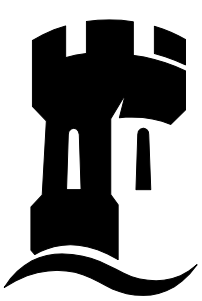

The University of Nottingham

Environmental exposure to metallic soil elements and risk of cancer in the UK population, using a unique linkage between THIN and BGS databases

Anwar Musah, MSc

Thesis submitted to the University of Nottingham for the degree of Doctor of Philosophy

August 2016

Abstract

Background: There have been many epidemiological studies into the influence of exposure to the most toxic elements on the risk of cancer in the workplace, mainly due to the exposure of certain occupational groups, or perhaps in populations near industrial sources. Toxic elements include arsenic, copper, nickel, and uranium; and many more of these elements have been shown to increase the risk of several different types of cancers in these highly-exposed groups. Many of these elements naturally exist in the soil, and the health impact of these levels of environmental exposures on the general population has received little attention to date possibly due to the belief that soil concentrations of these elements are too low to cause harm to the general population. Therefore, the long-term effect of such chronic exposure to metals in the soil remains unclear.

Aims and objectives: The goals are to utilise a new resource known as THIN-GBASE for conducting a series of environmental epidemiological studies to test the hypothesis that BCC, lung and GIT cancers are associated with high exposure to certain low-level metals in soil. We sought to use this resource in determining which soil metals should be tested for predicting each of the cancer outcomes.

Methods: For BCC, an ecological study was initially undertaken to assess the overall regional variation in BCC to provide national and contemporary breakdowns of incidence rates across the UK. The primary exposure of interest for BCC was low-level soil arsenic, and

we therefore quantified soil arsenic exposure levels based on the UK national safety limits for arsenic [i.e. *As*-C4SLs = 35 mg/kg]. A population-based cohort study was conducted to quantify the risks associated between the development of BCC and increasing levels of exposure to soil arsenic. For lung cancer, a two-stage process was adopted: 1) data mining analysis using the correlation-based filter selection model was used to find the restricted set of soil metals were best predictors for lung cancer; and 2) a prospective cohort study was use where these sets of elements were fitted together (adjusted for confounding variables) in a multivariable Cox proportional-hazards model to determine the risks associated between the development of lung cancer, with increasing levels of exposure to each specific element. For GIT cancers, a three-stage process was adopted: stages 1 and 2 used a similar methodology for the lung cancer study. In stage 3, all GIT cancers were divided into three broader outcomes i.e. upper GIT (includes mouth & oesophagus), stomach (as standalone) and colorectal (includes small, large, rectum and anal canal) cancers. A multivariate competing risk survival model was adjusted for the three different GIT cancers as competing events to identify associations between any of the selected group of metals found in stage 1 and GIT-specific cancers.

Results: For BCC, the findings for the ecological study show that overall EASRs & WASRs for BCC in the UK was 98.6 and 66.9 per 100,000 person-years, respectively. It indicates a large geographical variation in age-sex standardised incidence of BCC with the South East

having the highest incidence of BCC (202.7/100,000 person-years), followed by South Central (193.5/100,000 person-years) and Wales (185.7/100,000 person-years). Incidence rates of BCC were substantially higher in the least socioeconomically deprived groups. It was observed that increasing levels of deprivation led to a decreased rate of BCC ($p < 0.001$). In terms of age groups, the largest annual increase was observed among those aged 30-49 years. Assessment for soil arsenic indicated that individuals living in areas with concentrations ≥35mg/kg significantly had an increased hazard of developing BCC (35-70mg/kg: adjusted HR 1.08, 95% CI: 1.02-1.14; ≥70mg/kg: adjusted HR 1.17, 95% CI: 1.09-1.28). Urban residents with the highest exposure of soil arsenic had the greatest risk of developing BCC (≥ 70.0 mg/kg: HR 1.18, 95% CI: 1.06-1.36). For lung cancer, the correlation-based filter selection model identified aluminium, lead and uranium as the appropriate set of exposures for modelling lung cancer risk. Complete adjustments of hazards model showed evidence of an increased risk of developing lung cancer with elevated concentrations for only soil aluminium at medium levels ranging between 47,000-61,600mg/kg. Urban residents with the highest exposure of soil aluminium had the greatest risk of developing lung cancer (≥ 61,600mg/kg: HR 1.12, 95% CI: 1.04-1.22). For GIT cancers, the correlation-based filter selection model identified seven elements i.e. aluminium, phosphorus, zinc, uranium, calcium, manganese, and lead, as the appropriate set of exposures for predicting GIT cancer risk. The complete adjustment for hazards model indicated that the

risk of developing overall GIT cancers were significantly associated with elevated exposure levels of soil phosphorus only (873-1,127mg/kg: HR 1.08, 95% CI: 1.02-1.14; 1,127-1,456mg/kg: HR 1.07, 95% CI: 1.01-1.13; and ≥1,145mg/kg: HR 1.07, 95% CI: 1.01-1.13). There were no consistent relationships identified between any of the selected groups of elements and the GIT-specific cancer outcomes when adjusting for different GIT cancers as competing events.

Conclusion: There appears to be slight evidence of BCC, respiratory and GIT cancer risk with elevated exposure to soil arsenic, aluminium and phosphorus, respectively. The series of investigations conducted for this research are one of the first, if not, contemporary UK-based study to present novel estimates for a group of ill-defined pollutants. This research demonstrates that linking geochemical data with electronic primary care medical records can be a valuable approach of proving whether long term exposure to low-level soil contaminants may have a health consequence in the population.

List of publications:

i. **Musah A.**, Gibson JE., Leonardi-Bee J., Cave MR., Ander EL., Bath-Hextall F; (2013); Regional variations of Basal Cell Carcinoma incidence in the UK using THIN database (2004–10); British Journal of Dermatology; 169 (5): 1093-9; DOI: 10.1111/bjd.12446.

ii. **Gibson JE.**, Ander EL., Cave MR., Bath-Hextall F., Musah A., Leonardi-Bee J; (2016); Linkage of national soil quality measurements to primary care records in England and Wales: a new resource for investigating environmental impacts on human health; Population Health Metrics

The above paper has been accepted for publication

Press release(s):

i. South East coast has the highest rates of most common cancer, study reveals – For immediate release; (23rd May, 2013); British Association of Dermatologists; Press Release; URL: http://www.bad.org.uk/media/news?sitesectionid=154&from=01/01/2013%2000:00:00&to=01/01/2014%2000:00:00&range=2013

ii. Wales is UK nation with the highest rates of most common cancer, study reveals – For immediate release; (23rd May, 2013); British Association of Dermatologists; Press Release; URL: http://www.bad.org.uk/media/news?sitesectionid=154&from=0

1/01/2013%2000:00:00&to=01/01/2014%2000:00:00&range=2013

Conference attendance, poster and oral presentations:

i. Musah A, Gibson JE, Leonardi-Bee J, Cave MR, Ander EL, Bath-Hextall F. 2013. Environmental exposure to soil arsenic and risk of Basal cell carcinoma in England & Wales (2000 to 2012). **7th International Workshop on Chemical Bioavailability (3rd – 6th November 2013)** (oral presentation)

ii. Topic: Environmental exposure to soil arsenic and risk of Basal cell carcinoma in England & Wales (2000 to 2012). **EPH seminar 8th July 2013.** (oral presentation).

Acknowledgements

I would like to express special appreciation to my supervisors Professor Jo Leonardi-Bee and Dr Jack E. Gibson for conceiving this project, and for securing the funding for this project. I sincerely thank my supervisors for their continuous support, care of my wellbeing and the guidance throughout the work. They were always helpful in guiding me on the path of becoming an independent researcher. They are the greatest mentors a student could ever have.

I would like to express special thanks to Professor Fiona Bath-Hextall for her support and contributions in providing me with invaluable insights on clinical practice and knowledge of NMSCs and BCC. Special thanks to Dr Mark R. Cave and Dr E. Louise Ander for the technical support with the G-BASE database, and for their contributions and insights on principles of geochemistry and geology.

Also, I would like to thank Professor Richard Hubbard and Professor Joe West for providing me with technical support in managing lung and gastrointestinal cancer related Read codes, and Dr Yue Huang for the invaluable training and instructions on implementing data mining models for this research.

Finally, I would like to give special thanks to parents, my two sisters – Lalita and Hajra, and my wife Fuseina, for their encouragement, faith and support whilst I completed this thesis.

Table of contents

List of tables

List of figures

List of abbreviations

(95% CI) 95% confidence intervals

(BCC) basal cell carcinoma

(BGS) British Geological Survey

(C4SL) Category 4 Screening Levels

(CFS) correlation-based filter selection

(DEFRA) Department for Environmental Food and Rural Affairs

(Defra) Department of Environment, Food and Rural Affairs

(DL) detection limit

(EA) Environmental Agency

(EASR) European Age-Standardised Incidence Rates

(EMR) electronic medical records

(GAC) Generic Assessment Criteria

(G-BASE) Geochemical Baseline Survey of the Environment

(GIT) Gastrointestinal tract

(GP) general practices

(HR) Hazard Ratios

(ICRCL) Inter-Department Committee on the Redevelopment of Contaminated Land

(IQR) Interquartile range

(IR) incidence rates

(IRR) Incidence rate ratios

(LCC) large cell carcinoma

(NBC) normal background concentration

(NMSC) Non-melanoma skin cancers

(NSCLC) non-small cell lung carcinoma

(NSI-XRFS) National Soil Inventory x-ray fluorescence spectrometry

(OS) Ordnance Scale

(SCC) squamous cell carcinoma

(SCLC) small cell lung carcinoma

(SD) standard deviation

(SE) Standard Error

(SHA) strategic health authority

(THIN) The Health Improvement Network

(UV) Ultraviolet

(WASR) World Age-Standardised Incidence Rates

(WHO) World Health Organisation

(XRFS) x-ray fluorescence spectroscopy

(DNA) Deoxyribonucleic acid (DNA)

(ROS) reactive oxygen species (ROS)

Introduction

Chapter 1

1 Background

The earth's crust is the greatest source of all metals (metallic elements or compounds) that exists in the environment.[1,2] The vast majority of metals are ubiquitous, and are naturally occurring constituents of the lithosphere – the part of the earth's environment formed by crustal and uppermost solid mantle which is covered with soil.[3] Metallic elements in soil are heterogeneously distributed at varying concentrations,[4] and exist in a variety of chemical forms. Some, such as calcium, magnesium, iron, iodine, selenium and zinc, are essential to plants, animals (including humans) in trace amounts,[5-7] whilst others (such as cadmium, lead and mercury) have limited or no biological value.[7,8] In excess, all of these are typically naturally occurring elements can lead to toxicity in living organisms.[9-12]

Over the past decades, there has been interest in the assessment of geochemical and biological interactions characterised especially by the relationships between cancer and soil.[1,2,13-15] A concern has emerged among some public health practitioners and medical geologists that: (1) long-term low-level toxic effects of exposure to metals emerging from soil may have adverse health consequences,[16-18] and (2) that metals from soil entering the human bloodstream through long-term exposure may accumulate within tissues, leading to more severe adverse effects more typically associated with high-level exposure.[16-18]

There are three major routes for metals to leave the soil and enter the human body. The most prominent is the oral/ingestion pathway route of exposure, followed by the dermal and respiratory pathways.[18-20] Research in the UK and other middle to high income countries has used *in-vitro* methods to derive theoretical estimates of the oral bioaccessibility of potentially harmful soil metals[21-26] – that is, the fraction of the total concentration of a metal bound to soil (in various chemical forms) that is soluble in the gastrointestinal environmental and available for absorption (or uptake) into the bloodstream.[18] This group of studies has shown a close correlation between bioaccessible metal fractions and overall topsoil concentration levels of several metals,[27-29] and that these theoretical estimates are, in turn, positively associated with a range of biomarkers of human uptake (from hair, toe & finger nails and urine sample). [27-29] Furthermore, previous clinical studies have shown that these biomarkers are a proxy measure for the actual cause of adverse health conditions (including cancers).[22,27-31] This collection of studies has concluded that such harmful soil metals may potentially play an important role in adversely impacting human health and have advocated the need for further investigation to evaluate the direct associations between topsoil harmful metals and adverse health outcomes (including cancer).[22,27-31]

Direct epidemiological evidence supporting an association between potentially harmful soil metals and the development of cancers is substantially limited. The number of cases of cancer, and the time

that cases take to develop, are dependent on dose, duration and relative carcinogenicity of the exposure.[19,32,33] Except in areas of very severe soil contamination (where previous individual-level studies have tended to focus), human exposure to potentially harmful elements from soil is typically at relatively low levels (especially in high and middle-income countries where soil quality is assessed prior to granting permission for residential development), so a large cohort followed over a long period of time would be necessary to detect any increased risk. There is a lack of availability of individual-level datasets containing details of both soil exposures and health outcomes of a sufficient scale to provide the necessary statistical power[34] and area-level comparisons using overall diagnosis rates are problematic as they are especially vulnerable to ecological fallacies: areas which are highly polluted (where soils may contain relatively high levels of metallic elements of anthropogenic origin) are often inhabited by individuals who are relatively highly susceptible to cancer due to other factors (such as lower socio-economic status, tobacco use and poor diet).[19,32] This thesis will attempt to address this gap in knowledge by using a uniquely linked database between primary health care records from The Health Improvement Network (THIN) and geochemical data from the Geochemical Baseline Survey of the Environment (G-BASE) developed by the British Geological Survey (BGS).

This chapter introduces the key classifications of the most important metallic elements in soils and discusses their origins from soil &

minerals, and natural and anthropogenic metallic input into soil from external sources. A plausible framework for the major exposure pathways to soil metals, sources of exposure, and how they may lead to cancer is described. Finally, the importance of the UK soil guideline value – the Category 4 Screening Level (C4SL) - for specific soil metals is discussed, and the approach to testing the appropriateness of the current levels in the studies included in the remainder of the thesis is described.

1.1 Metallic elements and their classifications for carcinogenicity in humans

About 80.0% of the chemical elements occupying the periodic table are metallic in nature. Metallic elements are broadly classified as heavy[35–38] or light metals[39–41]. The majority of these elements are heavy metals, and they are found on the periodic table in groups III-XVI with periods ranging between IV and VII.[42] Heavy metals have atomic densities ranging from 3.5-5.0 g/cm^3. Common examples of heavy metals include arsenic, cadmium, lead, nickel, mercury and zinc. Light metals, on the other hand, are few in number having lower atomic densities than heavier metals. They are typically located outside groups III-XVI on the periodic table.[42] Examples of light metals are aluminium, silicon, magnesium and titanium.[39–41]

The carcinogenic potency metallic elements may exhibit in humans has been assessed using a variety of experimental studies on cancer in laboratory animals and through human health assessments and

epidemiological research.[9-12,43-45] Based on these findings, more than 900 agents with different properties (chemical, biological or physical in nature) have been evaluated by the International Agency for Cancer Research (IARC) and catalogued in monographs.[46] The main purpose of the IARC Monographs programme is to identify and evaluate potential environmental causes of cancer in humans.[46] Agents of focus include elements in pure, inorganic or organic forms;, chemicals (e.g. formaldehyde); complex mixtures present in air, soil and water (often arising from pollution, such as factory or automobile emissions); occupational exposures (e.g. asbestos fibres, sawdust); and physical agents (such as solar and ionising radiation).[46]

These agents (or carcinogens) are ranked accordingly into five categories (i.e. group 1, 2-A, 2-B, 3 and 4) ranging from agents that are deemed to be *carcinogenic to humans* (i.e. group I) to *probably not carcinogenic to humans* (i.e. group IV).[46] Agents categorised into group I are deemed *carcinogenic to humans* due to convincing evidence derived from epidemiological studies. [9-12,43-45] Group II agents are classed into two subcategories: group II-A refers to agents that are deemed *probably carcinogenic to humans*, whilst those in group II-B are *possibly carcinogenic to humans*. Agents are classified into the former group (II-A) when there is limited indication of carcinogenicity in humans and sufficient evidence in animal studies.[46] Limited evidence means that a positive association has been observed between the exposure of interest (i.e. agent) and cancer but that other explanations for the observations could not be ruled out.[46] The

latter (group II-B) comprises agents with limited evidence of carcinogenicity in humans and less than sufficient but acceptable evidence of carcinogenicity in experimental animals. [46] Group III comprises agents *not classifiable as to [their] carcinogenicity to humans* due to inadequate evidence of carcinogenicity found in most animal studies. When there is no evidence or a demonstrated lack of carcinogenicity in human and in experimental animals, they are classed as *probably not carcinogenic in humans*. [46] The IARC classification system for agents indicates only the strength of evidence that they may cause cancer; it is not intended to (and does not) indicate the degree of risk associated with exposure.[46]

Only a few metallic elements are found in group 1 and 2 (-A or -B) [9-12,43-45]. Inorganic arsenic, chromium (VI) and cadmium are examples of highly toxic metals classified as group 1 agents.[9,10,19,46,47] They are deemed *carcinogenic to humans* due to substantial evidence found from epidemiological studies focused on populations exposed such to metals largely from contaminated air and drinking water environments,[19,48,49] and from occupational hazards.[19,33,50] On the other hand, while inorganic lead, nickel and cobalt are also highly toxic to humans; they are classified as group 2 agents with the potential to be carcinogenic to humans due to the limited amount of evidence from clinical studies.[11,33,51,52] Overall, most metallic elements fall under group 3 (i.e. *not classifiable as to its carcinogenicity to humans*) because studies have not yet been carried out to ascertain the health effects.[46]

Metallic elements, including those that are classified as group 1 and 2 agent under the IARC monograph's classification system occur naturally in soil.[53] The concentrations of these metals in soil are dependent on the underlying minerology and processes of weathering which release metals into the soil, and on external (or anthropogenic) factors that influence the deposition of metallic elements on topsoil, but they are ubiquitous, and environmental exposure is therefore widespread among humans.[4,13,37,54,55] Long-term low-level uptake of these elements exposures in the UK and other middle to high income countries occurs primarily via inadvertent soil ingestion, inhalation and dermal absorption.[18-20] The effects on human health is not known – this thesis aims to address this knowledge gap.

1.2 Sources of metallic elements found in soil

1.2.1 Mineralogical sources of metallic elements

Soils contain a large range of metallic elements which are present at widely-varying concentrations. Most metals in soil are derived from mineral and bedrocks.[4,13,37,54,55] In soil, they may exist as inorganic substances, or be component of a complex compound in a mineral. They are locked in minerals and are released into the soil's environment naturally through chemical and physical weathering.[4,13,37,54,55]

Atmospheric processes are a major contributing factor to the chemical breakdown of minerals and release of their constituents into the soil.

In particular, atmospheric oxidation of iron (which is one of the commonest metals and part of the complex structure of most minerals) causes chemical decomposition through rusting, weakening the physical structure of the rocks and enabling the release of other metallic compounds into the soil via weathering.[4,13,37,54,55] Another major contributor is the action of acidic rain, which chemically dissolves minerals and washes them into soil.[4,13,37,54,55]

In addition, physical factors facilitate the mechanical breakdown of most minerals into soil particles without altering their geochemical composition. [4,13,37,54,55] Fluctuations in solar radiation and atmospheric temperature cause minerals to undergo thermal expansion and contraction. Expansion of the mineral occurs during the day as the outer layer is heated greatly, whilst contraction occurs during the night as it cooled. This repetitive process causes an alteration in the shape of the mineral, weakening the bonds between particles that form the mineral as a whole.[4,13,37,54,55] This eventually results in physical weathering.

Metallic elements are distributed throughout soil; however, certain metals may become highly accumulated in soils depending on the type of minerals that are geologically abundant in the area. For example, arsenopyrite is a mineral ore primarily composed of arsenic, iron and sulphide; therefore, chemical weathering by means of atmospheric oxidation in areas with high abundance of arsenopyrite (FeAsS) will lead to significant saturation of these three elements in the

accompanying soils.[4,47,56] The background concentrations of metallic elements in soil are therefore typically characterised by the mineral (or rock) that happens to geologically abundant (Table 1.1).

Table 1.1: Showing examples of a few mineral ores that metalliferous in nature with primary metals that appear to be native to them, as well as secondary metals that coexist but at constituent levels

Ore mineral[1]	Native metallic element[2]	Metallic elements contained in ore at constituent levels[3]
Arsenopyrite (FeAsS)	arsenic	antimony, copper, gold, mercury, molybdenum, silver, tin, uranium
Sphalerite (ZnS)	cadmium	copper, lead, zinc
Chromite (Fe, Cr_2O_4)	chromium	nickel, cobalt
Galena (PbS)	lead	antimony, cadmium, copper, selenium, silver, thallium, zinc
Cinnabar (HgS)	mercury	antimony, lead, selenium, silver, tellurium, zinc
Uraninite (UO_2)	uranium	arsenic, cobalt, copper, lead, molybdenum, selenium, vanadium

[1]Selected mineral ores that are metalliferous in nature

[2]Primary element (or metal that is most native to the mineral ore)

[3]Secondary elements (or metals known to be at constituent levels within ore)

Note: arsenic, selenium, tellurium are metalloids; uranium is a radioactive metal; and the rest are all heavy metals

Adapted from information in Alloway BJ. Sources of heavy metals and metalloids in soils. In: Chapter 2 Heavy Metals in Soils: Trace Metals and Metalloids in Soils and their Bioavailability; 2012. Section 2.3.1.1. Table 2.3. p. 23.

1.2.2 Other external and anthropogenic sources of metallic elements

The fast growth of urbanised and industrialised countries has led to significant changes in the environment and has resulted in environmental pollution due the heavy traffic-use and expanding industrial activities taking place in middle to high income countries.[4,13,37,54,55] Anthropogenic activities, which include industry and motor & vehicles usage, have caused widespread pollution of topsoil with a variety of metallic metals, such cadmium, copper and lead. Traffic or industrial activities involving with pyrometallurgy are major sources for emissions that contain airborne particulate-bound metallic elements.[4,13,37,54,55] Through atmospheric deposition, these elements are deposited on topsoil, and enter into our food chain (i.e. soil > plants > humans or soil > plants > animals > humans) directly via plant absorption.[4,13,37,54,55]

1.3 Soil contamination in the UK

1.3.1 Brief history on the occurrence of land contamination in UK

The accumulation of contaminants in UK soils can be attributed to anthropogenic practices that occurred since the start of Britain's industrial revolution,[57,58] and also to certain events linked to the World War II.[59] The legacy of Britain's industrialisation since the mid-18th century is not only one of economic prosperity, change in the

population's standards of living and scientific advancement;[60] but also, and unfortunately, the environmental impacts of industrial emissions and waste management practices that have occurred during that this time period.[57,58]

By the 1780's, Britain had already experienced a large growth in population density and was playing a global role in terms of international trade.[60] The growth in population and foreign trade created an environment where there was an increased demand for manufactured goods and services. [60] In order to meet this demand, Britain adopted a model that enabled the mass production of goods.[60] This feat was achieved by completely overhauling the traditional systems of manufacturing goods which relied mainly on man, animal and water power, and replacing them with mechanical technologies that were steam powered.[57,58,60] This was Britain's 'factory-age', [57,58,60] during which factories and furnaces responsible for the production and refinery of goods and raw materials, respectively, were built by the dozens, and functioned purely on technologies that were powered by steam.[57,58,60] From 1830 to 1922, a vast network of railways that relied exclusively on steam engine transportation was created connecting, for the first time, towns and major cities in England.[57,58,60] All these steam powered engines and technologies were dependent on coal. [57,58,61]

Coal had to be mined on a large-scale thus becoming one of the most important minerals.[60] While many deep and open mine pits were

already established across Britain for extracting coal, it was from the mid-18th to 21st century that the levels of coal production escalated most significantly, growing from 4.7 million to 250 million tonnes.[57,58,60,61] By the 1930s, the new anthropogenic activities related to mining, factories and transportation emerging from this industrial era had introduced new sources for air, water and land pollution.[57,58] Many studies that specifically addressed the problems of pollution in the UK have made irrefutable cases that the cause stemmed from practices during these centuries. A notable example is the Byker incinerator which was built in the late 1890s, and has been operating in Newcastle upon Tyne for over a century. This facility, which has recently been decommissioned, was responsible for the severe contamination of land situated in Newcastle upon Tyne due to the expulsion of waste ash into the atmosphere.[62–64]

Mining activities also generated vast amounts of landfill. In 1936, parliament passed the Public Health Act, which made initial provisions (in sections 101-105, 140 and 141) to mitigate the impacts of pollutants.[65] However, shortly afterwards, major cities such as London, Birmingham, Liverpool, Plymouth, Bristol, Glasgow, Southampton and Hull (and 18 other British cities) suffered further widespread contamination of air and land due to the war that emerged in Europe: between September 1940 and May 1941,[59] these cities suffered heavy aerial bombardment with incendiary and explosive devices from German warplanes, ensuring widespread contamination of soils with explosives (and bomb casings and

mechanisms), their by-products and combustion products from the ensuing fires.[59]

1.3.2 Formation of numerical values for soil safety limits and classification

After World War II, a group of policy makers and land developers formed the Inter-Department Committee on the Redevelopment of Contaminated Land (ICRCL) in 1976.[66] One of their initial technical notes published (i.e. ICRCL note 17/78 on landfill sites) advocated against the development of landfills. By 1980, a major conference focussed on curbing land contamination took place in Eastbourne, where a paper presented for the first time a set of numerical values for certain toxic metals and compounds intended for soil classification purposes that soon became used as screening values for chemically impacted *in situ* soils.[66]

In 1983, and in 1987, the ICRCL would made significant revisions to these values in order to use them for health evaluation and risk assessment – these sets of new values were published and named soil "trigger" and "intervention" values.[66] However, by 2002, these values had become outdated due to significant changes in concentrations of most geochemical elements. Most metals in UK soil exist at low-level concentrations not exceeding beyond 100.0 mg/kg. Previous values known as the Generic Assessment Criteria (GAC) were withdrawn in 2002, and the first set of Soil Guideline Values (SGV) was published[67] - these values were developed based on the regulations under the

statutory guidance (sections 4.1) which provides the definitions for what constitutes a significant harm caused by contaminated land; and Part 2A of the environmental protection act (1990) - a key piece of legislation for risk assessment of land contamination. Finally, the current system in use is an improved version of the SGVs, known as the category 4 screening levels, which were published in 2014.[68]

1.3.2.1 UK category 4 screening levels

The UK category-4 screening levels (C4SLs) were developed to monitor low contamination levels of a group of ill-defined elements.[68–72] They are used in the risk assessment of four different type of land-uses (i.e. includes garden allotments, residential, public use and industrial) to monitor low-level soil contamination in these settings. [68–72] In terms of soil and land contamination, these screening values were developed for the sole purpose of aiding land-users and developers to determine whether concentrations of an element have reached an unacceptable threshold in terms of the potential impact on public safety. [68–72]

C4SLs also describe a 4-stage warning system to inform the land-user or investigator to decide whether the land is contaminated. [68–72] Figure 1.1, shows how land can fall into one of four categories depending on the contamination levels of a particular contaminant: 1) Categories 1 and 2 describes land at which concentration of soil contaminants are present at exceedingly high levels, and exposure to such levels can cause acute adverse health outcomes;[73] 2) Categories 3 and 4 describe land at where concentrations of soil contaminants

are lower.[73] Although lands classed as 3 and 4 are deem uncontaminated areas, the focus on acute (and known) health impacts mean that the possibility that exposure to these low-level concentrations of soil elements may still pose a significant risk to human health over a longer term cannot be excluded.

According to section 4.1 of the UK statutory guidance, the definition of significant harm is graded into two levels, as ‘always’ or ‘may’, depending on whether the health impacts of land contamination: 1) always constitute significant harm which includes outcomes of death, life-threatening diseases (e.g. cancers) and other illness likely to result in serious health impacts such as physical deformity, impaired reproductive and congenital birth defects;[73] 2) may constitute significant harm which may include outcomes such as cancer, gastrointestinal disturbances, respiratory and cardiovascular effects, as well as skin ailments and effects on internal organs (e.g. liver or kidney).[73] Although the C4SLs were derived for the purposes of environmental monitoring of land contamination, they are also used as “trigger values” - meaning that where soil concentration of a metal exceeds recommended limit threshold value, there may be a cause for concern for the potential impact it may have on the general population.[73]

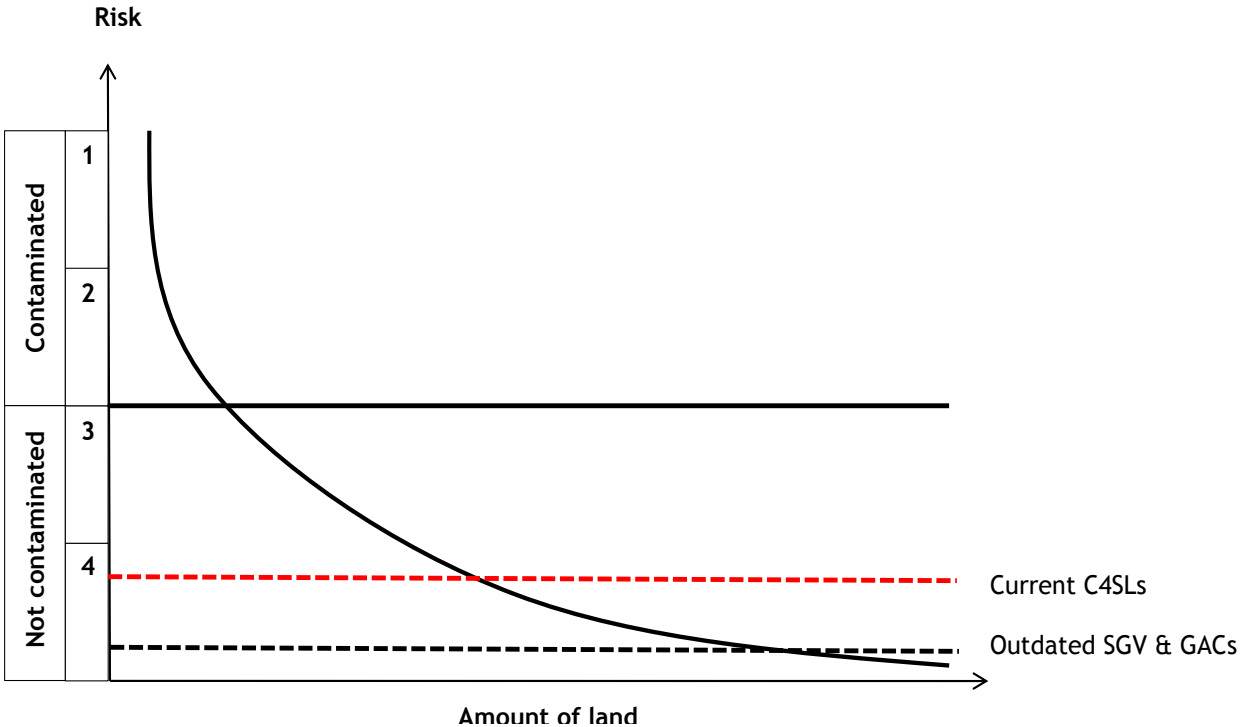

Figure 1.1: Illustrating the UK category-4 screening system. Straight-black (solid) line - point above which land is defined as 'contaminated land' under Part 2A Environmental Protection Act; Black (dashed) line – previous SGV & GAC values under category-4 to monitor low-level contamination; Red (dashed) line – current C4SLs under category-4 to monitor low-level soil contamination. Adapted information from Naima B et al. Essentials of environmental public health science (chapter 6, fig 6.1.); Original from - 'simplification of the contaminated land: impact assessment', Defra, and Cranfield University.

Derivations for all C4SLs or "trigger values" were based on the key assumption that health risks are attributable to long-term exposure to individual chemicals in soils.[68–72] Table 1.2 shows the current C4SLs values for residential areas of urban, rural and suburban built environments.

Table 1.2: List of current C4SLs values for 3 (out of 6) soil contaminants for UK residential settings; Previous guideline values from SGV and GAC have been withdrawn

Elements (in mg/kg)	C4SL[1]	SGV[2]	GAC[3]
Arsenic	35	32	-
Cadmium	87	84	
Chromium	21	-	4.3
Lead	130	450	-
Nickel	-	130	-
Selenium	-	350	-

For residential soil only
[1]Current UK category-4 screening levels (C4SLs) (2014)
[2]Soil guideline values (SGVs) (2009) Outdated
[3]Generic Assessment Criteria (GACs) Outdated

Although only three C4SLs (for arsenic, chromium and lead) are currently defined, they are important because evidence of widespread low-level contamination in many areas of England and Wales with soil concentrations exceeding these national safety limits.[68–72] According to the geochemical maps developed by the British Geological Survey – about 10% of land coverage has soil concentrations of arsenic and lead above national safety limits of 35 mg/kg and 130 mg/kg, respectively; while 80% of English and Welsh soil levels of chromium beyond 10,000 mg/kg.[53] The spatial variations in the soil concentration levels for these elements may suggest that a segment of the UK population may

have continuous low-level exposure which may have significant health implications, in particular, cancer. Furthermore, the derivation of the limits themselves incorporates only very limited population-level evidence (due to a lack of epidemiological studies in suitable settings) so it may be the case that increased risks are observable at supposedly "safe" concentrations. When assessing the health impacts of metallic elements for which C4SLs are defined, this thesis aims to determine whether increased risks of cancer are observable among those living in areas with soil concentrations both above and below the given limit, by defining a series of exposure categories based on multiples and fractions of the limit. Incorporating the limits into the exposure definitions in this way ensures that both the appropriateness of the current levels, and the impact of living in an area where they are exceeded, can be quantified. The implications of any observation of an increased risk on current policy and regulations can thereby be directly assessed by those responsible for setting and revising the limits.

1.4 The potential health impact of soil metals

There is emerging evidence that the presence of toxic elements such as arsenic, cadmium, chromium, nickel and lead in topsoil may pose a considerable risk towards human health. There are broad categories of exposure – 1) external exposure, which refers to individuals that are chronically exposed to moderate to high levels of contaminants that emerge from the environment (e.g. air, soil and water);[15,18,74–76]

2) internal exposure, which occurs within the body. This form of exposure relates to tissues and vital organs being in contact with excess amounts either of the contaminant itself or the derived metabolites remaining *in vivo* in the body's system when they are unable to be metabolised due to chemical over-saturation. This can occur in the serum of the target tissue, or in fluids produced by certain organs.[77–82] Alternatively, a tissue can be exposed to a particle that contains contaminants and becomes trapped in the associated organ during absorption. This may cause long-term irritation of a tissue resulting in adverse health effects.[15,18,74–76]

Chronic exposure to external contaminants that emerge from topsoil leads to elements gaining entry into the human body through the mouth, nostrils and skin at a continuous rate. Where these elements become absorbed into the body, internal susceptibility to toxicity occurs once the metabolic system reaches a critical point where it becomes incapable of breaking down any excess amount of either the soil metal itself or the derived metabolite that remain *in vivo* in the body's system. At this point there is a potential for these chemicals to accumulate within certain tissues.[77–82] The factors that influence bioaccumulation of chemical elements in tissues are usually dependent on the body's rate of eliminating any rogue substance through the processes of metabolism and excretion, as well as the overall propensity for tissues to store[77–82] or otherwise enable chemical elements to accumulate within them (i.e. tissue burden).[15,18,74–76]

1.4.1 The exposure pathways from soils to human

1.4.1.1 The external exposure from soil to human intake

There are 4 major ways through which external exposure to soil contaminants can occur to the human body: 1) ingestion; 2) inhalation of soil particulates; 3) dermal absorption (or skin contact) of soil particulates or entry through breakages of the skin; and 4) indirect contact through inhalation of dust which contains soil particles, or ingestion of drinking water contaminated with soil or from other sources.[15,18,74-76]

Consumption of soil is a widespread phenomenon. Children, especially those under three years of age, have a high propensity to accidentally ingest soil while playing outdoors. Young children are much more sensitive to contaminants and considered to be the group with the highest risk from being exposed to contaminants from soil (e.g. the absorption rate of lead via the digestion tract of children is 5 times efficient than that of adults).[74,76,83-85] In both adults and children, soils can be accidentally ingested by consuming fruit and vegetable with soil particulates attached to them, through lack of cleanliness after coming into contact with soils, or through swallowing of airborne soil particles which become trapped in the mucous of the mouth and nose. In addition, metals originating in soils may accumulate in the tissues of both plants and animals used as human foods. The ingestion of soil contaminants is considered to be the largest and most direct form of external exposure to soil metals due to their presence at all levels of

our food chain – e.g. plant-specific food chain: from soil > to plants > and to humans, and animal-specific food chain: from soil > to plants > to animals > and to humans[7]

Compared to the ingestion route, inhalation is considered to be a slightly less significant source for external exposure to soil metals, but nevertheless, may be important to those that experience repeated exposure over a long period of time. Any disturbance of the soil can lead to the release of soil particulates into the air where they remain ambient. Children that play on soil, as well as adults that practice gardening and other agricultural activities, and those living in areas where soils become dry and susceptible to wind erosion are at high risk of directly inhaling through the mouth and nostrils ambient soil particulates that may contain traces of heavy metals. [74,76,83–85]

Dermal absorption can also occur, although this process tends to be most significant for certain volatile and organic soil compounds (such as benzene, ethyl benzene, toluene and xylene), but may be less of a problem for most heavy metals,[10,11,44,45,52] since it will require long-term exposure for the skin to absorb sufficient amounts to cause skin damage, although certain forms of Cr(VI) and mercury are known exceptions and can become absorbed in significant quantities on only brief contact.[9,12]

Finally, soil contaminants have the ability to migrate from the soil to other environments such as the indoors of household to settle as household dust,[86] or into surface water. For examples, indirect

contact to high levels arsenic can occur from drinking water supplies which are often linked with arsenic-contaminated soil, although arsenic may also be naturally present at the source through which the drinking waters are exhumed.[47,48,56,87-89] Another example of indirect contact with contaminants may include inhalation of household dust which contains constituent elements in them.

