980). A recent record-based case-control study in Great Britain (GB, the United Kingdom minus Northern Ireland) of reasonable statistical power (~50%) found a statistically significant association of childhood leukaemia with natural background gamma radiation, with excess risks compatible with those derived from the Japanese atomic-bomb survivor data (Kendall et al. 2013). Exposures from natural background gamma radiation are characterized by low cumulative doses received at very low dose-rates, and the results of this case-control study provide support to the assumption that models of radiation-induced leukaemia risk derived from data observed at moderate and high doses and high dose rates may be appropriately applied to protracted red bone marrow (RBM) gamma-ray doses of about 1 mGy per annum. This is relevant to practical radiological protection in a number of situations, for example radiodiagnostic imaging procedures. The results of the study contradict the idea that there are no adverse radiation effects, or possibly beneficial effects, at these very low doses and dose-rates. It is important to note that this study involves not just low doses but also low dose-rates. Some medical imaging procedures, which have also been the subject of epidemiological studies (Pearce et al. 2012), involve similar doses but incurred briefly, not over years as with natural background radiation.

Naturally occurring gamma radiation is ubiquitous in the environment, including the domestic buildings in which most members of the British population, including children, generally spend the majority of their time. Natural background gamma radiation is usually taken to refer to photons from naturally occurring radioactive nuclides such as ^{40}K and members of the uranium and thorium decay chains. However, in the present context the term is used more generally to include the directly ionising component of cosmic-rays (at ground level, predominantly muons). We do this because measurements normally include both components and their radiological consequences may be taken to be similar (International Commission on Radiological Protection 2007).

The outdoor dose-rate is determined by the level of the relevant radionuclides in the environment and by the cosmic-ray dose-rate. The latter depends on altitude and latitude, but variations in British homes are small (Wrixon et al. 1988). Cosmic-ray dose-rates can be modelled over a wide range of latitude and altitude and, for example, the CARI-6 code (O'Brien et al. 2003) can be used to estimate doses to aircrew based on the parameters of the flights that they undertake. Cosmic-ray flux densities vary with the 11-year solar cycle, but at sea-level in temperate latitudes the muon variation is small (~5%) (UNSCEAR 1977) and can be disregarded. Gamma and cosmic-ray doses in homes are more complex and depend on a combination of several factors (Maduar and Hiromoto 2004): they will be reduced relative to those outside because of the shielding provided by the fabric of the house, but radionuclides in the building materials will provide an extra gamma-ray contribution. In addition, some superficial earth and rock is likely to be removed during groundwork in the construction process. The Radioactive Substances Advisory Committee (Radioactive Substances Advisory Committee 1971) suggested that a one-tenth-value thickness of about 50 cm is a fairly conservative value for estimating the attenuation due to shielding (i.e., assuming rather transmissible material), which implies that terrestrial gamma radiation in a house emanates predominantly from sources within the nearest metre or so of material. Gamma-ray levels within buildings do not substantially vary with time (unless changes are made to the building, its surroundings or shielding).

The indoor gamma-ray dose-rate will thus be partly determined by the local geology, and this influence will be stronger if building materials are from the same locality. If urban areas are built-up at roughly the same time with broadly similar house design and materials then this will also provide a mechanism whereby neighbouring houses have similar indoor gamma-ray dose-rates even if the building materials are not local. The UK population-weighted average outdoor dose-rate from the combination of terrestrial gamma-rays and the directly ionising component of cosmic-rays is about 80% of the average indoor dose-rate from these sources (see below).

UNSCEAR (United Nations Scientific Committee on the Effects of Atomic Radiation (UNSCEAR) 2010) estimated that the global mean effective dose in dwellings from naturally occurring gamma-rays (with the directly ionising component of cosmic-rays) is about 0.69 mSv/a (0.41 mSv/a from terrestrial gamma radiation and 0.28 mSv/a from cosmic-rays). For the UK, Watson *et al* (Watson et al. 2005) gave data implying that the total average indoor annual effective dose from gamma radiation is about 0.57 mSv (0.34 mSv from terrestrial gamma-rays and 0.23 mSv from directly ionising cosmic radiation). Fair and Howells also noted that outdoor directly ionising cosmic-ray doses are roughly equal to those from terrestrial gamma-rays (Fair and Howells 1955). All these figures relate to adult doses. The UK data are based on a National Survey of UK homes by Wrixon *et al* (Wrixon et al. 1988) which is discussed in more detail below.

Another source of information on indoor gamma-ray dose-rates is the measurements made for the UK Childhood Cancer Study (UKCCS) (United Kingdom Childhood Cancer Study Investigators 2002a; United Kingdom Childhood Cancer Study Investigators 2002b). These measurements were made on behalf of the UKCCS investigators by the National Radiological Protection Board (NRPB, now part of Public Health England), and were made in the homes of children who had developed cancer or in the homes of matched controls. These residences are broadly representative of those of the population of GB as a whole and the published data indicated that the UKCCS measurements appeared to be roughly consistent (United Kingdom Childhood Cancer Study Investigators 2002b) with the pattern of doses reported in the National Survey (Wrixon et al. 1988). A more detailed comparison is given below.

The level and variation of exposures to natural background radiation has intrinsic interest by, for example, setting in context other radiation exposures, but there are also much more specific uses for such data. One such use was as part of the large record-based case-control study of natural background radiation and childhood cancer in GB referred to above (Kendall et al. 2013). In the published analysis (Kendall et al. 2013) only the data of the National Survey were available to allow the estimation of the gamma-ray exposures of study subjects. This meant that doses had to be assigned as the mean for the county district in which the mother was resident at the time of the child's birth. [County districts are relatively small administrative areas, numbering 459 in GB, with a mean all-ages population of rather more than 110,000, of whom 24,000 are children, see Online Resources 1, 2.]. The gamma-ray dose estimation was thus relatively crude. Moreover, a degree of geographical matching between the maternal residences at birth of cases and controls meant that almost half the cases were assigned the same gamma-ray dose-rate as their control(s). It was clear that more locationally-specific gamma-ray dose estimates, ideally on an individual basis or at least for areas smaller than county districts, would improve the power and reliability of the record-based study. The second main purpose of this paper was to consider how such estimates might be made using the data of both the National Survey and that collected for the UKCCS.

Two approaches to address this problem were explored. In the first of these a geostatistical model was fitted, which assumes an underlying geologically-determined spatial variation, superimposed on which is a stochastic process that models the observed tendency of dose-rates in neighbouring houses to correlate, more than can be accounted for by purely geological factors (Diggle and Ribeiro 2007). For the second approach first a number of empirical dose-rate interpolation measures was derived, based on averages over geologically- or administratively-defined areas, or using distance-weighted averages of measurements at nearest-neighbour points, and a jackknife technique (Davison and Hinkley 1997) was used to validate and compare the various dose-rate interpolation techniques, using mean-square-error (MSE) as a measure of predictive accuracy. The MSE is a combination of the estimation error, which could be reduced if more data were available, and the irreducible error due to spatially unstructured inter-house variation. Ordinary least squares (OLS) techniques were then used to derive an optimal linear combination of these interpolation measures.

Finally, the results of the geostatistical and empirical (OLS) dose-rate estimation methods were compared by means of simulation studies.

Materials and Methods

Indoor measurement data

Both the measurements of the UKCCS (United Kingdom Childhood Cancer Study Investigators 2002a; United Kingdom Childhood Cancer Study Investigators 2002b) and those of the National Survey (Wrixon et al. 1988) were made by the National Radiation Protection Board (NRPB). For both the National Survey and the UKCCS, study participants were sent by post measurement packs, which included two TLD dosemeters with instructions to place one in the main living room and the other in the main bedroom. Measurements were made over a notional six months period, but inevitably there was some variation in the period actually measured. There were some differences between the dosemeters used in the National Survey and those used for the UKCCS, but as shown below the results were similar.

In the National Survey, invitations to participate were sent to a random sample of UK households, selected from the UK Post Office address file (Wrixon et al. 1988). A total of 2,500 accepted the invitation, an acceptance rate of about 50%; for various reasons gamma-ray measurements were completed in only 2,283 residences in GB. All these measurements were included in the present study. For the National Survey, the locations of the dwellings were determined from their postcodes because the ADDRESS-POINT system was not available at that time. The record-based case-control study (Kendall et al. 2013) found that the mean separation of the ADDRESS-POINT grid reference for a dwelling was less than 100 m from the centroid of its postcode. Relatively small uncertainties are thus introduced by the use of postcodes rather than the more precise ADDRESS-POINT locations.

The UKCCS provided the present investigators with a dataset containing measurements for 8,434 houses, 8,143 in which measurements were made in both bedroom and living room, and 291 in which a measurement was made in only one of the two rooms. For the UKCCS, interactions with study participants were carried out by the UKCCS investigators, and the NRPB had no knowledge of where dosemeters were being used. In the data used for the present study, case and control residences were not distinguished and locations were given only to a postcode for reasons of confidentiality. The published record-based case-control study (Kendall et al. 2013) found that gamma-ray doses-rates in the homes of cases and of controls differed little and that a detailed analysis in terms of cumulative dose was required to demonstrate an associated increase in the risk of childhood leukaemia. Consequently, UKCCS measurement data for both childhood cancer cases and controls are used in the present investigation.

Some of the measurements in the UKCCS dataset did not meet the criteria specified for analysis by the UKCCS investigators (United Kingdom Childhood Cancer Study Investigators 2002b). Records were excluded from further analysis if:

(a) the storage period was more than 14 days; or

(b) the transit period was more than 45 days; or

(c) the chip fading exceeded 30%; or

(d) the results of the individual chips differed from the mean by more than 20%

Outdoor gamma-ray dose-rates

A survey of outdoor gamma radiation levels in GB (Green et al. 1989) was conducted at the same time as the National Survey of radiation levels in homes (Wrixon et al. 1988). A total of 2,398 measurements were made using an instrument with an energy compensated Geiger-Müller tube over a period of ten minutes. Further details of the measurement protocol are to be found in the detailed report of the survey (Green et al. 1989). The principal published results were the double-smoothed dose-rates tabulated for each 10 km square of the National Grid in GB. However, estimates were sparse in some coastal squares and in islands (e.g. Orkney and Shetland), and for the present purposes some values were interpolated in these areas. Values were interpolated for 4% of the squares, containing 0.5% of the case-control subjects (see Online Resource 1; the extended map is shown as Online Resource 3). The population-weighted average outdoor dose-rate in GB was about 77 nGy/h, terrestrial gamma-rays and the directly ionising component of cosmic radiation contributing 34 nGy/h (Green et al. 1989) and 43 nGy/h, respectively; the cosmic-ray dose-rate is estimated from the indoor figure using the shielding factor of UNSCEAR (United Nations Scientific Committee on the Effects of Atomic Radiation (UNSCEAR) 2000).

Geological data

Two main types of geological classification were used in the present study

LEX-ROCK Bedrock codes

These are codes for the underlying geology as defined by the British Geological Survey (BGS) (Smith 2013), and provide the name of each rock unit or deposit (via its LEX or Lexicon code) and composition (via its ROCK or lithology code). The location of each indoor gamma-ray measurement in the National Survey and UKCCS database was determined in terms of the Eastings and Northings of the British National Grid (Ordnance Survey 2010) for the centroid of its postcode. LEX-ROCK codings for each location were extracted from BGS 1:250,000 bedrock geological maps for GB via EDINA's Digimap service (University of Edinburgh 2014). ArcGIS Desktop v10 (Ormsby et al. 2004) was used to 'Merge' and ‘Dissolve’ the separate tiles into a single seamless shapefile for GB. A 'Join' operation on the dose-rate shapefile and the dissolved BGS bedrock geology shapefile identified the bedrock geology polygon within which each measurement was located. This was done using the NEAREST bedrock polygon in order to avoid problems in coastal areas, where the 1:250,000 scale geological data was spatially imprecise.

The house measurements had been made on sites with a total of 379 LEX-ROCK classifications, which were combinations of 281 values of LEX and 51 values of Rock.

50k-BEDSUP combined simplified bedrock and superficial codes

These codes were developed by the BGS in the context of an investigation into the variation of indoor radon concentrations across the UK (Miles and Appleton 2005). Simplified bedrock and superficial geology classification systems were developed which grouped some geological units with similar characteristics (Appleton 2005). Thus the 4,500 named bedrock geological units in England and Wales on the BGS 1:50,000 (50k) scale DiGMapGB maps were grouped using a simplified bedrock classification comprising only 406 units (Appleton 2005). The 888 individually named 50k superficial geology units were grouped according to a simplified system based on permeability and genetic type (Goldthwait and Matsch 1989). The 50k-BEDSUP code at any location is derived by concatenating the 50k simplified BEDrock and SUPerficial geology codes,

The house measurements had been made on sites with a total of 488 50K-BEDSUP classifications. The 50K-BEDSUP classifications were combinations of 180 bedrock codes and 154 codes for surface geologies; for about 40% of the measurements, the 50K-BEDSUP surface geology was judged to be the same as the bedrock. As part of the present investigations use was also made of sixteen grouped Classes for the bedrock classifications. These are shown in the geological map of GB (Online Resource 4). These bedrock classes broadly correspond to geological periods. However, six rock types which correlate with patterns of gamma-ray dose rate are also included (see Results Section).