1.4.1.2 The internal exposure and uptake by tissues

Internal exposure can occur once soil contaminants have found entry via one or more of the major pathways for external exposure (i.e. ingestion, inhalation, dermal or indirect). Contaminants may become trapped in a tissue: for example, ingested soil particulates containing metallic elements will travel through the gastrointestinal tract (GIT) organs responsible for digestion (i.e. mouth, oesophagus, stomach, intestines etc.) of food before being absorbed through their walls. While these substances are absorbed into the body and transported to the liver, they are further processed into other micro-substances such as bile and metabolites[90-93] Where undigested particulates fail to have passed through the GIT during absorption remain trapped within the crevices of the associated digestive tissue,[77-82] at this point, the tissue with the trapped contaminant becomes exposed because it is in constant contact with the contaminant, potentially leading to irritation and consequent adverse health effects.[77-82]

1.4.1.3 Generalised framework of showing the biological mechanisms and carcinogenic effects of elements from soil

Data from animal studies (which have been extrapolated to humans) have suggested that tissues exposed to metallic elements can undergo induced genotoxicity – i.e. the production of harmful agents that damages tissues and cells causing mutations, thus leading to carcinogenesis. These harmful agents are known as reactive oxygen species (ROS) [i.e. hydroxyl (•OH), superoxide ($O_2•^-$) and hydrogen peroxide (H_2O_2)].[94–98]

These studies have indicated that the presence of metals triggers oxidation reactions in cells which lead to the rapid production of ROS.[94–98] There is a defence mechanism known as the 'antioxidant defence system' which is intrinsically initiated to prevent oxidation reactions, and to abate the number of free radicals produced in cells; [94–98] however, when oxidation reactions occur due to heavy metal-induced toxicity, the rate of oxidation reactions and rapid generation of ROS overwhelms this defence mechanism - this process in cells results in a condition known as oxidative stress.[94–98] Cells under duress of oxidative stress tend to malfunction as the ROS targets, and oxidises, the nucleic acidic part of the cell [Deoxyribonucleic acid (DNA)] – resulting in impairment of DNA reparation.[94–98] Continued impairment and lack of DNA repair eventually leads to cell mutations that are cancerous.[94–98]

It possible that the processes that we have illustrate in section 1.4.1.2 can trigger such genotoxic response which may lead to the development of cancer. A summarised view of the possible mechanisms involved with soil-based metal-induced oxidative stress and cancer occurrence are shown in Figure 1.2.

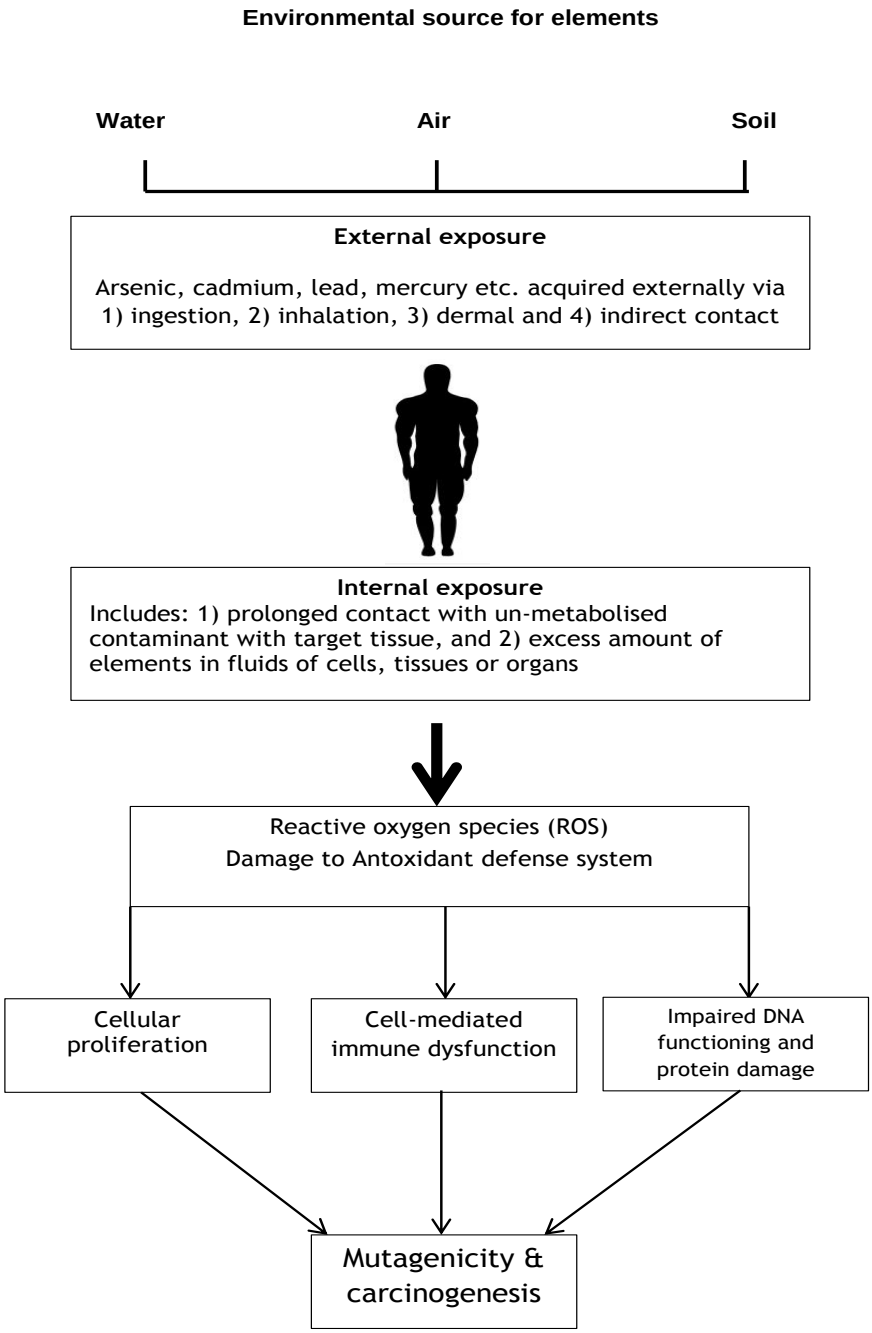

Figure 1.2: Illustrate the two broad categories of exposure to environmental contaminants - i.e. external and internal exposure; and how they play a crucial role on human carcinogenesis

1.5 Justification for assessing the health impacts of exposure to low-level soil concentration of metallic elements

In the UK, research has shown that naturally occurring metals are widely distributed at varying geologic concentrations which have implications for human health within the general population.[53] To my knowledge, in the past two decades, there has not been little research assessing the health impacts of soil elements in the UK. This is because most soil contaminants in many areas of today's environment remain effectively at low-level concentrations. This is, in part, due to positive efforts of the UK Environmental Agency (EA) and Department of Environment, Food and Rural Affairs (Defra). These agencies ensure that local authorities across the country monitor levels of contamination and make sure that they do not reach a threshold where they may pose as a significant health risk.[68-72] Unfortunately, this has brought about the common belief, as yet unproven, that soil contaminants at low-levels do not pose any health risks and thus this area of research have received little attention to date.

In contrast, before the 1980s (i.e. 1950-1980) in England and Wales, there was a considerable interest in assessing such links because a number of studies at the time found a modest association between certain soil elements (i.e. cadmium, chromium, cobalt, nickel and lead) with stomach, colon and ovarian cancers.[99-104] These studies are

historical, and were conducted at a time period where soils were more highly contaminated due to the many mining activities that took place at the time. There is therefore a need to update the evidence in relation to the modern environment. A further problem is that these studies were only focused on smaller and rural English and Welsh communities (i.e. in Cheshire, Devon, Anglesey and Montgomery), where the participants were most likely farmers and miners. In addition, the study designs adopted an ecological framework which is highly vulnerable to confounding by differences in cancer susceptibility (due to behaviour, other carcinogenic exposures, socioeconomic status and other factors) among the populations compared.

The absence of up-to-date and high-quality epidemiological evidence; and consequent reliance on extrapolations from laboratory studies, when setting safety limits for soil contaminants urgently requires to be addressed if confidence in the system of regulation is to be maintained, and the zoning of potentially harmful environments for housing development is to be avoided. Highly detailed, direct-contact, individual-level studies are inherently time-consuming and costly due to the time taken for cancer to develop, therefore there is a need to develop methodologies to identify the mostly likely sources of any increased risk and the populations most susceptible (and therefore most suitable for detailed further investigation), and to provide an evidence base to support further research. Therefore, a unique linkage between primary care medical records from The Health

Improvement Network (THIN) and the British Geological Survey's Geochemical Baseline Survey of the Environment (GBASE) has been created, enabling this avenue of research to be explored. With this database, this thesis uses contemporary and novel approaches to establish evidence for adverse carcinogenic effects from environmental soil exposures.

Aims & Objectives

Chapter 2

2 Overview of research objectives

This PhD research used data from two major sources: (1) The Health Improvement Network (THIN) database, and (2) the Geochemical Baseline Survey of the Environment (G-BASE). In collaboration with the research community at the British Geological Survey (BGS), experts from the University of Nottingham had used data from these databases and linked them together creating a new resource for investigating environmental impacts of soil metals on human health.

G-BASE database contains geochemical information on the normal background concentrations for different trace elements in UK topsoil. Soil samples were collected throughout England and Wales from urban and rural soils within 1-2km of an individual's home. Soil samples were analysed using the X-ray fluorescence spectroscopy to detect geochemical composition and concentrations levels for each element. The soil concentrations for fifteen elements were spatially interpolated over a continuous raster shapefile for England and Wales to produce point estimates at a pixel-level. Spatially referenced point estimates that overlapped the street postcode of patient registered to a general practice using the THIN were merged to produce the THIN-GBASE database.

My role as the investigator is to utilise this resource for my PhD thesis. Broadly, the aim of this PhD is to test a series of hypotheses that the risk of site specific cancers is associated with higher exposure to heavy metals in soils in England and Wales. The site-specific cancers

were chosen based on the routes of exposure - basal cell carcinoma (skin cancer) from dermal contact, lung cancer from inhalation, and gastrointestinal cancers from ingestion.

2.1 Study 1: Basal cell carcinoma

Specific objectives are:

1. To perform an ecological study to provide contemporary estimates of the overall and regional variation in the incidence of basal cell carcinoma (BCC) skin cancer in England and Wales over the last 10 years.
2. To use an individual level population-based cohort study design to determine the association between arsenic levels in residential soils and the risk of developing a BCC.

2.2 Study 2: Lung cancer

The second study will consist of 2-stage process with the following objectives:

1. To apply data mining techniques to identify an optimal subset of metallic elements which independently predict the risk of the development of lung cancer.
2. To use an individual level population-based cohort study design to determine the magnitude of association between exposure to the subset of elements identified and the risk of developing lung cancer.

2.3 Study 3: Gastrointestinal tract cancer

The third study consists of 3-stage process with each stage having the following objectives:

1. To apply data mining techniques to identify an optimal subset of metallic elements in residential soils which independently predict the risk of the development of gastrointestinal cancers.
2. To use an individual level population-based cohort study to determine the association between exposure to the subset of elements identified and the risk of developing gastrointestinal cancer.
3. To apply multivariate analysis to the individual level population-based cohort study using competing risk modelling to determine the association between exposure to the subset of elements and the risk of developing site-specific gastrointestinal cancer (upper gastrointestinal cancer, stomach cancer, colorectal cancer).

THIN & G-BASE

Chapter 3

3.1 The Health Improvement Network

The Health Improvement Network (THIN) is a large database for storage of electronic medical records (EMRs) collected from primary care general practices (GP) throughout the UK. The information stored in THIN database contains substantial information on all diagnoses made by or reported to general practitioners. It includes, not only data on medical events, but essential details of a patient's demographic characteristics, prescriptions, important data relating to their lifestyle and area-level information linked his their postcode.[105]

The THIN database is an important resource for academics and medical practitioners in epidemiology to study health outcomes, it provides researchers access to four major unique structured data files which are harmoniously linked to one another via identifiers (i.e. patient's identification code and their practice of registration).

Currently, there are at least 587 participating GPs across the UK contributing to the THIN database of which the percentage coverage is at most 6.0% of the total practices in the country. The computerised information contained in THIN comprises a total of 12,374,853 patients of which 30.0% (i.e. 3,609,061 patients) are alive and actively contributing to the data in THIN. The overall number of patients contributing computerised data in THIN is 85,803,247 person-years.[106]

Despite the coverage of practices contributing to THIN is relatively small; several studies have shown that data derived from THIN are

generalizable. Many health outcomes have been validated, including mortality, demonstrating the data representativeness of the UK population[105,107-110]

3.2 Geochemical Baseline Survey of the Environment

The British Geological Survey's (BGS) joint project - Geochemical Baseline Survey of the Environment (G-BASE) and BGS re-analysis of the National Soil Inventory x-ray fluorescence spectrometry [NSI(XRFS)] samples, is a nationwide project aimed for determining the normal background concentration (NBC) levels of several geochemical (or trace) elements in UK topsoil. The programme, which was initiated in the late 1960s, seeks to establish a baseline (i.e. a point of reference) to monitor variations in NBCs of elements in topsoil.[34,111-113]

The project has achieved complete coverage of the English and Welsh land surface through routine soil sampling of areas with stream sediments, and topsoil situated in urban and rural regions.[111,114] The G-BASE project has currently determined NBCs for 48 different elements in which their information are spatially referenced and stored electronically in large database.[111,114] This information is generated on high resolution maps to represent its distribution across the English and Welsh land surface.[53]

Usage of the data greatly supports the exploration for beneficial mineral resource for UK economic development. It aids in the planning

and suitability of land-use for agricultural and social purposes. It helps to identify contaminated land and affected areas to quickly prioritise & mobilise measures of land remediation. It is a resource to advance our understanding of the potential health impacts of soil elements on humans.[34,112]

3.3 Linkage of THIN to soil quality data in G-BASE

To re-emphasis, the linkage was performed through the joint efforts from experts at the University of Nottingham and the BGS. Furthermore, I did not have any involvement initial the linkage process. However, my role was focused on using the derived database to build a formatted dataset that will enable the analyses of the soil data through a series of epidemiological studies that are presented in this thesis. Section 3.3 presents a detailed explanation of the soil sampling process carried out by the BGS, the linkage procedure, and an overall descriptive analysis for the distribution of the soil concentration levels for metals in linked database.

3.3.1 The primary soil data sets

The three primary soil datasets used as part of the G-BASE project are the G-BASE rural, G-BASE urban and NSI(XRFS) soil data. The G-BASE rural and urban soil samples are collected in a consistent manner that provides total soil concentration estimates for forty-eight different elements at a very high density (i.e. 1 site per 0.25km^2 for urban and

1 site per 2.0km^2 for rural) to enable interpretations to be made down to a local scale.[115]

The G-BASE rural and urban soil data only has coverage for central and eastern England. However, the areas that are not covered by G-BASE are supplemented by the NSI(XRFS) soil dataset which has complete coverage for the whole of England and Wales.[115-117] The NSI(XRFS) topsoil samples, which were collected and prepared in a similar fashion to the G-BASE rural or urban samples, has been reanalysed at BGS to determine total concentrations for each element. The NSI(XRFS) sampling sites were established in non-urban lands at a density of 1 site per 25.0km^2.[115-117] The combined number of G-BASE rural and urban sampling points in England are 37,269 (with 23,686 established in rural and 13,583 in urban areas); and for NSI(XRFS) sample sites are 6,127 (with 4,864 collected from English soil and 1,263 derived from Welsh soils).[115-117]

3.3.2 Soil sampling at G-BASE (rural & urban) and NSI(XRFS) sites

This section provides a detailed description of how the soil samples were collected. G-BASE soil sampling strategy was based on a regular grid across England and Wales using a standard British Ordnance Scale of 1:25,000. Sampling locations the for G-BASE database were determined at different densities depending on the type of environment and as well as the origin of the surface [i.e. urban terrain (concrete surface) or rural terrain (stream sediment, surface

soil or deep soil)] in which soil sample was collected from.[118] Whereas, samples from the NSI(XRFS) provided completeness of coverage for England and Wales for areas that have yet to be surveyed by the G-BASE project.

For G-BASE rural zones, suitable sampling points were identified in rural areas whereby soil samples were collected at a density of 1 per 2.0km^2, at alternating grids, across the rural terrain. The rural G-BASE data are mostly concentrated in eastern and central areas of England, with additional samples collected from the Tamar catchment of South West England. G-BASE urban are classed in urban centres whereby soil samples were collected at a much finer, and higher resolution compared to G-BASE rural samples at a density of 1 per 0.25km^2 (or 4 per 1.0km^2). The NSI(XRFS) sampling points for collecting soil samples were conducted at a much coarser resolution when compared to both G-BASE urban and rural soils, at a density of 1 per 25.0km^2 (Figure 3.1).[118]

At all sampling points on the grid (Figure 3.1), irrespective of whether soils were defined as G-BASE rural, urban or NSI(XRFS), an area with dimensions of 20.0m x 20.0m was carved for soil collection. A total of five samples were collected from the vertices and at centre of the area with augur flights at a depth of 2-15cm (>0-2cm for concrete terrains; 2-5cm for topsoil sampling; 5-15cm for deep soil sampling) (Figure 3.2).[118]

The soil materials were aggregated and taken to BGS laboratories for analyses using x-ray fluorescence spectroscopy (XRFS) as the standard method to detect the geochemical composition, as well as total and extractable concentrations for major and trace elements present in the soil sample. These measurements were spatially referenced to the location of collection and stored electronically in BGSs G-BASE database; whilst aggregated soil samples were preserved and archived in BGS storerooms for future reference.[118]

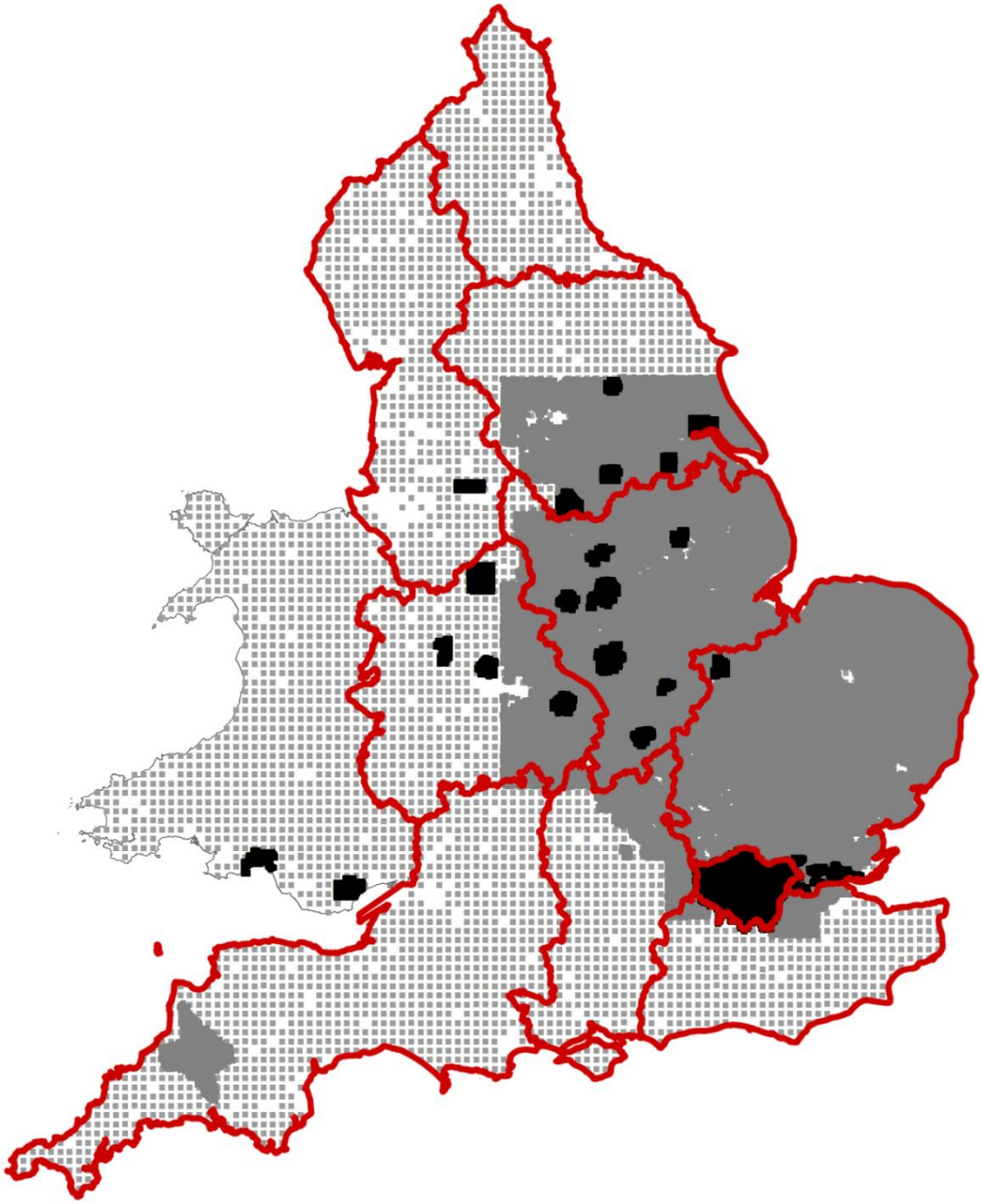

Figure 3.1: Map of G-BASE & NSI(XRFS) sample sites across England & Wales[1,2]

[1] Black colour denotes G-BASE urban regions with sampling points are at a density of 1 per 0.25km^2. Grey colour denotes G-BASE rural regions with sampling points at a density of 1 per 2km^2. Grey dots indicate locations of NSI(XRFS) at a density of 1 per 25km^2

[2] Permission was granted by to use above material. Figure 3.1 was originally published by Jack E. Gibson et al., (2016)

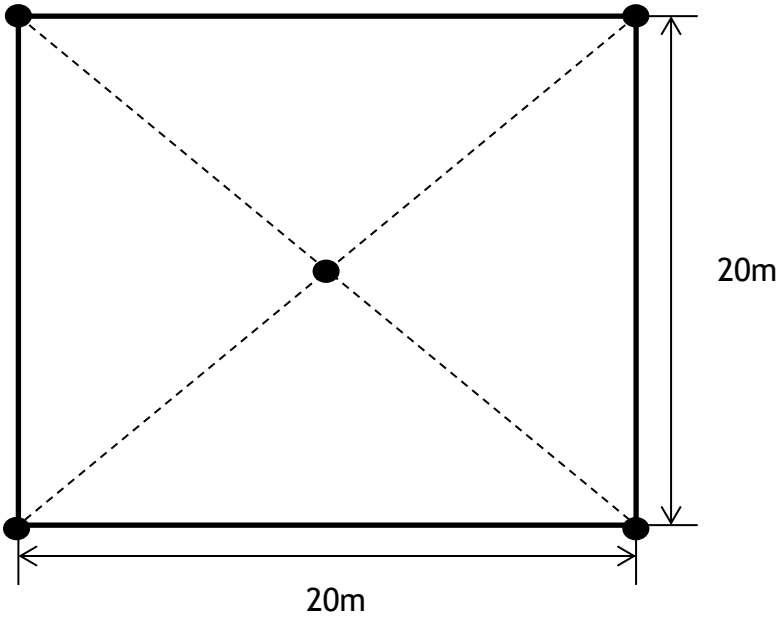

Figure 3.2: Shows an area marked at a sampling location for soil collection.

3.3.3 **Linkage procedure**

The overall number of sampling points from G-BASE and NSI(XRFS) described in section 3.1.1 were treated as a point-layer, where each site contains the total concentrations measured for each metal. The point-layer is loaded into GIS software [ArcGIS desktop 10.3 (ESRI, Redland, California, USA)] and mapped over the cartographies of England and Wales which contains the spatial features i.e. residential postcodes areas.

For each metal, interpolated layers were produced, with the inverse distance weighting (IDW) option (using a search radius of 5.0km and an output cell size of 1.0km) to predict values for substances at locations that were unsurveyed. The estimates derived for each metal covered the entire cartographies for England and Wales, where a predicted value(s) intersected the centroid of a feature that defined a residential postcode – the summary estimate would be linked to that specific postcode.

15 out of 48 elements from G-BASE were selected for the linkage with EMRs; these were aluminium, arsenic, calcium, chromium, copper, iron, lead, manganese, nickel, phosphorous, selenium, silicon, uranium, vanadium and zinc. Elements in the linkage were outputted with one of the three indicators: as -2 if linked XRFS measurements were missing; as -1 if constituent measurements for elements from a soil sample collected at a sampling area were below the detection

limit (DL) (Table 3.1); and a positive continuous value – which indicates a concentrations level.

Table 3.1: Shows the lower limits of detection for the 15 soil elements as part of the GBASE sampling strategy used to distinguish the presence or absence of substance in a soil sample

Soil element (symbol)	XRFS detection limits (mg/kg)
Aluminium (Al)	2,000.0
Arsenic (As)	0.5
Calcium (Ca)	500.0
Chromium (Cr)	5.0
Copper (Cu)	1.3
Iron (Fe)	70.0
Lead (Pb)	1.3
Manganese (Mn)	70.0
Nickel (Ni)	1.3
Phosphorus (P)	220.0
Silicon (Si)	470.0
Selenium (Se)	0.2
Uranium (U)	0.5
Vanadium (V)	5.0
Zinc (Zn)	7.0

The spatially interpolated dataset was passed on to THIN for linking the concentration values to a patient's postcode of residence. Prior to the linkage THIN had to ensure that practices falling within the G-BASE coverage area had to provide formal consent, after the linkage, THIN had to ensure that patient confidentiality was preserved by rendering a patient's private details such as name, the exact location and registered practice anonymous. The data used for this PhD project was kept completely anonymised.

At the time of the linkage, 377 out of 395 English & Welsh GPs contributing to THIN database agree to participate in the linkage. The

total number of patients ever registered participating practices in THIN was 7,137,365 (includes alive or deceased patients). Of these, 6,825,382 patients were merged successfully to G-BASE soil data. From the cohort of patients with linked data about 6.24 million had complete NBC values for all 15 elements. Those that were alive and actively contributing to THIN amounted to approximately 2.88 million patients with G-BASE data.

For each of the 15 soil elements the observed NBC levels of the linkage measurements were validated by comparing the distribution among the THIN population with those expected among the general UK population, which has shown to be similar. This section of this thesis regarding the methodology for linking THIN & G-BASE and validation has been described extensively elsewhere.[119]

From this point, and onwards, we will be referring to the linked database as THIN-GBASE. It should be noted that the dataset used for this research contained approximately 2.3 million patients with G-BASE soil data. Depending on the type of cancer outcome, this dataset was further formatted according to specific set of inclusion and exclusion criteria which are discussed more in subsequent chapters. In subsequent analysis, the applied formats will result in sample size of at least 1.8 million participants.

The next section provides a descriptive account for each of the fifteen soil elements that are present in THIN-GBASE. The descriptive analysis was limited to participants having all data across fifteen elements

(regardless of it being coded as the following: concentration value, below detection limit (-1), not available (-2)) – these patients were typically at the ages of 18 years or above without any history of any cancer diagnosis before 1st January 2004. The descriptive analysis was restricted to participants who were registered at general practice for at least one year before the 1st of January 2004, and their general practices having an Acceptable Mortality Recording before 1st of January 2004.

3.3.4 **Descriptive analysis of geochemical data in THIN-GBASE**

Descriptive analyses were performed accordingly on 1,742,205 patients in the THIN-GBASE who have soil data across all 15 elements. Table 3.2 provides the statistical summaries in a form of median, interquartile ranges (IQRs) and maximum value observed among the cohort of participants who are contributing data to the THIN-GBASE database. For instance, 1,664,155 (out of 1,742,205) (95.52%) participants have a concentration value for aluminium – among the participants the estimates typically ranged between 2,000 to 116,700 mg/kg with a median of 51,000 mg/kg (IQR: 40,300-59,300 mg/kg). The remaining participants (4.48%) either had concentrations for aluminium that were deemed as estimates below detection limits (i.e. less than 2,000 mg/kg (coded as -1)) or not available (coded -2).

The soil concentrations for aluminium, calcium, iron and silicon were vastly higher than those measured for the remaining elements – arsenic, chromium, copper, lead, nickel, manganese, phosphorus,

uranium, vanadium and zinc. The concentrations in soil samples for these constituent elements that were detected at large abundance typically ranged from measurements exceeding levels of 10,000 mg/kg. Overall, the elements with highest median concentrations were aluminium (median: 51,000 mg/kg, IQR: 40,300-59,300 mg/kg) followed by silicon (median: 29,900 mg/kg, IQR: 26,500-32,900 mg/kg), iron (median: 27,406 mg/kg, IQR: 21,575-34,830 mg/kg) and calcium (median: 3,300 mg/kg, IQR: 670.0–14,300 mg/kg). Soil elements with median concentrations between 100.0-1,000.0 mg/kg included phosphorus (median: 995.0 mg/kg, IQR: 730.0-1,358.0 mg/kg) and manganese (median: 496.0 mg/kg, IQR: 380.0-771.0 mg/kg) (Table 3.2). The remaining soil elements in this scenario are considered as low-level trace elements typically because their median concentration levels detected in soil samples did not exceed 100.0 mg/kg.

Table 3.2: Descriptive soil analysis was performed on all 15 elements in linked database (THIN-GBASE) to derive the following summary statistics – median concentration levels, interquartile ranges (IQR) and maximum value observed

	Median (IQR) [mg/kg]	Maximum value [mg/kg]	Total (1,742,205) (%)
Soil element			
Aluminium	51,000 (40,300-59,300)	116,700	1,664,155 (95.52%)
Arsenic	15.6 (11.5-19.8)	1032.2	1,671,441 (95.94%)
Calcium	3,300 (670.0-14,300)	214,400	1,654,706 (94.98%)
Chromium	70.0 (58.0-83.0)	1,141.0	1,671,486 (95.94%)
Copper	27.2 (17.9-50.1)	1,320.9	1,671,486 (95.94%)
Iron	27,406 (21,575-34,830)	137,888	1,670,842 (95.90%)
Lead	74.0 (47.0-162.0)	3,045.0	1,671,486 (95.94%)
Manganese	496.0 (380.0-771.0)	19,159	1,671,713 (95.95%)
Nickel	23.9 (17.7-32.1)	178.8	1,671,486 (95.94%)
Phosphorus	995.0 (730.0-1,358.0)	6,110.0	1,660,318 (95.30%)
Silicon	29,900 (26,500-32,900)	467,000	1,659,074 (95.23%)
Selenium	0.5 (0.37-0.70)	11.3	1,658,164 (95.18%)
Uranium	1.97 (1.57-2.38)	61.7	1,668,515 (95.77%)
Vanadium	75.0 (61.0-95.0)	653.0	1,671,486 (95.94%)
Zinc	100.0 (68.0-168.0)	4,681.0	1,671,486 (95.94%)

The overall distribution for all soil elements in terms of their concentration levels was non-normal. A skewed-right distribution was observed whereby a large proportion of patients in the linkage living on residential soil significantly contained low concentration. A series of two-way histograms with cumulative proportions (percentiles) were generated to illustrate the skewedness of the concentration levels for each soil element (Figure 3.3-3.17). For some trace elements, such copper and lead – it depicts that the mounded part of their distribution falls below the 75th percentile value. The tailed part or group of observations above this value were the remaining 25% of patients in the cohort that had somewhat medium to very high concentrations of copper (Figure 3.7) or lead (Figure 3.9).

We use soil copper in this context to further illustrate such feature - the mounded part of the distribution falling 75th percentile score corresponds to patients in the linkage with soil copper levels that ranged between the lowest detection (1.3 mg/kg) to 50.1 mg/kg. The remaining 25% of the distribution above this score corresponds to those residing in areas with soil copper concentration ≥ 50.1 mg/kg. Please note that the few in the 99th percentile bracket have the highest exposure ranging from 170.0 to 1,320.9 mg/kg (Figure 3.7). Other trace metals such as arsenic (Figure 3.4), chromium (Figure 3.6), nickel (Figure 3.11), selenium (Figure 3.13), uranium (Figure 3.15), vanadium (Figure 3.16) and zinc (Figure 3.17) – the mounded part of the distribution is the below 90th percentile point. The tailed part of observations that fall above this value were the remaining 10%

of patients in the linkage living in areas with residential soil having higher concentrations of these elements. For instance - the mounded part of the distribution below the 90th percentile as illustrated for soil chromium ranged from the lowest detection limit (5.0 mg/kg) to 92.0 mg/kg. Those with residing in areas with soil chromium above this value corresponded to patients ranked in the 90th percentile bracket with chromium levels ranging ≥ 92.0 mg/kg to the maximum observed value of 1,141.0 mg/kg (Figure 3.6).

When accounting for recent C4SLs for special elements that are toxicological concern in the UK, we observed the following: only 4.0% of patients contributing to the THIN-GBASE database lived in areas with arsenic concentration in English and Welsh above the C4SL soil guideline value of 35.0 mg/kg (Figure 3.4); 45.0% of patients in the cohort lived on soil with lead levels exceeding the C4SL soil guideline value of 86.0 mg/kg (Figure 3.9); and 98.0% of patients resided in areas with soil chromium above the CS4L of 21.0 mg/kg (Figure 3.6).