Case-Control data

For simulation studies use was made of case-control data based on the analysis file developed by Kendall et al (Kendall et al. 2013), augmented by the inclusion of cases and controls in GB from the National Registry of Childhood Tumours (NRCT) for an expanded range of calendar years (1962-2010) to give a total of 126,817 study subjects. These data were also used to illustrate in broad geographical terms the distribution of the population of GB (see Online Resource 4).

Statistical methods

Empirical spatial interpolation variable selection

A number of candidate dose-interpolation measures were selected, predictive performance of which was evaluated using jackknife methods (Davison and Hinkley 1997). Dose-rate at a given point was generally assumed to be interpolated based on measurements of dose-rate at “nearby” points. Three basic types of interpolation were used: (a) weighting of neighbouring dose-rates by an inverse power of distance; or (b) an average over a given number of nearest neighbours; or (c) an average over some administratively defined area (e.g., county district). Jackknife methodology was used for determining the possible bias or estimating error, whereby single measurements are successively removed from the full set of N measurements and the remaining ($N-1$) values are used to estimate the measurement that had been withdrawn, as above. The mean-square-error (MSE) of the difference of these jackknife estimates from the actual dose-rates was estimated over all $N = 10,199$ points for which gamma dose-rates were measured. A Monte Carlo method was used to estimate the variance of the MSE, using a slight modification of the method of Efron and Stein (Efron and Stein 1981) – see Appendix C for further details.

Empirical ordinary least squares model selection and fitting, and Gaussian process maximum likelihood fitting

One of the fundamental means of evaluating spatial correlations is the semi-variogram, which is defined for a pair of points as half of the mean-square-difference of some random variable (in the prsent case, the dose-rate) at the two points. It can be shown that if, as often happens, the spatial correlation between dose-rates reduces with increasing distance, the dose-rate semi-variogram will increase with distance, with the asymptotic (large distance) value given by the variance of the dose-rate at any point (assumed to be constant) (Diggle and Ribeiro 2007) (and see also Appendix C). Figure 1 plots the semi-variogram for the raw indoor gamma-ray dose-rate.

In order to model the observed spatial correlation in dose-rates two approaches were considered. In the first a spatial geostatistical model was fitted to the gamma-ray dose-rate, which assumes an underlying geologically-determined spatial variation, superimposed on which is a stochastic process that models the observed tendency of dose-rates in neighbouring houses to correlate, more than can be accounted for by purely geological factors (Diggle and Ribeiro 2007). Specifically, it was assumed that the gamma-ray dose-rate at a given location is given by a sum of: (a) a "mean" process, with distinct values at each of the 16-level Classes of the 50KBedSup-Bedrock variable; (b) a stochastic spatial correlation process; and (c) a Gaussian "noise" term. The spatial correlation is assumed to result from unobserved variables that would be correlated between nearby areas (with greater correlation the closer the areas are), and that might affect gamma-ray dose-rate. The spatial correlation process is that developed by Matérn (Matérn 1960); further details are given in Appendix C. The Gaussian "noise" term is assumed to be the result of random measurement error in assessing the dose-rates. The parameters of the model were estimated by maximum likelihood, using the geoR package (Diggle and Ribeiro 2007) in R.

Another strategy to model (and eliminate) the spatial correlation in dose-rate was to construct a more highly parameterized empirical model, based on linear combinations of the interpolation measures described in the previous section. In order to construct an empirical model that satisfactorily explained the spatial variation of mean dose-rate, linear regression with ordinary least squares (OLS) (Rao 2002) was used. This model assumes a standard Normal error, and attempts to model the spatial correlation in dose-rate using combinations of explanatory variables. In order to avoid over-parameterised models, the Akaike Information Criterion (AIC) (Akaike 1973; Akaike 1981) was minimized to select the optimal set of interpolation variables (the calculation of which is outlined above). AIC penalises against overfitting by adding 2 x [number of fitted parameters] to the model deviance (residual sum of squares). An iterative mixed-forward-backward stepwise procedure was used to minimise AIC for OLS using R (R Project version 3.1.1 2014).

The fit of all these models was tested using a standard cross-validation process. The models were fitted to a randomly selected 70% of the data, and the indicated models were then used to predict gamma-ray dose-rates in the remaining 30% of the data.

The predicted variance for the full location data (n=126,817) were calculated using the Gaussian-Matérn model and the optimal OLS model fitted to the original dose-rate measurement data (n=10,199). The methods by which the model predicted mean and variance are derived for each model are set out in Appendix E.

Simulation study to test Gaussian-Matérn maximum-likelihood fitted model and optimal empirical OLS-fitted model

The predictive performance of each the two candidate models was evaluated, Gaussian-Matérn (Diggle and Ribeiro 2007) using the 16-level Classes of the 50KBedSup-Bedrock variable, and empirical optimal (via OLS fitting) linear-combination model via simulation. Dose-rate data from each of the two models were simulated and then case-control data for these two underlying sets of doses were simulated using the standard linear-logistic model (Preston et al. 1998). Each model was then refitted to the two sets of doses for the measurement set of 10,199 datapoints, and these refitted models were used to predict the dose-rates for the full set of 126,817 measurement locations. Then it was assessed how well these re-estimated dose rates predicted the underlying (simulated) disease process by fitting a linear-logistic model to the four sets of data (and the case-control sets predicted as above). Further details are given in Appendix C.

Results

Indoor Measurement Data

After the exclusions detailed above, there were valid measurements of indoor gamma-ray dose-rates for 7,916 residences in the UKCCS database. In 7,270 instances these were based on measurements in both bedroom and living room; in 646 instances they were based on measurements in only one of the two rooms. As noted above, the National Survey contributed measurements of indoor gamma-ray dose-rates in 2,283 homes. The total number of measurements was thus 10,199.

Table 1 gives the number of measurement records from the National Survey and from the UKCCS, together with the minimum and maximum dose-rates recorded and the means and standard deviations. Similar data are given for the distribution of means for postcode districts, for county districts and for counties. The results for the National Survey are essentially the same as those presented by Kendall *et al* (Kendall et al. 2006). Dosimetric data for county districts, based on all 10,199 measurements, are presented in On-Line Resource 2. Table 2 summarises the main features of the indoor and outdoor dose-rates.

Online Resource 7 gives the distribution of indoor gamma-ray dose-rates from the National Survey and from the full dataset. Both are approximately normally distributed, but as expected, the larger number of measurements in the full dataset has noticeably reduced the random deviations from a smooth curve.

A map of the county districts in GB, coloured to indicate the mean dose-rates within each county district is given as Fig 2. This map also shows the geographical density of the measurement points. The latter can be compared with Online Resource 5, a map that shows the locations of the maternal residence at birth of the controls for the case-control study of Kendall *et al* (Kendall et al. 2013); as expected, this closely follows the density of the general population of GB.

Influence of geology

The National Survey (Wrixon et al. 1988) reported that indoor gamma-ray dose-rates were found to vary both with outdoor dose-rates and geology. However, it is difficult to identify many specific geological associations on the 10 km grid map (see Fig 7 of (Wrixon et al. 1988), see also Online

Resource 6) apart from the high dose-rates associated with the granites in south-west England and consistently low dose-rates in London and the surrounding area. More detailed information is available to the present study and this is summarized in the geological map (Online Resource 4).

The geological unit (50k-BEDSUP) that is associated with the highest average indoor dose-rate is granite from south-west England (142 nGy/h) (Online Resource 4). Twenty-nine of the 37 geological units with five or more indoor dose-rate measurements and relatively high average dose-rates >104 nGy/h belong to the Carboniferous Coal Measures together with the adjacent Carboniferous sandstones and shales, and Permian-Triassic sandstones, mudstones and dolomites. The Carboniferous and Permian-Triassic sandstones and dolomites would be expected to have relatively low outdoor dose-rates as they are not characterised by high K, Th or U concentrations. It seems likely, therefore, that the spatial association of relatively high indoor dose-rates with the coalfields and adjacent areas of England, Wales and Scotland (Online Resource 4) results from the use of bricks made from colliery waste in a large proportion of the older houses in this area. Other clay resources used for brick manufacture include the Triassic Mercia Mudstone (average 101 nGy/h), the Jurassic Oxford Clay, and the Cretaceous Weald and Wadhurst Clay Formations (average 80-90 nGy/h) (Online Resource 4). The majority of the geological units associated with relatively low average indoor dose-rates (<85 nGy/h) are found in southern England and include (1) the Cretaceous Chalk, Wealden Group sandstones and clays, Gault Clay, and the Lower and Upper Greensand Formations, (2) the Tertiary (Palaeogene) sands and clays of the Thanet Sand Formation, Lambeth Beds, London Clay Formation, Bracklesham Group, and Bagshot Formation (Online Resource 4).

Results of empirical spatial interpolation variable selection

The default values of the parameters used in the jackknife calculations are given in Appendix A Table A1. Unless otherwise stated these values were always employed. All runs were based on the full dataset with 10,199 measurements of indoor gamma-ray dose-rate. However, it is important to note that while the parameters of Appendix A Table A1 are used for the base calculations they do not represent the optimum model finally identified.

Appendix A Table A2 demonstrates that moving much beyond the default maximum radius of 30 km gives little improvement in fit of the model. If one does not impose any distance limit, as can be seen from Appendix A Tables A2 and A4, there is a slight, non-significant, reduction in MSE in moving from a limit of 30 km (MSE = 388.69) to no distance limit (MSE = 385.87). Appendix A Table A3 shows that increasing the number of nearest neighbours decreases the MSE, although beyond

100 neighbours the improvement in MSE is slight: using a maximum of 100 nearest neighbours within 30 km yields an MSE of 388.69, compared with an MSE without a limit on the number of neighbours of 388.67. It should be noted that with no limit on distance or the maximum number of neighbours the MSE increases quite substantially, to 423.29.

Table 3 gives the MSE for various values of the exponent, k, and Table 4 explores the variation in MSE as k is varied for measurements on the same and on different geologies. Several of the MSE values are similar and fall within the respective standard deviations. However, the lowest MSE is found for $k = 1$ (inverse distance weighting) with little evidence of differential weighting by geology.

Table 5 gives variation of the MSE (and its standard deviation) for estimates based on small area averages or averages over geological categories. Differences in MSE between estimates in this table are not always statistically significant, but it is clear that county district averaging performs best. The original county district-averaged gamma-ray dose-rates based only on the results of the National Survey, i.e. on the 2,283 measurements used in the published study (Kendall et al. 2013), perform reasonably well when compared with the MSE obtained using other areal units, although the MSE is somewhat lower for a number of the area-based measures (for example, using either county district, 10 km grid square, LEX-ROCK, LEX, or 50K-BEDSUP bedrock) based on the full dataset of 10,199 measurements. Analysis of variance indicates that the proportion of the variance explained by grouping the dose-rate data by 10 km grid square, county district, LEX-ROCK and 50k-BEDSUP is 27.8%, 25.5%, 21.8% and 20.5%, respectively. There are very few measurements for certain geological categories but repeating the calculations for categories with >4 or >9 measurements has virtually no effect on the results

When the linear regression model, as given by Appendix C expression (C4), was fitted to the internal (d_i) and external $(d_{i,ext})$ dose-rate data (shown in Fig. 3) a value of α of 0.806 resulted. Using this value of α indoor dose-rates were calculated as $\alpha\, d_{i,ext}$, and using these estimates gives a MSE of 463.54 (Table 5), worse than many of the other models considered – the GB average indoor gamma-ray dose-rate was worst with MSE of 512.76, followed by LEX-ROCK-Rock with MSE of 499.43, then postcode district with MSE of 465.20 (Table 5).

Using the optimal nearest neighbour model, based on the 100 nearest measurements weighted by the inverse of the separation from the measurement point, generates estimates with a standard deviation from the true value of 19.6 nGy/h (Tables 3, 4). The use of averages for the relevant county district is only a little less reliable, with a standard deviation of about 20.1 nGy/h; however, these standard deviations are statistically significantly different (p<0.0001) (Table 5). Both these approaches are significantly better (p<0.0001) than any of the other averaging methods using the full dataset of 10,199 measurements (Table 5), and better than county district averages based only on the 2,288 measurements of the National Survey, for which the standard deviation is 21.3 nGy/h.

Results of empirical linearly-combined (OLS) model selection and fitting, and Gaussian-Matérn maximum likelihood fitting

The optimal model using stepwise OLS applied to the gamma-ray dose-rate data, selecting from the list of variables shown in Appendix B Table B1, is given in Table 6; the associated parameter values and their uncertainties are given in Appendix B Table B2. As Fig. 4 demonstrates, there is little variation in the semi-variance of the residuals from this optimal model with distance, suggesting that the model very largely accounts for the observed spatial correlation in gamma-ray dose-rate (Fig. 1). As can be seen from evaluation of the MSE from the 70%:30% cross-validation in Table 7, the optimal OLS model of Table 6 performs much better than any other model. The next best model is the maximum-likelihood fitted Gaussian-Matérn model, using the 16-level 50K-BEDSUP-Bedrock derived Classes, but the MSE is larger than that of the optimal OLS model (410.92 *vs* 377.64) (Table 7).