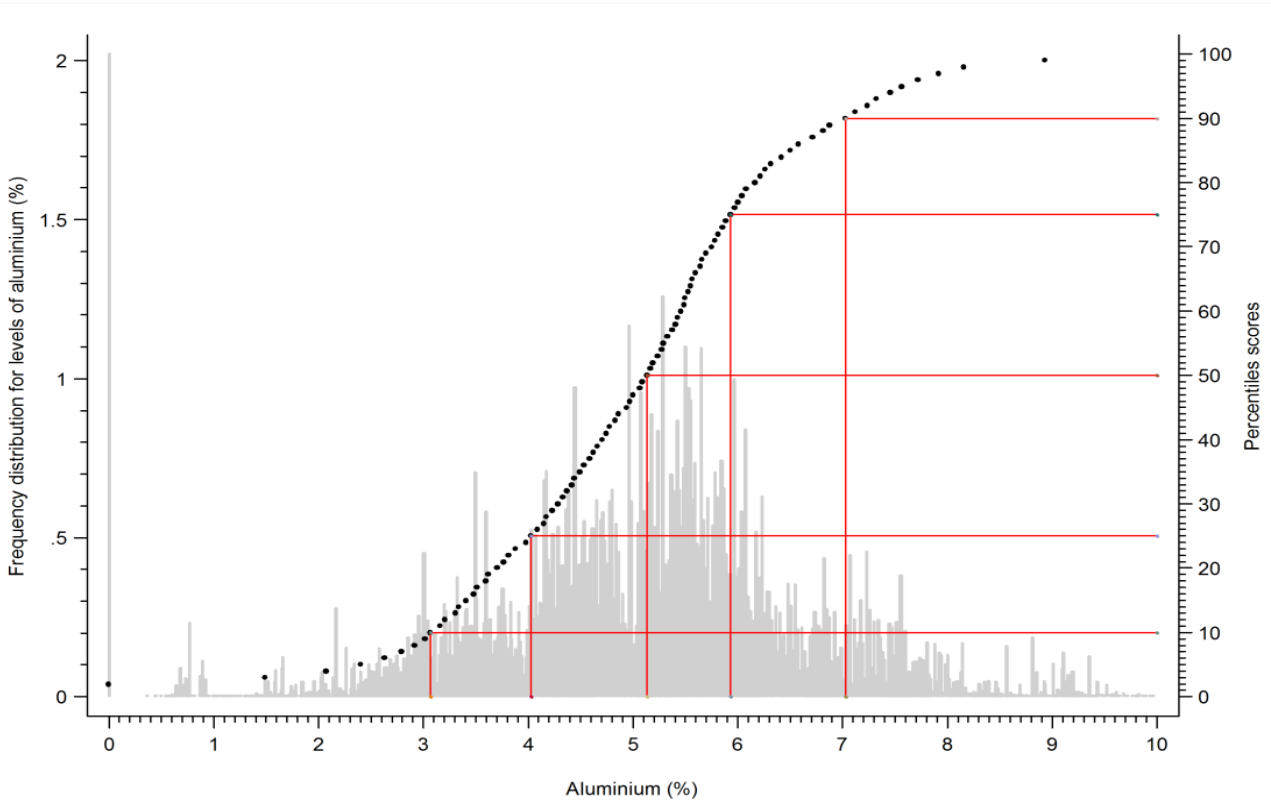

Figure 3.3: Two-way histogram with cumulative proportions showing the overall distribution of patients in THIN-GBASE with specific soil concentration levels for aluminium. Left y-axis: corresponds to the observed proportion of patients with specific soil levels of aluminium. Right y-axis: Black dots correspond to a percentile score – i.e. the proportion of patients that fall under specific soil concentration value for aluminium; Red line indicates: 10th, 25th, 50th, 75th and 90th percentiles (i.e. 31,000, 40,000, 51,000, 59,300 and 70,500 mg/kg respectively). The concentrations for aluminium were converted to a weight percentage (mg/kg÷10,000), whereby 1.0% = 10,000 (of aluminium) parts-per million.

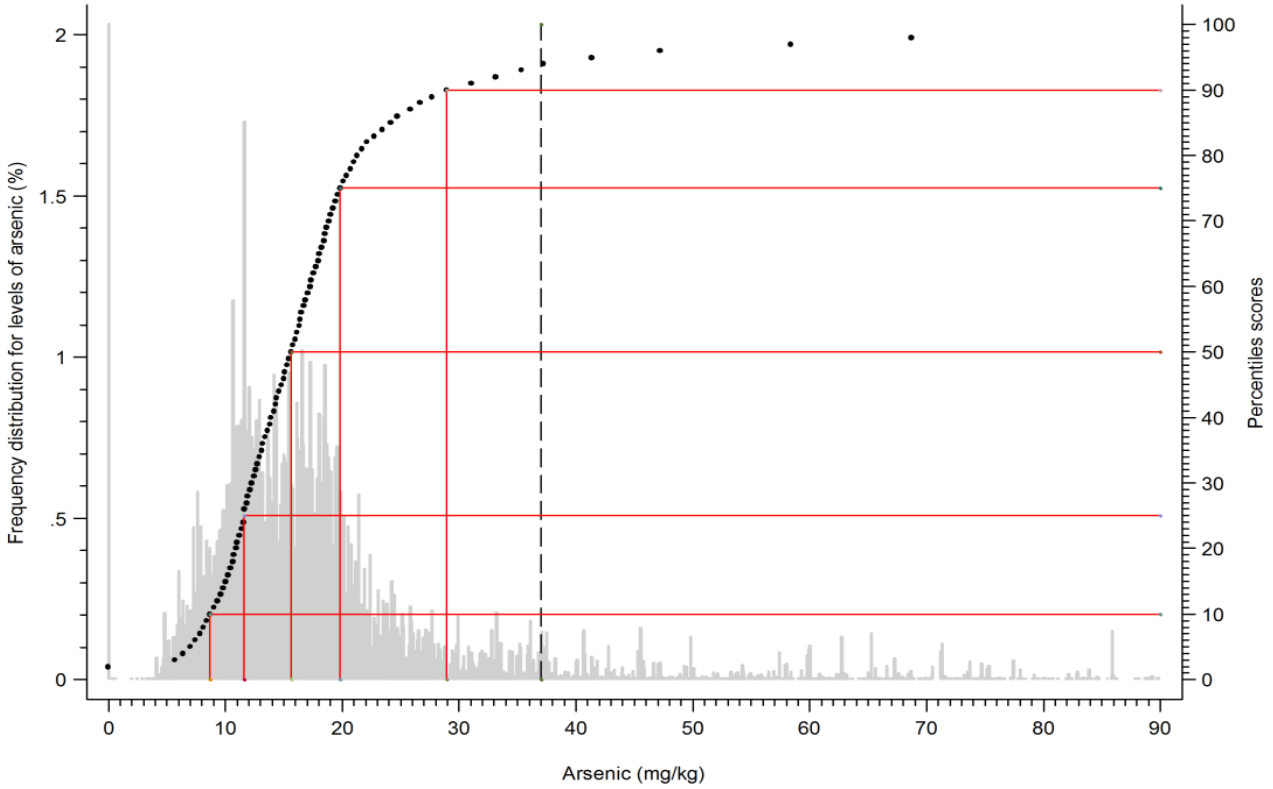

Figure 3.4: Two-way histogram with cumulative proportions showing the overall distribution of patients in THIN-GBASE with specific soil concentration levels for arsenic. Dashed black line represents the UK arsenic C4SL soil guideline value (35.0 mg/kg). Left y-axis: corresponds to the observed proportion of patients with specific soil levels of arsenic. Right y-axis: Black dots correspond to a percentile score – i.e. the proportion of patients that fall under specific soil concentration value for arsenic; Red line indicates: 10th, 25th, 50th, 75th and 90th percentiles (i.e. 8.5, 11.6, 15.6, 19.8 and 29.0 mg/kg respectively)

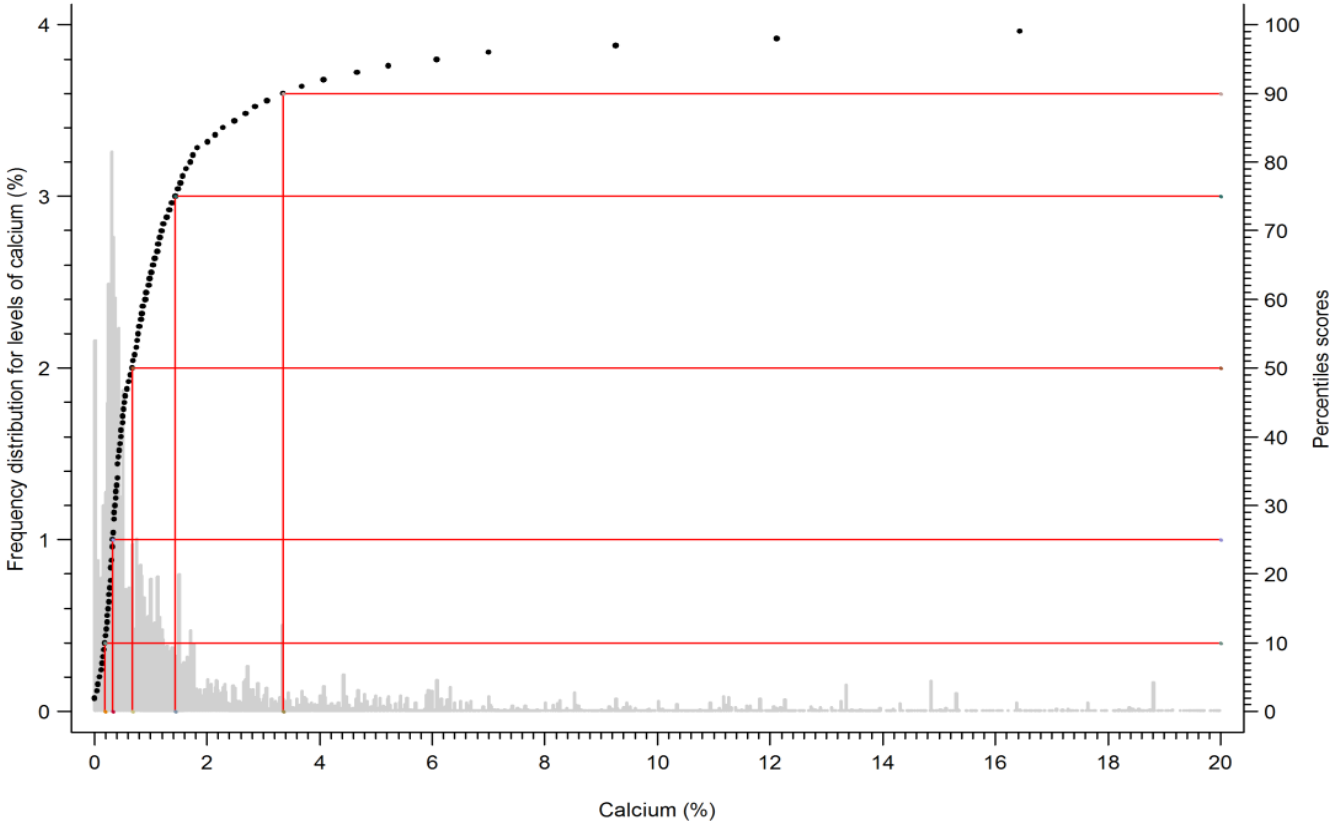

Figure 3.5: Two-way histogram with cumulative proportions showing the overall distribution of patients in THIN-GBASE with specific soil concentration levels for calcium. Left y-axis: corresponds to the observed proportion of patients with specific soil levels of calcium. Right y-axis: Black dots correspond to a percentile score – i.e. the proportion of patients that fall under specific soil concentration value for calcium; Red line indicates: 10th, 25th, 50th, 75th and 90th percentiles (i.e. 200.0, 670.0, 3,300, 14,300 and 34,000 mg/kg respectively). The concentrations for calcium were converted to a weight percentage (mg/kg÷10,000), whereby 1.0% = 10,000 (of calcium) parts-per million

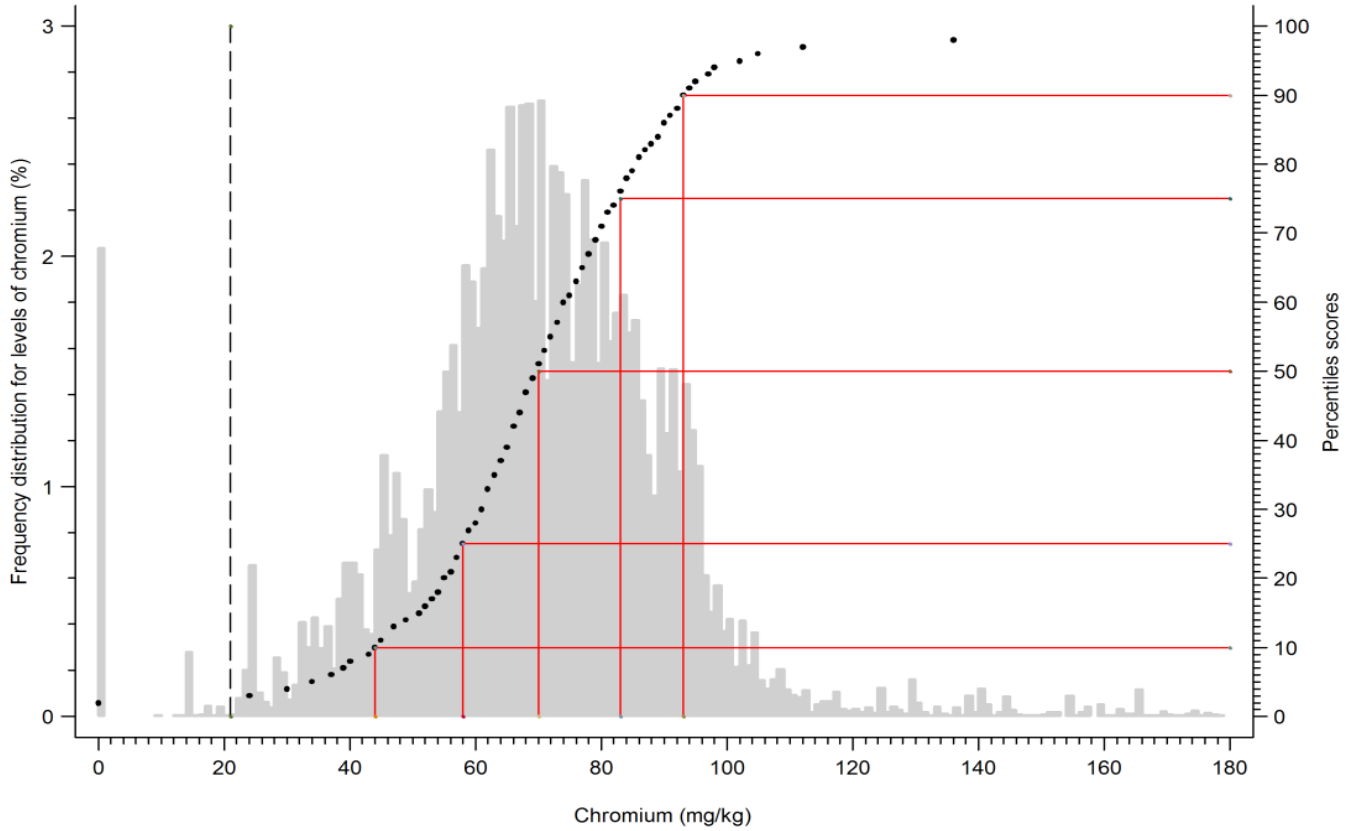

Figure 3.6: Two-way histogram with cumulative proportions showing the overall distribution of patients in THIN-GBASE with specific soil concentration levels for chromium. Dashed black line represents the UK chromium C4SL soil guideline value (21.0 mg/kg). Left y-axis: corresponds to the observed proportion of patients with specific soil levels of chromium. Right y-axis: Black dots correspond to a percentile score – i.e. the proportion of patients that fall under specific soil concentration value for chromium; Red line indicates: 10th, 25th, 50th, 75th and 90th percentiles (i.e. 44.0, 58.0, 70.0, 83.0 and 93.0 mg/kg respectively)

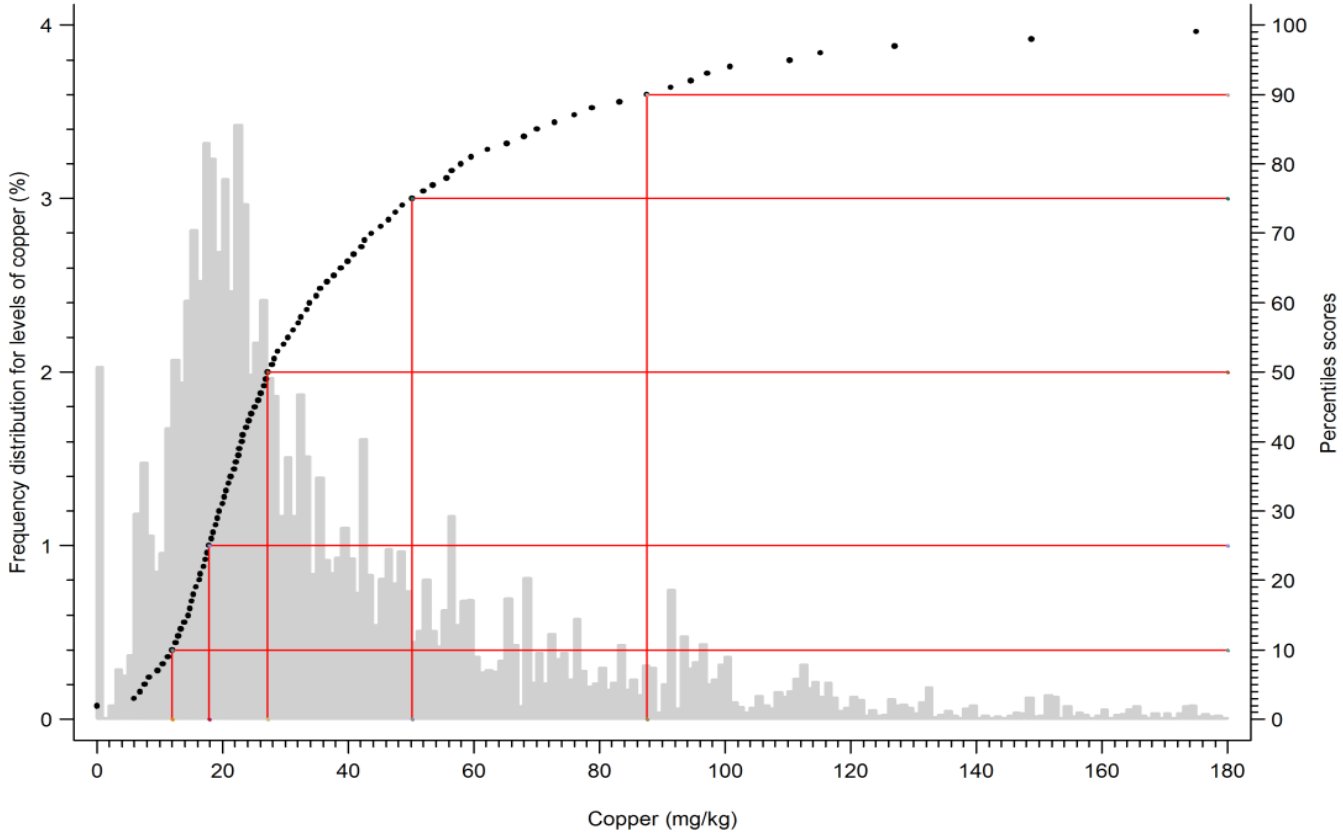

Figure 3.7: Two-way histogram with cumulative proportions showing the overall distribution of patients in THIN-GBASE with specific soil concentration levels for copper. Left y-axis: corresponds to the observed proportion of patients with specific soil levels of copper. Right y-axis: Black dots correspond to a percentile score – i.e. the proportion of patients that fall under specific soil concentration value for copper; Red line indicates: 10th, 25th, 50th, 75th and 90th percentiles (i.e. 13.0, 17.9, 27.2, 50.1 and 88.0 mg/kg respectively)

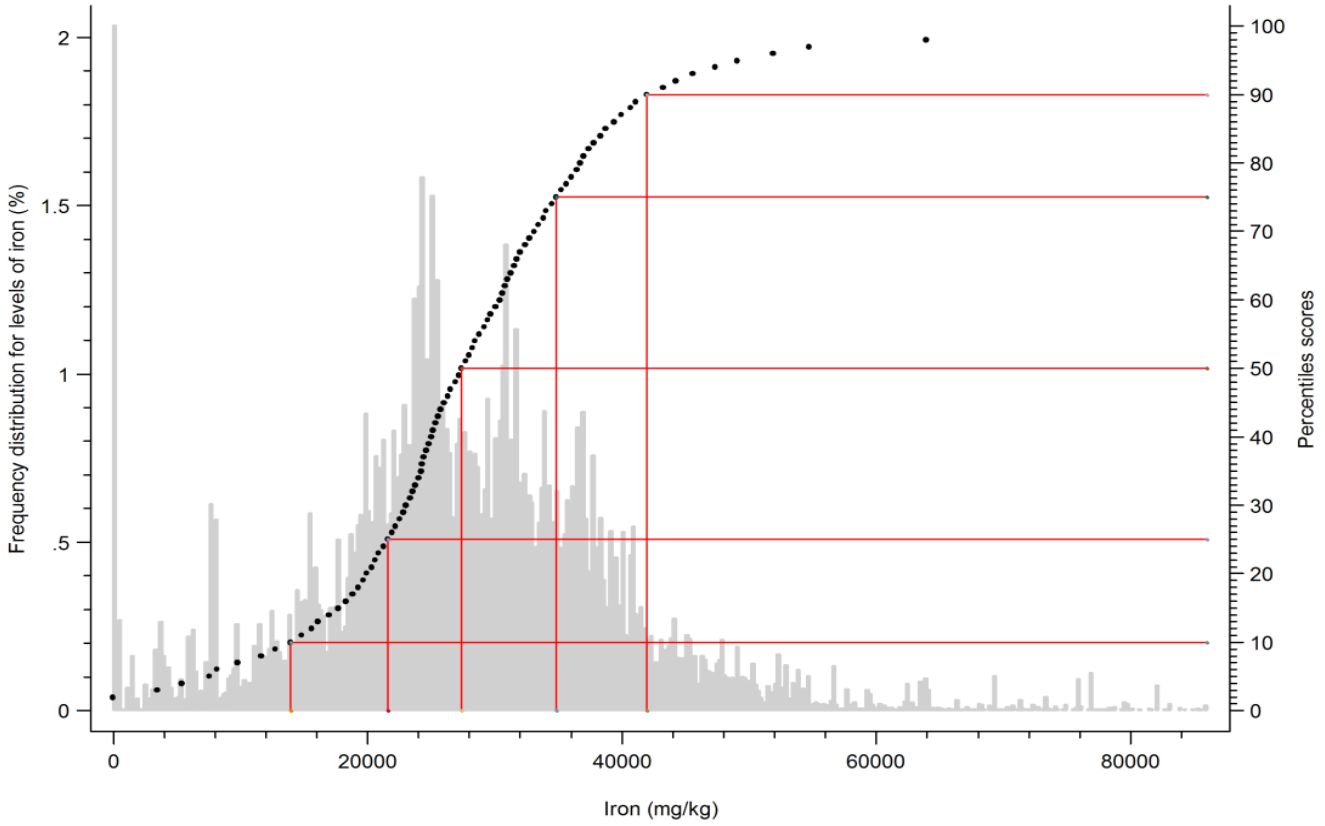

Figure 3.8: Two-way histogram with cumulative proportions showing the overall distribution of patients in THIN-GBASE with specific soil concentration levels for iron. Left y-axis: corresponds to the observed proportion of patients with specific soil levels of iron. Right y-axis: Black dots correspond to a percentile score – i.e. the proportion of patients that fall under specific soil concentration value for iron; Red line indicates: 10th, 25th, 50th, 75th and 90th percentiles (i.e. 13,000, 21,575, 27,406, 34,830 and 43,000 mg/kg respectively)

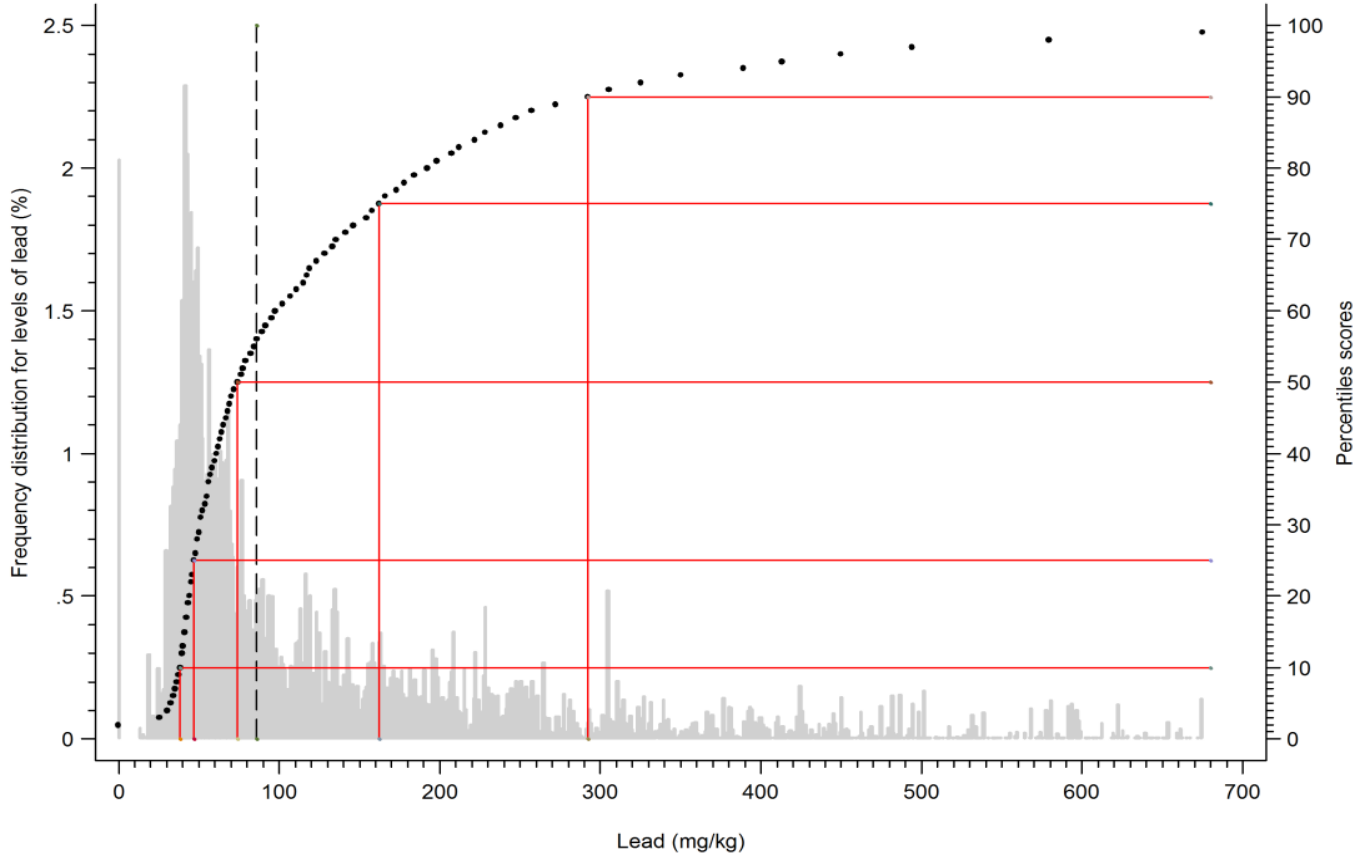

Figure 3.9: Two-way histogram with cumulative proportions showing the overall distribution of patients in THIN-GBASE with specific soil concentration levels for lead. Dashed black line represents the UK lead C4SL soil guideline value (86.0 mg/kg). Left y-axis: corresponds to the observed proportion of patients with specific soil levels of lead. Right y-axis: Black dots correspond to a percentile score – i.e. the proportion of patients that fall under specific soil concentration value for lead; Red line indicates: 10th, 25th, 50th, 75th and 90th percentiles (i.e. 36.0, 47.0, 74.0, 147.0 and 290.0 mg/kg respectively)

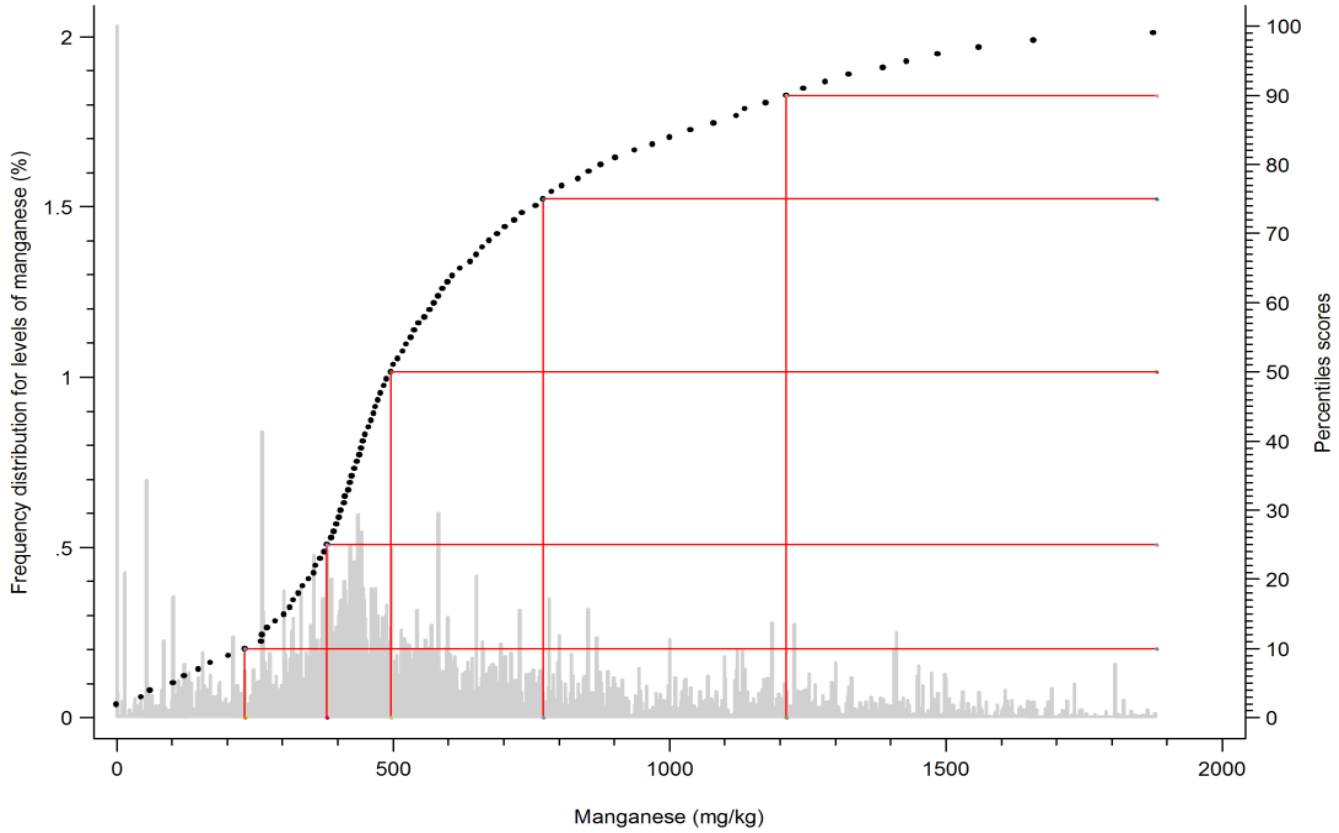

Figure 3.10: Two-way histogram with cumulative proportions showing the overall distribution of patients in THIN-GBASE with specific soil concentration levels for manganese. Left y-axis: corresponds to the observed proportion of patients with specific soil levels of manganese. Right y-axis: Black dots correspond to a percentile score – i.e. the proportion of patients that fall under specific soil concentration value for manganese; Red line indicates: 10th, 25th, 50th, 75th and 90th percentiles (i.e. 223.0, 380.0, 496.0, 771.0 and 1200.0 mg/kg respectively)

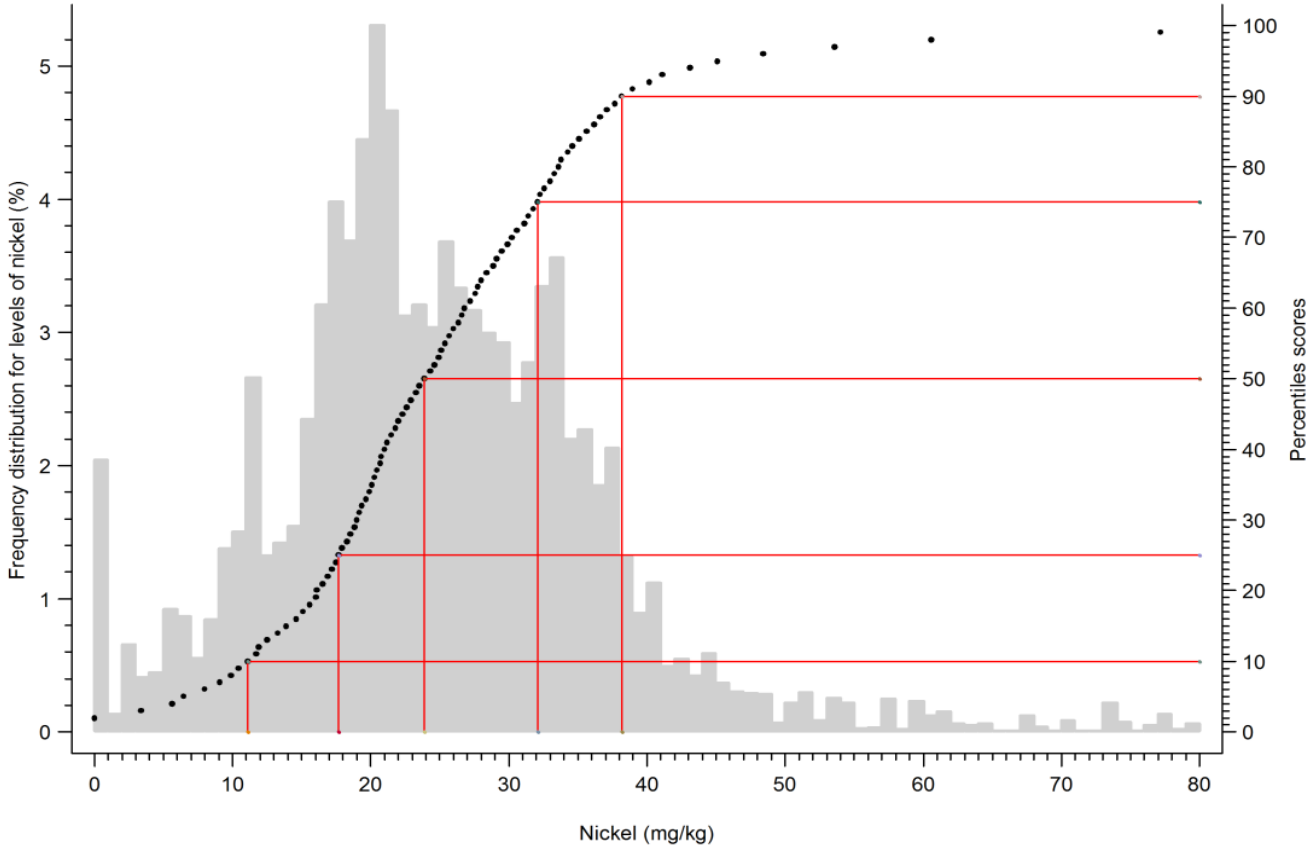

Figure 3.11: Two-way histogram with cumulative proportions showing the overall distribution of patients in THIN-GBASE with specific soil concentration levels for nickel. Left y-axis: corresponds to the observed proportion of patients with specific soil levels of nickel. Right y-axis: Black dots correspond to a percentile score – i.e. the proportion of patients that fall under specific soil concentration value for nickel; Red line indicates: 10th, 25th, 50th, 75th and 90th percentiles (i.e. 11.0, 17.7, 23.9, 32.1 and 38.4 mg/kg respectively)