The results of fitting the model described in Appendix C (expression (C11)) suggest that there are no significant trends of standard deviation *vs* mean (p=0.139) for the optimal OLS model, implying that there is no statistically significant heteroscedasticity (i.e., there is no statistically significant evidence that the data fall into subpopulations as might arise, for example, if geology played a strong role).

Table 8 and Fig. 5 show the predicted variance for the full location data (n=126,817) using the Gaussian-Matérn model and the optimal OLS model fitted to the original dose-rate measurement data (n=10,199). As can be seen the OLS model predicts on average a smaller variance than the Gaussian-Matérn model, by a factor of nearly 3 (ratio of means = 5.48/1.85).

Results of simulation study to test optimal empirical OLS-fitted model and Gaussian-Matérn maximum-likelihood fitted model

Table 9 details the fits of logistic models to simulated case-control datasets based on various sets of doses estimated for the putative full 126,817 cases and controls from the extended NRCT childhood cancer data. As can be seen the linearly-combined (OLS) model gets creditably close to the generating excess odds ratio (EOR) of 120 Gy^{-1}: 126.0 Gy^{-1} (95% CI 97.4, 161.1). The refitted linearly-combined (OLS) model also generates an acceptably close EOR: 116.6 Gy^{-1} (95% CI 25.4, 348.9). However, the Gaussian-Matérn model refitted to the linearly-combined-(OLS)-generated data is badly biased: -24.1 Gy^{-1} (95% CI -54.7, 120.5). The Gaussian-Matérn model refitted to the Gaussian-Matérn dose-rate data produced a downwardly biased estimate, an EOR of 45.4 Gy^{-1} (95% CI -16.0, 217.3), compared with the EOR in relation to the "true" EOR (generated by the Gaussian-Matérn model), of 92.2 Gy^{-1} (95% CI 69.0, 120.7), and even this figure is somewhat lower than the generating EOR (120 Gy^{-1}). The OLS refit to the Gaussian-Matérn data is also downwardly biased, with an EOR of 27.8 Gy^{-1} (95% CI -25.6, 200.6). The refitted Gaussian-Matérn model had parameters somewhat different from those used to generate the data, as can be seen by comparing the theoretical variograms of the generating and refitted models (Fig. 6).

Discussion

The present dataset has over 10,000 measurements of indoor gamma-ray dose-rates (including the directly ionising component of cosmic-rays). This is over four times as many measurements as the original National Survey of natural background radiation exposures in UK dwellings (Wrixon et al. 1988) on which estimates of population dose-rates have been based (Watson et al. 2005)

It is reassuring that the new dataset confirms the mean and range of the previously published national distribution (Table 1). The larger dataset demonstrates more clearly that the distribution of measurements is broadly normal with a mean of about 96 nGy/h and with a standard deviation about one quarter of the mean (Online Resource 7). In contrast, the distribution of radon exposures is typically log-normal and a significant fraction of the population receives relatively high (lung) doses from this source (Kendall et al. 1994). For indoor gamma-rays, only a small proportion of dwellings, perhaps one in a thousand, has dose-rates more than twice the mean.

It is noted that data on UK exposures to terrestrial gamma-rays and to cosmic-rays has been missing from the UNSCEAR reviews and the analysis presented here allows this omission to be rectified (Table 2). This requires the separation of the contributions from directly ionising cosmic-rays and terrestrial gamma-rays. This was done here with reasonable confidence for the population means, but the ranges of exposure, particularly the lower end of the range of indoor exposures, is more complex. We have followed the National Survey in assuming that cosmic-ray doses outdoors are effectively constant across the country; this approximation is probably reasonable for GB given the limited population variation in latitude and altitude. However, the indoor cosmic-ray dose will be significantly affected by the shielding provided by the material of the house. This will vary between buildings and in some cases, for example the lower floors of large blocks of flats, shielding will be more substantial than the average and the general estimate of cosmic-ray dose, 34 nGy/h, will not apply. Wrixon et al estimated that the minimum indoor gamma-ray dose rate (without the cosmic-ray component) is probably around 10 nGy/h and this value is adopted here. The enhanced dataset is particularly useful in providing information on small-area variation in gamma-ray levels and these are described in the on-line appendix (Online Resource 1)

The jackknife calculations reported in Tables 3-5 suggest that interpolating dose-rates with an inverse power of distance works best of the empirical measures evaluated there. Perhaps surprisingly, county district averaging based on the full dataset of 10,199 measurements did best of the various purely areal measures, and these values were markedly better than averages based only on the National Survey, as used in the published record-based case-control study (Kendall et al. 2013). Although there was no indication of a better fit, based simply on the jack-knife error estimate, if measurements on the same geology as the test point were differentially weighted (Table 4), the step-AIC analyses using OLS-fitted linearly-combined models reported in Table 6 suggests that nevertheless there is some contribution made by measures that differentially-weight nearby points by the similarity or otherwise of geology to the index point.

Postcode districts had poor predictive power if the averages were based only on measurements within the postcode district itself (Table 5). Both the LEX-ROCK and 50K-BEDSUP geological codings performed reasonably well. Small differences between the predictive power of the various estimation approaches should not be overinterpreted. Nevertheless, estimates based on the measured outdoor gamma-ray dose-rate did fairly badly in the comparison exercise. However,the outdoor data used here are limited: they are based on only 2,398 measurements and the estimates are means for 10 km grid squares which have been subject to double smoothing (Green et al. 1989). Figure 3 suggests that the smoothing process may have reduced the range of the estimates (see also Online Resource 3). It is possible that more precise outdoor gamma-ray estimates might provide more accurate predictions of indoor levels. This is important because, while indoor gamma-ray dose-rates are more likely to be important for epidemiological purposes than outdoor dose-rates, the former are often less available, as noted by UNSCEAR (UNSCEAR 2008)

It is likely that these broad findings are a consequence of two competing factors: the smaller the area the better the reflection of local circumstances, but on the other hand larger areas have more measurements and thus less variable estimates of mean dose-rates. In this respect it is noted that the averaging areas that perform best are those with several hundred measurements rather than those with smaller or larger numbers. The explanation is further evidenced by the improvement in MSE as neighbouring measurements are included in the mean for postcode districts (Appendix A Table A5), this effect being not nearly so apparent when county districts are used (Appendix A Table A6)), and this is likely to be due to the different approach to the division of populations used in the construction of these two sets of areas.

The cross-validation performed in Table 7 and the results of the simulation exercise reported in Table 9 both suggest that the optimal OLS-fitted model performs rather better than the Gaussian-Matérn geostatistical model. The poorer performance of the Gaussian-Matérn model is perhaps surprising. It is very likely that geology must account for most of the variation in outdoor dose-rate, but there is a substantial difference between outdoor and indoor gamma dose-rate (Figure 3), and the latter is the quantity of most relevance to assessing cancer risk (Kendall et al. 2013). Evidence presented here suggests that indoor gamma-ray dose-rates in neighbouring houses have a greater tendency to be similar to each other than is implied simply by the underlying geology. As discussed above, the gamma-ray dose-rate inside a house can be regarded as the outdoor gamma-ray dose-rate attenuated by the fabric of the house, but augmented by a contribution from radionuclides in the building materials. The second of these contributions is generally dominant (UNSCEAR, 2000). Choice of building materials and of house construction are likely to vary from place to place so that they are influenced by geography, but only to a limited extent by local geology. Further, the contribution from cosmic-rays will be determined by the outdoor flux as modified by the material of the house and thus entirely unaffected by geology. These factors could account for the greater success in this instance of models involving simple proximity rather than geological boundaries. As noted above, the geological classifications available for the present study account for about 20-22% of the variance in indoor gamma-ray dose-rates observed here, and this is somewhat lower than the corresponding proportion for radon, 25% (Appleton and Miles 2010). Local geology therefore appears unlikely to be the dominant factor directly driving the similarity in gamma-ray levels in neighbouring houses.

One matter that must be taken into consideration is the rather larger variance for the predicted gamma-ray dose-rates estimated using the Gaussian-Matérn model compared with those for the optimal OLS model (Table 8, Fig. 5). It is likely *a priori* that the OLS model will not estimate the variance correctly. Set against that, the Gaussian-Matérn model clearly does not fit the data very well, and little weight should be attached to the variance estimate produced by a poorly-fitting model. This has some implications for propagating these individual error estimates (the method of calculation of which is outlined in Appendix E) to the regression model (relating cancer to radiation dose) in the enlarged case-control dataset, which if the empirical OLS method were to be used would require inflating. The errors are likely to be a mixture of Berkson and classical in form, with both shared and unshared errors, whichever method is used to derive the estimated doses. Methods for taking account of dose errors are well known (Carroll et al. 2006). When errors are not too large regression calibration can be used (Carroll et al. 2006). Full likelihood methods such as Bayesian Markov Chain Monte Carlo (Bennett et al. 2004; Little et al. 2008) and Monte Carlo maximum likelihood (Little et al. 2014; Stayner et al. 2007) have both been employed in other radiation studies. All these methods could be employed here.

Notwithstanding the fact that the particular geostatistical model that was employed here performed much worse than the OLS model (Tables 7, 9), it is possible that other such models, based on different geological classification and other potentially relevant data, perhaps with an alternative geostatistical model, would perform better. It is a feature of such geostatistical models that they can statistically explain the observed tendency of dose-rates in neighbouring houses to correlate (Fig. 1), more than can be accounted for by purely geological factors. Purely geostatistical models have other advantages over the particular OLS models fitted here; in particular there is a more adequate likelihood-based framework for inference, so that model selection and estimates of prediction error can be more reliably performed. However, the poor predictive performance of the particular Gaussian-Matérn model used here suggests that it should not be used for the estimation of doses for the purposes of epidemiological analysis. We are engaged in ongoing work to explore other geological data and models that could be employed in order to construct a better geostatistical model.

Conclusion

It is concluded that the best model for estimating indoor gamma-ray dose-rate at a point at which no measurement has been made is obtained using a combination of geological codes (50K-BEDSUP, and its surface and bedrock components) and inverse power of distance weights of the dose-rates at neighbouring points. This model has superior predictive performance to a purely geostatistical Gaussian-Matérn model based on currently available geological information. Reasonable (if less precise) estimates may also be obtained by taking the mean for the county district in question. A second analysis of the UK-NRCT data (Kendall et al. 2013) is planned using this model to derive new gamma-ray dose-rate estimates and an expanded set of cases and controls covering a longer calendar period.

Acknowledgements We are very grateful to Jill Simpson of the University of York and to the other UKCCS investigators for making available the indoor gamma ray measurements made for the United Kingdom Childhood Cancer Study and for advice on the interpretation of the data. We are also very grateful to J D Appleton for advice on geological matters generally and in particular for suggesting geological classification schemes. We are grateful to Phil Gilvin, Luke Hager and Rick Tanner at Public Health England (PHE) for advice on the dosimetry of the National Survey and the UKCCS, and to Simon Bouffler and other colleagues at PHE for allowing the use of the National Survey data and for advice on its interpretation. They also made helpful comments on the draft. Helpful comments were also made by the two referees, and by Bernd Grosche, Michael Murphy and Graham Smith. This work was supported by Children with Cancer (UK) and by the Intramural Research Program of the National Institutes of Health, the National Cancer Institute, Division of Cancer Epidemiology and Genetics. Much of the work by two of the authors (GMK, TJV) was undertaken at the Childhood Cancer Research Group, whose loss is much regretted. The research work of CCRG was supported by the Department of Health for England and Wales, Scottish Government and Children with Cancer (UK).

Table 1 Comparison of indoor gamma-ray dose-rate estimates (nGy/h) from the National Survey, UKCCS and combined dataset

		Indoor gamma-ray dose-rate (nGy/h)				
		Number of estimates	Mean	SD	Min	Max
Individual House Measurements	National Survey	2,283	94.8	24.5	25	278
	UKCCS	7,916	95.8	22.1	26	271
	Total	10,199	95.6	22.7	25	278
Means for postcode districts	National Survey	1,454	94.5	22.4	25	225
	UKCCS	2,076	94.9	17.2	38	212
	Total	2,255	94.3	17.1	25	225
Means for county districts	National Survey	440	93.6	17.4	38	169
	UKCCS	455	94.1	12.2	61	143
	Total	458	94.0	12.7	64	169
Means for counties	National Survey	65	95.2	13.1	63	128
	UKCCS	65	94.4	10.6	71	111
	Total	65	94.3	10.9	70	116

Note: No measurements were available for the county district "City of London" or for some postcode districts.

Table 2 Exposures to naturally occurring terrestrial gamma-rays and directly ionising cosmic-rays (mainly muons) in Great Britain (absorbed dose-rates in air (nGy/h)). This follows the format of UNSCEAR (United Nations Scientific Committee on the Effects of Atomic Radiation (UNSCEAR) 2010).

Outdoors						Indoors					
Cosmic radiation		Terrestrial Radiation		Total		Cosmic radiation		Terrestrial Radiation		Total	
Mean	Range[1]	Mean	Range	Mean	Range	Mean	Range[1]	Mean	Range	Mean	Range[2]
43	--	34	10-90	77	50-130	34	--	62	10-240	96	25-280

[1] Cosmic-ray dose-rate taken to be invariant across the UK

[2] Based on means for 10 km squares. The true range may be somewhat greater.