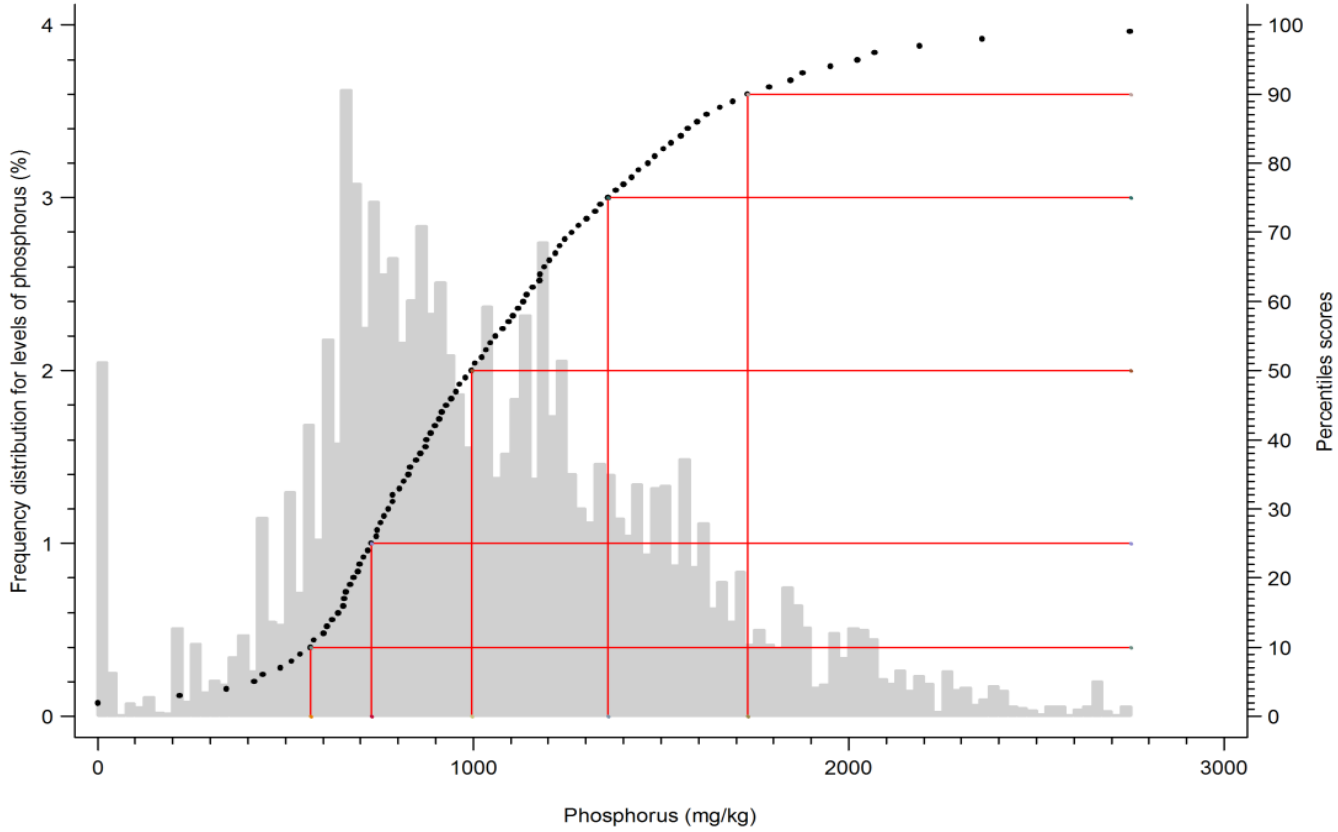

Figure 3.12: Two-way histogram with cumulative proportions showing the overall distribution of patients in THIN-GBASE with specific soil concentration levels for phosphorus. Left y-axis: corresponds to the observed proportion of patients with specific soil levels of phosphorus. Right y-axis: Black dots correspond to a percentile score – i.e. the proportion of patients that fall under specific soil concentration value for phosphorus; Red line indicates: 10th, 25th, 50th, 75th and 90th percentiles (i.e. 570.0, 730.0, 995.0, 1,358.0 and 1,730.0 mg/kg respectively)

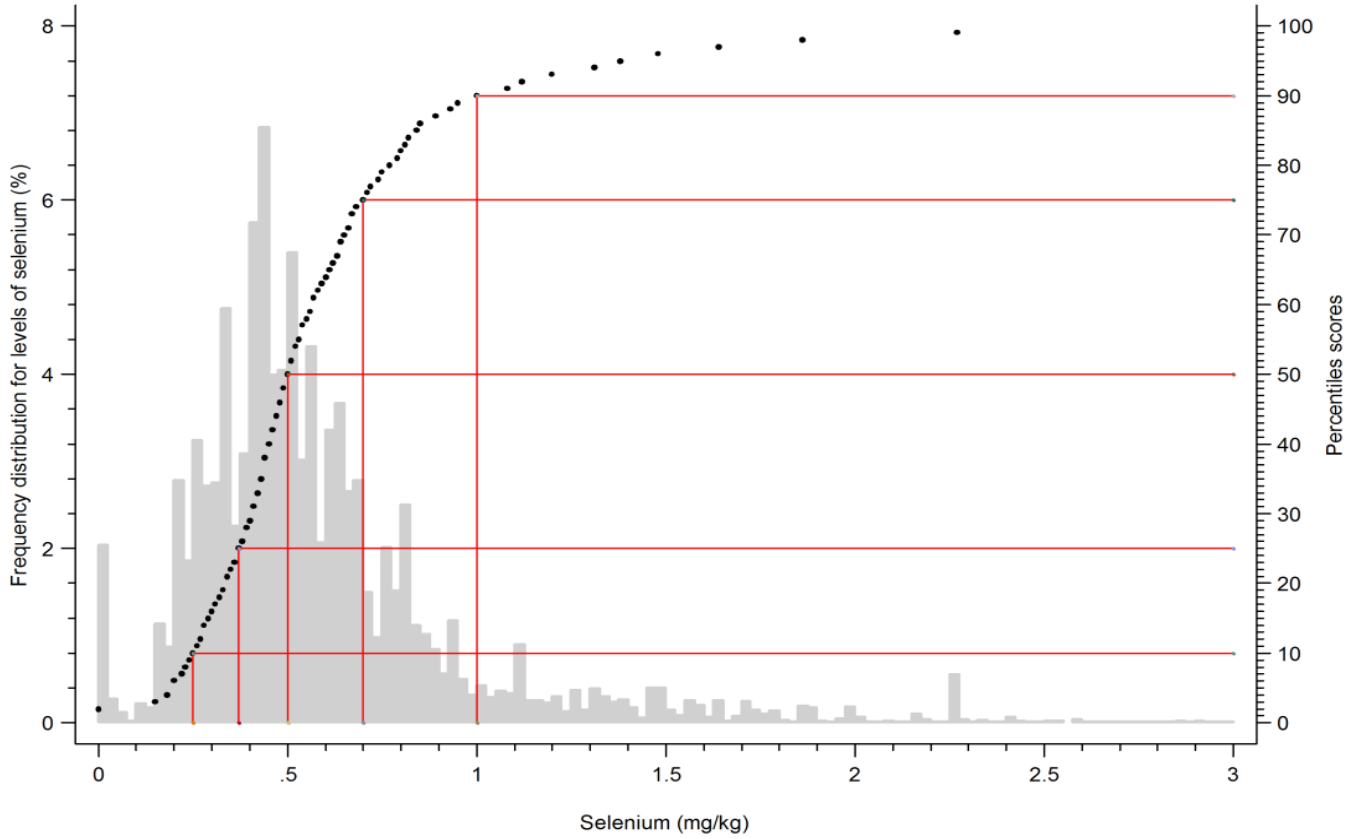

Figure 3.13: Two-way histogram with cumulative proportions showing the overall distribution of patients in THIN-GBASE with specific soil concentration levels for selenium. Left y-axis: corresponds to the observed proportion of patients with specific soil levels of selenium. Right y-axis: Black dots correspond to a percentile score – i.e. the proportion of patients that fall under specific soil concentration value for selenium; Red line indicates: 10th, 25th, 50th, 75th and 90th percentiles (i.e. 0.25, 0.37, 0.5, 0.7 and 1.0 mg/kg respectively)

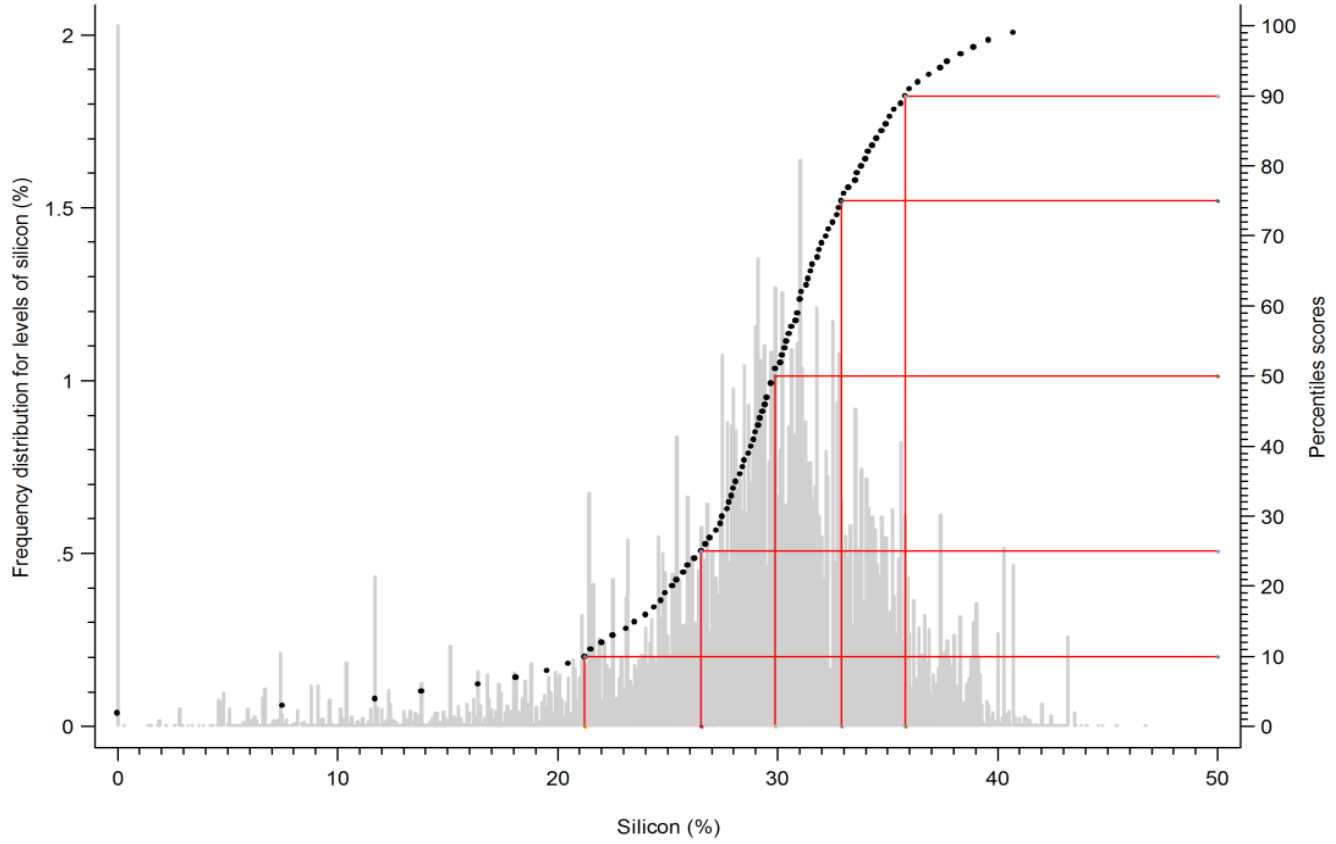

Figure 3.14: Two-way histogram with cumulative proportions showing the overall distribution of patients in THIN-GBASE with specific soil concentration levels for silicon. Left y-axis: corresponds to the observed proportion of patients with specific soil levels of silicon. Right y-axis: Black dots correspond to a percentile score - i.e. the proportion of patients that fall under specific soil concentration value for silicon; Red line indicates: 10th, 25th, 50th, 75th and 90th percentiles (i.e. 20,100, 26,500, 29,900, 32,900 and 35,900 mg/kg respectively). The concentrations for silicon were converted to weight percentage (mg/kg÷10,000), whereby 1.0% = 10,000 (of silicon) parts-per million

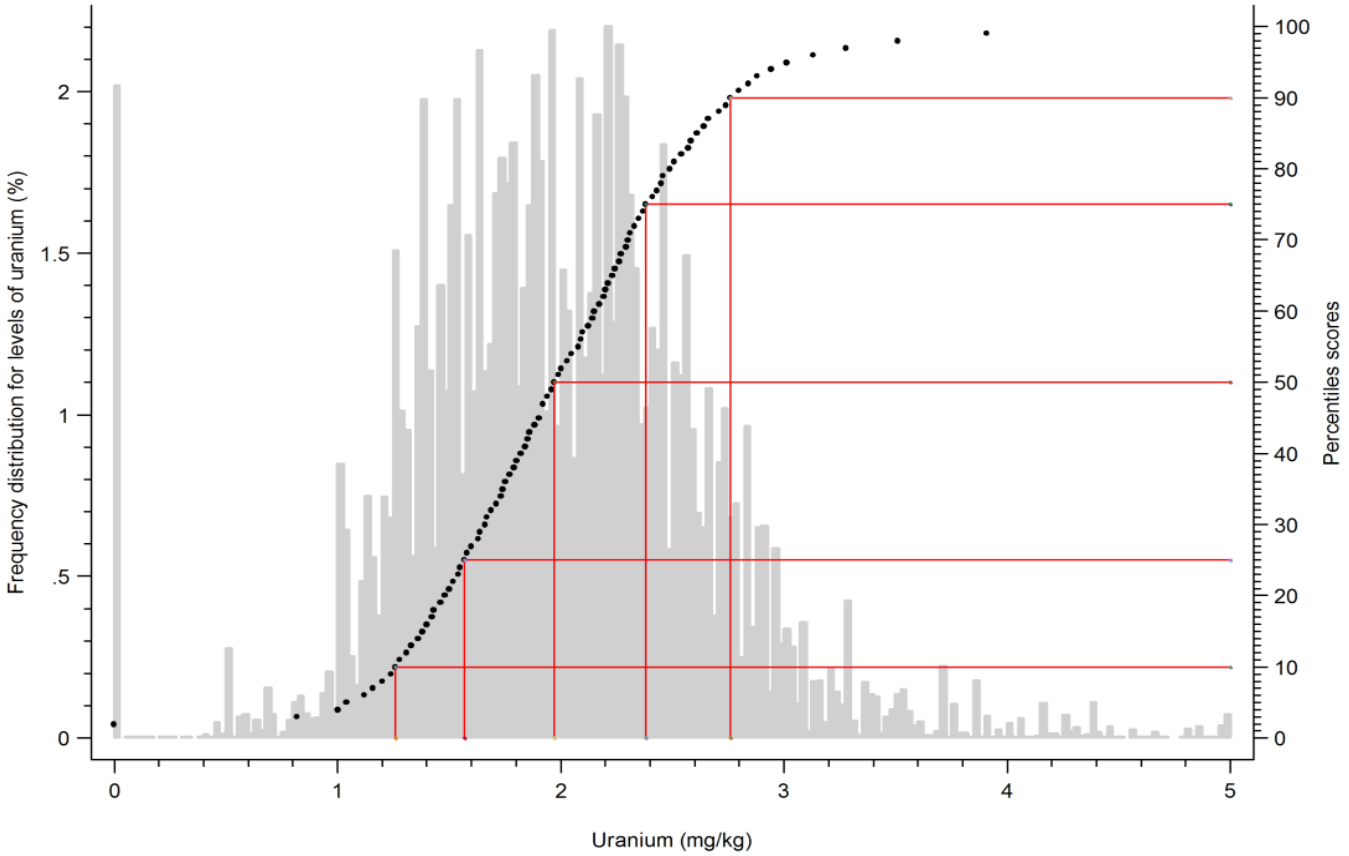

Figure 3.15: Two-way histogram with cumulative proportions showing the overall distribution of patients in THIN-GBASE with specific soil concentration levels for uranium. Left y-axis: corresponds to the observed proportion of patients with specific soil levels of uranium. Right y-axis: Black dots correspond to a percentile score – i.e. the proportion of patients that fall under specific soil concentration value for uranium; Red line indicates: 10th, 25th, 50th, 75th and 90th percentiles (i.e. 1.23, 1.57, 1.97, 2.38 and 2.78 mg/kg respectively)

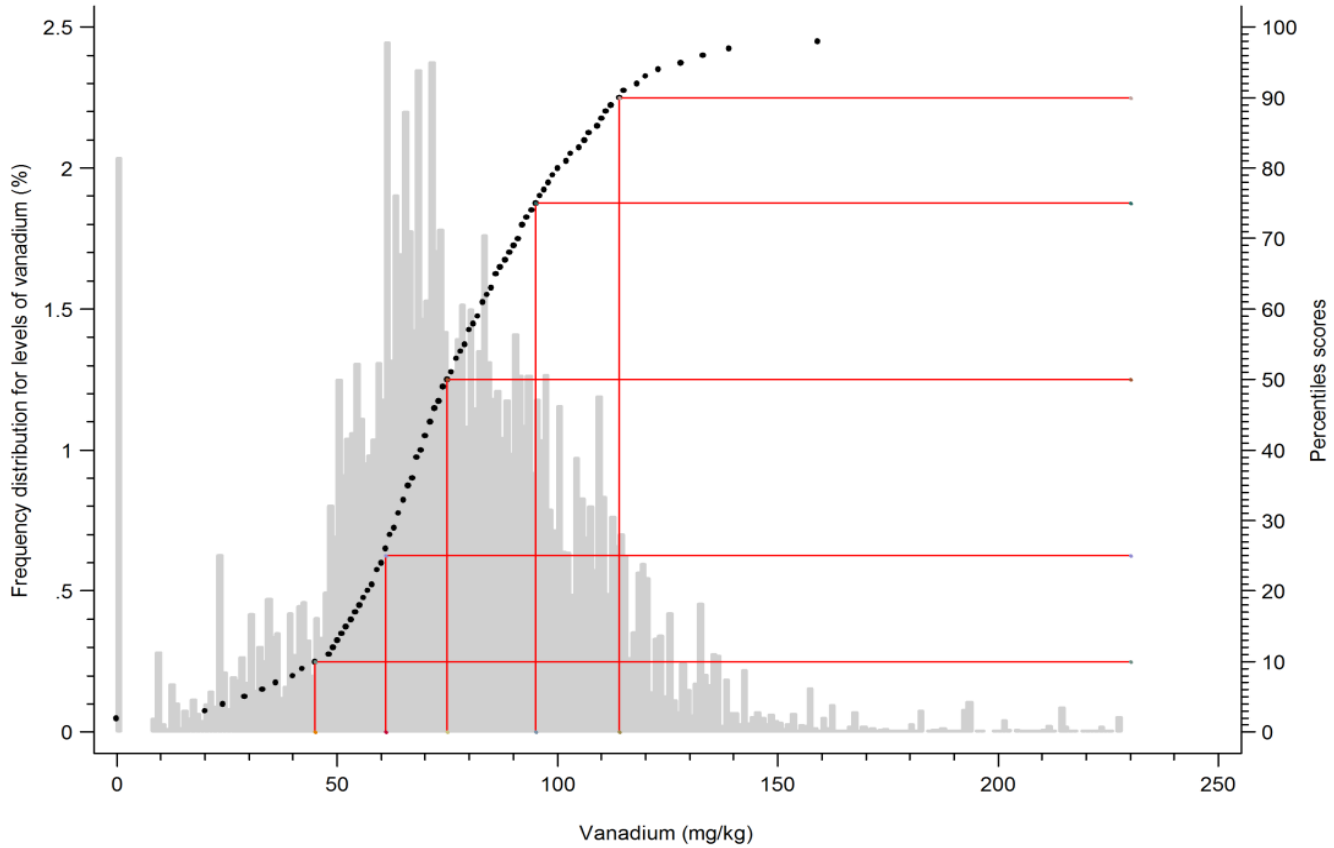

Figure 3.16: Two-way histogram with cumulative proportions showing the overall distribution of patients in THIN-GBASE with specific soil concentration levels for vanadium. Left y-axis: corresponds to the observed proportion of patients with specific soil levels of vanadium. Right y-axis: Black dots correspond to a percentile score – i.e. the proportion of patients that fall under specific soil concentration value for vanadium; Red line indicates: 10th, 25th, 50th, 75th and 90th percentiles (i.e. 45.0, 61.0, 75.0, 95.0 and 114.0 mg/kg respectively)

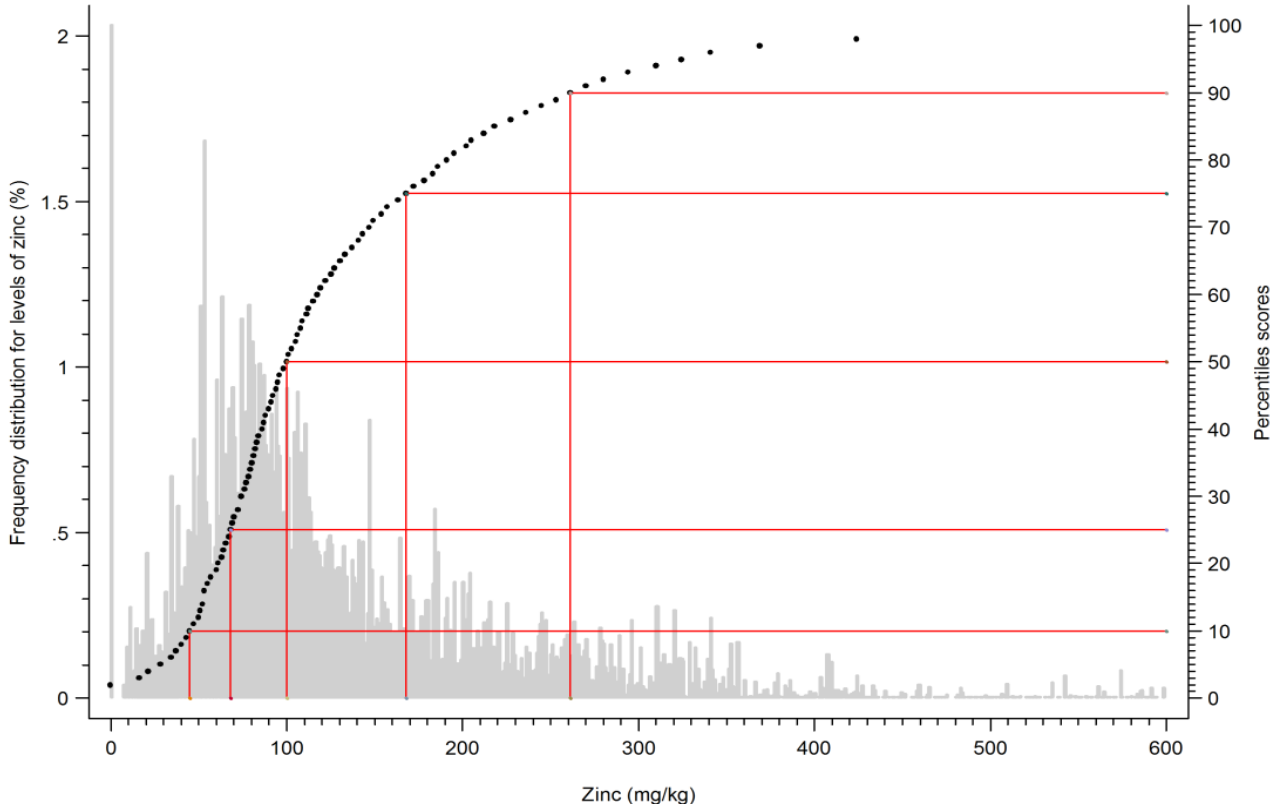

Figure 3.17: Two-way histogram with cumulative proportions showing the overall distribution of patients in THIN-GBASE with specific soil concentration levels for zinc. Left y-axis: corresponds to the observed proportion of patients with specific soil levels of zinc. Right y-axis: Black dots correspond to a percentile score – i.e. the proportion of patients that fall under specific soil concentration value for zinc; Red line indicates: 10th, 25th, 50th, 75th and 90th percentiles (i.e. 46.0, 68.0, 100.0, 168.0 and 260.0 mg/kg respectively)

Basal Cell Carcinoma

Chapter 4

4 Summary

Basal cell carcinoma (BCC) is one of the most common types of non-melanoma skin cancer in the UK. There is a well-established link between environmental arsenic exposure and the development of BCC, but at considerably higher levels of exposure than those likely to be observed in the UK. We therefore carried out a study to determine whether there is evidence that more modest levels of arsenic in soil increase the risk of BCC, as an example of testing a specific, evidence-informed hypothesis using the new data source (other elements were therefore not considered, except where there was concern they might modify the effect of arsenic). As little is known about how the incidence of BCC varies across the UK, we first took the opportunity to quantify the variation. Therefore, this chapter describes two studies:

1. Population-based ecological study design:

We sought to determine the variation in BCC throughout the UK. All adults with a first recorded diagnosis of BCC between 1-January-2004 and 31-December-2010 were extracted from the THIN-GBASE. European and world age-standardised incidence rates (EASRs and WASRs were obtained for country-level estimates and levels of socioeconomic deprivation, while strategic health authority (SHA)-level estimates were directly age-sex standardised to the UK standard population. Incidence-rate ratios were estimated using multivariable Poisson regression models to specifically assess the effects of calendar

year of diagnosis, socioeconomic deprivations, and geographical location (i.e. ten English SHAs, Wales, Northern Ireland and Scotland).

The overall EASR and WASR of BCC in the UK were 98.6 and 66.9 per 100,000 person-years, respectively. Regional-level incidence rates indicated a significant geographical variation in the distribution of BCC, which was more pronounced in the southern parts of the UK. The South East Coast had the highest BCC rate followed by South Central, Wales and South West. Incidence rates were substantially higher in the least socioeconomically deprived groups. It was observed that increasing levels of deprivation led to a decreased rate of BCC ($p < 0.001$). In terms of age groups, the largest annual increase was observed among those aged 30-49 years.

2. Population-based cohort study design:

In accordance with the UK category 4 screen levels (C4SLs) for soil arsenic, we sought to determine whether residential soil arsenic exposure above the soil guideline value (i.e. arsenic-C4SL = 35.0 mg/kg) was associated with an increased risk of developing BCC in the UK population.

All patients with a first diagnosis of BCC between 1-January-2004 and 31-December-2011, and with X-ray fluorescence spectroscopic measurements of total soil arsenic levels were extracted from THIN-GBASE. Multivariable Cox regression models were used to quantify associations between BCC and soil arsenic.

The adjusted hazard ratio (HR) for participants with exposures of 35.0-70.0 mg/kg (HR 1.08, 95% CI: 1.02-1.14) and ≥70.0 mg/kg (HR 1.17, 95% CI: 1.09-1.28) had increased hazards of developing BCC compared to those with exposures <17.5 mg/kg. Urban residents with the highest exposure significantly had the greatest risk of developing BCC (≥ 70.0 mg/kg: HR 1.18, 95% CI: 1.06-1.36).

***Conclusion*:** BCC is an increasing health problem in the UK. In our ecological study, it is established that southern regions of the England; and those in the least deprived socioeconomic group have a much higher incidence rate of BCC. In addition, our findings from our cohort study provides evident that exposure to soil arsenic is an important risk factor. Residential soil concentration levels for arsenic should be considered for further investigation of BCC aetiology.

4.1 Background

Non-melanoma skin cancers (NMSC) are a group of neoplasms that develop in the epidermis of the skin. NMSCs are named according to the skin cell-type from which the tumour originates.[120] There are two major types of NMSCs: basal cell carcinoma (BCC) and squamous cell carcinoma (SCC).[120–123] BCCs are tumours that originate from the basal cells situated in the lowest layer of the skin's epidermis. SCCs originate from the squamous cells located in the mid-section of the epidermis. Unlike BCCs, SCCs are more life threatening as their tumours can potentially spread to other organs if left untreated. Other rarer forms of NMSCs include Merkel cell carcinoma, Kaposi sarcoma and T-cell lymphoma of the skin.[120–123]

BCCs are the most common form accounting for 75% of cases diagnosed with NMSC, which globally affects close to 1.5 million individuals annually.[122] Past studies have shown individuals at risk of BCC are those of Caucasian descent living in US, Canada, Australia and most European countries.[124–128] The main risk factor for NMSC is exposure to Ultraviolet (UV) light, through sun exposure or the use of tanning lights, and individual tanning behaviour.[129–131] Other major risk factors are advancement of age,[132] gender,[132,133] skin type (i.e. fair, white or freckled skin)[132,134–136] and a prior history of skin cancer.[123]

4.2 Regional variations of Basal cell carcinoma incidence in UK

4.2.1 Background

In the UK, BCC incidence is increasing at an unprecedented rate. The overall incidence of NMSC is estimated to be well over 100,000 cases per year, with BCC accounting for 75% of cases, SCC for 20% and other rare skin cancer types (i.e. Merkel cell carcinoma, Kaposi's sarcoma and T-cell lymphoma of the skin) for 5%.[121]

Recent studies have shown that whilst the incidence of BCC varies geographically in the UK,[124,137] rapidly increasing incidence has been observed in many areas. For instance, in West Glamorgan (Wales) the incidence rate increased by 60% between 1988 and 1998.[138] Similarly, in England, the incidence of BCC in North Humberside tripled over the 13 year period from 1978 to 1991.[124,139] Scotland and Northern Ireland have lower incidence of BCC relative to England and Wales, however, within the past two decades the incidence of BCC among men has risen approximately by 16% in Scotland and 18% in Northern Ireland.[140,141] The elderly population contributes substantially to the disease burden, with risk of BCC increasing after 40 years of age, however, we are now seeing an increased incidence among younger people and in particular those of ≤ 30 years of age.[142]

Socioeconomic status and deprivation are also known to modify the risk of BCC. Some studies suggest that BCC appears to be more

common in those belonging to higher social class;[136,143,144] however, such associations are not well understood and the distribution of BCC in terms of levels socioeconomic deprivation in the UK population is unknown.

We therefore used data from a UK-wide database of primary care medical records to derive contemporary regional breakdowns of the incidence of BCC in UK. We present novel incidence rate estimates stratified by level of socioeconomic deprivation in the UK and additional analyses examining whether BCC incidence has continued to increase in recent years, particularly in the younger age groups.

4.2.2 Methods

4.2.2.1 Study design

We conducted a large population-based study using data from THIN. THIN is a large database comprised of anonymised primary care electronic medical records of more than 10 million patients from across all regions of the UK. The information contained within THIN includes details on all diagnoses made by or reported to the general practitioner, as well as other additional health information relevant to primary care. THIN is recognised for its completeness and accuracy of data recording, and has been validated for its suitability for use in medical research,[107] including specific validation of diagnoses of BCC.[145] In addition, a range of socio-demographic indicators are

available in THIN, including quintiles of Townsend Deprivation Index[146] in each patient's postcode of residence.

4.2.2.2 Case definition for Basal cell carcinoma

The medical histories and deprivation indicators of all adults aged 18 years or above with a first recorded diagnosis of BCC between January 1st, 2004 and December 31st, 2010, were extracted from THIN. Subjects diagnosed with Basal cell nevus syndrome (or Gorlin's syndrome), organoid naevi or other genetic syndromes were excluded from the study. Patient ages were categorised into 10-year bands (18–29, 30–39, 40–49, 50–64, 65–79 and 80+ years). Patients were categorised into thirteen regional based (in England) on the SHA or devolved government administration (Wales, Scotland and Northern Ireland) to which each patient's primary care practice belongs as follows: North East, North West, Yorkshire and Humber, East Midlands, West Midlands, East of England, London, South East Coast, South Central, South West, Wales, Scotland and Northern Ireland. This is the only spatially referenced data available from the anonymised patient records.

4.2.2.3 Statistical analysis

4.2.2.3.1 Methodology for calculating incidence rates

The primary outcome measures were incidence rates (IR) of BCC in the whole UK, in its constituent countries and principalities, and each English SHA region. We also estimated the incidence of BCC across

quintiles of socioeconomic deprivation groups. IRs was calculated as the number of patients receiving their first diagnosis of BCC divided by the total number of person-years at risk. Diagnoses within the first year of registration with a participating primary care practice were excluded as such recordings can relate to prevalent, rather than incident, events (being an artefact of back-entry of records from a previous practitioner). Second or subsequent diagnoses were also excluded as these are difficult to differentiate from recurrences and follow-up consultations in primary care records. Population denominators were mid-year (1st July) total numbers of persons registered for at least one year at a primary care practice enrolled in THIN. IRs was presented as rates per 100,000 person-years. We derived estimates for European and World Age-Standardised incidence Rates (EASRs and WASRs, respectively) using the direct standardisation method to allow direct comparisons between country-level incidence rates (i.e. UK, England, Scotland, Northern Ireland and Wales) with other populations.[147] IRs of BCC at a regional level were directly age-sex standardised to the UK standard population.[148]

4.2.2.3.2 **Multivariable Poisson regression modelling**

Poisson multivariable regression model was used to examine the effects of all factors (i.e. calendar year of diagnosis, socioeconomic deprivation and regions) on the incidence of BCC adjusted for sex and age groups. Stratified Poisson multivariable analyses were used to determine whether associations between all factors and the incidence

of BCC were modified by sex, while controlling for age groups. For secondary analyses, we further used stratified models by age groups to assess calendar years as a continuous variable to determine the average change (per year) in incidence of BCC. Incidence rate ratios (IRR) were estimated with 95% confidence intervals (CI). All statistical analyses were carried out using STATA version 12 (STATA Corporation, College Station, TX, USA).

4.2.3 **Results**

There were 38,121 incident cases of BCC were identified from 546 GPs in the THIN database. Mean age was 64 years (standard deviation, SD 13 years) with slightly more men (52.4%) diagnosed with a BCC.

4.2.3.1 **BCC incidence at country and regional level**

The crude IR of BCC between 2004 and 2010 in our THIN database was 171.9 per 100,000 person-years (95% CI: 170.1–173.6). The crude IR of BCC was higher among men (183.1 per 100,000 person-years; 95% CI: 180.5–185.6) than women (161.0 per 100,000 person-years; 95% CI: 158.7–163.4) (Table 4.1). When comparing the overall figures between 2004 and 2010, we found that there was an increase from 154.0 per 100,000 person-years to 182.0 per 100,000 person-years. Our Poisson multivariable regression model show that the overall significant 16% increase in incidence in 2010 as compared to 2004 (IRR 1.16, 95% CI: 1.11–1.20), which equates to an average increase of 2.5% per year (95% CI: 1.9% – 3.0%; *p* for trend = 0.001) (Table 4.3).

Comparatively, at a country-level, Wales has significantly the highest overall crude rate of BCC (IR 196.4, 95% CI: 189.2–203.8) followed by England (IR 178.5, 95% CI: 176.5–180.5). We observed the incidences were low in Scotland (IR 128.7, 95% CI: 124.6–133.0) and Northern Ireland (IR 131.6, 95% CI: 123.3–140.3) (Table 4.1). We observed the incidences are low, and similar for Scotland and Northern Ireland.

We observed important geographical variations in the distribution of BCC (Figure 4.1). The age-sex standardised incidence rate of BCC in South East Coast, South West, South Central and Wales are significantly higher than other regions of the UK (Table 4.2).

Our models show that the incidence of BCC were significantly lower in West Midlands (IRR 0.92, 95% CI: 0.88–0.97), Northern Ireland (IRR 0.92, 95% CI: 0.85–0.98) and Scotland (IRR 0.87, 95% CI: 0.83–0.91) than in London (the referent). Conversely, we found that the incidence of BCC was significantly higher in South East Coast (IRR 1.28, 95% CI: 1.22–1.34), Wales (IRR 1.23, 95% CI: 1.16–1.29), South Central (IRR 1.21, 95% CI: 1.15–1.27), South West (IRR 1.15, 95% CI: 1.09–1.21) than in London. Our results show no substantial sex-specific difference in the incidence of BCC in any regions (Table 4.3).

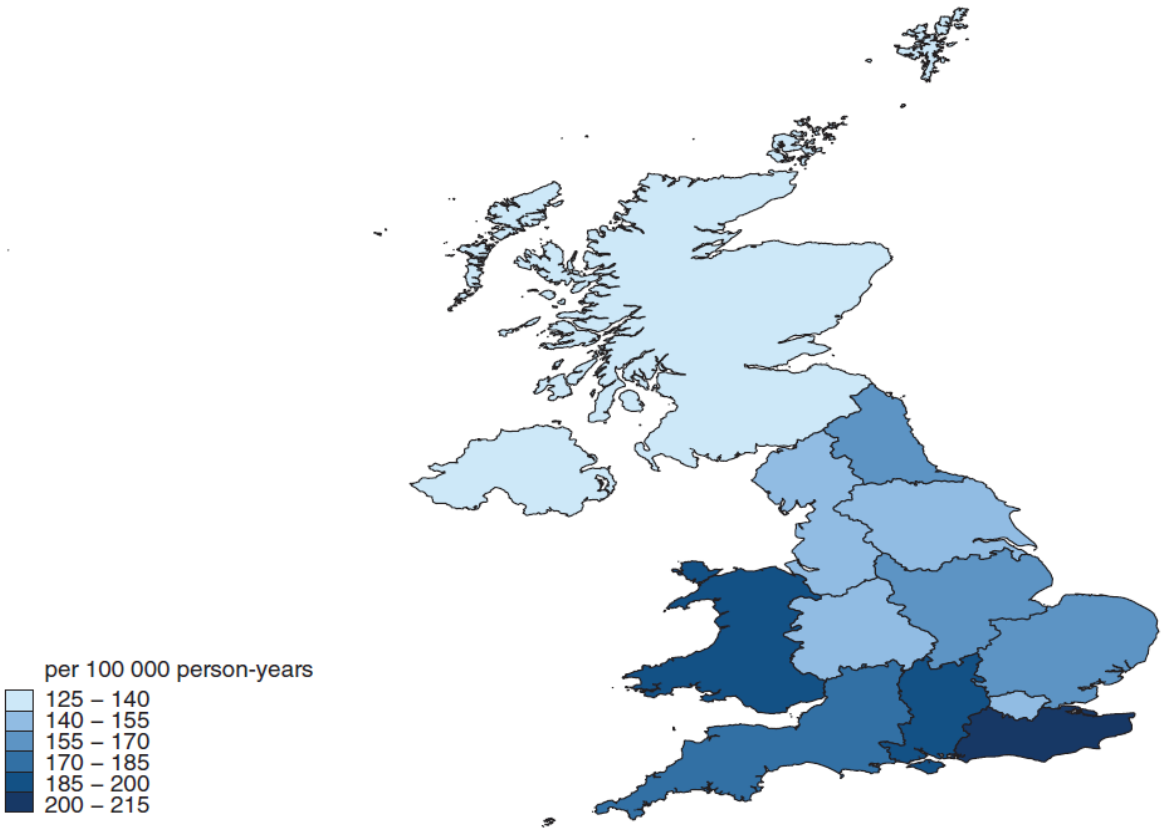

Figure 4.1: Thematic map showing direct age & sex- standardised incidence rates of BCC in the UK standard population (THIN database) 2004-2010

Table 4.1: Crude and sex-specific age-standardised incidence rates of Basal cell carcinoma in UK and countries, THIN database (2004–2010)

	Men; IR (*n*)	Women; IR (*n*)	Overall; IR (*N*)
United Kingdom			
Crude	183.1 (19,960)	161.0 (18,161)	171.9 (38,121)
EASR[a]	112.2	88.1	98.6
WASR[b]	74.8	60.7	66.9
England			
Crude	189.9 (16,079)	167.5 (14,671)	178.5 (30,750)
EASR	114.9	91.4	101.5
WASR	76.6	63.1	69.0
Northern Ireland			
Crude	144.7 (502)	119.3 (439)	131.6 (941)
EASR	99.6	67.5	81.6
WASR	66.2	45.2	54.6
Scotland			
Crude	137.9 (1,904)	119.8 (1,704)	128.7 (3,608)
EASR	89.3	65.6	75.9
WASR	59.1	44.7	51.1
Wales			
Crude	208.1 (1,475)	185.0 (1,347)	196.4 (2,822)
EASR	128.7	103.1	114.4
WASR	86.4	71.3	78.1

[a]EASR, European age-standardised rate.
[b]WASR, World age-standardised rate.

Table 4.2: Regional-level estimates for sex-specific and age-sex standardised incidence rates of Basal cell carcinoma in UK, THIN database (2004-2010)

	Age-sex standardised rate[a]	Sex-specific age standardised rate[b]	
Regions	Overall (*N*)	Men (*n*)	Women (*n*)
Scotland	127.9	139.5 (1,904)	116.8 (1,704)
Northern Ireland	138.4	155.4 (502)	122.2 (439)
London	144.0	148.6 (1,432)	139.6 (1,415)
West Midlands	144.1	152.2 (1,537)	136.3 (1,422)
North West	146.5	156.6 (1,785)	136.8 (1,627)
Yorkshire & Humber	151.4	163.3 (686)	140.0 (618)
North East	156.0	165.2 (539)	147.2 (503)
East Midlands	158.6	166.4 (776)	151.2 (718)
East of England	161.1	170.7 (1,457)	151.8 (1,325)
South West	180.2	196.2 (2,438)	165.0 (2,123)
Wales	185.7	197.6 (1,475)	174.4 (1,347)
South Central	193.5	208.7 (2,998)	178.9 (2,645)
South East Coast	202.7	215.0 (2,431)	191.0 (2,275)

[a]Estimates are directly age-sex standardised using UK as the standard population.
[b]Sex-specific estimates are directly age-standardised using the UK as the standard population.

Table 4.3: Overall & sex-specific incidence rate ratio (IRR) estimates showing associations between incidence of BCC and risk factors

	Men[a]	Women[a]	Overall[b]
	IRR (95% CI)[c]	IRR (95% CI)	IRR (95% CI)
Years			
2004	1	1	1
2005	1.07 (1.02 – 1.14)	1.00 (0.94 – 1.06)	1.04 (0.99 – 1.08)
2006	1.10 (1.04 – 1.16)	1.07 (1.07 – 1.13)	1.09 (1.04 – 1.13)
2007	1.14 (1.08 – 1.21)	1.16 (1.10 – 1.23)	1.15 (1.10 – 1.20)
2008	1.16 (1.10 – 1.22)	1.16 (1.10 – 1.23)	1.16 (1.12 – 1.20)
2009	1.15 (1.09 – 1.22)	1.13 (1.07 – 1.20)	1.15 (1.10 – 1.19)
2010	1.12 (1.06 – 1.18)	1.21 (1.14 – 1.27)	1.16 (1.12 – 1.21)
Annual increase	1.8% (1.1% – 2.5%)	3.2% (2.5% – 4.0%)	2.5% (1.9% – 3.0%)
(*p* for trend)	$p = 0.003$	$p = 0.008$	$p < 0.001$
Socioeconomic deprivation[d]			
5th (Most deprived)	1	1	1
4th	1.13 (1.06 – 1.20)	1.01 (0.95 – 1.08)	1.07 (1.02 – 1.12)
3rd	1.28 (1.21 – 1.36)	1.13 (1.07 – 1.20)	1.21 (1.16 – 1.26)
2nd	1.47 (1.39 – 1.56)	1.26 (1.19 – 1.33)	1.37 (1.31 – 1.43)
1st (Least deprived)	1.62 (1.53 – 1.72)	1.36 (1.28 – 1.44)	1.50 (1.44 – 1.56)
Unknown	1.22 (1.11 – 1.35)	1.12 (1.01 – 1.23)	1.17 (1.09 – 1.25)
(*p* for trend)	$p < 0.001$	$p < 0.001$	$p < 0.001$

Regions						
London	1		1		1	
Scotland	0.91	(0.85 – 0.98)	0.82	(0.76 – 0.88)	0.87	(0.83 – 0.91)
Northern Ireland	0.99	(0.89 – 1.10)	0.84	(0.76 – 0.94)	0.92	(0.85 – 0.98)
West Midlands	0.93	(0.87 – 1.00)	0.92	(0.85 – 0.99)	0.92	(0.88 – 0.97)
North West	0.98	(0.91 – 1.05)	0.93	(0.87 – 1.00)	0.96	(0.91 – 1.01)
Yorkshire & Humber	1.04	(0.96 – 1.15)	0.98	(0.89 – 1.08)	1.01	(0.95 – 1.08)
East Midlands	1.02	(0.93 – 1.11)	1.02	(0.93 – 1.12)	1.02	(0.96 – 1.09)
North East	1.08	(0.97 – 1.19)	1.03	(0.93 – 1.14)	1.05	(0.98 – 1.13)
East of England	1.05	(0.98 – 1.13)	1.02	(0.95 – 1.10)	1.04	(0.98 – 1.09)
Wales	1.25	(1.17 – 1.35)	1.19	(1.11 – 1.29)	1.23	(1.16 – 1.29)
South Central	1.24	(1.16 – 1.32)	1.18	(1.10 – 1.26)	1.21	(1.15 – 1.27)
South West	1.19	(1.12 – 1.28)	1.10	(1.03 – 1.18)	1.15	(1.09 – 1.21)
South East Coast	1.30	(1.21 – 1.39)	1.27	(1.18 – 1.36)	1.28	(1.22 – 1.34)

[a]Models were stratified by sex, includes all covariates and adjusted for age groups: i.e. 18–29, 30–39, 40–49, 50–64, 65–79 and 80+ years.

[b]Overall model includes all covariates and adjusted for sex and age bands: i.e. 18–29, 30–39, 40–49, 50–64, 65–79 and 80+ years.

[c]IRR, incidence rate ratio; CI, 95% confidence interval.

[d]Quintiles of Townsend deprivation index.

4.2.3.2 BCC trends over time by age group

Models were stratified by age groups to determine the effects of calendar years on the incidence of BCC. Our results show a small average increase of 0.4% per year in the age group of 18–29 years, however, this failed to reach statistical significance (95% CI: -8.0% to 9.3%; $p = 0.91$). In particular, the largest average increase in incidence was observed for those in 30–39 years (3.9% per year, 95% CI: 0.2% to 7.7%; $p = 0.04$) and 40–49 (4.0% per year, 95% CI: 2.0% to 6.1%; $p < 0.001$) age groups (Figure 4.2).

4.2.3.3 BCC incidence by socioeconomic deprivation

The crude incidence of BCC was significantly highest for those living in areas with the lower levels of deprivation, with estimates of 222.5 per 100,000 person-years (95% CI: 218.5–226.5) and 203.2 per 100,000 person-years (95% CI: 199.1–207.3) for those in the 1st quintile (least deprived) and 2nd quintile, respectively (Table 4.4). We observed that the incidence of BCC was lowest for those living in the most deprived areas (IR 110.6, 95% CI: 106.8–114.7). Using our models, there appeared to be a linear effect of decreasing incidence of BCC with increasing levels of deprivation (p for trend < 0.001). We found that those living in the least deprived areas were 50% significantly more likely to have a BCC than those with the highest levels of deprivation (IRR 1.50, 95% CI: 1.44–1.56). Our models also show substantial difference in magnitude of the incidence of BCC between men and

women, where the IRRs for socioeconomic deprivation were higher in men than in women (Table 4.3).

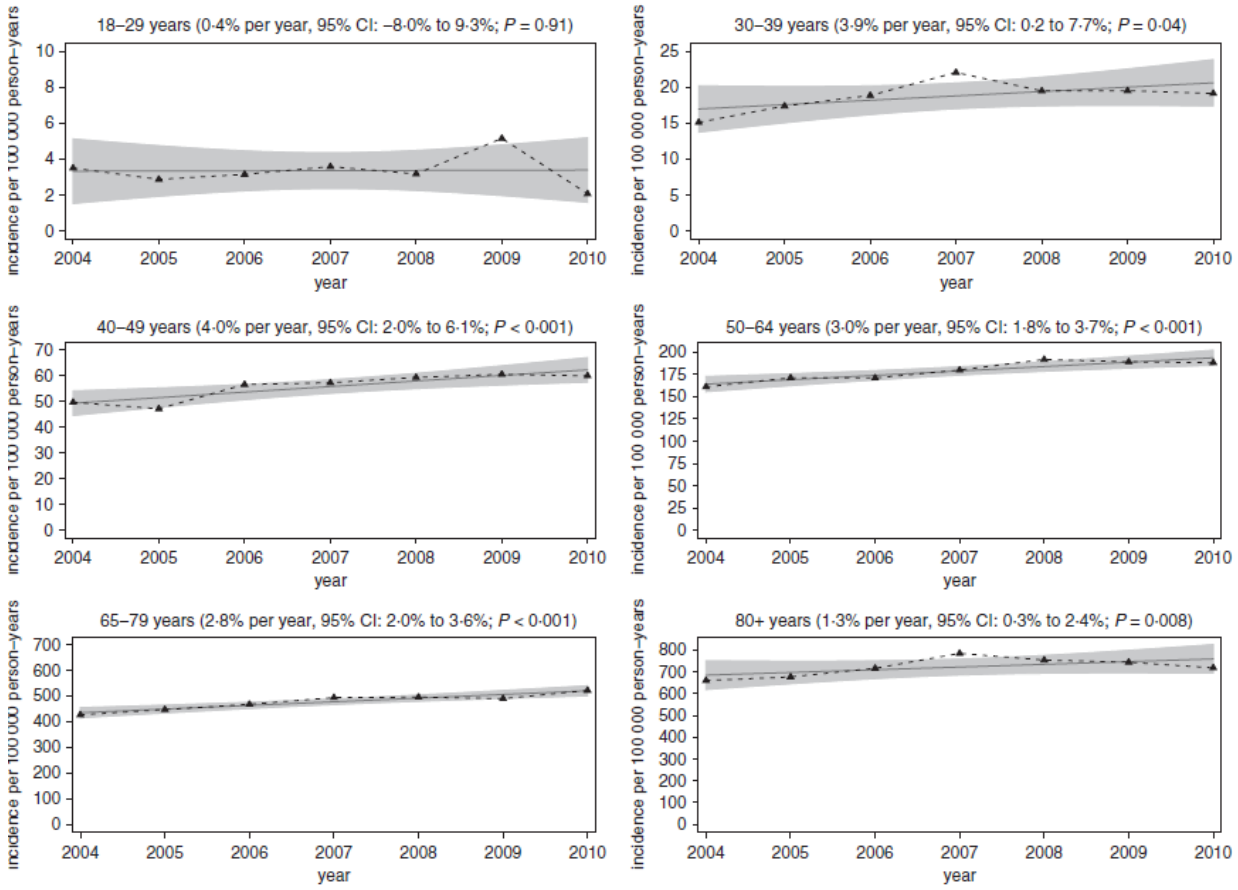

Figure 4.2: Average change in incidence of BCC in the UK stratified age-groups (18-29, 30-39, 40-49, 50-64, 65-79 & 80+) (THIN database) 2004 – 2010. Grey shaded area represent 95% confidence intervals for the year-on-year change in incidence rates.

Table 4.4: Crude and sex-specific age-standardised incidence rates of Basal cell carcinoma in the UK, by quintiles of Townsend deprivation index (THIN database 2004–2010)

Deprivation index	1	2	3	4	5	Unknown
Overall						
Crude	222.5 (12,070)	203.2 (9,575)	162.1 (7,175)	131.7 (5,173)	110.7 (3,013)	115.6 (1,115)
EASR[a]	120.2	106.9	92.2	79.4	70.6	-
WASR[b]	82.1	72.7	62.4	53.5	47.2	-
Men						
Crude	246.7 (6,602)	220.8 (5,097)	169.7 (3,677)	134.3 (2,583)	106.0 (1,441)	119.0 (560)
EASR	137.5	122.0	104.7	90.4	77.4	-
WASR	92.3	81.5	69.6	59.7	50.9	-
Women						
Crude	198.9 (5,468)	186.2 (4,478)	154.7 (3,498)	129.2 (2,590)	115.3 (1,572)	112.4 (555)
EASR	105.7	94.9	83.1	71.9	66.2	-
WASR	73.2	65.7	57.1	49.2	44.7	-

[a]EASR, European age-standardised rate.
[b]WASR, World age-standardised rate.
Incidence rates (per 100,000)

4.2.4 **Discussion**

Our results indicate that the incidence rate of BCC is increasing in the general population, in particular amongst those aged 30 to 49 years. It shows that Wales and the southern parts of England have the highest recorded rates of skin BCC. For socioeconomic deprivation, incidence of BCC was consistently higher in the least deprived groups. Based on our estimates (i.e. EASRs), they show that approximately 61,500 new cases of BCC are diagnosed annually in the UK population. Previous reports using EASRs have estimated that 53,000 cases of BCC were reported yearly using a cohort between 1996 and 2003,[137] comparatively; this represents an overall increase of 16% in diagnosis rates in the past decade.

This study has several strengths; to our knowledge, it uses the largest sample size of incident cases of BCC compared to previous research conducted in the UK.[136–141,143] Due to our large sample size means that our findings are unlikely to be due by chance. Also, the data was obtained from a national database and prospectively recorded by GPs, thus excluding the possibility of recording or recall bias in both our exposure and outcome. The major limitation is our inability to account for important factors such as history of sun exposure during childhood and adolescence (i.e. frequency of sunburns and overseas holidays),[149,150] latitudinal position (i.e. proximity to the equator),[132,151] settings of occupation (i.e. indoor, mixed or outdoor)[152] and skin type (i.e. fair, white or freckled skin).[136] In

addition, we were unable to classify adults according to subtypes of BCC (i.e. superficial, nodular or infiltrative).

Our results for country-level incidence rates were consistent with previous studies showing escalating rates in England, Northern Ireland, Scotland and Wales.[138–141] The most likely explanation for the rise in incidence may be linked to previous behaviour with regards to sun exposure during childhood or adolescence. Exposure to UV radiation during this stage plays a significant role in the future development of BCC. Previous studies have shown that subjects to have reported to travel frequently and spend more than 4 or 5 weeks (per year) at the beach before the age of 20 years were more likely to have developed the skin malignancy in their adulthood.[149,150] Although we were unable to account for this factor, history of sun exposure through frequent holidays to sunnier places has been a strong predictor for BCC. Another likely explanation may be possibly due to UK's aging population. BCC is highly prevalent in the older age groups; in our cohort, the number of cases diagnosed with the skin malignancy was consistently high among those aged 50+ years.

We found a significant increase among those aged 30–39 and 40–49 years. A previous study has shown similar findings, where the annual increase in incidence was estimated to be approximately 3.9% and 5.2% for 30–39 and 40–49 years, respectively, although these estimates did not reach statistical significance.[137] Interestingly, we found an increase among those aged 18–29 years is increasing, although our

models showed no statistical significance. The incidence of BCC in this particular age group have risen to approximately 5.2 per 100,000 person-year in 2009 (Figure 4.2), which is consistent with escalating rates observed by others.[142]

We observed especially high incidence in areas of South East Coast, South Central, South West and Wales. Compared to London, we found there were significant increases in the risk of developing BCC in these areas. This observation may be linked to several environment factors. The most prominent is the latitudinal position of a location.[132,151] Areas in proximity to the equator, but situated in the temperate zone usually experience prolonged duration of sunlight in the summer season. In the UK, the hours of sunshine normally last longer in the south than the northern regions of UK, especially during the summer season, the southern parts of England and Wales are usually known to receive the greatest hours of annual sunshine.[153]

Our findings for socioeconomic deprivation showed that the incidence was high among the least deprived groups, and that the risks for BCC tends to decrease as the level of deprivation increased. Our results are consistent with previous studies conducted in the UK and Netherlands.[136,144] It is interesting to note the wide difference in incidence of BCC between the least and most deprived groups which may be an indication that socioeconomic status or deprivation is risk factor for BCC. This observation may be linked to higher levels of income for frequent holidays overseas to sunnier places, thereby

exposing the skin to sunlight, or having available funds for pursuing other lifestyle habits which are risk factors for BCC, for instance, the frequent use of tanning beds[131–133] or consumption of alcohol.[154] Interestingly, we also observed that the incidences differ substantially between sexes, perhaps, this may possibly be due to differences in behaviour in terms of sun exposure, clothing habits and tanning behaviour.[136,144]

BCC is an increasingly important health problem in the UK, with extremely high levels observed in the least deprived groups, and in the southern parts of the UK. Due to the multi-factorial nature of BCC, further work is warranted to identify causes, as well as, investigate the detailed reasons of these findings. Our results demonstrate that the incidence of BCC will continue to rise much higher in all age bands if it remains unchecked, which will have a significant impact on the workload and costs for health services. Better strategies are required to inform the public of the risk factors associated with the skin malignancy, as well as, which preventive measures can be implemented to avoid future development of BCC.

4.3 Potential exposure to soil arsenic and risk of Basal cell carcinoma in UK

4.3.1 Background

Arsenic is well-known for its notoriety as a result of its presence in groundwater which has led to widespread contamination of drinking water in more than 70 countries throughout the world, including the USA, some parts of Eastern Europe and Bangladesh.[47,56] Arsenic is an environmentally persistent element which is found naturally occurring in soil and groundwater.[19] The primary pathways for exposure for arsenic are inhalation and ingestion of particulate matter that are transported from the earth (or soil) to the atmosphere as a result of anthropogenic activities or wind erosion, ingestion of contaminated foods or drinking water obtained from contaminated areas, and through direct contact with the skin.[32,33,155]

Arsenic is a recognised carcinogen shown to be linked with several types of cancers.[48,156,157] The most prominent are the effects of arsenic on the risk of BCC and SCC.[87,158–160] Epidemiological studies have been documented extensively, and focused on populations exposed through drinking water emerging from arsenic-contaminated underground aquifers.[47,56] However, there has been little or no focus on the potential health impacts of arsenic emerging from lithological sources, in particular soils, due to the belief that exposures at environmental concentrations are low and insufficient to have an effect.[15,76]

It has been implied that the potential mechanism for cancer development in consequence of geochemical arsenic exposures from topsoil are somewhat similar to those which apply to exposures from occupational settings, except that the processes are ongoing and long-term.[1,2,13] The individual is exposed to arsenic emerging from the environment (soils) via the primary pathways; continued exposure eventually leads to accumulation of such toxicants which are broken down to metabolites causing toxicity to the body which, in turn, have genotoxic effects on various tissues such as the lungs, kidney or skin thereby resulting in significant DNA damage which leads to cancer.[38]

In the UK, there are widespread areas with elevated concentrations of arsenic, with specific geologic formations yielding especially high concentrations in some areas of England and Wales.[161–163] These increased concentrations resulted mainly from the extensive mining activities that took place before the 1970s, and also due to natural lithological processes occurring in soils such as mineralisation and rock formation.[161] DEFRA have provided screening values known as the C4SLs for assessing soil safety for arsenic, and recommends concentration levels in residential soils should not exceed 35 mg/kg.[71] However, there are residential soils in many urban and rural areas of England and Wales with arsenic concentrations exceeding 100 mg/kg;[164] whether this kind of long-term exposure for people living on soils with elevated arsenic concentration is contributory to the BCC incidence in UK remains unclear. A small number of studies have found concentrations of soil arsenic in the environment to correlate

with biomarkers of exposure (urine, hair and toe or fingernail samples),[27,31,92], and that these biomarkers are, in turn, a proxy measurements for increased risk of skin cancer.[30,158]

To our knowledge, there has been no research directly examining the relationships between geochemical arsenic exposure, at typical and widespread environmental concentrations, and BCC risk in the UK. Thus, the aim of the study was to quantify the risk of BCC with increasing soil arsenic concentration through applying specified cut-off values which included the UKs C4SLs for soil arsenic using a population based cohort.

4.3.2 **Methods**

4.3.2.1 **Geochemical soil arsenic data**

The geochemical data used for this population based cohort study were derived from the BGSs G-BASE database,[113,165] and from the NSI-XRFS project which have been re-analysed to extend the number of elements and upgrade the analytical quality to match the G-BASE data.[53,166] Based on a systematic grid across England and Wales using the British Ordnance Scale (OS) of 1:25,000. Soil samples were collected at different densities depending on the type of environment at a fixed time point. For G-BASE rural areas, suitable sampling points (or sites) were identified at a density of 1 per 2km^2 alternate-grid across England and Wales, while sample points from G-BASE urban areas were chosen at a density of 1 per 0.25km^2. Most of G-BASE data

were concentrated in the eastern and central parts of England with additional samples from the Tamar catchment of South England. Existing data from the NSI collected at a density of 1 per 25km^2 was re-analysed by the BGS, and was augmented with the G-BASE database to achieve completeness of coverage (Figure 3.1).

For the G-BASE samples, a composite of five soil samples was taken from the topsoil at each location, at a depth of 2–15cm using the augur flights. For NSI samples, 25 cores of soil were taken at the nodes of a 4m grid within a 20m by 20m square centred on each OS 5-km grid point across England and Wales. The aggregated soil materials were analysed using x-ray fluorescence spectrometry (XRFS) to detect the geochemical composition, and measure the concentration levels for arsenic present in soil. The measurements for soil arsenic were spatially referenced, and loaded into ArcGIS desktop 10.3 (ESRI, Redlands, California, USA) to create an interpolated surface map for arsenic so as to derive estimates at finer resolution (at a patient postcode level). These estimates were uniquely linked to the primary care electronic medical records (EMRs) of eight million individuals throughout England and Wales. Each individual was registered at a one of 377 GPs contributing to The Health Improvement Network database (THIN) that agreed to participate in the linkage project.[119]

4.3.2.2 Study design

The population based cohort study was between the period of 1st January, 2004, and 31st December, 2011. Participants were extracted

from the linked database (from GPs contributing to THIN with G-BASE coverage) if they were registered with a GP for at least one year before 1st January, 2004, and if their GP has an AMR date before the 1st January 2004. The overall selection criterion for participants was they must be at least the age of 18 years or above without any previous history of any cancer diagnosis before the start date of the study.

4.3.2.3 Case definition for Basal cell carcinoma

The case definition for BCC at follow-up was based on our previously validated code list for THIN.[107,145] Cases were excluded if they have history (or any recurrent form) of NMSCs or malignant melanoma cancer prior to the start date, or during the follow-up period of the cohort study. Patients found with Basal cell nevus syndrome (or Gorlin's syndrome), organoid naevi or other BCC-related genetic disease were also excluded from the study population. Patients that experienced the outcome between the time periods were right-censored at the date of their initial BCC recording in THIN. Participants with a death record within the duration of the study were censored right, at the recorded death date. Those with no failures throughout the time course of the study were automatically right-censored on the 31st December 2011.

4.3.2.4 Exposure and confounding variables

Soil concentration levels for arsenic were classified in four major groups using the UK C4SL 35 mg/kg,[71] where soil levels up to half the C4SL (18.0 mg/kg) were classified as the lowest exposure (group I, referent exposure); soil levels between half & up to the C4SL (18.0-34.9 mg/kg) were group II; soils levels that were 1-2 times the C4SL (35.0-69.9 mg/kg) were group III, and concentrations that were 2 times (and beyond) the C4SL (≥ 70 mg/kg) were classed as the highest exposure (group IV).

Potential confounding variables included the age (at baseline), gender, socioeconomic deprivation (i.e. Townsend index categorised as quintiles) status of the participants and lifetime sunlight exposure. The monthly data on the total duration of bright sunlight hours spanning from the 1890s to 2013 was obtained from 35 UK Meteorological Office (Met Office) climate stations that were based in several locations across England and Wales.[167] Sunlight data from climate stations that was specific to SHA and Wales were aggregated to derive monthly daylight averages at a English SHA, and Welsh-level. These monthly daylight averages were used as a proxy to quantify the cumulative sunlight exposure for each participant from their date of birth, and up to the baseline (or start date) of the cohort study. Soil concentration of iron and phosphorus (in mg/kg) was included as environmental confounders, as these were elements are known to

influence the geochemical distribution of arsenic in soil, and it's presence edible plants.

4.3.2.5 Statistical analysis

Univariable Cox regression analysis was used to compute the crude association between BCC and increasing levels of soil arsenic, while a multivariable Cox regression model was used to allow for confounding variables. A stratified analysis was conducted to assess the impact by type of residential setting (urban, suburban or rural) to determine whether associations between arsenic exposure and BCC were significantly modified.

The results were presented as Hazard ratios (HR) with 95% confidence intervals (CI), where statistical significance was determined by p-values of 0.05 or less. P-value for linear trend was derived to assess whether patterns of dose relationships were significant. The chosen method for testing the proportional hazards assumption was the Schoenfeld's residual, which provided p-values for the global model and exposure categories vs. referent group (i.e. for any variable) to assess whether there was non-proportionality in the overall model fit and exposure groups, respectively.

In the case where a variable (or category) is in violation of the proportional hazards assumption, we used Aalen plots as a visual diagnostic tool to identify the time points where the hazard function changed significantly so as to impose cut points on those changes in

order to create interval specific variables (or categories), and refit them into the regression model. The advantage of using this procedure eliminates the time-vary artefact present in the data.[168] All statistical analyses were carried out using Stata/MP 12.0 (Stata Corporation, College Station, Texas, US).

4.3.3 Results

4.3.3.1 Descriptive results of study population

The study was based on a prospective UK cohort of 1,812,372 persons in the English and Welsh areas. The total follow-up time in our study population between 2004 and 2011 was 13.7 million years. Among the participants, 28,783 developed BCC with an average duration of follow-up time of 3.97 years (SE: 0.013, median: 3.92 years, IQR: 1.97–5.94). Demographic characteristics of the cohort population are presented in (Table 4.5). The proportion of men was greater among those with BCC than for those without. Participants developing BCC were more likely to be in the older age groups and from the least deprived groups (Group I: 35.2%; Group II: 25.6%).

The overall median was 16.0 mg/kg (IQR: 11.6-20.0 mg/kg) and the maximum observed exposure amongst participants was 1039.2 mg/kg. Approximately 96% of the study population have soil arsenic concentration below the C4SL value (35.0 mg/kg). Only 1.0% of the population have soil arsenic concentrations exceeding 70.0 mg/kg. The median residential soil arsenic concentration was 14.9 mg/kg

among BCC cases and 15.6 mg/kg among those without BCC. Visual comparisons of outcome do not show any stark difference (Figure 4.3).

Table 4.5: Baseline demographic characteristics of the study population, using The Health Improvement Network (THIN) database from 2004 to 2011

Characteristics	Without BCC (N_1 = 1,783,589)		With BCC (N_2 = 28,783)		Total (N =1,812,372)	
	n_1	%	n_2	%	N	%
Sex						
Male	878,768	49.3	15,058	52.3	893,826	49.3
Female	904,821	50.7	13,725	47.7	918,546	50.7
Age group						
18-29	288,533	16.2	134	0.5	288,667	15.9
30-39	352,712	19.8	899	3.1	353,611	19.5
40-49	347,099	19.5	2,409	8.4	349,508	19.3
50-59	310,579	17.4	5,302	18.4	315,881	17.4
60-69	224,946	12.6	7,586	26.4	232,532	12.8
70-79	159,263	8.9	7,975	27.7	167,238	9.2
80+	100,457	5.6	4,478	15.6	104,935	5.8
Socioeconomic deprivation						
1st (least deprived)	502,316	28.2	10,143	35.2	512,459	28.3
2nd	393,881	22.1	7,363	25.6	401,244	22.1
3rd	361,753	20.3	5,354	18.6	367,107	20.3
4th	310,608	17.4	3,749	13.0	314,357	17.3
5th (most deprived)	200,396	11.2	1,971	6.8	202,367	11.2
Unknown	14,635	0.8	203	0.7	14,838	0.8

Residence classification						
Urban	1,422,102	79.7	21,723	75.5	1,443,825	79.7
Suburban	210,428	11.8	4,232	14.7	214,660	11.8
Rural	137,184	7.7	2,660	9.2	139,844	7.7
Unknown	13,875	0.8	168	0.6	14,043	0.8
SHA						
London	228,984	12.8	2,577	9.0	231,561	12.8
East Midlands	57,273	3.2	915	3.2	58,188	3.2
East England	134,315	7.5	2,107	7.3	136,422	7.5
West Midlands	205,640	11.5	2,885	10.0	208,525	11.5
North East	59,602	3.3	1,017	3.5	60,619	3.3
North West	219,276	12.3	3,254	11.3	222,530	12.3
Yorkshire Humber	40,558	2.3	592	2.1	41,150	2.3
South Central	284,786	16.0	5,212	18.1	289,998	16.0
South East	225,137	12.6	4,004	13.9	229,141	12.6
South West	195,500	11.0	3,811	13.2	199,311	11.0
Wales	132,518	7.4	2,409	8.4	134,927	7.4

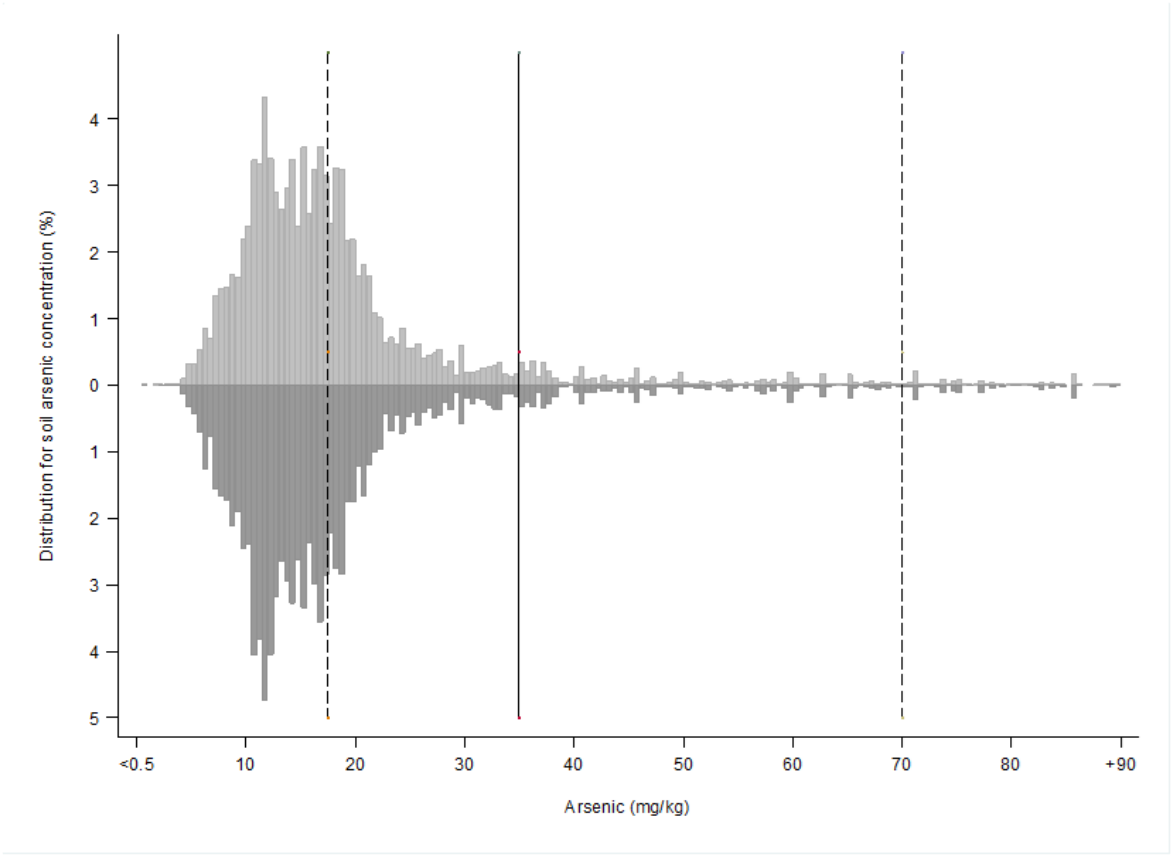

Figure 4.3: Joint histograms plotted on the same axes to show the observed distribution (proportion) of participants registered to practices contributing to THIN their soil arsenic concentration levels. Upper-light grey histogram corresponds to BCC patients; lower-dark grey histogram corresponds to controls. The lower and upper dashed bars were used to mark off soil arsenic concentrations levels at points 18.0 mg/kg and 70.0 mg/kg, respectively. The solid bar corresponds to the current UK C4SL for soil arsenic (35.0 mg/kg)

4.3.3.2 Multivariable Cox regression analysis

Results from unadjusted analyses found evidence of a J-shaped relationship where the lowest risk of developing a BCC was in participants living in areas with arsenic concentrations between 18.0 and 35.0 mg/kg (HR 0.82, 95% CI: 0.81–0.85). The hazard for developing BCC was similar between those living in areas with arsenic concentrations between 35.0 and 70.0 mg/kg and the reference group (< 18.0 mg/kg). However, we found that those living on residential soils with exposure levels above 70.0 mg/kg had a significant increased risk of developing BCC (HR 1.08, 95% CI: 1.01–1.17) (Table 4.6).

After including adjustments for confounding, a clearer pattern emerged (p-value for trend < 0.01) where participants with residential exposures between 35.0 and 70.0 mg/kg and greater than 70.0 mg/kg had a 9% and 20% significant increased hazards of developing BCC, respectively (35–70 mg/kg: HR 1.09, 95% CI: 1.04–1.16; > 70.0 mg/kg: HR 1.20, 95% CI: 1.09–1.32), compared with those living in areas with arsenic concentrations < 18.0 mg/kg (Table 4.6).

4.3.3.3 Stratified analysis based on residential settings

The pattern of BCC risk in relation to the four soil arsenic exposure categories differed substantially across urban, suburban and rural residents (Figure 4.4). Urban residents with the highest exposure of arsenic (≥ 70 mg/kg) were the only group to have a significant

increase in risk of BCC (HR 1.18, 95% CI: 1.06–1.36). The pattern of association amongst urban residents indicates a significant positive trend across urban exposure groups ($p < 0.001$). Even though there appears to be positive dose-response relationship between increasing levels of arsenic exposure and the risk of BCC in suburban residents, our trends test indicates that such linear patterns were not significant ($p = 0.06$), and individual comparisons for exposures in suburban areas were not significant at the 5% level. There was no evidence of an effect of arsenic concentrations on the development of BCC for rural residents.