Table 3 Nearest neighbour analysis: variation of mean-square-error (and its standard deviation) with the inverse power of the separation distance between the point of interest and other measurement locations, k.[a]

Inverse power of distance, k	Mean-square-error $(nGy/h)^2$ (standard deviation of mean-square-error)
0.0	394.88 (5.68)
0.5	388.30 (5.26)
1.0	385.87 (4.34)
1.5	407.98 (4.43)
2.0	444.10 (6.04)
2.5	478.01 (7.53)
3.0	505.70 (8.69)

[a]Default analysis parameters were used as per Appendix Table A1.

Table 4 Nearest neighbour analysis: variation of mean-square-error (and its standard deviation) with the inverse power of the separation distance between the point of interest and other measurement locations, k, by whether the points are on the same or different geological categorisation.[a]

Mean-square-error $(nGy/h)^2$ (standard deviation of mean-square-error)			
Inverse power of distance, k, for points on different geology	Inverse power of distance, k, for points on same geology		
	0.5	1.0	1.5
0.5	388.30 (5.26)	386.47 (5.23)	389.82 (5.32)
1.0	386.71 (4.50)	385.87 (4.34)	394.63 (4.63)
1.5	392.62 (3.94)	394.11 (3.73)	407.98 (4.43)

[a]Default analysis parameters were used as per Appendix Table A1.

Table 5 Variation of mean-square-error (and its standard deviation) for estimates based on area averages or averages over geological categories. In every case the *p*-value for difference from the optimal model (with inverse distance weighting, as per Table 3) is <0.0001

Area used for averaging	Mean-square-error $(nGy/h)^2$ (standard deviation of mean-square-error)
County district	403.74 (6.44)
LEX-ROCK	417.44 (5.84)
LEX-ROCK-Lex	420.30 (5.51)
50K-BEDSUP bedrock	423.93 (7.78)
50K-BEDSUP	427.98 (8.68)
10 km grid square	440.10 (5.96)
50K-BEDSUP surface	447.47 (8.13)
National Survey county district mean[d]	452.91 (6.40)
Outdoor gamma-ray dose-rate[c]	463.54 (9.97)
Postcode district	465.20 (8.27)
LEX-ROCK-Rock	499.43 (11.05)
GB average indoor gamma-ray dose-rate[b]	512.76 (10.89)

[a]Default analysis parameters were used as per Appendix Table A1.

[b] Average gamma dose-rate for Great Britain (over all 10,199 sample points).

[c] Dose-rate estimated using mean outdoor gamma-ray dose-rate in 10 km square, using Eq. 4.

[d] County district-averaged dose-rate based on the 2,283 measurements of the National Survey only

Table 6. Optimal model fitted by ordinary least squares to gamma-ray dose-rate data. The optimal model is derived via the step-forward-backward Akaike Information Criterion (AIC), using the variables listed in Appendix B Table B1.

Model variables (in order of selection)
Average gamma-ray dose-rate by 50K-BEDSUP (25 groups)
Average gamma-ray dose-rate by 50K-BEDSUP
Easting
Easting^2
Average gamma-ray dose-rate by 50K-BEDSUP Surface
Average gamma-ray dose-rate by 50K-BEDSUP Bedrock
Average gamma-ray dose-rate weighted by $1/\text{distance}^{3.0}$
Average gamma-ray dose-rate weighted by $1/\text{distance}^{0.5}$ (different geology), $1/\text{distance}^{1.0}$ (same geology)
Average gamma-ray dose-rate weighted by $1/\text{distance}^{0.0}$
Average gamma-ray dose-rate weighted by $1/\text{distance}^{0.5}$ (different geology), $1/\text{distance}^{1.5}$ (same geology)

Note: a linear term in Easting has been added to the OLS optimal model to render it algebraically complete.

Table 7 Mean-square-error in randomly chosen 30% test sample, resulting from fitting various models to the remaining 70% of the data.

Model used	Mean-square-error
Ordinary least squares fits	
Optimal model as in Table 6	377.64
0^{th} order spatial sub-model of expression (C8)	446.27
1^{st} order (linear) spatial sub-model of expression (C8)	1276.37
2^{nd} order (quadratic) spatial sub-model of expression (C8)	744.56
3^{rd} order (cubic) spatial sub-model of expression (C8)	308,756.22
4^{th} order (quartic) spatial sub-model of expression (C8)	81,791.42
Maximum-likelihood fits	
Gaussian-Matérn model using 16-level code derived from 50K-BEDSUP-Bedrock variable	410.92

Table 8 Centile distribution and mean of individual variances for full location dataset (n=126,817) estimated using Gaussian-Matérn model and ordinary least squares linearly-combined model fitted to dose-rate measurement data (n=10,199).

Centile (%)	Gaussian-Matérn	Ordinary least squares
1	1.10	0.60
5	1.23	0.69
10	1.51	0.74
25	2.59	0.87
50	3.86	1.15
75	6.31	1.87
90	10.44	3.67
95	14.89	5.12
99	28.38	13.97
Mean	5.48	1.85

Table 9 Model fits to the simulated dose-rate data, generated using an excess odds ratio (EOR) of 120 Gy^{-1}.

Dataset	EOR / Gy (95% CI)
Empirically generated dose-rate data using "true" empirical doses to generate data for 126,817 sample	126.0 (97.4, 161.1)
Empirical-model refitted to jittered 10,199 empirical data used to regenerate data for 126,817 sample ($M_{OLS-OLS,F}$)	116.6 (25.4, 348.9)
Gaussian-Matérn model fitted to jittered 10,199 empirical data used to regenerate data for 126,817 sample ($M_{OLS-MLE,F}$)	-24.1 (-54.7.120.5)
Gaussian-Matérn generated dose-rate data using "true" Gaussian-Matérn doses to generate data for 126,817 sample	92.2 (69.0, 120.7)
Empirical model fitted to jittered 10,199 Gaussian-Matérn data used to regenerate doses for 126,817 sample ($M_{MLE-OLS,F}$)	27.8 (-25.6, 200.6)
Gaussian-Matérn model refitted to jittered 10,199 dose-rate data used to regenerate doses for 126,817 sample ($M_{MLE-MLE,F}$)	45.4 (-16.0, 217.3)

Appendix A. Supplementary Tables Relating to Dose-Interpolation Measures

This Appendix contains details of the nearest neighbour calculations (Tables A1 to A4) and of small area averaging in which measurements within an area are augmented by those outside but close to it (Tables A5, A6). Table A5 shows that the MSE in postcode district-based area averages can be considerably improved by augmenting the averaging areas by neighbouring points which are close to the centroid of the postcode district but lie outside it. Table A6 shows no such improvement when the same exercise is carried out for county districts, and the MSE for county districts is always lower than the MSE for augmented postcode districts.

Table A1 Default parameter values.

Parameter	Value
Inverse power of distance, k	1
Minimum number of neighbours, $N_{\min}$, below which area-based average used	2[a]
Maximum number of nearest neighbours used, $N_{\max}$	100
Area-averaged measure used when number of neighbours $< N_{\min}$ [a]	LEX-ROCK
Minimum distance below which distance is limited	0.1 km
Maximum distance used for selecting nearest neighbours	∞ km

[a]This parameter is only of relevance to those models (Appendix A Tables A2, A3, A4) in which a distance limit was imposed.

Table A2 Variation of mean-square-error with maximum distance used for selecting nearest neighbours.[a]

Maximum distance used for nearest neighbour selection (km)	Mean–square-error $(nGy/h)^2$
5	421.81
10	407.49
15	397.80
20	393.94
25	390.09
30	388.69
35	387.99
40	387.71
45	387.38
50	387.29
∞	385.87

[a]Default analysis parameters were used as per Appendix A Table A1.

Table A3 Variation of mean-square-error with maximum number of nearest neighbours used, $N_{\max}$.[a]

Maximum number of nearest neighbours, $N_{\max}$	Mean-square-error (nGy/h)^2
10	419.82
20	402.89
30	396.93
40	393.72
50	392.04
100	388.69
∞	388.67

[a]Default analysis parameters were used as per Appendix A Table A1, with 30 km maximum distance.

Table A4 Variation of mean-square-error with relaxation of various neighbourhood parameters.[a]

Model criteria used	Mean square error
No maximum nearest neighbour limit, 30 km distance limit	388.67
Maximum 100 nearest neighbour limit, no distance limit	385.87
No maximum nearest neighbour limit, no distance limit	423.29

Table A5 Variation of mean-square-error (and its standard deviation) when postcode district-averaging is used throughout, augmented by averaging over neighbouring points a given distance from the centroid of the index area, but lying outside the area.[a]

Minimum number of measurements in an area below which augmentation by points within 7 km[b] of the centroid, but lying outside the area, is used	Postcode district mean-square-error $(nGy/h)^2$	*p*-value for difference from optimal model (with inverse distance weighting, as per Table 3)
unaugmented average	465.20 (8.20)	<0.0001
3	440.84 (7.02)	<0.0001
4	432.94 (7.77)	<0.0001
5	427.56 (6.98)	<0.0001
6	422.87 (6.91)	<0.0001
7	421.07 (6.86)	<0.0001
8	418.89 (7.18)	<0.0001
9	417.10 (6.98)	<0.0001
10	416.00 (7.43)	<0.0001
30	413.28 (6.46)	<0.0001

[a]Default analysis parameters were used as per Appendix A Table A1.

[b] Selected as the mean distance between measurement locations within the postcode district and the centroid of the postcode district.

Table A6 Variation of mean square error when county district-averaging is used throughout, augmented by averaging over neighbouring points a given distance from the centroid of the index area, but lying outside the area.[a]

Minimum number of measurements in an area below which augmentation by points within 7 km of the centroid, but lying outside the area, is used	County district mean square error $(nGy/h)^2$
unaugmented average	403.74
3	403.80
4	403.67
5	403.80
6	403.90
7	403.90
8	403.94
9	403.77
10	403.46
30	402.17

[a]Default analysis parameters were used as per Appendix A Table A1.

Appendix B. Supplementary table relating to ordinary least squares (OLS) model fits

Table B1 Variables used in ordinary least squares (OLS) regressions as part of the forward-backwards stepwise Akaike Information Criterion (AIC) variable selection.

Variable
Northing
Easting
Northing^2
Easting^2
Grouping into 16 geological groups *a priori* by 50K-BEDSUP Bedrock codes
Average gamma-ray dose-rate by 50K-BEDSUP (25 groups)
Average gamma-ray dose-rate by 50K-BEDSUP (50 groups)
Average gamma-ray dose-rate by 50K-BEDSUP (100 groups)
Average gamma-ray dose-rate by 50K-BEDSUP (200 groups)
Average gamma-ray dose-rate by county district
Average gamma-ray dose-rate by Post Code District
Average gamma-ray dose-rate by 10 km square
Average gamma-ray dose-rate by 50K-BEDSUP
Average gamma-ray dose-rate by LEX-ROCK
Average gamma-ray dose-rate by LEX-ROCK-Lex
Average gamma-ray dose-rate by LEX-ROCK-Rock
Average gamma-ray dose-rate by 50K-BEDSUP Bedrock
Average gamma-ray dose-rate by 50K-BEDSUP Surface
Average gamma-ray dose-rate weighted by 1/distance$^{0.0}$
Average gamma-ray dose-rate weighted by 1/distance$^{0.5}$
Average gamma-ray dose-rate weighted by 1/distance$^{1.0}$
Average gamma-ray dose-rate weighted by 1/distance$^{1.5}$
Average gamma-ray dose-rate weighted by 1/distance$^{2.0}$
Average gamma-ray dose-rate weighted by 1/distance$^{2.5}$
Average gamma-ray dose-rate weighted by 1/distance$^{3.0}$
Average gamma-ray dose-rate weighted by 1/distance$^{0.5}$ (different geology), 1/distance$^{1.0}$ (same geology)
Average gamma-ray dose-rate weighted by 1/distance$^{0.5}$ (different geology), 1/distance$^{1.5}$ (same geology)

Average gamma-ray dose-rate weighted by 1/distance$^{1.0}$ (different geology), 1/distance$^{0.5}$ (same geology)
Average gamma-ray dose-rate weighted by 1/distance$^{1.0}$ (different geology), 1/distance$^{1.5}$ (same geology)
Average gamma-ray dose-rate weighted by 1/distance$^{1.5}$ (different geology), 1/distance$^{0.5}$ (same geology)
Average gamma-ray dose-rate weighted by 1/distance$^{1.5}$ (different geology), 1/distance$^{1.0}$ (same geology)

Table B2. Parameter values and 95% CI for optimal model fitted by ordinary least squares.[a]