Table 4.6: Using Cox regression model to estimate hazard ratios (HR) for BCC in association with potential exposure to soil arsenic, using THIN linked G-BASE database from 2004 to 2011

Cox regression model	Unadjusted[1]		Adjusted[2]	
	Hazard ratio	95% CI	Hazard ratio	95% CI
Arsenic (*As*) (mg/kg)				
As < 18.0 mg kg^{-1}	1.00	referent	1.00	referent
18.0 ≤ *As* < 35.0 mg/kg	0.82	(0.81–0.85)	0.96	(0.96–1.00)
35.0 ≤ *As* < 70.0 mg/kg	0.99	(0.95–1.04)	1.08	(1.02–1.14)
As ≥ 70.0 mg/kg	1.08	(1.00–1.17)	1.17	(1.09–1.28)
p-value for trend[3]		-		p < 0.001

[1] Proportion hazards (PH) assumption test was based on Schoenfeld residuals – using the overall global test, we find no evidence that our specification for arsenic violates the PH assumption (p=0.87). A category-by-category test shows no violation of the PH assumption: category 2 vs. category 1 (p=0.93), category 3 vs. category 1 (p=0.43) and category 4 vs. category 1 (p=0.89)

[2] Adjusted for age at baseline, gender, lifetime sunlight exposure (hours) prior to start date, socioeconomic deprivation (quintiles of Townsend index), soil concentration levels of iron (mg/kg) and phosphorus (mg/kg)

[3] Orthogonal polynomial contrasts were fitted to test for significant linear trends across exposure categories. Parameter estimates for the unadjusted model shows both a significant linear and non-linear trend indicating a more complex pattern.

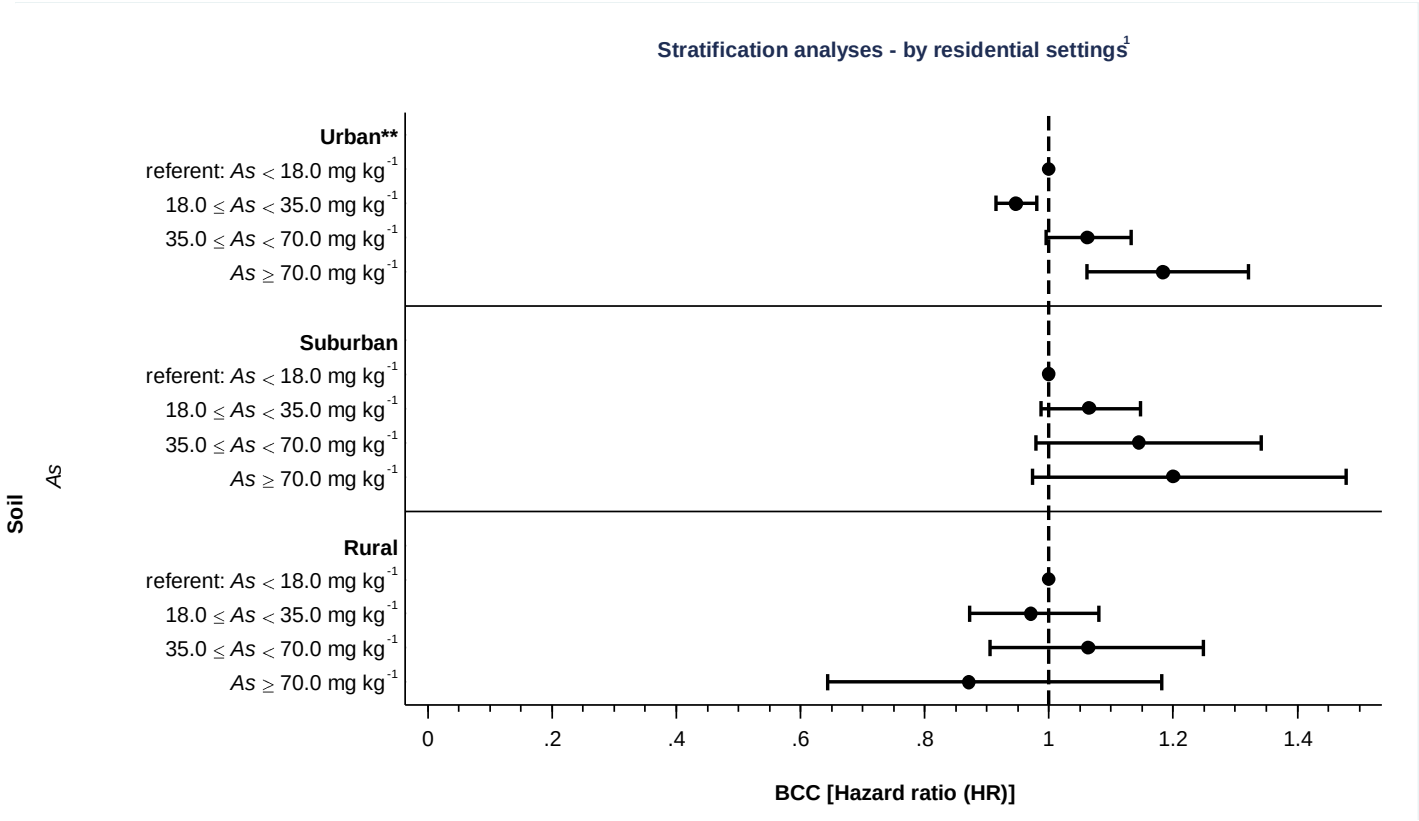

Figure 4.4: Modified scatter plot with range capped spikes were used to represent HR estimates derived from a stratified Cox regression model based on residential classification to show the association between BCC risk and soil arsenic. Models were adjusted for sex, age group, cumulative sunlight exposure, socioeconomic deprivation, soil concentrations for iron and phosphorus.

4.3.4 **Discussion**

This is the first study to use a UK-based population cohort to assess the effect of soil arsenic levels and the development of BCC. Our findings show evidence of an increased risk for BCC in association with potential exposure to soil arsenic. We found that residents living in areas with soil arsenic concentration above the current UKs 2014 provisional category for screening levels (C4SL) (*As* ≥ 35.0 mg/kg) have approximately 9.0% (35.0–70.0 mg/kg) and 20.0% (> 70.0 mg/kg) increased risk of BCC. We also found that elevated risk appeared to be confined to residents with the highest exposure in urban settings.

By adopting a population-based retrospective cohort design and using the THIN database, we believe that we minimised the potential for selection bias. We extracted incident records of BCC that were recorded in THIN by the time this study was being conducted. Selection bias is not a major concern as we attempted to use the entire cohort of eligible patients in the THIN-GBASE, we have a clearly defined inclusion and exclusion criteria that limits our sample of interests to adults (aged 18 years and above) registered at a GP practice at least 1-year before start date of study (i.e. January 1st, 2004), who have measurements for soil arsenic. However, a source of selection bias can arise from the way we had chosen the study periods, since we are only observing incident cases of BCC that have occurred after the beginning of 2004. We must acknowledge, the study periods in which the participants were observed retrospectively

was short (2004-2011). We chose this date because of its significance, which marks the introduction of the Quality and Outcomes Framework (QOF) which encourages GPs to record all new cases of clinical outcomes (including cancers and non-melanoma) when they are examined.

In terms of completeness, the BCC records used for this study are from a reliable source. We know from previous studies that all BCC patients were referred to hospitals or dermatology clinics, with 93.0% diagnosis of BCCs being confirmed either by letters from hospitals or histological (or pathology) reports.[107] Although BCC records are not defined according to the different subtypes of BCC (superficial, nodular or infiltrative); however, validation studies for THIN have confirmed that recordings (as well as the medical read codes) for BCC in THIN are complete and sufficiently accurate for medical research.[105,107] The only issue that was beyond our ability to account for was the differences in terms of BCC ascertainment rates at a GP-level across England and Wales, which may introduce some bias in the effects detected for arsenic, if BCC ascertainment rates at a practice-level are associated with geographical variation in soil arsenic exposure.[169] Although, we attempted to control for this with the inclusion of the Strategic Health Authorities' variable; however, a proficient adjustment would have been such of the inclusion of a performance indicator at a practice-level measured as a proxy for denoting how well practices tend to record clinical outcomes; however, this information is seldom available in THIN.

The linkage between THIN and G-BASE has been validated to ensure that the soil arsenic concentration levels for the observed distribution of patients in THIN are representative of the general UK population.[119] By modelling the hazards of soil arsenic using a multivariable, and stratified Cox regression, we were able to take into account the implicit time factor that exists for participants in terms of their potential exposure to soil arsenic from the beginning of the study period until BCC development. Additionally, we attempted to reduce the effects of confounding by adjusting for meaningful covariates such as the inclusion of sunlight data as a proxy for UV-exposure. Due to the paucity of appropriate UV-exposure indicators in THIN, we therefore obtained area-level monthly sunshine data (in hours) from the UK meteorological office[167] to quantify a person's cumulative sunlight exposure. While, it is correct to adjust for sunlight; however, we must acknowledge that these estimates are not as robust as they do not account for: .1) the actual time spent outside in the sun by an individual, .2) the surface area of a body exposed to sunlight, and .3) individual behaviour (e.g. holiday habits etc.).[130,149,152,170] The estimates quantified were based on area-level measurements, and therefore, this may introduce ecological bias in our risk assessment.

The major limitation to this study was our inability to use in our analysis the exact location of a patient where their soil concentration samples for arsenic were determined, due to ethical implications. A recent study has shown using G–BASE data, spatially, the locations (*As*-domains) for where arsenic are naturally mineralised, and where they

are highly concentrated due to the presence of ironstone and phosphorus – these *As*–domains are specifically located in major areas of North West, South West and parts of central regions of England, and southern Wales.[161,163] The background concentrations for topsoil in those environments exceed the 2014 C4SL for soil arsenic.[71,164] It could be possible that the observed risk found in this study might be related to exposures within those specified *As*–domains; however, it is difficult to determine whether this assertion is true because we were unable to cluster participants according to those *As*–domains. To verify this claim, a new cohort study would be needed as currently the data is not available in our linked database. Our inability to account for important factors such as occupational settings;[33,156] household measurements of arsenic from drinking water;[87,160] and other biomarkers for long-term exposures (finger nails, toenails and hair),[27,31] as well as not being able to adjust for the amounts of arsenic that can be potentially ingested via the digestive tract (which can be calculated with an exposure assessment model) may give rise to residual confounding in our analysis.

From a spatial epidemiologic point of view, another major limitation was our inability to utilise the following data in our analysis: 1.) Information related to a patient's addresses; 2.) the location of general practices he/she attended; and 3.) the georeferenced sampling points for the soil samples that were collected at different densities. We mentioned earlier that this was a limitation due ethical and legal implications outlined by THIN. The geospatial details of

patients were anonymised, and therefore, we could not utilise this data to ascertain the distribution of participants that fall within, or on to a sampling point; let alone, determine the distribution of those that lived on soils classified as either G-BASE urban, G-BASE rural, or NSI(XRFS) areas. If this information were available, we could have stratified the population at risk of BCC in accordance to these three zones. We attempted to rectify this issue of differentiating participants that lived on urban and rural terrain (i.e. G-BASE urban, rural and NSI(XRFS)) by using a stratified Cox regression analyses based on the type of residential setting indicators recorded in THIN, to minimise the possibility of information bias that may affect our risk estimates for BCC because of the systematic differences in potential exposure to arsenic due to the samples being collected at densities.

Our study showed those in the higher exposure groups (35.0-70.0 mg/kg and above 70.0 mg/kg) significantly had an increased risk of BCC. Similar findings were established in other studies for skin cancer (i.e. melanoma and NMSCs) although the risks were quantified using different sources or biomarkers for exposure, as well as different cell-types for skin cancer.[30,158,160] For instance, previous findings in Hungary, Romania and Slovakia found a significant association for BCC and exposure to arsenic through drinking water with concentration at modest levels.[160] The most likely explanation for our overall findings could be attributed to the fact that elevated levels of arsenic in the soil tends to positively correlate with other environments (groundwater, drinking–water and air quality) giving rise to multiple

pathways for exposure. The long-term exposure (directly or indirectly) to such environmental concentrations could possibly ensure continued absorption of arsenic into the body, which will eventually cause to some degree genotoxic effects leading to BCC development.

The results from our subgroup analysis defined by residential classification suggested that the elevated risks of BCC in relation to higher soil arsenic exposures were confined to urban residents. An explanation for this may be related to the relative mobility of arsenic in soils from different environments. Recent studies have shown that urban soils had higher bioaccessibility than rural soils (the fraction of arsenic released from the soil into solution in the gastro-intestinal tract in a form that can potentially be absorbed into the bloodstream).[21] The bioaccessible fraction in urban domains was 19-28% compared to 5-9% in rural domains which are dominated by naturally occurring arsenic. Another explanation could be that urban environments have higher arsenic concentrations than rural environments due to anthropogenic contamination. Another study, however, has clearly shown that naturally occurring arsenic derived from geogenic mineralisation and from underlying ironstone formations found in rural locations were typically an order of magnitude higher than those found in the urban domain.[161]

The most significant relationship with BCC risk was found in urban residents with arsenic exposures of 70 mg/kg and above. Our findings indicate that potential exposures are may be important risk factors,

and must be taken to under consideration in the investigation of cancer aetiology. These results warrant further investigations, but nevertheless, it is important to minimise human exposure to arsenic, especially for those emerging from lithological sources such as topsoil.

Respiratory tract cancer

Chapter 5

5 Summary

Lung cancer is one of the deadliest forms of malignancies in the UK. It is the third most common type of cancer in the UK. Lung cancer has a plethora of risk factors – an individual's risk of developing the malignancy depends mostly on the advancement of age, genetics and family history of cancer. However, avoidable lifestyle factors such as smoking, certain occupational exposures and ionising radiation are the biggest causes of lung cancer in the UK. Metallic elements such as arsenic, chromium and aluminium, as well as radioactive elements such uranium and radon have been linked to lung cancer. These are typical examples of elements that are naturally occurring in soils and remain ubiquitous in our environment; however, little is known of the health impacts of such soil exposure to most of these elements on lung cancer incidence in the UK.

The goal of this chapter is address this gap in knowledge by using a two-stage process to conduct the following: (1) using data mining to determine which group of soil metals are the best predictors of lung cancer, and (2) using a population-based cohort design for quantifying the effect size for lung cancer risk for individuals living in a residential area with high concentration levels for the selected soil elements.

All patients with a first diagnosis of lung cancer between 1-January-2004 and 31-December-2014, including their X-ray fluorescence spectroscopic measurements for total concentration levels of all soil

elements in THIN-GBASE were extracted. The correlation-based filter selection was the chosen data mining method to determine which restricted group of soil elements were highly corrected with lung cancer. Two multivariable Cox regression models were used to determine the association between the subset of selected elements and lung cancer. Model one, termed as mutually adjusted model, consists of only subset of elements found from the data mining analysis. Model two is the corrected version which includes confounding variables such as age, gender, smoking status and socioeconomic deprivation. Additionally, a stratified analysis was carried out to assess the impact by type of residential setting (urban, suburban and rural).

1,823,312 participants with complete soil concentration estimates for all 15 elements were extracted for THIN-GBASE database. Of these 10,740 (0.6%) participants developed lung cancer (median survival time 4.8 years, IQR: 2.4 – 7.3 years). The correlation-based filter selection data mining identified aluminium [median 51,400 mg/kg, IQR: 40,500-59,300mg/kg (max = 116,700 mg/kg)], lead [median 72.0 mg/kg, IQR: 47.0-158.0 mg/kg (max = 3,045 mg/kg)] and uranium [median 1.98 mg/kg, IQR: 1.58-2.40 mg/kg (max = 61.7 mg/kg)] as the most appropriate subset of soil elements to be modelled as risk factors for lung cancer with an error rate of selection = 1.47%.

The mutually adjusted model has shown that the risk of developing lung cancer was significantly among individuals living in areas with

elevated soil concentrations for aluminium (47,200-54,700 mg/kg: HR 1.21, 95% CI: 1.13-1.29; 54,700-61,600 mg/kg: HR 1.19, 95% CI: 1.11-1.28; and ≥ 61,600 mg/kg: HR 1.18, 95% CI: 1.09-1.27), lead (60.0-95.0: HR 1.12, 95% CI: 1.06-1.20; 95.0-184.0 mg/kg: HR 1.11, 95% CI: 1.04-1.18) and uranium (≥ 2.50 mg/kg: HR 1.13, 95% CI: 1.05-1.21). For the corrected model, the patterns of association were maintained for medium exposure groups for aluminium (47,200-54,700 mg/kg: HR 1.09, 95% CI: 1.02-1.17; and 54,700-61,600 mg/kg: HR 1.10, 95% CI: 1.02-1.18) while there were marginal reductions in risk for lead (60.0-95.0: HR 1.12, 95% CI: 1.06-1.19; 95.0-184.0 mg/kg: HR 1.08, 95% CI: 1.01-1.15). However, the observed risks (i.e. initial model) were abrogated for those living in areas with highest concentrations of aluminium (≥ 61,600 mg/kg: HR 1.06, 95% CI: 0.98-1.15) and uranium (≥ 2.50 mg/kg: HR 1.05, 95% CI: 0.97-1.13). The risk of lung cancer was only isolated to urban residents on residential soil with aluminium concentrations above 47,200 mg/kg.

***Conclusion*:** There was not enough evidence to conclude that soil uranium and lead were associated with lung cancer due to the ambiguous patterns found in both mutually adjusted and corrected model. However, this study suggests that residents living on soil with aluminium greater than 47,200 mg/kg have a higher risk of developing lung cancer, and such risks are confined to urban residential areas. Further research is required to firmly establish this apparent association.

5.1 Background

Lung cancer in humans refers to a group of malignant neoplasms that develop within the air passages (trachea, bronchi or bronchial branches) of the respiratory tract and internal regions of the right lung (upper, middle & lower lobe) or left lung (upper or middle lobe, and lingula).[171–173] Lung cancer is classified according to the type of cells in which the tumour originates: small cell lung carcinoma (SCLC) and non-small cell lung carcinoma (NSCLC). SCLCs are miniature tumours with growth patterns of a sporadic nature, which can potentially to spread to other vital organs giving rise to secondary tumours. NSCLCs usually grow into a single large and dense tumour which can potentially create blockages limiting gaseous exchange. The common types of NSCLCs are squamous cell carcinoma (SCC), adenocarcinoma and large cell carcinoma (LCC).[171–173]

Lung cancer is a global health problem; where approximately 1.82 million lung cancers were diagnosed globally in 2012.[174–176] Additionally, lung cancers are the most prevalent type of solid cancer (12.9% of all solid cancers), possibly due to it being caused by several risk factors which work together in a multifactorial manner. Globally, lung cancer has higher mortality rates than any other solid cancer equating approximately 1.6 million per year in 2012; the high mortality rate is due to it being difficult to treat, and the majority of patients are often diagnosed late.[174–176]

The innate aetiological factors which increase the risk of developing lung cancer include the advancement of age, heredity and biological factors at cellular levels typically influenced by the functional decline of the lungs.[171,174,176] However, the biggest risk factor for lung cancer is lifestyle-related - the most significant is the inhalation of tobacco smoke (chiefly from cigarettes, but also pipes and cigars).[171,174,176] The most established environmental factor is attributed to long-term exposure to high levels of environmental radon that have emerged for soil and diffused to indoor environments[177–180] - the World Health Organisation (WHO) have labelled radon as the second most important cause of lung cancer after smoking.[181]

In the context of environmental exposure to soil metallic elements – there are limited studies directly examining potential association between metallic elements (including arsenic, chromium, nickel, uranium, lead and zinc) that emerge from soil and lung cancer, particularly in relation to exposure of low-dose or low-level soil metallic elements. The purpose of this chapter is to establish a plausible explanation for the attribution of soil metals and how they may cause lung cancer, before carrying out a two-stage approach using data mining and a population-based cohort study to explore their association.

5.2 Soil elements and lung cancer incidence in the UK

The most important and likely exposure route to consider is the inhalation of soil particulates through the respiratory system. The vast

majority of particulates that are airborne are composed of elemental constituents, and while there are various sources, the significant amount of particulates that are ambient originate from soil through diffusion, wind erosion and soil disturbances due to anthropogenic processes.[4] Pollution-based studies have quantified the mass contributions of various sources that give rise to airborne particulates and have analysed their constituents,[182–184] while a few epidemiological studies have identified associations between lung cancer and concentrations of inhalable particulate matter. The former had shown that soils were the largest contributor to most particulates that were ambient compared to other mediums (traffic, motor vehicles and industries). The latter have established that elevated concentration levels of ambient particulates (with their elemental constituents) have a modest association with increased incidence of lung cancer.[185,186]

Thus, this establishes a plausible mechanistic framework for linking soils with lung cancer, by showing how soils, air particulates and cancer are states related to one another (Figure 5.1).

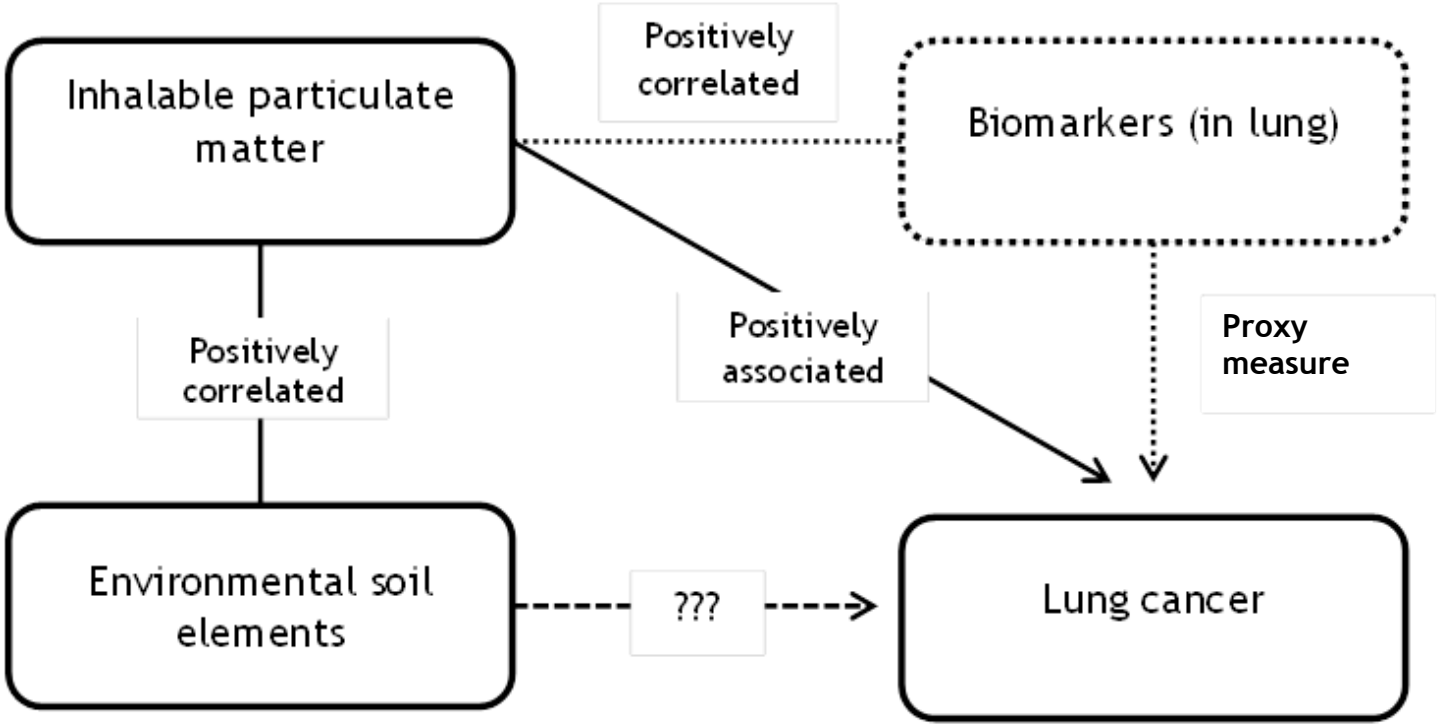

Figure 5.1: Diagram depicting a plausible mechanistic framework for causal pathway for lung cancer in relation soil elements

Lung cancer in the UK is currently a pressing public health issue. It is the second most common type of malignancy diagnosed in the primary health care followed after breast cancer.[187] A recent study using the THIN database had indicated that the overall incidence of lung cancer was estimated as 41.4 per 100,000 person-years.[188] It has suggested that the incidence of lung cancer in the UK will increase by 4.0% for every three-year periods.[188] According to Cancer Research UK 89.0% of lung cancer cases occurring in the UK are strongly related to major modifiable lifestyle factors – mainly smoking – while they are convinced that 11.0% of the remaining cases are caused by environmental and occupational factors such as exposure to arsenic (and inorganic arsenic compounds); production of aluminium, iron and steel; chromium (VI) compounds; and outdoor air pollution (& particulate matter).[187] In the context of soil metal exposure in relation to lung cancer - the literature is limited.

Currently, there have been two studies that strictly considered examining the associations between low-level exposure to soil metals and risk of lung cancer. The first study was conducted in Taiwan which found that elevated chromium concentrations in soil were associated with increased incidence of SCC;[157] while nickel and zinc were found to be positively associated with both adenocarcinoma and SCC.[157] The other study was performed in Northern Ireland, which has shown that elevated levels for the total arsenic content in soil has a positive correlation with lung cancer incidence – estimated at a ward-level.[189] However, what these two studies have in common is that

their study design utilises an ecological framework. By virtue of the fact that the soil exposure in these studies were aggregated at a ward-level (or geographical unit) makes them prone to ecological fallacy. Thus, the individual-level association and the relative contribution of concentration levels of soil elements on lung cancer remain unclear.

Therefore, the purpose of this chapter is to address this research gap. By using the THIN-GBASE database, we used series of epidemiological analysis using a two-stage approach to assess the impact of soil metallic elements on lung cancer risk.

5.3 Methods

5.3.1 Study design

A 2-stage process was used to determine which of the 15 soil elements within the linked database were associated with lung cancer. The initial process involved the usage of data mining techniques as an exploratory exercise to find which of the element(s) (or best subset of elements) were predictive of lung cancer. The next stage used a population-based cohort design for the purpose of quantifying the effect size for lung cancer risk for individuals living in a residential area with high concentration levels for the selected soil elements.

5.3.2 Study population

5.3.2.1 Case definition for lung cancer

The case definition for lung cancer patients was any form of malignant neoplasm found in the internal sites of the respiratory tract (i.e. trachea, bronchi and lung). In THIN, lung-related neoplasms are coded mainly as site-specific cancers rather than cell-specific (i.e. SCLCs or NSCLCs). Clinical experts in respiratory medicine at the University of Nottingham were consulted with to determine which of lung cancer Read Codes were the most appropriate to be used. Patients with malignant neoplasms of the lung were identified using Read Codes under the following hierarchies: malignant neoplasms of trachea, bronchus and lung *B22..00*; carcinoma *in situ* of the respiratory tract *B81..00*; respiratory tract adenomas and adenocarcinomas *BB5S.00*; and any lung-related malignant neoplasm otherwise specified *By…00*, were extracted from the Medical and AHD records of the THIN-GBASE database.

Patients found with any rare lung-related genetic syndrome or disease, neoplasms of the respiratory tract exhibiting uncertain behaviours or mesotheliomas were excluded. Patients coded with lung cancers located on the walls of the thoracic or chest cavity were also excluded from the analysis.

5.3.2.2 Inclusion and exclusion criteria

Eligible participants were those aged 18+ years, which were registered at their GP practice and contributing data for at least one year before the study start date (1-January-2004). Participants were excluded if they had a lung cancer diagnosis before the start date of the study or did not have complete exposure data on the 15 soil elements. Patients were only included if the practice at which they were registered had an acceptable mortality recording (AMR) date before the start of the study. The end date for the study was 31-December-2013.

Participants diagnosed with lung cancer during the course of the study (i.e. between 1-January-2004 and 31-December-2013) were right-censored at the date of their first lung cancer recording. Participants with a death record within the duration of the study were right-censored at the date they were recorded to have died. Participants with no lung cancer event throughout the duration of the study were automatically right-censored on 31-December-2013.

5.3.2.3 Exposure and confounding variables

Exposures for all 15 soil elements were classified into four groups using the CS4L, where soil levels up to half the CS4L were classified as the lowest exposure (Group I, referent group); soil levels between half & up to the C4SL were group II, soil levels with 1-2 times the C4SL were group III, and soil levels 2 times above the C4SL were the highest exposure (group IV). Where a soil guideline value was not available, or

if the guideline value was out of range of our data (i.e. not found between lowest and highest observation for specific element), or using the above method produced a category containing no cases; we used quintiles to categorise the exposure.

Age, gender, smoking status, and socioeconomic deprivation were considered as confounding variables, due to their association with both the outcome and potential exposures of interest. Age was categorised as below 40, 41-50, 51-60, 61-70, 71-80 and 81+ years. Smoking status before the start of the study was extracted from the database using validated Read Codes,[190] and categorised as never smoked, non-smoker, ever-smoked or unknown. Quintiles of Townsend indices of deprivation were used to measure socioeconomic deprivation. We also extracted data on the type of residential settings, which were categorised as urban, suburban and rural.

5.3.3 Statistical analyses

5.3.3.1 Filter method for feature selection (Stage 1)

We have applied feature selection data mining techniques to our database to generate new hypothesis that may aid in determining the relationship between soil elements from GBASE and clinical outcomes in THIN. Feature selection is a very useful data mining tool which has a suite of wrapper, filter and embedded methods for searching potential exposures used for optimising risk predictions of certain outcome variables in a large database.[191,192] The filter methods are

especially useful in determining which exposure, or subset of exposures that are relevant for building a predictive model. When applying filter methods to a large database, they act as filters for identifying relevant exposures needed for building and optimising risk models, whilst, at the same time removing any exposures that are redundant and do not contribute to the accuracy of the predictive model.[191,192] This technique is certainly helpful in building our own risk models because there is paucity in literature that establish any direct relationships between specific soil elements and lung cancer. Contemporary studies that have shown associations between soil elements and lung cancer have conflicting results,[157,189] and so it would be inappropriate to rely on their results to build our predictive models for risk of lung cancer. We therefore relied on these filter-based methods to select the relevant soil elements. The technique optimises risk prediction of lung cancer based on the selected group of soil element and thus do not guarantee any statistical significance thereby limiting the potential of a type-I error occur in our results.

The correlation-based filter selection (CFS) method was chosen as the most appropriate data mining method to use for this study because it generates a restricted group of potential soil exposures which were highly correlated with the outcome. The advantages are it can analyse continuous or categorical variables, and during the selection process it takes into account the collinearity between the subgroup of chosen elements and other elements that are least correlated to the outcome. [191,192,193] The default search algorithm used for CFS was the

forward (or backwards) greedy stepwise technique whereby exposures were progressively incorporated in a successive order of importance to form larger and larger subsets. The search process ends once the generated subset of exposures was reached, which produces a summary statistic known as the merit score. This information shows the overall strength of error in adding attributes into the subset – a summary score of 15% and below is usually preferred. [191,192,193] Note that the CFS algorithm does not guarantee model estimates for exposure categories in selected attributes to be statistically significant; rather, it ensures that the model constructed is fully optimised by returning the best log-likelihood score. All data mining analyses were performed in Weka 3.7.12 (University of Waikato, Hamilton & Tauranga, New Zealand).

5.3.3.2 **Multivariable Cox regression modelling (Stage 2)**

Multivariable Cox regression analysis was used to model the association between the subset of selected soil elements (derived from data mining) and lung cancer, allowing for confounding variables (section 5.3.2.3). Our initial analysis comprised of a mutual adjusted model containing only the subset of elements derived from the data mining in stage 1. Afterwards, a corrected model was fitted which contained both subset of elements from stage 1, and potential confounding variables. A stratified analysis was carried out to assess the impact by type of residential setting (urban, suburban and rural). Results are presented as Hazard Ratios (HR) with 95% confidence

intervals (95% CI), where statistical significance was deemed if 95% CI excluded the null value of 1.

A trend test using orthogonal polynomial contrasts was carried out to assess whether lung cancer hazard increased linearly with higher exposure categories for the subset of elements, whereby a p-value < 0.05 indicates that the increase hazard ratios across increasing exposure categories adequately (or crudely) follows a linear pattern.[194]

We assessed whether the hazards of exposures and confounding variables were proportional using a test based on the Schoenfeld's residuals, which provides p-values to determine non-proportionality in the overall model and exposure group (vs. referent group).[195,196] Where violation of the assumption was identified (p-values less than 0.05) Aalen plots were used as a diagnostic tool to identify the time points where the hazard function changed significantly.[168] Variables were then created on the cut points of where there is an obvious change in the pattern of the hazard function, and re-fitted into the multivariable regression model. This procedure eliminates the time-varying effect that is present in the exposure.

All statistical modelling was conducted in Stata 12.0 MP for Windows 7.0 (Stata Corporation, Station College, Texas, USA).

5.4 Results

5.4.1 Demographic characteristics

A complete cohort of up to 2.3 million patients was identified before 1-January-2004. Of these, 20.7% (n = 144,307) of patients were excluded from the study because they did not meet the inclusion criteria. Those excluded from the analysis were patients with partial or completely missing soil data (69.0%, n = 99,575), those registered at a GP practice with an AMR date after 1-January-2004 (18.2%, n = 26,202) and those with ages below 18 years (8.1%, n = 11,705). The final extract contained a sample of 1,823,312 participants (Figure 5.2).

During the study, 10,740 (0.6%) participants developed lung cancer. The overall median survival time for these participants was 4.8 years (Interquartile range (IQR): 2.4-7.3 years). Participants who developed lung cancer were more likely to be older (51-60 years: 19.5%; 61-70 years: 31.5% and 71-80 years: 31.8%), a male (57.3%) current smoker (48.0%) from the south of England (36.2%), and from the deprived groups (group IV: 21.2%) (Table 5.1).

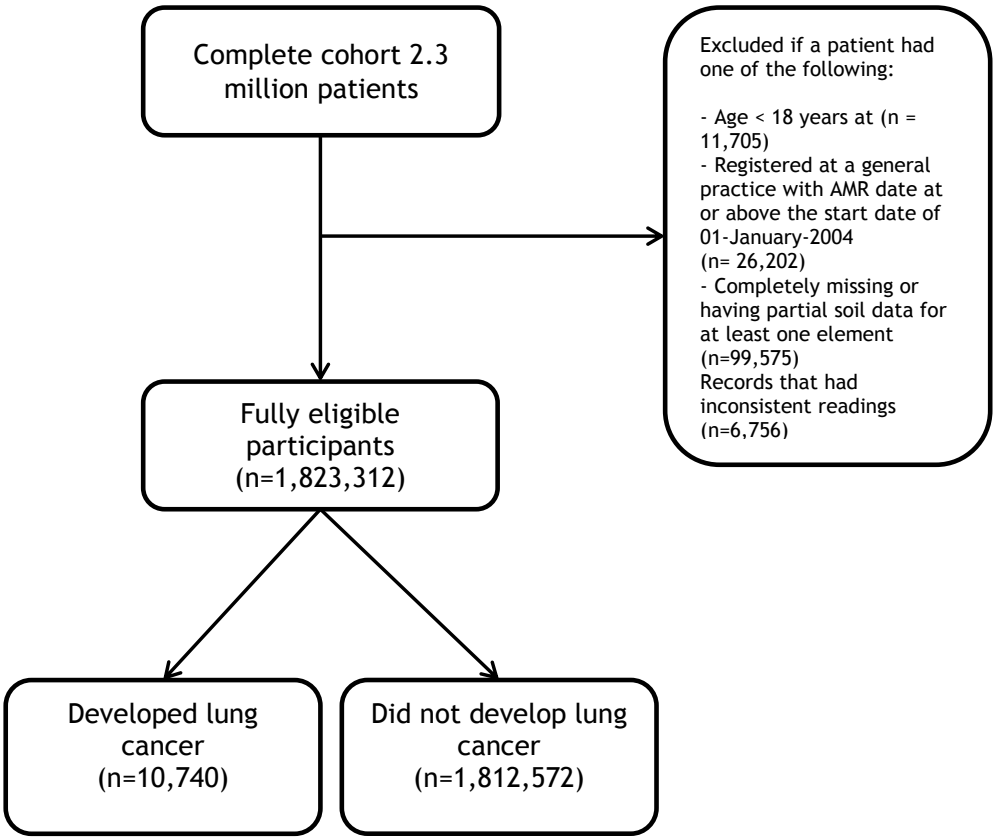

Figure 5.2: Schematic diagram representing how participants were included or excluded from cohort study

Table 5.1: Baseline demographic characteristics of participants for lung cancer study, using THIN-GBASE database from 01-Jan-2004 to 31-Dec-2013

Characteristics	Without lung cancer		With lung cancer		Total	
	N	%	N	%	N	%
Sex						
Male	892,740	49.3	6,155	57.3	898,895	49.3
Female	919,832	50.7	4,585	42.7	924,417	50.7
Age group						
≤ 40	670,191	37	117	1.1	670,308	36.8
41-50	338,083	18.7	566	5.3	338,649	18.6
51-60	312,896	17.3	2,099	19.5	314,995	17.3
61-70	224,962	12.4	3,385	31.5	228,347	12.5
71-80	167,256	9.2	3,420	31.8	170,676	9.4
81+	99,184	5.0	1,153	10.8	100,337	6.0
Smoking status						
Never	568,821	31.4	978	9.1	569,799	31.3
Non	284,919	15.7	773	7.2	285,692	15.7
Ex-smoker	250,403	13.8	3,006	28	253,409	13.9
Current smoker	383,965	21.2	5,102	47.5	389,067	21.3
Unknown	324,464	17.9	881	8.2	325,345	17.8
Socioeconomic deprivation						
Group I	518,606	28.6	2,291	21.3	520,897	28.6
Group II	402,128	22.2	2,157	20.1	404,285	22.2
Group III	365,795	20.2	2,238	20.8	368,033	20.2
Group IV	311,153	17.2	2,282	21.2	313,435	17.2

Group V	200,108	11	1,719	16	201,827	11.1
Unknown	14,782	0.8	53	0.5	14,835	0.8
Residential classification						
Urban	1,438,529	79.4	8,702	81	1,447,231	79.4
Suburban	217,767	12	1,317	12.3	219,084	12
Rural	142,280	7.8	680	6.3	142,960	7.8
Unknown	13,996	0.8	41	0.4	14,037	0.8
Health authority						
London	233,028	12.9	1,099	10.2	234,127	12.8
East Midland	59,182	3.3	379	3.5	59,561	3.3
East of England	138,380	7.6	758	7.1	139,138	7.6
West Midlands	205,309	11.3	1,092	10.2	206,401	11.3
North East	61,494	3.4	528	4.9	62,022	3.4
North West	220,919	12.2	1,713	15.9	222,632	12.2
Yorkshire & Humber	40,326	2.2	298	2.8	40,624	2.2
South Central	294,553	16.3	1,531	14.3	296,084	16.2
South East	228,556	12.6	1,165	10.8	229,721	12.6
South West	198,389	10.9	1,191	11.1	199,580	10.9
Wales	132,436	7.3	986	9.2	133,422	7.3

5.4.2 **Exploratory analysis of geochemical data in THIN-GBASE**

The greedy stepwise search algorithm for CFS found the following metals to be the most appropriate subset for modelling the risk of lung cancer: aluminium, lead and uranium. The soil elements were selected in successive order of importance – the result indicated that aluminium was the most important attribute present in the subset, followed by uranium, and then lead – which was the attribute flagged as being the lowest important variable. The overall merit score for the chosen subset of attributes was 0.0147 (or 1.47%) which signed that the error rate for selecting, and including attributes in the subset for model construction was low. The remaining elements not included were arsenic, calcium, chromium, copper, iron, nickel, manganese, phosphorus, selenium, silicon, vanadium and zinc (Table 5.2).

The overall median concentrations for the selected soil heavy metals were: aluminium 51,400 mg/kg (IQR: 40,500-59,300mg/kg with highest value = 116,700 mg/kg); lead 72.0 mg/kg (IQR: 47.0-158.0 mg/kg with highest value = 3,045 mg/kg); and uranium 1.98 mg/kg (IQR: 1.58-2.40 mg/kg with highest value = 61.7 mg/kg). By visual inspection of the observed distribution for soil concentration levels of aluminium (Figure 5.3), lead (Figure 5.4) and uranium (Figure 5.5) among those with, and without lung cancer – the outputs indicate that the unadjusted association in terms of soil exposure do not substantial differ by outcome.

Table 5.2: Showing the sequence in which soil elements were selected for model construction of lung cancer risk, subsets were generated using the Correlation-based Filter Selection (CFS) method

Sequence	Order of selected attribute	Subset generated	
	Start set:	-	[start]
1	Aluminium	{Aluminium}	
2	Uranium	{Aluminium, Uranium}	
3	Lead	{Aluminium, Uranium, Lead}	[End]
		Overall merit score (i.e. error rate) = 0.0147	
	Excluded attributes:	**Arsenic, Cadmium, Chromium, Copper, Iron, Nickel, Manganese, Selenium, Phosphorus, Silicon, Vanadium and Zinc**	

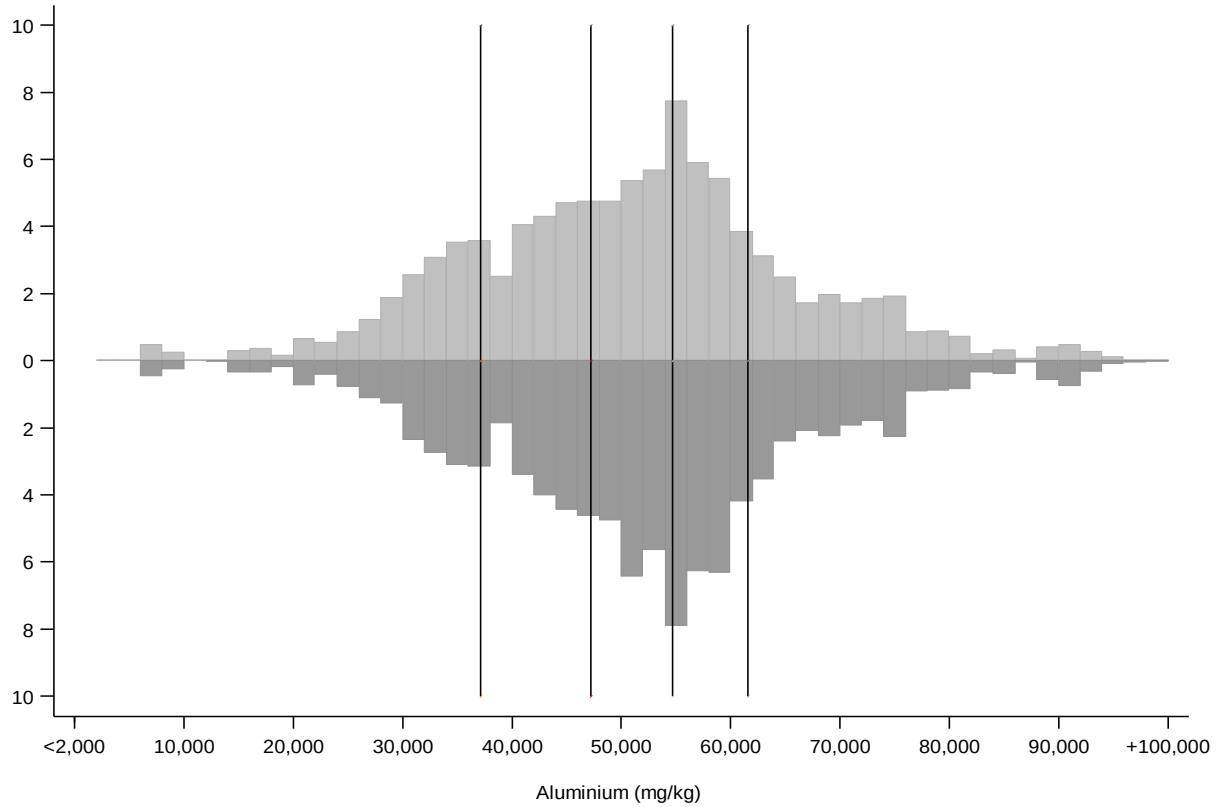

Figure 5.3: Joint histograms plotted on the same axes to show the observed distribution (proportion) of participants registered to practices contributing to THIN-GBASE their soil aluminium concentration levels. Upper-light grey histogram corresponds to lung cancer patients; lower-dark grey histogram corresponds to controls (without lung cancer). Each black vertical line corresponds to soil aluminium value that falls on quintile to create categories: <37,100, 37,100-47,200, 47,200-54,700, 54,700-61,600 and ≥61,600

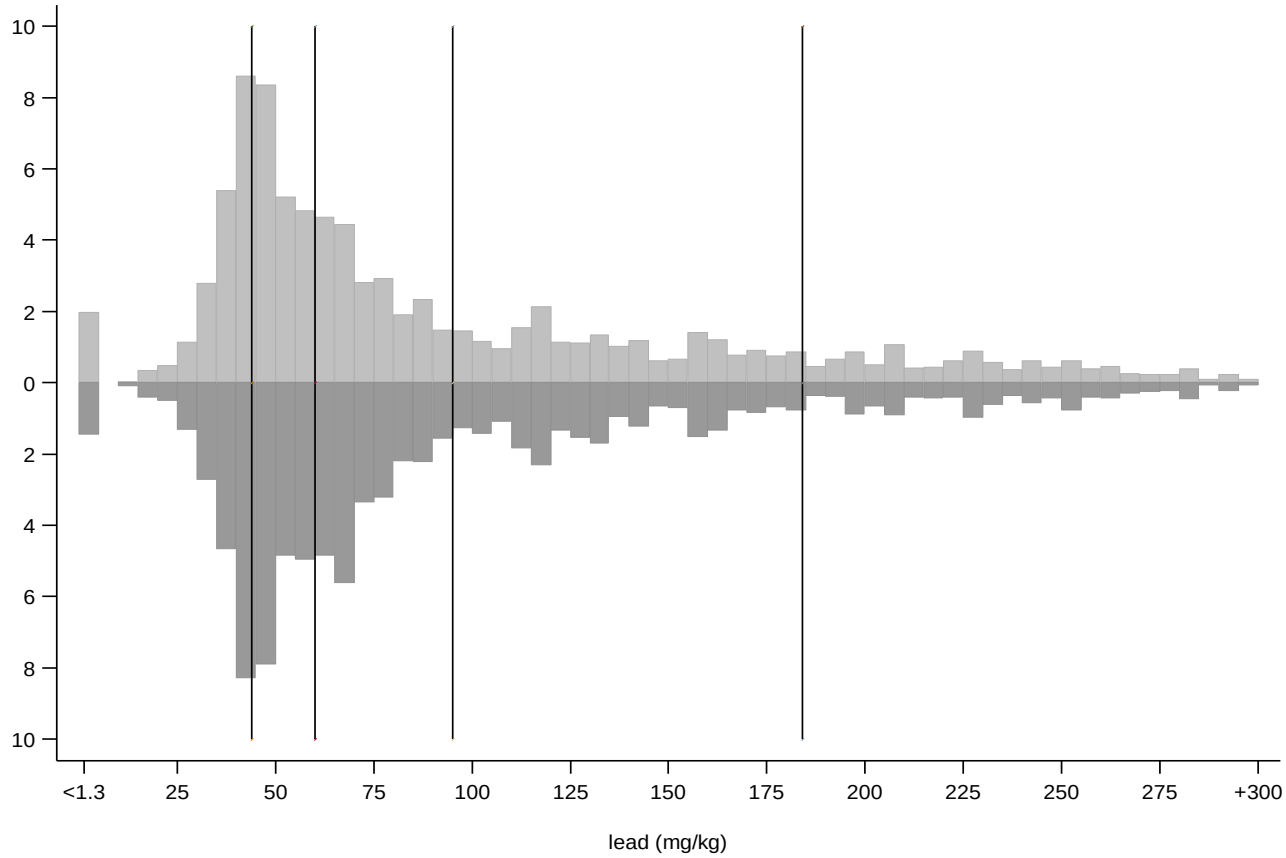

Figure 5.4: Joint histograms plotted on the same axes to show the observed distribution (proportion) of participants registered to practices contributing to THIN-GBASE their soil lead concentration levels. Upper-light grey histogram corresponds to lung cancer patients; lower-dark grey histogram corresponds to controls (without lung cancer). Each black vertical line corresponds to soil lead value that falls on quintile to create categories: 44.0, 44.0-60.0, 60.0-95.0, 95.0-184.0 and ≥184.0

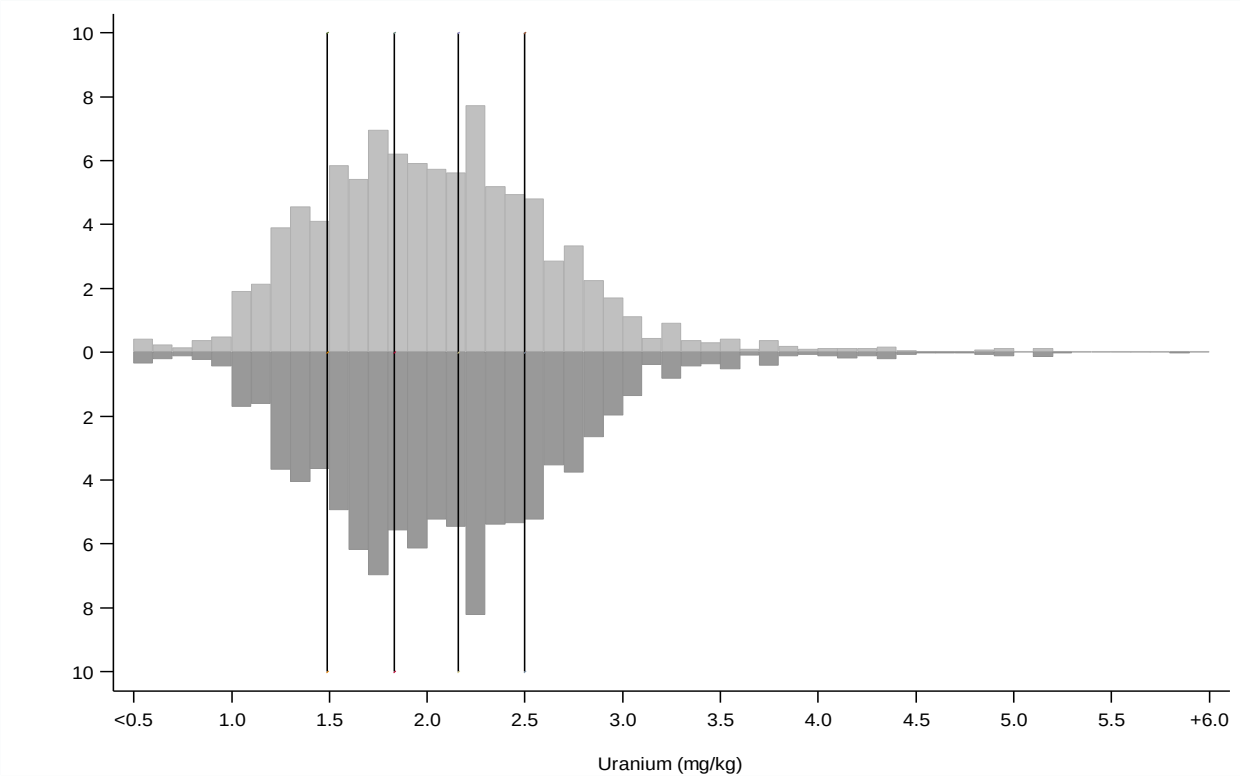

Figure 5.5: Joint histograms plotted on the same axes to show the observed distribution (proportion) of participants registered to practices contributing to THIN-GBASE their soil uranium concentration levels. Upper-light grey histogram corresponds to lung cancer patients; lower-dark grey histogram corresponds to controls (without lung cancer). Each black vertical line corresponds to soil uranium value that falls on quintile to create categories: <1.49, 1.49-1.83, 1.83-2.16, 2.16-2.50 and ≥2.50

5.4.3 Cox regression model

5.4.3.1 Mutually adjusted Cox multivariable regression model

A pre-diagnostic test using the Schoenfeld's residuals was used for checking the proportional-hazards assumption before reporting the any hazard ratios derived for the subset of soil elements. In the mutually adjusted model (without adjustment for confounding factors), our test found no evidence that our model specification for aluminium, lead and uranium violated the proportional-hazards assumption (global test: p-value = 0.60), and p-values for each category were non-significant (Table 5.3).

The mutually adjusted model indicated that the risk of developing lung cancer was confined to individuals who lived in areas with higher soil aluminium concentrations above 47,200 mg/kg (Table 5.4). Our test seems to indicate a significant linear trends relationship for increased exposure of aluminium and risk of lung cancer ($p < 0.001$) (Figure 5.6). However, the patterns of risk for lung cancer in relation to aluminium show a plateau effect which are unclear.

There was an increased risk of lung cancer for participants living in areas with medium soil concentration levels of lead (Group III: HR 1.12 95% CI: 1.06-1.29; Group IV: HR 1.11 95% CI: 1.04-1.18) (Table 5.4). However, the pattern of association for lead remains unclear as there

is an apparent ambiguity in the order of direction for the hazard ratios across exposure categories, as well as statistically non-significant linear trend (p = 0.54).

There was no association with lung cancer for uranium exposure between groups I to IV. However, there was a significant increase in risk for those with the highest uranium exposure (group V: HR 1.13 95% CI: 1.05-1.21) (Table 5.4). Although the result indicated a significant linear trend across the exposure groups (p = 0.01); however, visual inspection showed that the patterns across exposure categories crudely captures a linear trend (Figure 5.6).

Table 5.3: Test of proportional-hazards assumption for mutually adjusted model for assessing risk of lung cancer with the selected group of soil metals using the Schoenfeld's residual test

Soil element	Aluminium[1]	Lead[2]	Uranium[3]
Exposure groups	p-value	p-value	p-value
Group I	-	-	-
Group II	0.61	0.28	0.70
Group III	0.39	0.62	0.25
Group IV	0.62	0.77	0.28
Group V	0.73	0.91	0.88

[1] Soil aluminium [Al] (mg/kg) was categorised as quintiles: group I (Al < 37,100), group II (37,100 ≤ Al < 47,200), group III (47,200 ≤ Al < 54,700), group IV (54,700 ≤ Al < 61,600) and group V (Al ≥ 61,600)

[2] Soil lead [Pb] (mg/kg) was categorised as quintiles: group I (Pb < 44.0), group II (44.0 ≤ Pb < 60.0), group III (60.0 ≤ Pb < 95.0), group IV (95.0 ≤ Pb < 184.0) and group V (Pb ≥ 184.0)

[3] Soil uranium [U] (mg/kg) was categorised as quintiles: group I (U < 1.49), group II (1.49 ≤ U < 1.83), group III (1.83 ≤ U < 2.16), group IV (2.16 ≤ U < 2.50) and group V (U ≥ 2.50)

[4] Global test: p-value = 0.60

Table 5.4: Using mutually adjusted multivariable Cox regression model to estimate hazard ratios (HR) for lung cancer in association with aluminium, lead and uranium, using THIN-GBASE database from 01-Jan-2004 to 31-Dec-2013

Soil element	Aluminium[1,3]	Lead[1,4]	Uranium[1,5]
Exposure groups	HR (95% CI)[2]	HR (95% CI)	HR (95% CI)
Group I	1.00	1.00	1.00
Group II	1.01 (0.94-1.08)	0.95 (0.89-1.01)	1.08 (0.94-1.07)
Group III	1.21 (1.13-1.29)	1.12 (1.06-1.20)	0.98 (0.92-1.06)
Group IV	1.19 (1.11-1.28)	1.11 (1.04-1.18)	1.06 (0.98-1.14)
Group V	1.18 (1.09-1.27)	0.94 (0.88-1.01)	1.13 (1.05-1.21)

[1] Cox model contains aluminium, lead and uranium only

[2] Hazard ratio (HR); 95% confidence interval (95% CI)

[3] Soil aluminium [Al] (mg/kg) was categorised as quintiles: group I (Al < 37,100), group II (37,100 ≤ Al < 47,200), group III (47,200 ≤ Al < 54,700), group IV (54,700 ≤ Al < 61,600) and group V (Al ≥ 61,600)

[4] Soil lead [Pb] (mg/kg) was categorised as quintiles: group I (Pb < 44.0), group II (44.0 ≤ Pb < 60.0), group III (60.0 ≤ Pb < 95.0), group IV (95.0 ≤ Pb < 184.0) and group V (Pb ≥ 184.0)

[5] Soil uranium [U] (mg/kg) was categorised as quintiles: group I (U < 1.49), group II (1.49 ≤ U < 1.83), group III (1.83 ≤ U < 2.16), group IV (2.16 ≤ U < 2.50) and group V (U ≥ 2.50)

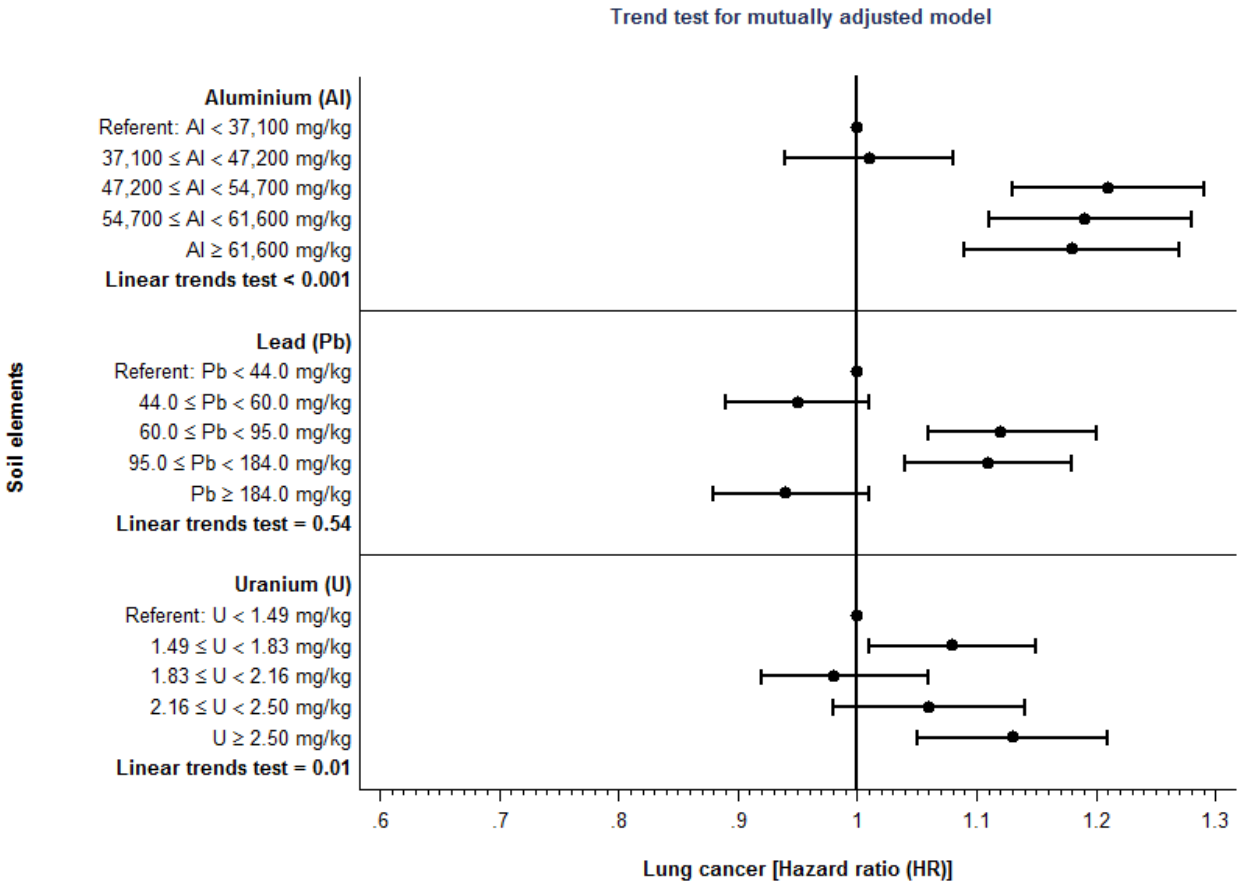

Figure 5.6: Modified scatter plot with range capped spikes showing patterns of hazard ratio as seen from our mutually adjusted model in table 5.4. P-value for trends test was used to determine if hazard ratios increased linearly across increasing exposure groups for each element.

5.4.3.2 Corrected Cox multivariable regression model

The Schoenfeld's residuals test and Aalen plots were carried out as pre-diagnostic test for examining the proportional hazards assumption after correcting for potential confounding through including age, sex, smoking status and socioeconomic deprivation in the corrected model. Overall, we found significant evidence that the assumption of the proportional-hazards was violated (i.e. global test: p-value < 0.0001). The test showed significant evidence of non-proportional hazards and time dependent effects for categories in the following covariates: gender (female), age groups (71-80 and +81 years) and smoking status (unknown) (Table 5.5).

We identified the form of the time-varying effects for each category affected by non-proportional hazards through generating Aalen plots to show the patterns of the estimated cumulative hazards regression coefficients so as to enable visual interpretation of the parameterisation of the time-varying effects found for these categories. These were used to identify the important cut-points where the patterns for the cumulative regression hazard function changed significantly. Based on the cut-points – interval-specific covariates for the affected categories were generated (Figure 5.7-5.10).

The original categories were replaced with the time-based interval-specific covariates generated in Aalen analyses (Figure 5.7-5.10), and were refitted into the corrected model to ensure that the

proportional-hazards assumptions have been met, as well as the removal of any time-varying effects. After Aalen diagnostic testing, the re-analyses using Schoenfeld's residuals shows all p-values that were significant previously for problematic categories have been abrogated successfully (Table 5.6).

The inclusion of confounders, along with time-based interval-specific variable derived from the Aalen plots resulted in marginal reductions in the estimated hazard ratios for all subset of soil elements (Table 5.7). The significant risk derived from the mutually model for those living in areas with the highest soil concentrations levels for aluminium (≥ 61,600 mg/kg) and uranium (≥ 2.50 mg/kg) was abrogated. Furthermore, the patterns of association for the subset of soil elements in terms of significant trends differed from the previous analysis, where aluminium was the only element that retained its significant linear trend pattern ($p = 0.01$) as opposed to lead and uranium whereby their patterns appeared to be abrogated after including confounding variables (Figure 5.11).

Table 5.5: Test of proportional-hazards assumption for the corrected model which includes confounding factors for assessing the risk of lung cancer with the selected group soil elements using Schoenfeld's residuals

Confounding variables	Proportional-hazards assumption test
Sex	p-value
Male (referent)	-
Female	< 0.0001
Age group (years)	
≤40 (referent)	-
41-50	0.54
51-60	0.38
61-70	0.07
71-80	<0.0005
+81	<0.0001
Smoking status	
Never (referent)	-
Non-smoker	0.37
Ever-smoked	0.25
Unknown	<0.0001
Socioeconomic deprivation	
Group I (referent)	-
Group II	0.32
Group III	0.49
Group IV	0.07
Group V	0.14
Unknown	0.12

Global test: p-value < 0.0001

P-values for aluminium, lead and uranium were not significant (p>0.05) (data not included)

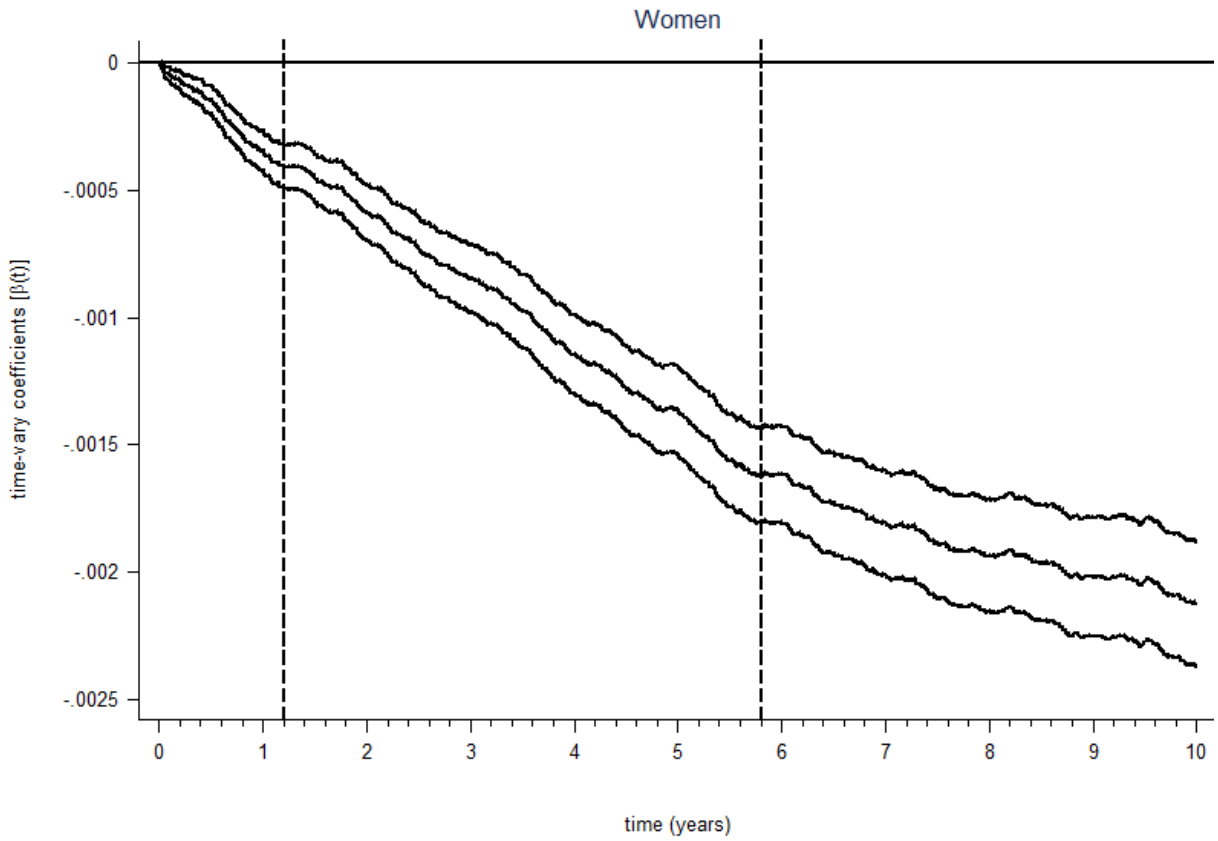

Figure 5.7: Aalen plot showing the estimated cumulative regression coefficients for lung cancer patients (with 95% confidence interval) who were women (versus men). The vertical dashed lines are cut-points at 1.2 & 5.8 which show the change in slope of the cumulative hazard function. The following three time-based interval-specific effects for the female category were generated using the above cut-points: Early effects ($t \leq 1.2$), Middle effects ($1.2 < t \leq 5.8$) and late effects ($t > 5.8$)

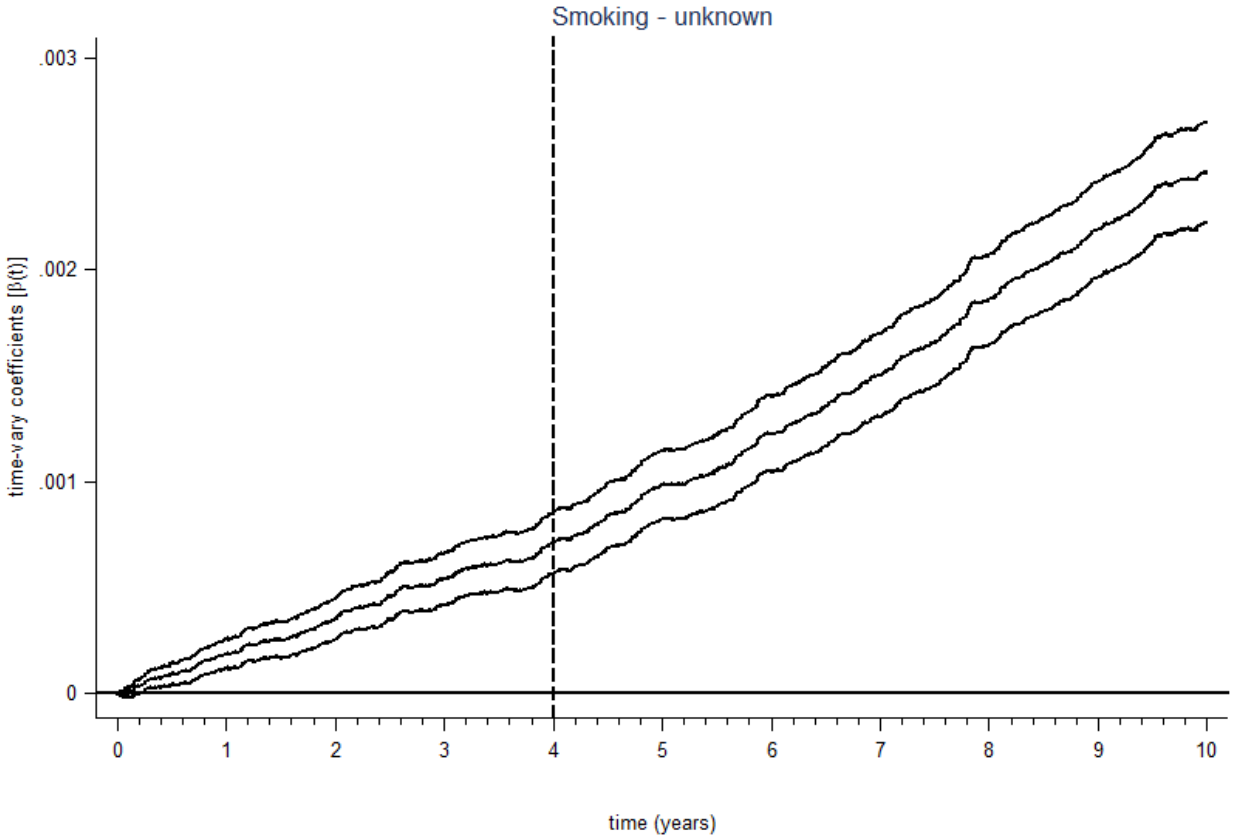

Figure 5.8: Aalen plot showing the estimated cumulative regression coefficients for lung cancer patients (with 95% confidence interval) whose smoking status was unknown (versus those who had never smoked). The vertical dashed line at the cut-point 4 is the change in the slope's direction of the cumulative hazard function. Only two time-based interval-specific effects for the unknown smoking status category was generated using the above cut-point: Early effects (t ≤ 4) & late effects (t > 4)