Constant (Intercept)	93.347 (81.048, 105.645)
Average gamma-ray dose-rate by 50K-BEDSUP (group 2 vs group 1)	-16.194 (-24.630, -7.757)
Average gamma-ray dose-rate by 50K-BEDSUP (group 3 vs group 1)	-28.032 (-34.821, -21.244)
Average gamma-ray dose-rate by 50K-BEDSUP (group 4 vs group 1)	-32.360 (-38.861, -25.860)
Average gamma-ray dose-rate by 50K-BEDSUP (group 5 vs group 1)	-33.448 (-39.874, -27.022)
Average gamma-ray dose-rate by 50K-BEDSUP (group 6 vs group 1)	-35.084 (-41.533, -28.636)
Average gamma-ray dose-rate by 50K-BEDSUP (group 7 vs group 1)	-35.531 (-41.965, -29.098)
Average gamma-ray dose-rate by 50K-BEDSUP (group 8 vs group 1)	-37.138 (-43.562, -30.714)
Average gamma-ray dose-rate by 50K-BEDSUP (group 9 vs group 1)	-37.528 (-44.057, -30.998)
Average gamma-ray dose-rate by 50K-BEDSUP (group 10 vs group 1)	-39.104 (-45.587, -32.621)
Average gamma-ray dose-rate by 50K-BEDSUP (group 11 vs group 1)	-39.679 (-46.393, -32.964)
Average gamma-ray dose-rate by 50K-BEDSUP (group 12 vs group 1)	-43.458 (-50.082, -36.834)
Average gamma-ray dose-rate by 50K-BEDSUP (group 13 vs group 1)	-44.531 (-51.761, -37.300)
Average gamma-ray dose-rate by 50K-BEDSUP (group 14 vs group 1)	-43.825 (-51.024, -36.626)
Average gamma-ray dose-rate by 50K-BEDSUP (group 15 vs group 1)	-45.344 (-52.603, -38.086)
Average gamma-ray dose-rate by 50K-BEDSUP (group 16 vs group 1)	-46.748 (-53.511, -39.986)
Average gamma-ray dose-rate by 50K-BEDSUP (group 17 vs group 1)	-47.565 (-54.314, -40.815)
Average gamma-ray dose-rate by 50K-BEDSUP (group 18 vs group 1)	-48.382 (-55.086, -41.678)
Average gamma-ray dose-rate by 50K-BEDSUP (group 19 vs group 1)	-49.446 (-56.194, -42.698)
Average gamma-ray dose-rate by 50K-BEDSUP (group 20 vs group 1)	-50.423 (-57.534, -43.312)
Average gamma-ray dose-rate by 50K-BEDSUP (group 21 vs group 1)	-51.325 (-58.379, -44.272)
Average gamma-ray dose-rate by 50K-BEDSUP (group 22 vs group 1)	-51.789 (-58.897, -44.681)
Average gamma-ray dose-rate by 50K-BEDSUP (group 23 vs group 1)	-57.186 (-64.902, -49.471)
Average gamma-ray dose-rate by 50K-BEDSUP (group 24 vs group 1)	-59.352 (-67.005, -51.700)
Average gamma-ray dose-rate by 50K-BEDSUP (group 25 vs group 1)	-72.423 (-82.245, -62.600)
Average gamma-ray dose-rate by 50K-BEDSUP group	-0.208 (-0.303, -0.113)
Easting (km)	-0.011×10^{-3} (-302.555×10^{-3}, -112.838×10^{-3})
Easting (km)^2	23.263×10^{-6} (-13.300×10^{-6}, 59.826×10^{-6})
Average gamma-ray dose-rate by 50K-BEDSUP surface group	0.097 (0.025, 0.169)
Average gamma-ray dose-rate by 50K-BEDSUP bedrock group	-0.096 (-0.204, 0.011)
Average gamma-ray dose-rate weighted by 1/distance$^{3.0}$	0.082 (0.039, 0.126)
Average gamma-ray dose-rate weighted by 1/distance$^{0.5}$ (different	2.137 (1.526, 2.747)

geology), 1/distance$^{1.0}$ (same geology)	
Average gamma-ray dose-rate weighted by 1/distance$^{0.0}$	-0.792 (-1.109, -0.476)
Average gamma-ray dose-rate weighted by 1/distance$^{0.5}$ (different geology), 1/distance$^{1.5}$ (same geology)	-0.764 (-1.107, -0.422)

[a]Note: a linear term in Easting has been added to the OLS optimal model to render it algebraically complete.

Appendix C. More detailed description of statistical methods

Empirical spatial interpolation variable selection

A number of candidate dose-interpolation measures were selected, the predictive performance of which was evaluated using jackknife methods (Davison and Hinkley 1997). The dose-rate at a point x, was interpolated based on measurements of dose-rate $\{d_i\}$ at spatial coordinates $\{x_i\}$, $i = 1,...,N$. Estimates based on some sort of inverse power k of distance were defined as:

$$D_{est,dist}(x \mid \{(d_i, x_i)\}, S) = \frac{\sum_{i \in S(x)} \frac{d_i}{\max[\mid x - x_i \mid, d_{\min}]^k}}{\sum_{i \in S(x)} \frac{1}{\max[\mid x - x_i \mid, d_{\min}]^k}} \tag{C1}$$

Special cases include $k = 0$ (constant), $k = 1$ (inverse distance) and $k = 2$ (inverse distance squared). There is inevitably some uncertainty in the spatial resolution of the data, and for this reason and because there are distinct points with the same grid coordinates, which would otherwise cause the expression to be indeterminate, the distance between candidate points was limited to some minimum distance, $d_{\min}$, generally taken to be 0.1 km.

The set function $S(x)$ yields a set of "eligible" points from the available points $i = 1,...,N$. Two alternative choices for eligible points were used:

a) one based on being within a certain fixed distance: $S(x) = \{i : \mid x_i - x \mid < r_{\max}\}$, or

b) one based on the m nearest neighbours $S(x) = \{i \in \{j_1, j_2, ..., j_m\} : \max_{i = j_1, j_2, ..., j_m} \mid x_i - x \mid \leq \min_{i \in \{1,...,N\} \setminus \{j_1, j_2, ..., j_m\}} \mid x_i - x \mid\}$.

For the first type of model, in a small number of instances relatively few measurements were available within the specified radius, and because estimates based on a few nearest neighbours may prove unreliable, they were replaced by the average over an area in which the reference point fell. This was generally done whenever there were fewer than two neighbours within the appropriate distance band ($N_{\min}$ =2). Four such areas were tested:

(a) county district;

(b) postcode district;

(c) LEX-ROCK geological area; and

(d) 10 km square grid.

However, the results using county district, postcode district or 10 km square grid are very similar to those presented, using LEX-ROCK, in Tables A2, A3, and are not shown further.

A jackknife methodology was used for determining the possible bias or estimating error, as follows:

a) Take the ith point in the dataset, (d_i, x_i), the point "omitted", leaving the remainder points in the dataset $\left\{(d_j, x_j)\right\}_{j\neq i}$.

b) Using the data "left in" $\left\{(d_j, x_j)\right\}_{j\neq i}$ construct an estimate for a general point x: $D_{est,dist}(x \mid \left\{(d_j, x_j)\right\}_{j\neq i}, S)$ as above.

c) Compute the estimating squared-error $ESR_i = \left[d_i - D_{est,dist}(x_i \mid \left\{(d_j, x_j)\right\}_{j\neq i}, S)\right]^2$ (squared error based on sample d_i "omitted").

d) Repeat a)-c) "omitting" all N (=10,199) datapoints in turn and average the estimating mean-square-error (*MSE*) sum of squares $MSE(d_1, d_2, ..., d_N) = \frac{1}{N}\sum_{i=1}^{N} ESR_i$.

In order to determine the standard deviation of the MSE a modification of the standard jackknife estimate was used. Efron and Stein (Efron and Stein 1981) demonstrate that under certain regularity conditions, the variance estimate of a statistic $S(X_1, X_2, ..., X_N)$, a function of some sample $\{X_1, X_2, ..., X_N\}$, can be estimated via $\frac{N-1}{N}\sum_{i=1}^{N}\left[S_{(-X_i)} - S_{(-)}\right]^2$, where $S_{(-X_i)} = S(X_1, X_2, ..., X_{i-1}, X_{i+1}, ..., X_N)$ is the statistic obtained by leaving out X_i, and $S_{(-)} = \frac{1}{N}\sum_{i=1}^{N} S_{(-X_i)}$. Because of the excessive computational burden in estimating the statistic $\frac{N-1}{N}\sum_{i=1}^{N}\left[S_{(-X_i)} - S_{(-)}\right]^2$ for $MSE(d_1, d_2, ..., d_N)$, a Monte Carlo sampling strategy was adopted to approximate this quantity, choosing $n = 20$ points, $d_{k_1}, d_{k_2}, ..., d_{k_{20}}$, at random from the $N = 10{,}199$ and estimating the variance in the MSE by:

$$\frac{(N-1)^2}{N}\left\{\frac{1}{n-1}\sum_{i=1}^{n}\left[MSE_{(-d_{k_i})}-MSE_{(-)}\right]^2\right\} \qquad \text{(C2)}$$

The same technique can be applied to a statistic comprising the difference of MSE from two different methods, $S_{a,b}(d_1,d_2,...,d_N)=MSE_a(d_1,d_2,...,d_N)-MSE_b(d_1,d_2,...,d_N)$. The above jackknife method was used to assess the variance of this difference statistic in precisely the same way. Having obtained the SD of the difference, $\sigma_{a,b}=SD\left[S_{a,b}(d_1,d_2,...,d_N)\right]=\left(\mathrm{var}\left[S_{a,b}(d_1,d_2,...,d_N)\right]\right)^{0.5}$, the *p*-value associated with it was then computed being at least as large as the observed value, $S_{a,b}(d_1,d_2,...,d_N)=MSE_a(d_1,d_2,...,d_N)-MSE_b(d_1,d_2,...,d_N)$, in the standard way *via*:

$$P\left[N(0,1)\geq S_{a,b}(d_1,d_2,...,d_N)/\sigma_{a,b}\right] \qquad \text{(C3)}$$

i.e., the probability of a standard Normal random variable being this size or bigger. In particular this formula was used to assess the extent to which the various area-based measures considered in Tables 5 and A5 compare with the optimal distance-based measure given in Tables 3 and 4.

Use of outdoor gamma-ray dose-rates

A linear regression model was fitted to the internal (indoor) (d_i) and external (outdoor) $(d_{i,ext})$ dose-rate data:

$$d_i=\alpha d_{i,ext}+\varepsilon_i \qquad \text{(C4)}$$

where $\varepsilon_i \sim N(0,\sigma^2)$ are assumed to be independent and identically distributed. If for each point in the test sample, the external dose-rate at the nearest measurement point is taken, and the model-estimated internal dose-rate calculated by equation (C4), this value is then used as the estimated dose-rate (instead of that produced by model (C1)), and the MSE calculated as previously.

Empirical ordinary least squares model selection and fitting, and Gaussian-Matérn maximum likelihood fitting

Having constructed the necessary set of candidate interpolation and other variables as above, an empirical model using them was then constructed. A more standard geostatistical model was also fitted. Both sets of models were fitted to the indoor gamma-ray dose-rate data. The geoR R library was used (Diggle and Ribeiro 2007) for fits of the geostatistical model. One of the fundamental means of evaluating spatial correlations is the semi-variogram, which is defined for a spatial point process $Y(.)$, at a pair of (spatial) points, x, x', by:

$$\begin{aligned} V(x,x') &= 0.5E\left[Y(x)-Y(x')\right]^2 \\ &= 0.5E\left[(Y(x)-E[Y(x)])-(Y(x')-E[Y(x')])+(E[Y(x)]-E[Y(x')])\right]^2 \\ &= 0.5\left\{\mathrm{var}[Y(x)]+\mathrm{var}[Y(x')]-2\,\mathrm{cov}[Y(x),Y(x')]+(E[Y(x)]-E[Y(x')])^2\right\} \end{aligned} \tag{C5}$$

Assuming that the means are spatially constant (so that the last term vanishes), and that covariance structure depends only on the vector difference $x-x'$ and that correlations are spatially isotropic, this has the additional property that:

$$V(x,x') = V(\| x-x' \|) \tag{C6}$$

Now assume that $S(x)$ is a spatial Gaussian process with mean $\mu(x)$ and variance $\sigma^2(x) = \mathrm{var}[S(x)]$ and correlation function $\rho(x,x') = Corr[S(x),S(x')] = \rho(\| x-x' \|)$. One can assume that $Y(x) = S(x) + Z(x)$ where $Z(x)$ are Gaussian iid $N(0,\tau^2)$, and independent of all $S(x)$. Notice that then $\mathrm{var}[Y(x)] = \sigma^2(x) + \tau^2$ and $\mathrm{cov}[Y(x),Y(x')] = \sigma(x)\sigma(x')\rho(\| x-x' \|) + \tau^2 1_{x=x'}$, so that when the variance of $S(x)$ is spatially uniform ($\sigma^2(x) \equiv \sigma^2$), and $S(x)$ has constant mean ($\mu(x) \equiv \mu$), then (C5) reduces to:

$$\begin{aligned} V(x,x') &= \sigma^2 + \tau^2 - [\sigma^2 \rho(\| x-x' \|) + \tau^2 1_{x=x'}] \\ &= \sigma^2 [1-\rho(\| x-x' \|)] + \tau^2 1_{x \neq x'} \end{aligned} \tag{C7}$$

Since the correlation coefficient would often be expected to decrease with distance, going to 0 at large distances, this means that the variogram will generally be an increasing function of distance, with asymptotic (large distance) value given by $\mathrm{var}(Y(x))$. In Fig. 1 the semi-variogram corresponding to expression (C6) for the raw gamma dose rate is plotted; the Monte Carlo confidence intervals are derived by randomly permuting the residuals, and give a measure of the statistical significance of the observed changes of correlation with distance (Diggle and Ribeiro 2007).

In order to construct an empirical model that satisfactorily explains the variation of mean dose-rate, the framework of ordinary least squares was initially adopted, ignoring the spatial correlation in dose-rate. An iterative procedure was used based on minimising the Akaike Information Criterion (AIC) (Akaike 1973; Akaike 1981) to determine the optimal model. AIC penalizes against overfitting by adding 2 x [number of fitted parameters] to the model deviance. The iterative mixed-forward-backward stepwise procedure method that was employed selects at each stage the best variable to enter the model as that which results in the largest reduction in AIC. At each stage candidate variables are also selectively dropped from the model, until a set of candidate variables remain none of which can be dropped from the model without increasing AIC. These iterative cycles of adding variables to the model and selectively dropping them are repeated until there is a final set of variables selected, none of which can be dropped without increasing AIC, and none of the remaining variables can be added to the model without increasing AIC. The mixed-forward-backward step-AIC routine using ordinary least squares (OLS) was performed in R (R Project version 3.1.1 2014). The candidate variables that could be used are listed in Appendix B Table B1, and the optimal model variables selected in Table 6. Figure 4 plots the semi-variogram for the residuals from this optimal OLS model. It should be noticed that a number of quantities was constructed based on the 50K-BEDSUP geologic measure. In particular, the mean gamma-ray dose-rate was computed by 50K-BEDSUP code, and having sorted 50K-BEDSUP codes by mean gamma-ray dose-rate, this was used to group these codes into 25, 50, 100, or 200 similar (by mean dose-rate) areas. The groupings into 50 areas are nested within those for 25, the groupings for 100 are nested within those for 50 (and *a fortiori* therefore also within the groupings into 25), and the groupings for 200 are nested within those for 100 (and *a fortiori* therefore also within the groupings into 50 and 25). The resulting groups of 50K-BEDSUP codes were then used in the models as factor variables in the standard way. An *ad hoc* grouping of the 50K-BEDSUP Bedrock codes into 16 geological groups chosen *a priori* as a candidate classifier variable was also used.

Various simple purely spatial models were also fitted to the data, via OLS. A 6^{th} order polynomial model in the spatial coordinates (Easting, Northing), of the form was first fitted:

$$
\begin{aligned}
D_i &= \alpha_0 + \sum_i \alpha_i 1_{\text{COM2006Bedrock}=i} + Easting + Northing \\
&+ Easting^2 + Easting \cdot Northing + Northing^2 \\
&+ Easting^3 + Easting^2 \cdot Northing + Easting \cdot Northing^2 + Northing^3 \\
&\ldots \\
&+ Easting^6 + Easting^5 \cdot Northing + Easting^4 \cdot Northing^2 + \ldots + Northing^6 \\
&+ \varepsilon_i
\end{aligned}
\qquad \text{(C8)}
$$

The optimal model was selected via a mixed forward/backward step-AIC algorithm (using OLS) from the full 6th order polynomial model (C8). A pure-linear-spatial sub-model of (C8) (omitting the terms $Easting^2 + Northing^2 + Easting \cdot Northing$ and all higher order terms) was also employed, as was a quadratic-spatial sub-model of (C8) (omitting the terms $Easting^3 + Easting^2 \cdot Northing + Easting \cdot Northing^2 + Northing^3$ and all higher order terms), and pure cubic (3rd order) and quartic (4th order) spatial polynomials in the same manner.

Finally, a more standard spatial geostatistical model was fittedto the gamma-ray dose-rate via maximum likelihood (Diggle and Ribeiro 2007). Specifically, it was assumed that the gamma-ray dose-rate Y_i at spatial location x_i was given by:

$$Y_i = \mu(x_i) + S(x_i) + Z_i \qquad \text{(C9)}$$

where $\mu(x_i) = E[Y_i]$ is the mean of Y_i, $S(.)$ is a stationary Gaussian process with mean 0 and variance σ^2, and spatial correlation $\rho(u) = Corr\left[S(x), S(x-u)\right]$, and Z_i are independent Gaussian random variables with mean 0 and nugget variance τ^2. The Matérn model (Matérn 1960) was employed, which assumes that:

$$\rho(u \mid \phi, \kappa) = 2^{\kappa-1}(u/\phi)^{\kappa} K_{\kappa}(u/\phi) \qquad \text{(C10)}$$

where $K_{\kappa}(.)$ is a modified Bessel function of order κ (Diggle and Ribeiro 2007). The Matérn model was chosen because of its flexibility. The integer part of the parameter κ determines the mean-square differentiability of the process $S(.)$ – the process is mean-square differentiable when $\kappa > 1$, and mean-square twice differentiable when $\kappa > 2$. Formal estimation of the parameter κ is difficult and was not attempted here. As a working model of the mean process, $\mu(x)$, the fit was made using the 16-level classification of the 50K-BEDSUP-Bedrock variable, as above. Examination of the empirical variogram with this mean model suggested a differentiable process (so with $\kappa > 1$) with measurement error (nugget) variance of about $\tau^2 = 0.04$. Preliminary curve fitting, by eye, suggested that $\kappa = 2.5$ gave a reasonable fit; this value was used in all subsequent analyses. With this parameter fixed the remaining parameters in the model (16 levels of $\mu(x)$, ϕ, τ^2) were estimated by maximum likelihood.

The fit of all these models was tested using a standard cross-validation process. The models were fitted to a randomly selected 70% of the data, and the indicated models were then used to predict gamma-ray dose-rate in the remaining 30% of the data. Table 7 shows the mean-square-error estimated in the 30% test sample.

In order to assess possible model heteroscedasticity, from the various optimal models the residual SD were estimated, and regressed against the model predicted mean, averaged by 50K-BEDSUP area. A linear model was then fitted via linear regression (OLS), of the form:

$$SD_i = \alpha_0 + \alpha_1 Mean_i + \varepsilon_i \qquad \text{(C11)}$$

Simulation study to test optimal empirical OLS-fitted model and Gaussian-Matérn maximum-likelihood fitted model

From the given dose-rate measurement set $M = \{(x_i, y_i, D_i) : i = 1,...,10,199\}$ a jittered set of measurement locations $L = \{(x_i', y_i') : i = 1,...,10,199\}$ was constructed (very close to the actual measured locations) and the predicted dose-rate measurements were constructed for this and for the full set of case-control measurement locations (n=126,817) under the two models (ordinary least squares (OLS), Gaussian-Matérn (Diggle and Ribeiro 2007) maximum-likelihood estimated (MLE)) (using for each model parameters derived from fitting to the original 10,199 dose-rate data):

$$M_{OLS} = \{(x_i', y_i', D_{i,OLS}) : i = 1,...,10,199\}$$

$$M_{OLS,F} = \{(x_{i,F}, y_{i,F}, D_{i,OLS,F}) : i = 1,...,126,817\}$$

$$M_{MLE} = \{(x_i', y_i', D_{i,MLE}) : i = 1,...,10,199\}$$

$$M_{MLE,F} = \{(x_{i,F}, y_{i,F}, D_{i,MLE,F}) : i = 1,...,126,817\}$$

Then the case-control sets for the two full datasets were imputed:

$$C_{OLS,F} = \{(c_{i,OLS,F}, x_{i,F}, y_{i,F}, D_{i,OLS,F}) : i = 1,...,126,817\}$$

$$C_{MLE,F} = \{(c_{i,MLE,F} x_{i,F}, y_{i,F}, D_{i,MLE,F}) : i = 1,...,126,817\}$$

subject to the assumed standard semi-linear logistic model (Preston et al. 1998). An excess odds ratio (EOR) per Gy of $\beta = 120\ \text{Gy}^{-1}$ was assumed, corresponding to the central estimate of Kendall *et al.* (Kendall et al. 2013) for leukaemia, using the given ages of diagnosis. Specifically, the assumed logistic model for the matched case-control set k, comprising $n_k + 1$ individuals (with case indexed as $i = 0$, controls as $i = 1,...,n_k$), with ages at diagnosis $(a_{ik})_{i=0}^{n_k}$, was:

$$P[C_{0k} = 1 \mid (D_{ik})_{i=0}^{n_k}, (a_{ik})_{i=0}^{n_k}, \beta] = \frac{1 + \beta a_{ik} D_{0k}}{\sum_{i=0}^{n_k} [1 + \beta a_{ik} D_{ik}]} \quad \text{(C11)}$$

Appendix D. Outline of cross-validation method of comparing the Gaussian-Matérn and ordinary least squares (OLS) methods of model fitting

Using the optimal OLS and Gaussian-Matérn MLE models refitted to the predicted measurement set based on the two models, $M_{OLS} = \{(x_i', y_i', D_{i,OLS}) : i = 1,...,10,199\}$ and $M_{MLE} = \{(x_i', y_i', D_{i,MLE}) : i = 1,...,10,199\}$, the predicted dose-rates were derived for the full dataset of 126,817 measurement locations in all four possible ways:

D1) $M_{OLS-OLS,F} = \{(x_{i,F}, y_{i,F}, D_{i,OLS-OLS,F}) : i = 1,...,126,817\}$

D2) $M_{OLS-MLE,F} = \{(x_{i,F}, y_{i,F}, D_{i,OLS-MLE,F}) : i = 1,...,126,817\}$

D3) $M_{MLE-OLS,F} = \{(x_{i,F}, y_{i,F}, D_{i,MLE-OLS,F}) : i = 1,...,126,817\}$

D4) $M_{MLE-MLE,F} = \{(x_{i,F}, y_{i,F}, D_{i,MLE-MLE,F}) : i = 1,...,126,817\}$

Then the logistic models were fitted to the resulting four datasets:

Da1) $C_{OLS-OLS,F} = \{(c_{i,OLS,F}, x_{i,F}, y_{i,F}, D_{i,OLS-OLS,F}) : i = 1,...,126,817\}$

Da2) $C_{OLS-MLE,F} = \{(c_{i,OLS,F}, x_{i,F}, y_{i,F}, D_{i,OLS-MLE,F}) : i = 1,...,126,817\}$

Da3) $C_{MLE-OLS,F} = \{(c_{i,MLE,F}, x_{i,F}, y_{i,F}, D_{i,MLE-OLS,F}) : i = 1,...,126,817\}$

Da4) $C_{MLE-MLE,F} = \{(c_{i,MLE,F}, x_{i,F}, y_{i,F}, D_{i,MLE-MLE,F}) : i = 1,...,126,817\}$

A linear-logistic model was fitted to the two simulated matched case-case-control datasets using the four re-estimated sets of dose-rates (D1)-(D4) above by maximum-likelihood (McCullagh and Nelder 1989) using EpiWin (Preston et al. 1998); all confidence intervals were estimated by means of the profile likelihood (McCullagh and Nelder 1989). Results are presented in Table 9 and Fig. 6.

Appendix E. Methods used to estimate model-predicted means and variance for individual data points for the Gaussian-Matérn and ordinary least squares (OLS) models.

Gaussian-Matérn model

As outlined by Diggle and Ribeiro (Diggle and Ribeiro 2007), and at greater length in Appendix C, a model was assumed for gamma-ray dose-rate Y_i at spatial location x_i given by:

$$Y_i = Y(x_i) = \mu(x_i) + S(x_i) + Z_i \tag{E1}$$

in which $\mu(x_i) = E[Y_i]$ is the mean of Y_i, which may be (and will be in the present case) spatially non-uniform, $S(.)$ is a stationary Gaussian process with mean 0 and variance σ^2, and spatial correlation $\rho(u) = Corr\left[S(x), S(x-u)\right] = \rho(\| u \|)$ (i.e., isotropic, depending only on the distance apart of the two measured points $x, x-u$), and Z_i are independent Gaussian random variables with mean 0 and nugget variance τ^2. This implies that $(Y_i)_{i=1}^N$ is multivariate Gaussian with mean $\mu(x_i)$ and variance $\sigma^2 V = \sigma^2[R + \upsilon^2 I] = \sigma^2 R + \tau^2 I$ where I is the NxN identity matrix, and $R = (r_{ij})_{i,j=1}^N = \left(\rho(\| x_i - x_j \|)\right)_{i,j=1}^N$. If now one has an arbitrary point x, then $(Y(x), (Y(x_i))_{i=1}^N)$ is multivariate Gaussian with mean $(\mu(x), (\mu(x_i))_{i=1}^N)$ and variance $\begin{bmatrix} \sigma^2 & \sigma^2 r^T \\ \sigma^2 r & \sigma^2 V \end{bmatrix}$ where the vector r is given by $r = (r_i)_{i=1}^N = \left(\rho(\| x - x_i \|)\right)_{i=1}^N$. Then it is easily shown (e.g., p.136 in (Diggle and Ribeiro 2007)) that the conditional distribution of $Y(x)$ given $(Y(x_i))_{i=1}^N$ is also multivariate Gaussian with mean:

$$\mu(x) + r^T V^{-1} \left[Y(x_i) - \mu(x_i)\right]_{i=1}^N \tag{E2}$$

and variance:

$$\sigma^2 \left[1 - r^T V^{-1} r\right] \tag{E3}$$

In estimating these the maximum-likelihood estimates of $\mu(), V, \sigma^2$, namely $\hat{\mu}(), \hat{V}, \hat{\sigma}^2$ were substituted.

Ordinary least squares model

It is assumed here that the model for the dose-rate at Y_i at spatial location x_i is given by:

$$Y_i = (X\beta)_i + \varepsilon_i \tag{E2}$$

where X is the design matrix (composed of columns of distance weighted sums of dose-rates, COM2006-weighted dose-rates, spatial coordinates, etc), β are the (unknown) model parameters, and ε_i are independent and identically distributed $N(0,\sigma^2)$ errors. One may write this vectorially as:

$$Y = X\beta + \varepsilon \tag{E2'}$$

Using standard linear model theory β can be estimated via:

$$\hat{\beta} = (X^T X)^{-1} X^T Y \tag{E3}$$

and σ^2 via:

$$\hat{\sigma}^2 = \frac{1}{N-p}[Y - X\hat{\beta}]^T [Y - X\hat{\beta}] = \frac{1}{N-p} Y^T [1 - X(X^T X)^{-1} X^T] Y \tag{E4}$$

where p is the number of model parameters (in the present case 49). Assuming that one wants to predict the gamma-ray dose-rates $Y^{pred} = (Y_j^{pred})_{j=1}^M$ at a new vector of spatial locations $(x_j^{pred})_{j=1}^M$ (this will of course be associated with a new design matrix X^{pred}) the best estimate of the Y^{pred} will be:

$$\hat{Y}^{pred} = X^{pred} (X^T X)^{-1} X^T Y \tag{E5}$$

which has variance:

$$\mathrm{var}[\hat{Y}^{pred}] = \sigma^2 X^{pred} (X^T X)^{-1} X^{pred,T} \tag{E6}$$

Fig 1 Semi-variogram of gamma-ray dose-rate, with Monte Carlo 95% confidence interval (CI) derived by resampling residuals 99 times.

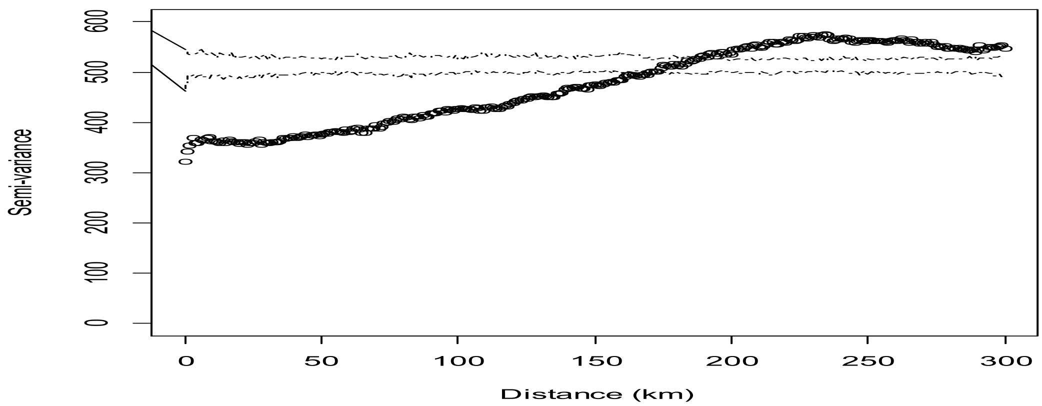


Fig 2 Map showing the county districts of Great Britain and the district-averaged dose-rates (nGy/h), and the locations of the individual dose-rate measurement locations (red dots).

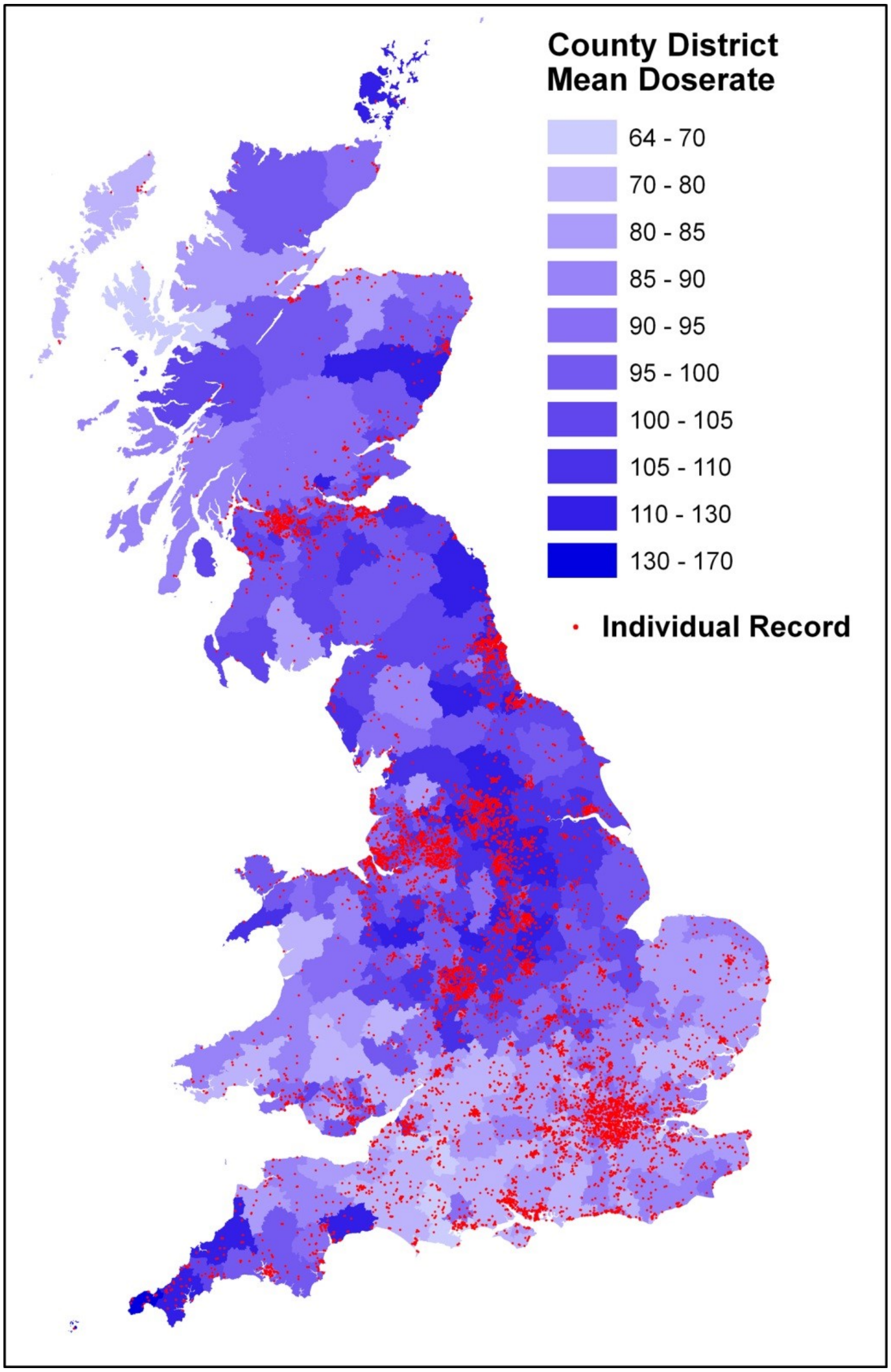


Fig 3 Internal (indoor) gamma-ray dose-rate vs external (outdoor) gamma-ray dose-rate

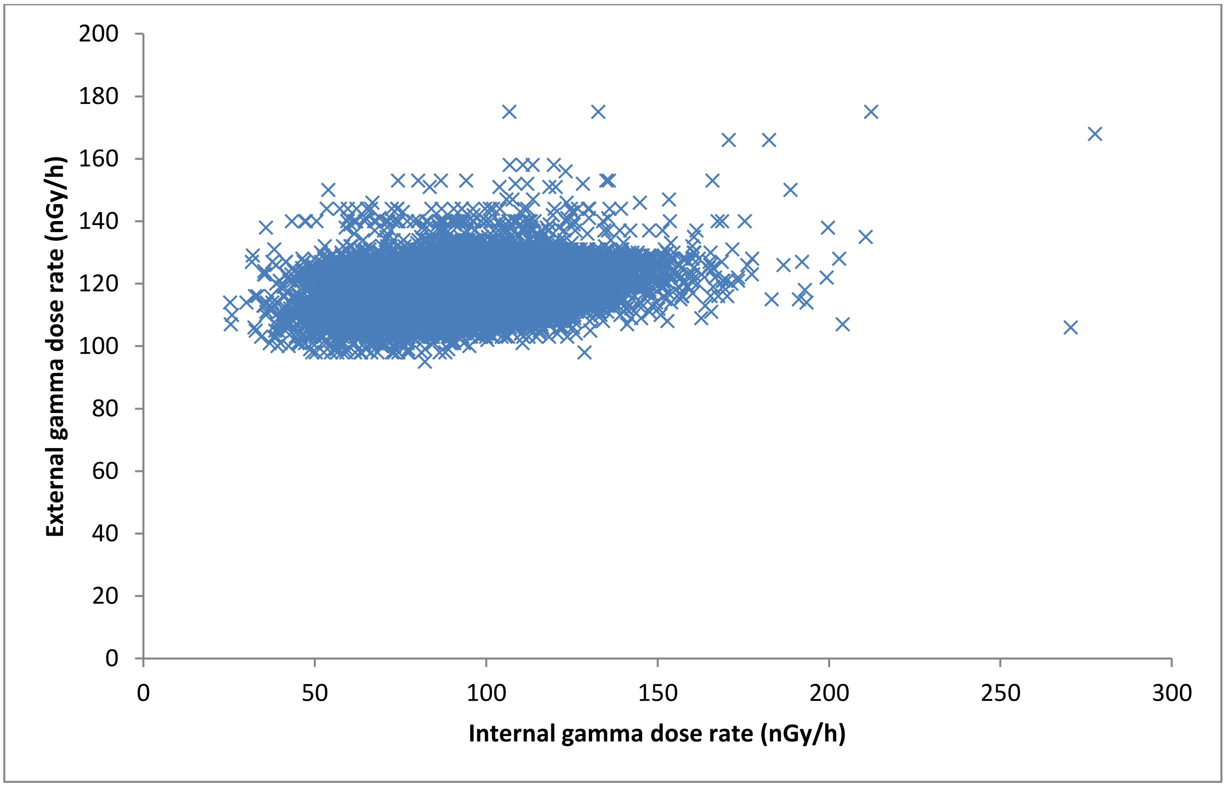


Fig 4 Semi-variogram for residuals from gamma-ray dose-rate model fitted by ordinary least squares (optimal model as given in Table 6), with Monte Carlo 95% CI derived by resampling residuals 99 times.

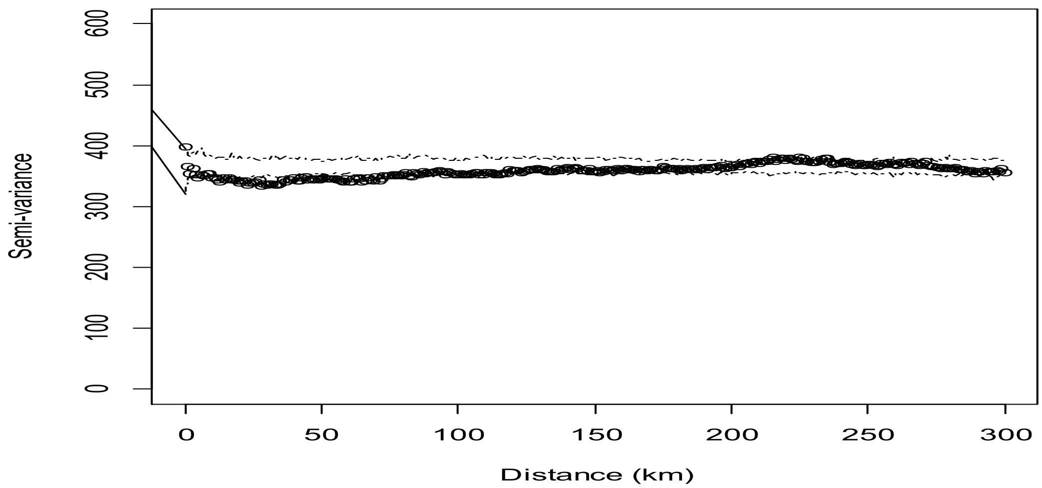


Fig 5 Individual variances for full location dataset (n=126,817), estimated using Gaussian-Matérn model and ordinary least squares model fitted to dose-rate measurement data (n=10,199). Solid red line is the line X=Y.

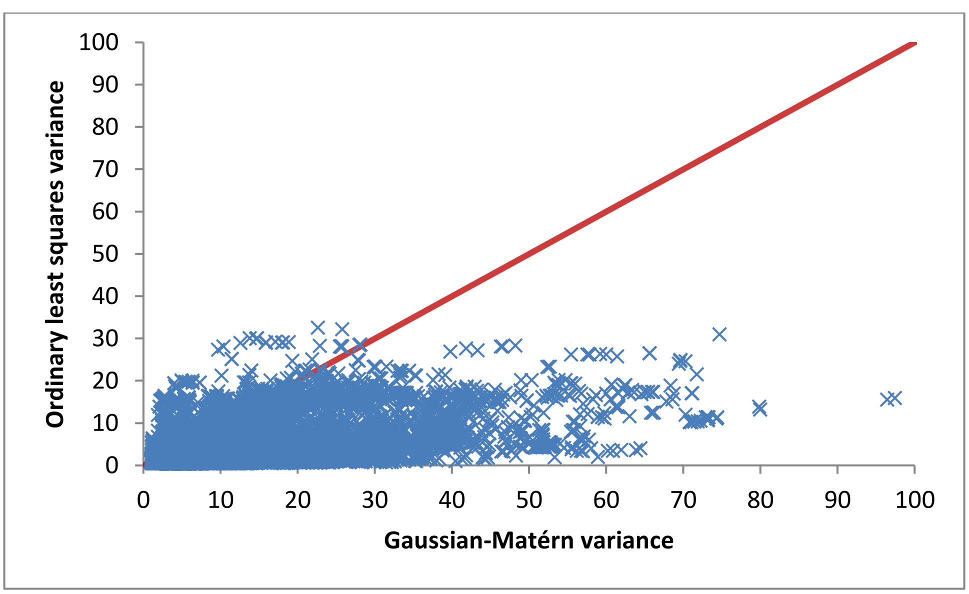
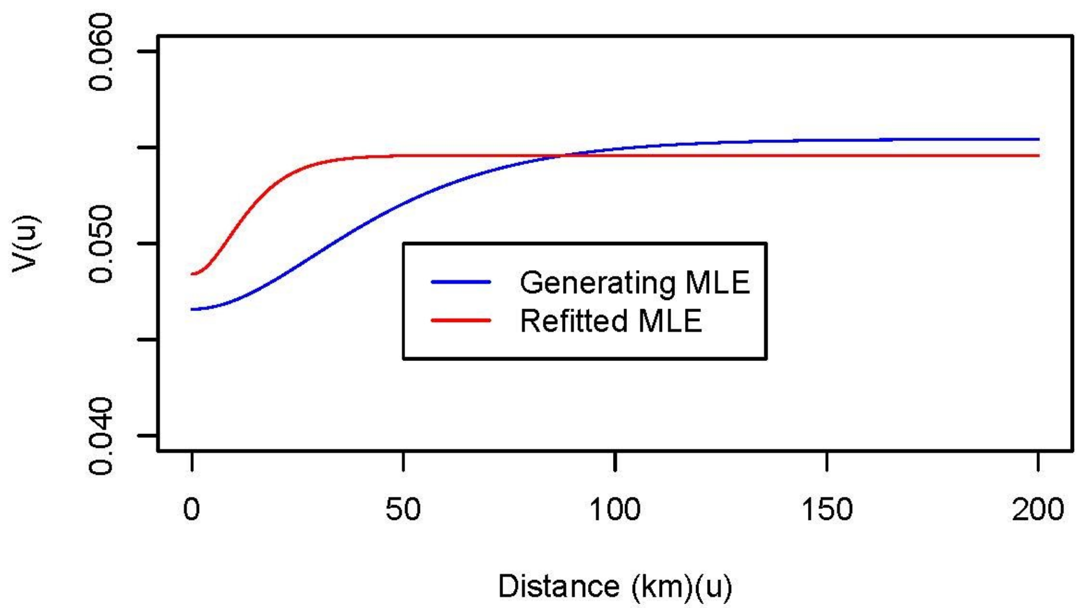


Fig 6 Comparison of variograms of generating Gaussian-Matérn model and of model refitted to jittered 10,199 spatial data.



Reference List

Akaike H (1973) Information theory and an extension of the maximum likelihood principle. Akadémiai Kiadó, Budapest

Akaike H (1981) Likelihood of a model and information criteria J Econometrics 16:3-14

Appleton JD (2005) Simplified geological classification for radon potential mapping in England and Wales (based on DiGMapGB-50 V3.12) vol Internal Report IR/05/153R. British Geological Survey,

Appleton JD, Miles JC (2010) A statistical evaluation of the geogenic controls on indoor radon concentrations and radon risk Journal of environmental radioactivity 101:799-803 doi:10.1016/j.jenvrad.2009.06.002

Axelson O, Fredrikson M, Akerblom G, Hardell L (2002) Leukemia in childhood and adolescence and exposure to ionizing radiation in homes built from uranium-containing alum shale concrete Epidemiology 13:146-150

Belson M, Kingsley B, Holmes A (2007) Risk factors for acute leukemia in children: a review Environ Health Perspect 115:138-145

Bennett J, Little MP, Richardson S (2004) Flexible dose-response models for Japanese atomic bomb survivor data: Bayesian estimation and prediction of cancer risk Radiat Environ Biophys 43:233-245 doi:10.1007/s00411-004-0258-3 [doi]

Carroll RJ, Ruppert D, Stefanski LA, Crainiceanu CM (2006) Measurement error in nonlinear models. A modern perspective. Chapman and Hall/CRC, Boca Raton, FL

Davison AC, Hinkley DV (1997) Bootstrap methods and their application. Cambridge University Press, Cambridge

Diggle PJ, Ribeiro PJ, Jr (2007) Model-based geostatistics. Springer, 233 Spring Street, New York, NY 10013

Efron B, Stein C (1981) The jackknife estimate of variance vol 9.

Fair DRR, Howells H (1955) A survey of the natural gamma radioactivity in the west Cumberland area J Nuclear Energy 1:274-279

Goldthwait RP, Matsch CL (1989) Genetic Classification of Glacigenic Deposits. Final Report of the Commission on Genesis and Lithology of Glacial Quaternary Deposits of the International Union for Quaternary Research (INQUA). International Union for Quaternary Research, Commission on Genesis and Lithology of Glacial Quaternary Deposits. CRC Press/A.A. Balkema, Rotterdam

Green BMR, Lomas PR, Bradley EJ, Wrixon AD (1989) Gamma-radiation levels outdoors in Great Britain vol NRPB-R191. National Radiological Protection Board,, Chilton

Didcot

Oxon

OX11 0RQ

International Commission on Radiological Protection (2007) The 2007 Recommendations of the International Commission on Radiological Protection. ICRP publication 103 Ann ICRP 37:1-332 doi:S0146-6453(07)00031-0 [pii];10.1016/j.icrp.2007.10.003 [doi]

Kendall GM, Hughes JS, Oatway WB, Jones AL (2006) Variations in radiation exposures of adults and children in the UK JRadiolProt 26:257-276 doi:S0952-4746(06)23763-7 [pii];10.1088/0952-4746/26/3/001 [doi]

Kendall GM et al. (2013) A record-based case-control study of natural background radiation and the incidence of childhood leukaemia and other cancers in Great Britain during 1980-2006 Leukemia 27:3-9 doi:leu2012151 [pii];10.1038/leu.2012.151 [doi]

Kendall GM et al. (1994) Exposure to radon in UK dwellings vol NRPB-R272. NRPB Publications, Chilton, UK

Land CE (1980) Estimating cancer risks from low doses of ionizing radiation Science 209:1197-1203

Little MP, Hoel DG, Molitor J, Boice JD, Jr., Wakeford R, Muirhead CR (2008) New models for evaluation of radiation-induced lifetime cancer risk and its uncertainty employed in the UNSCEAR 2006 report Radiat Res 169:660-676 doi:RR1091 [pii];10.1667/RR1091.1 [doi]

Little MP et al. (2014) Impact of uncertainties in exposure assessment on estimates of thyroid cancer risk among Ukrainian children and adolescents exposed from the Chernobyl accident PLoS ONE 9:e85723 doi:10.1371/journal.pone.0085723

Little MP, Wakeford R, Lubin JH, Kendall GM (2010) The statistical power of epidemiological studies analyzing the relationship between exposure to ionizing radiation and cancer, with special reference to childhood leukemia and natural background radiation Radiat Res 174:387-402 doi:10.1667/RR2110.1 [doi]

Maduar MF, Hiromoto G (2004) Evaluation of indoor gamma radiation dose in dwellings Radiat Prot Dosimetry 111:221-228 doi:10.1093/rpd/nch329

Matérn B (1960) Spatial variation. In: Meddelanden fran Statens Skogsforsknings institut, vol Band 49. vol 5. Stockholm,

McCullagh P, Nelder JA (1989) Generalized linear models. 2nd edition. Chapman and Hall/CRC, Boca Raton, FL

Miles JCH, Appleton JD (2005) Mapping variation in radon potential both between and within geological units Journal of radiological protection : official journal of the Society for Radiological Protection 25:257-276 doi:10.1088/0952-4746/25/3/003

Muirhead CR, Cox R, Stather JW, MacGibbon BH, Edwards AA, Haylock RGE (1993) Estimates of late radiation risks to the UK population Documents of the NRPB 4:15-157

O'Brien K, Smart DF, Shea MA, Felsberger E, Schrewe U, Friedberg W, Copeland K (2003) World-wide radiation dosage calculations for air crew members Advances in Space Research 31:835-840

Ordnance Survey (2010) A guide to coordinate systems in Great Britain: An introduction to mapping coordinate systems and the use of GPS datasets with Ordnance Survey mapping. Ordnance Survey, Southampton

Ormsby T, Napoleon E, Burke R, Groessl C, Feaster L (2004) Getting to Know ArcGIS Desktop: The Basics of ArcView, ArcEditor, and ArcInfo Updated for ArcGIS 9 Getting to Know series, 2nd updated edn. ESRI Press, Redlands, CA

Pearce MS et al. (2012) Radiation exposures from CT scans in childhood and subsequent risk of leukaemia and brain tumours, an historical cohort study Lancet 380:499-505

Preston DL, Lubin JH, Pierce DA, McConney ME (1998) Epicure release 2.10. Hirosoft International, Seattle

R Project version 3.1.1 (2014) R version 3.1.1 http://www.r-project.org/.

Radioactive Substances Advisory Committee (1971) Handbook of Radiological Protection, Part 1: Data. Her Majesty's Stationery Office, London

Rao CR (2002) Linear statistical inference and its applications. 2nd edition. John Wiley & Sons, Inc, Singapore

Richardson S, Monfort C, Green M, Draper G, Muirhead C (1995) Spatial variation of natural radiation and childhood leukaemia incidence in Great Britain StatMed 14:2487-2501

Smith A (2013) Digital Geological Map of Great Britain, information notes, 2013. British Geological Survey Open Report OR/13/007. British Geological Survey, Keyworth, Nottingham, UK

Stayner L, Vrijheid M, Cardis E, Stram DO, Deltour I, Gilbert SJ, Howe G (2007) A Monte Carlo maximum likelihood method for estimating uncertainty arising from shared errors in exposures in epidemiological studies of nuclear workers Radiat Res 168:757-763 doi:RR0677 [pii];10.1667/RR0677.1 [doi]

United Kingdom Childhood Cancer Study Investigators (2002a) The United Kingdom Childhood Cancer Study of exposure to domestic sources of ionising radiation: 1: radon gas Br J Cancer 86:1721-1726 doi:10.1038/sj.bjc.6600276 [doi]

United Kingdom Childhood Cancer Study Investigators (2002b) The United Kingdom Childhood Cancer Study of exposure to domestic sources of ionising radiation: 2: gamma radiation Br J Cancer 86:1727-1731 doi:10.1038/sj.bjc.6600277 [doi]

United Nations Scientific Committee on the Effects of Atomic Radiation (UNSCEAR) (2000) Sources and effects of ionizing radiation. UNSCEAR 2000 report to the General Assembly, with scientific annexes. Volume I: Sources. vol E.00.IX.3. United Nations, New York

United Nations Scientific Committee on the Effects of Atomic Radiation (UNSCEAR) (2008) UNSCEAR 2006 Report. Annex A. Epidemiological Studies of Radiation and Cancer. United Nations, New York

United Nations Scientific Committee on the Effects of Atomic Radiation (UNSCEAR) (2010) UNSCEAR 2008 Report. Annex B. Exposures of the public and workers from various sources of radiation. United Nations, New York

University of Edinburgh (2014) http://digimap.edina.ac.uk/webhelp/geology/data_information/bgs_products_available/50k.htm . University of Edinburgh. http://digimap.edina.ac.uk/webhelp/geology/data_information/bgs_products_available/50k.htm . Accessed 10/2014 2014

UNSCEAR (2008) United Nations Scientific Committee on the Effects of Atomic Radiation Report to the General Assembly, Volume 1: Effects of Ionising Radiation with Scientific Annexes A and B. United Nations, New York

Wakeford R (2013) The risk of childhood leukaemia following exposure to ionising radiation-a review JRadiolProt 33:1-25 doi:10.1088/0952-4746/33/1/1 [doi]

Wakeford R, Tawn EJ (2010) The meaning of low dose and low dose-rate J Radiol Prot 30:1-3 doi:10.1088/0952-4746/30/1/E02 [doi]

Watson SJ, Jones AL, Oatway WB, Hughes JS (2005) Ionising radiation exposure of the UK population: 2005 review vol HPA-RPD-001. Health Protection Agency, Centre for Radiation, Chemical and Environmental Hazards, Radiation Protection Division, Chilton, Didcot, Oxon OX11 0RQ, UK

Wrixon AD et al. (1988) Natural radiation exposure in UK dwellings vol R190. National Radiological Protection Board, Chilton, Didcot, Oxon OX11 0RQ, UK