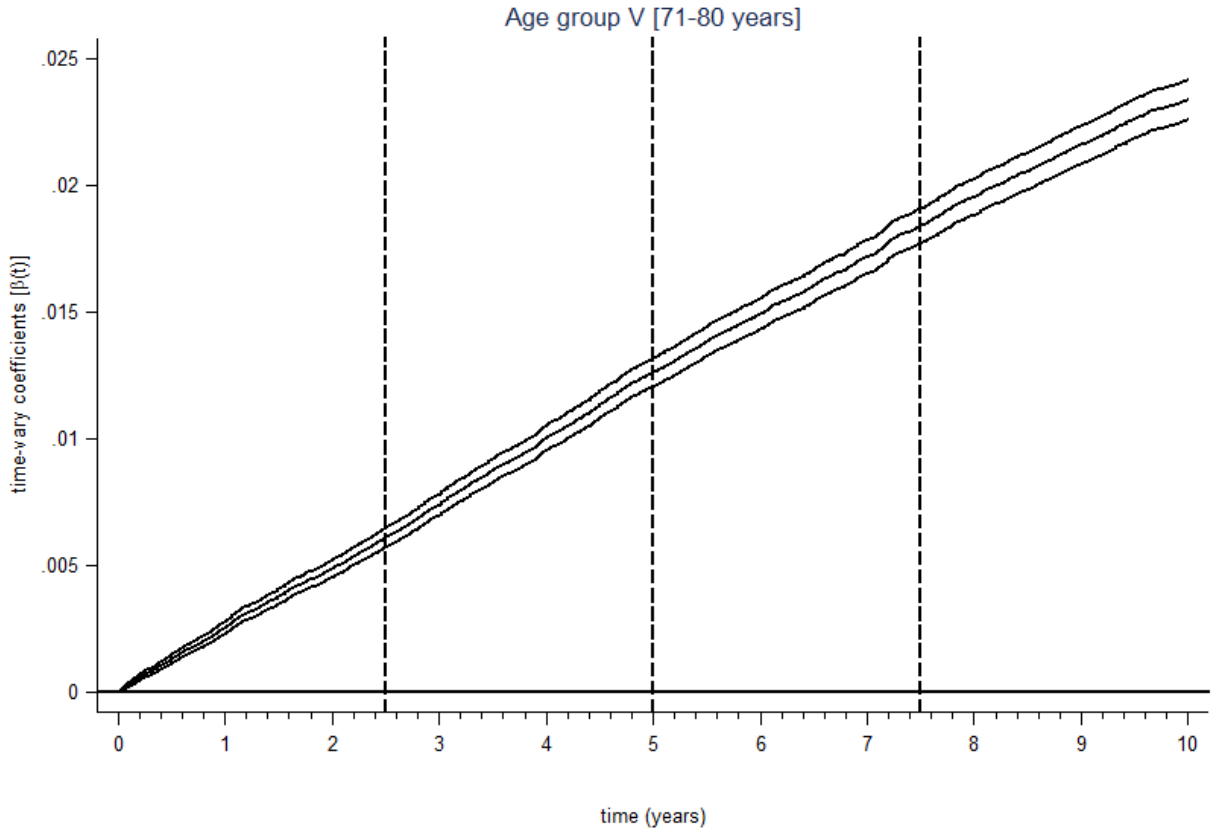

Figure 5.9: Aalen plot showing the estimated cumulative regression coefficients for lung cancer patients (with 95% confidence interval) who were 71-80 years of age (versus age groups ≤ 40 years). The changes in the hazard function in the above output were inconspicuous therefore cut-points were specified at a set interval of 2.5 years to eliminate the time-varying effects. The following four time-based interval-specific effects for the age group were generated using the proposed cut-points: Early effects ($t \leq 2.5$), early-middle effects ($2.5 < t \leq 5.0$), middle-late effects ($5.0 < t \leq 7.5$) and late effects (> 7.5)

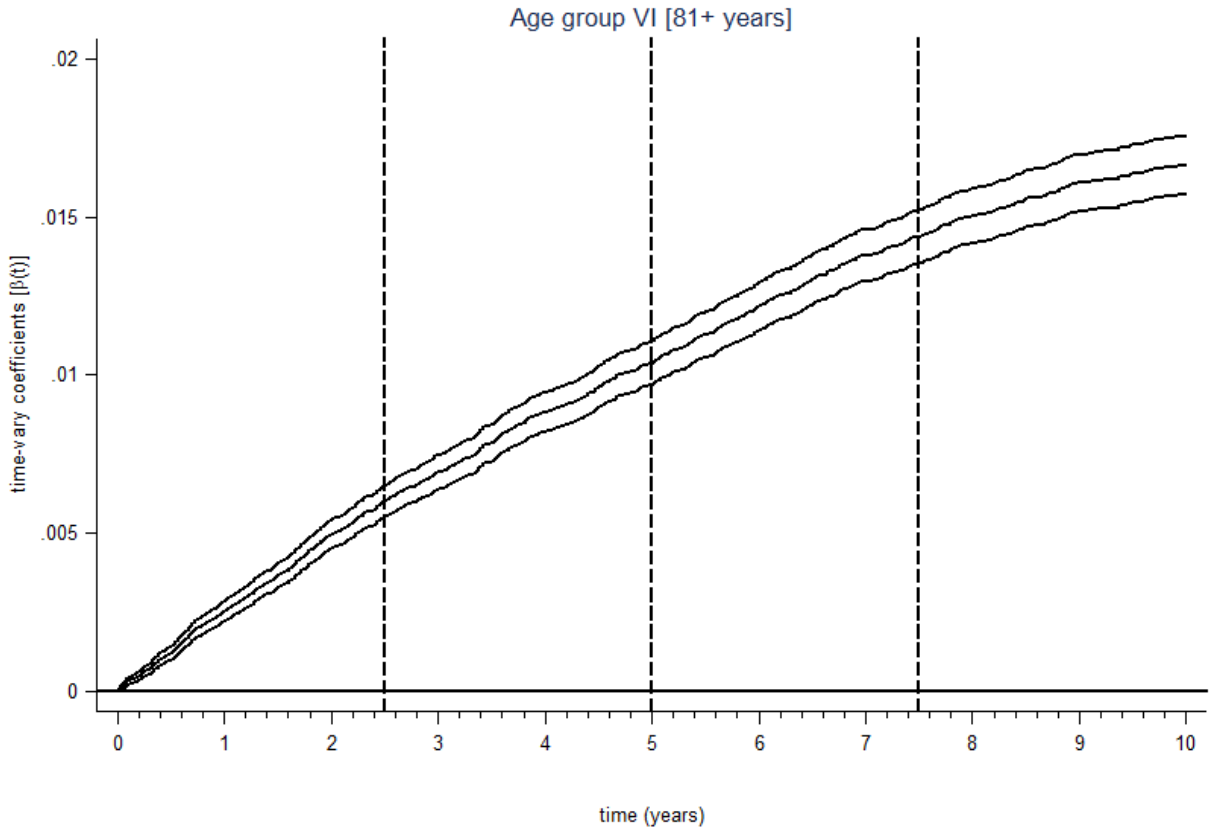

Figure 5.10: Aalen plot showing the estimated cumulative regression coefficients for lung cancer patients (with 95% confidence interval) who were 71-80 years of age (versus age groups ≤ 40 years). The changes in the hazard function in the above output were inconspicuous therefore cut-points were specified at a set interval of 2.5 years to eliminate the time-varying effects. The following four time-based interval-specific effects for the age group were generated using the proposed cut-points: Early effects (t ≤ 2.5), early-middle effects (2.5 < t ≤ 5.0), middle-late effects (5.0 < t ≤ 7.5) and late effects (> 7.5)

Table 5.6: Test of proportional-hazards assumption for confounding factors in the corrected multivariable Cox regression model using Schoenfeld's residuals after using Aalen plots to remove time-varying effects for sex, age group and smoking status

Confounding variables	Proportional-hazards assumption test
Sex	p-value
Male (referent)	-
*Female	
Early	0.91
Mid	0.79
Late	0.73
Age group (years)	
≤40 (referent)	-
41-50	0.55
51-60	0.38
61-70	0.08
*71-80	
Early	0.42
Early to mid	0.58
Mid to late	0.48
Late	0.48
*+81	
Early	0.45
Early to mid	0.51
Mid to late	0.37
Late	0.35
Smoking status	
Never (referent)	-
Non-smoker	0.38
Ever-smoked	0.25
*Unknown	
Early	0.37
Late	0.34
Socioeconomic deprivation	
Group I (referent)	-
Group II	0.31
Group III	0.43
Group IV	0.08
Group V	0.12
Unknown	0.12

Global test: p-value = 0.761

P-values for aluminium, lead and uranium were not significant. Data not included in table

*Category with reformatted to interval-specific variables as suggested by the Aalen plots

Table 5.7: Using a corrected multivariable Cox regression model to estimate hazard ratios (HR) for lung cancer in association with aluminium, lead and uranium, using THIN-GBASE database from 01-Jan-2004 to 31-Dec-2013

Soil element	Aluminium[1,3]	Lead[1,4]	Uranium[1,5]
Exposure groups	HR (95% CI)[2]	HR (95% CI)	HR (95% CI)
Group I	1.00	1.00	1.00
Group II	0.99 (0.92-1.06)	1.00 (0.93-1.06)	1.01 (0.94-1.07)
Group III	1.09 (1.02-1.17)	1.12 (1.06-1.19)	0.94 (0.88-1.01)
Group IV	1.10 (1.02-1.18)	1.08 (1.01-1.15)	1.00 (0.93-1.08)
Group V	1.06 (0.98-1.15)	0.97 (0.91-1.04)	1.05 (0.97-1.13)

[1] Cox model contains aluminium, lead and uranium corrected for confounding variables: gender, age group, smoking status and socioeconomic deprivation. As suggested by the Aalen plots the following categories were fitted as time-based interval-specific variables: female, unknown smoking status, age groups 71-80 & +81 years.

[2] Hazard ratio (HR); 95% confidence interval (95% CI)

[3] Soil aluminium [Al] (mg/kg) was categorised as quintiles: group I (Al < 37,100), group II (37,100 ≤ Al < 47,200), group III (47,200 ≤ Al < 54,700), group IV (54,700 ≤ Al < 61,600) and group V (Al ≥ 61,600)

[4] Soil lead [Pb] (mg/kg) was categorised as quintiles: group I (Pb < 44.0), group II (44.0 ≤ Pb < 60.0), group III (60.0 ≤ Pb < 95.0), group IV (95.0 ≤ Pb < 184.0) and group V (Pb ≥ 184.0)

[5] Soil uranium [U] (mg/kg) was categorised as quintiles: group I (U < 1.49), group II (1.49 ≤ U < 1.83), group III (1.83 ≤ U < 2.16), group IV (2.16 ≤ U < 2.50) and group V (U ≥ 2.50)

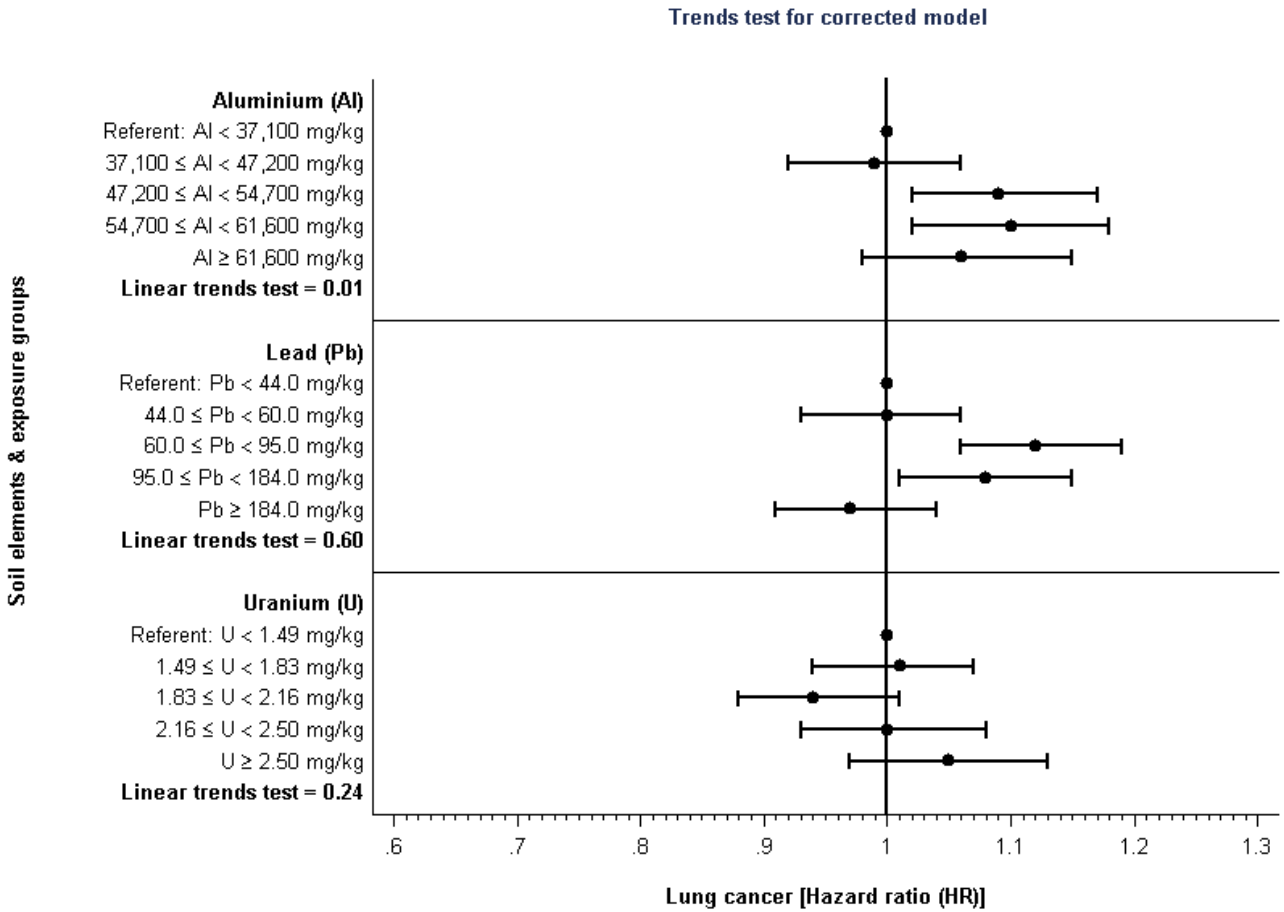

Figure 5.11: Modified scatter plot with range capped spikes showing patterns of hazard ratio as seen from our corrected model in table 5.7. P-value for trends test was used to determine if hazard ratios increased linearly across increasing exposure groups for each element.

5.4.3.3 Stratified analysis based on residential settings

The patterns of lung cancer risk in relation to soil aluminium, lead and uranium differed across different types of residential environment. Aluminium appeared to be the only element which significantly increased the risk of lung cancer among urban residents only (Figure 5.12). The patterns seem mimic those seen in the mutual adjusted model; however, the stratified model suggests that the increased risks for lung cancer were between 12-16% for urban residents with soil aluminium concentrations above 47,200 mg/kg. It appears that there was a significant dose-response pattern within the urban residential areas, whereby the risks increased non-linearly between aluminium exposure groups I to III, and then the effects plateau from group III onwards ($p < 0.001$). The remaining exposure groups for all other elements – lead (Figure 5.13) and uranium (Figure 5.14) - were non-significant in all different environmental settings, with unclear patterns.

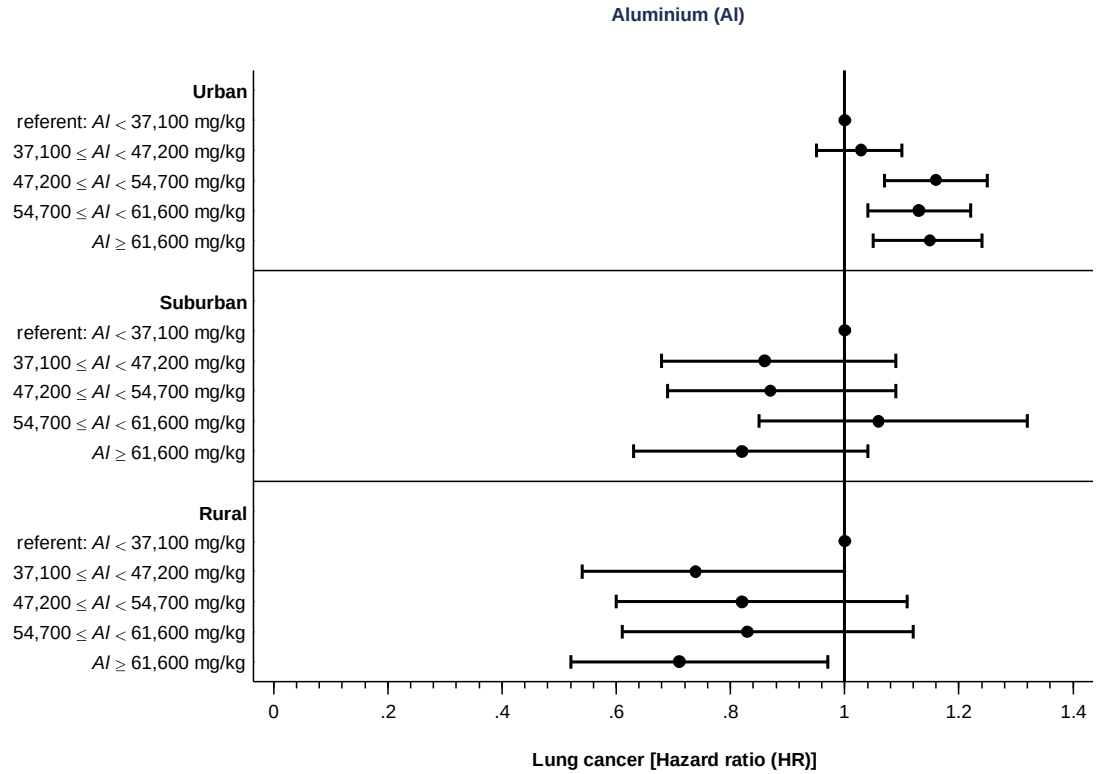

Figure 5.12: Modified scatter plot with range capped spikes were used to represent HR estimates derived from a stratified Cox regression model based on residential classification to show the association between BCC risk and soil aluminium, using THIN-GBASE database 01-01-2004 to 31-12-2013. The stratified model included other primary exposures - soil lead and uranium, and were adjusted for sex, age group, smoking status, socioeconomic deprivation and categories fitted as time-based interval-specified covariates derived from earlier Aalen plots

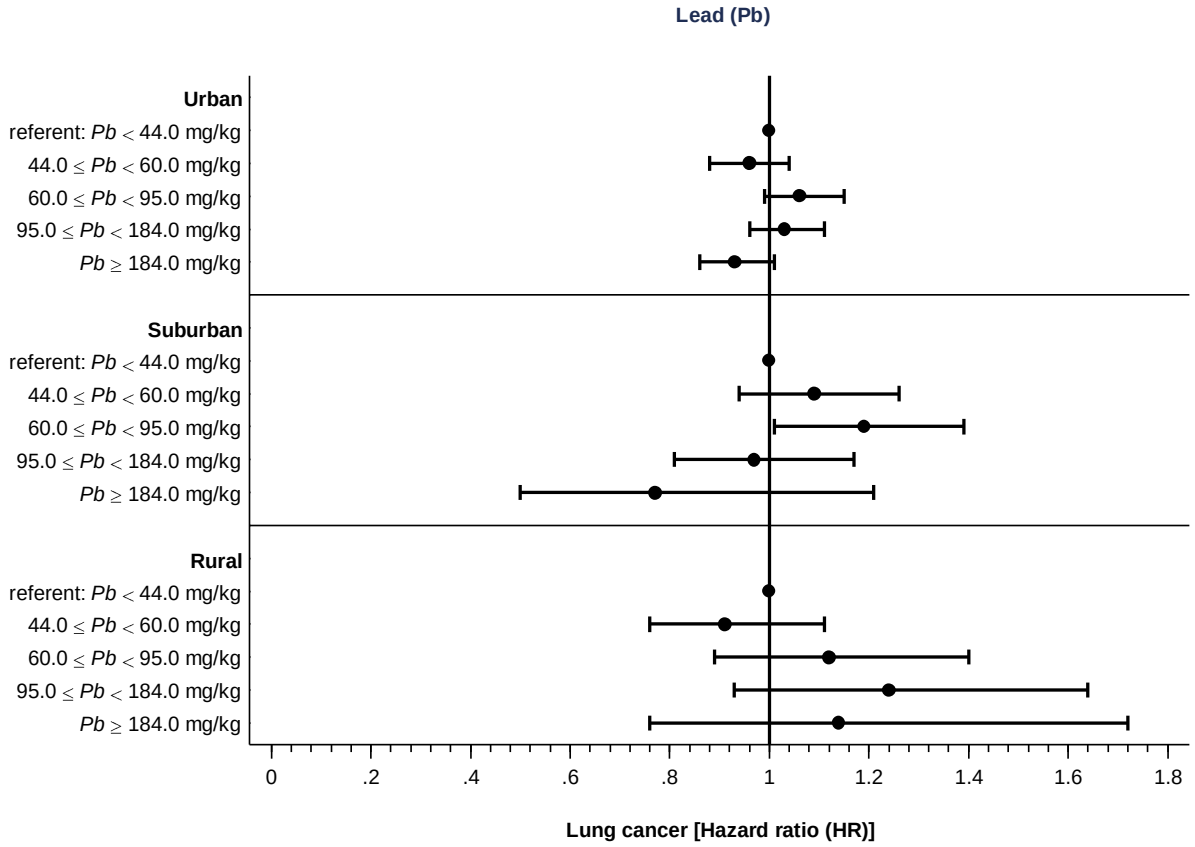

Figure 5.13: Modified scatter plot with range capped spikes were used to represent HR estimates derived from a stratified Cox regression model based on residential classification to show the association between BCC risk and soil lead, using THIN-GBASE database 01-01-2004 to 31-12-2013. The stratified model included other primary exposures - soil aluminium and uranium, and were adjusted for sex, age group, smoking status, socioeconomic deprivation and categories fitted as time-based interval-specified covariates derived from earlier Aalen plots

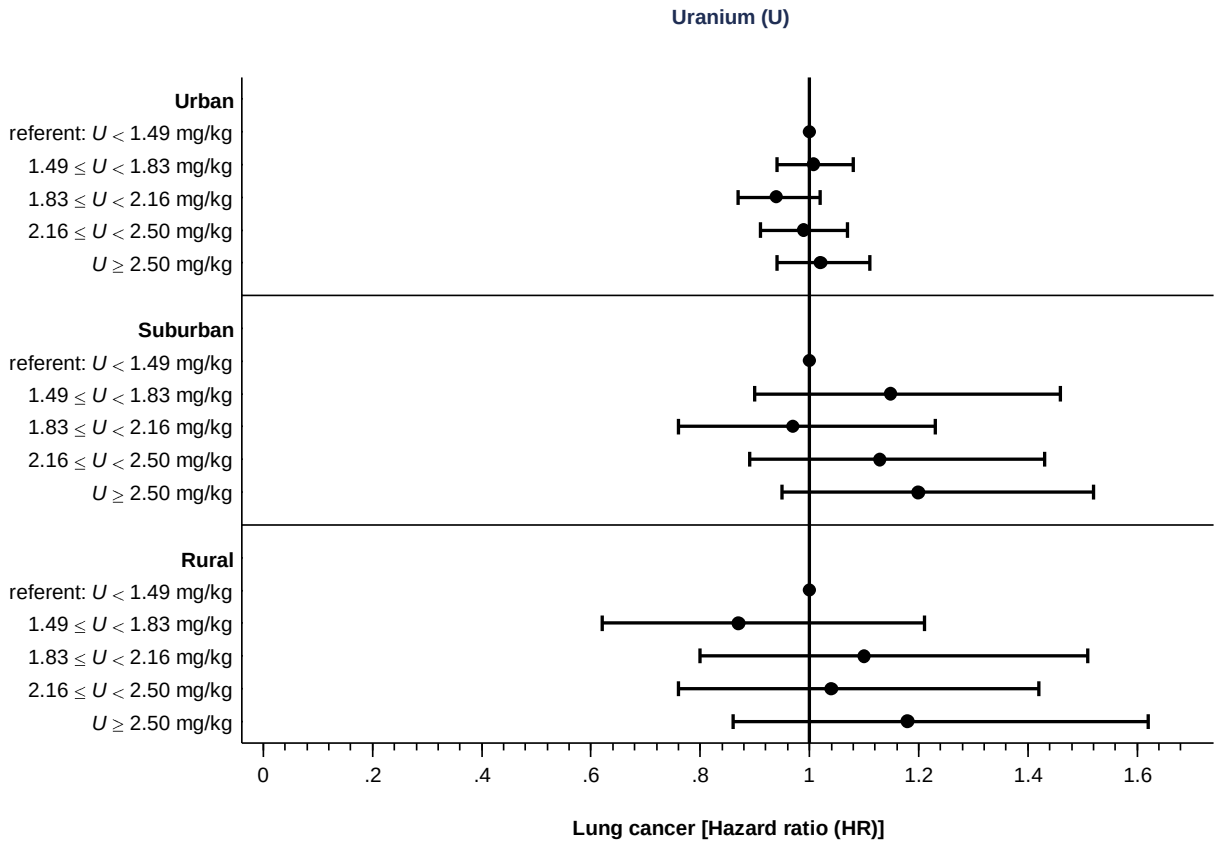

Figure 5.14: Modified scatter plot with range capped spikes were used to represent HR estimates derived from a stratified Cox regression model based on residential classification to show the association between BCC risk and soil uranium, using THIN-GBASE database 01-01-2004 to 31-12-2013. The stratified model included other primary exposures - soil aluminium and lead, and were adjusted for sex, age group, smoking status, socioeconomic deprivation and categories fitted as time-based interval-specified covariates derived from earlier Aalen plots

5.5 Discussion

This is the first study to use a 2-stage approach to evaluate the association between environmental soil elements and lung cancer in the UK. Our data mining approach identified soil aluminium, lead and uranium as the most important subset of metals for lung cancer risk prediction. Our initial findings showed a modest increase in hazards for lung cancer in participants residing in areas with high concentrations for aluminium and uranium, and medium levels for soil lead. However, after controlling for confounding, the modest effects can only be seen for potential exposure to aluminium (i.e. above 47,200 mg/kg). The risk of lung cancer appeared to be pronounced in urban residential areas.

One of the key advantages of using the CFS method was that it allowed us to approach the THIN-GBASE database without any prior hypotheses, and to discover which restricted group of soil elements may potentially be related to lung cancer. The main purpose of using such filter-based search method was to ensure that the selected group of soil elements were the most relevant variables for model construction, yielding a model that would optimally predict lung cancer risk. However, there is the potential for this approach to create subsets which could include elements which were not biologically plausible to be associated with the risk of lung cancer. Lead and uranium are established agents known to be carcinogenic in humans.[11,46,197,198]

Uranium is widely treated as a proxy for radon, which, in turn, is strongly associated with cancer.[177–181] The assertion that uranium being a proxy radon was derived from previous geological studies which showed soil concentrations with the most abundance of uranium were positively associated with higher levels of radon emissions into the air.[199,200] While radon's radioactive gas is produced from a series of decays; its lithological origin stems from radium which in turn is derived from uranium-enriched soil and rocks.[199,201] The World Health Organisation (WHO) have estimated that life-long inhalation of low-level radon is the second most important cause of lung cancer and have concluded that exposure to uranium-enriched soil can potentially be a contributing factor to lung cancer incidence, as it is responsible for radon emissions.[181] Aluminium was the least expected to appear among the list of variables relevant for lung cancer prediction due to the IARC having not classified aluminium as a human carcinogen due to the limited amount of data;[202] however, the IARC have acknowledged that exposure occurring within aluminium refinery (or reduction) industry is a major risk factor for lung and bladder cancer.[46,202]

The results in the second stage of our analysis showed that potential exposure to medium to high concentrations of aluminium may be associated with lung cancer. A possible explanation for this finding can be drawn from previous occupational-based studies that have reported an increased risk of lung[203] and bladder[204] cancer among workers in the aluminium industry;[43] however, most of these studies

asserted that the risks were attributed to exposure to other dangerous substances (polycyclic aromatic hydrocarbons, sulphur dioxide, fluorides etc.) involved with aluminium refinery.[202–204] Despite this notion, these studies have acknowledged that long-term exposure to aluminium can impair the normal functioning of the respiratory tract, which, in turn, may lead to lung carcinogenesis.[202,204–207] In the context of our study, particulate matter emitted from soil represents an important contribution to air quality in the UK. Aluminium has the largest abundance in UK top soil in terms of it having the highest concentration levels,[53] it may be plausible that aluminium-bound particulates (or dust) emerging from soil and may remain airborne – and the general population may come into long-term contact through inhaling it.

This study was unable to establish soil lead as significant risk factor for lung cancer. To the best of our knowledge no study within the domain of environmental exposures have found an association between increased exposure to soil concentrations of lead and risk of lung cancer. However, past studies found an association between lead and lung cancer[208,209] however, these studies have later been contradicted by recent research. These studies were conducted within the industrial-occupational settings, and were focused on workers within the lead smelters and lead battery industry[11,19,33]

Our study found that individuals living in areas with the highest levels of soil uranium concentration above 2.50 mg/kg have an increased risk

of developing lung cancer; however, this result did not persist following adjustment for potential confounding factors. Several studies have shown sites with localised soil uranium concentrations are associated with some of the highest levels of indoor radon. Uranium is ubiquitous in soils and rock which serves as a source for radon emission outdoors[199,201,210] in a residential environment were uranium-enriched soil are in abundance leads to radon diffusing from an outdoor environment where radon particles are of higher concentration, to an indoor environment where concentrations are lower potentially leading to accumulation in houses (especially in basements).[28,199,210,211] Long-term exposure to radon through inhalation leads to severe cellular DNA damage within the respiratory tract resulting in lung cancer. Several epidemiological studies have shown that environmental levels of radon were associated to lung and GIT cancer,[177,178,197-199,212] and that concentrations of soil uranium were key contributor to indoor air contamination with radon.

By using a retrospective cohort design for our lung cancer study, we believe that we managed to reduce the possibility of selection bias. This study uses incident records of lung cancer that were recorded in THIN by the time this study was being conducted. Similar to the studies in chapter 4, selection bias is not a major concern as we attempted to use the entire cohort of eligible patients in the THIN-GBASE, we have provided a clear definition for our inclusion and exclusion criteria that limits our population of interests to adults (aged 18 years and above) registered at a GP practice at least 1-year

before start date of study (i.e. January 1st, 2004). One source of selection bias may occur from the dates we chose to begin the study which only captures incident cases of lung cancer after 2004, with a short study period limited to ten years (i.e. 2004-2014). However, we chose this date because of its significance because it marks the introduction of the QOF encouraging GPs to report new cases of lung cancers as soon as they are diagnosed. Prior to 2004, recordings of lung and other cancers were much lower in THIN than expected when compared on the national cancer registry data.[188,213] Therefore, to avoid any inconsistencies in the recording rates of lung cancer that occurred before, and after 2004, we therefore limited the analysis to events that occur after 2004.

In terms of completeness, the THIN-GBASE are reliable source of lung cancer data. Recent studies have shown that the recording of lung cancer in the THIN database consistent, especially after the introduction of QOF whereby the recording rates had increased to approximately 80.0%.[188,213] Validation studies have indicated that lung cancer data from THIN captures most cases from other cancer registries, and therefore its complete and representative.[188] One problem that was beyond our ability to resolve was the differences in lung cancer ascertainment rates at a GP-level across England and Wales. If case ascertainment rates at a practice-level is associated with geographical variation in exposures to the selected group soil element, then are our risk estimates for lung cancer was be biased.[169] We attempted to control for this by making adjustments with the SHA

variable; however, a better approach would have been the inclusion of some performance indicator at a practice-level that could be treated as a proxy for measuring how well practices tend to record clinical outcomes.

We have incorporate data-driven approaches to determine initially which soil elements are likely predictors of lung cancer with data mining,[193] before deriving detailed hazard estimates using Cox survival regression analysis, where we were able to remove any time-varying effects through using the Aalen additive survival model.[168]

One of the major shortcomings was our inability to incorporate into our analyses the exact location of where a patient lived, and the spatial point of where soil samples were measured. The inability to display spatially by means of usage of high resolution maps could have potentially provided a convincing picture as to whether these soil elements (where they are highly concentrated at) are linked to large clusters of incident lung cancer cases. Furthermore, by incorporating that spatial component into the analysis would have reduced any error (spatial variability) that exists between samples measured. Another limitation was that although we were able to make adjustments for meaningful confounding variables, we unable to include other potentially important confounders; for example, the type of industrial occupation of a patient (i.e. in aluminium production), ambient levels of heavy metal particulates, and environmental levels of radon, which are highly correlated with soil levels of uranium, and may serve as a

better effect modifier (i.e. for stratification) for lung cancer than urban vs. rural. In addition, we cannot be certain that residential soil element levels as measured by the THIN-GBASE link directly translate into increased exposure and/or bioavailability at individual level.

Just like our previous study for arsenic and BCC, another shortcoming for this study was our inability to use the following data: 1.) Information related to a patient's addresses; 2.) the location of general practices he/she attended; and 3.) the georeferenced sampling points for the soil samples that were collected at different densities. We mentioned these limitations were due ethical and legal implications outlined by THIN. The geospatial details of lung cancer patients were anonymised, and therefore, we could not utilise this resource to directly ascertain the distribution of those that fall within a sampling point; let alone, determine the distribution of those that lived on soils classified as either a G-BASE urban or rural terrain, or BGS reanalysed NSI(XRFS) areas. If this information were available, we could have stratified the population at risk of lung cancer in accordance to these three areas. Our approach to this issue of differentiating participants that lived on urban and rural terrain (i.e. G-BASE urban, rural and NSI(XRFS)) by using a stratified Cox regression analyses was based on the type of residential setting indicators recorded in THIN (i.e. urban, suburban and rural), so as to reduce the possibility of any information bias that may affect our risk estimates for lung cancer because of systematic differences in exposure in soil

aluminium, lead and uranium as a result of samples being collected at densities.

For aluminium, specifically, our results show a positive trend which indicates that there could be a potential dose-response relationship between the increased exposure groups for aluminium and lung cancer incidence. This has implications for the priority with which element should be monitored by environmental or geological agencies. Other sources of aluminium that contribute to soil and atmospheric contamination, especially within urban residential areas, should be monitored. Further research is required to conclusively establish whether soil elements in general are linked to lung cancer.

In conclusion, the current study suggests that those living in areas with soil aluminium levels above 47,200 mg/kg may have a greater risk of developing lung cancer. The result suggests that aluminium exposure among urban residents may be the cause of lung cancer for this group. While, the results indicate statistical significance, they need to be interpreted with caution due to the limitations that are present for this study. Further studies will be needed to validate the findings made in this investigation.

Gastrointestinal tract cancer

Chapter 6

6 Summary

In the past (1950-80s), there has been much epidemiological research into the influence of exposure to soil metals on the risk of various GIT-related cancers; however, these studies are historical and outdated, because they were conducted at a time period where soils were highly contaminated due to the many mining activities that took place. There is a need to update the evidence in relation to the modern environment. Therefore, we aim to examine the relationships between low-level soil contaminants and risk GIT cancers.

A three-stage process was used to conduct the following: (1) using data mining to determine which group of soil metals are the best predictors of GIT cancer; (2) using a population-based cohort design for quantifying the effect size for GIT cancer risk for individuals living in a residential area with high concentration levels for the selected soil elements; and (3) multivariate analysis using competing risks survival models were applied to determine the association between the subset of selected elements and the risk of developing following cancers: a. upper GIT cancers, b. stomach cancer, and c. colorectal cancer.

Our data mining model identified the following soil elements as appropriate predictors for overall GIT cancers: aluminium, calcium, lead, phosphorus, manganese, uranium and zinc. In stage 2, our mutually adjusted model had initially shown residents living on soils with elevated levels of aluminium, phosphorus and uranium had an

increased risk of developing GIT cancers, while soil levels for zinc were protective effect against GIT cancers. Only soil phosphorus remained significantly associated with overall GIT cancer while all remaining elements were rendered non-significant after making adjustments for confounding variables. In stage 3, our competing risk models had not shown any meaningful relationship for any of the selected group of elements while taking into account of potential competing events.

The results for soil phosphorus authentically remained consist throughout the regression analysis, when compared to elements used for this analysis. Therefore, we concluded that increased exposure to low-levels of phosphorus may have an impact on overall GIT cancer incidence in the UK.

6.1 Background

Gastrointestinal tract (GIT) cancer mainly refers to a group of internal malignant neoplasms occurring on primary organs actively involved with digestive processes – mouth, oesophagus, stomach, ileum (or small intestine), colon (or large intestine), rectum and anal canal. GIT cancers may also include the appearance of solid tumours on secondary or accessory glands such as the liver, bile, gallbladder and pancreas, which provide aid to the digestive organs.[214,215] Primary organs of the GIT can be classified into three major groups: (1) the mouth and oesophagus comprise the upper GITs, (2) the stomach (standalone term); and (3) the duodenum, ileum, colon, rectum and anal canal comprise of the bowel (or colorectal area).

GIT cancers are among the most frequently diagnosed cancers in the world[176]. There is significant geographical variability in the incidence of GIT cancers, especially by site and cell-type, with North America, Western Europe, Australia and Japan documented as having the highest occurrence of oesophageal and colorectal cancer[175,214,216].whereas Africa and Asia (especially India and Iran) have the highest relative prevalence of stomach cancer relatively the highest prevalence of stomach cancers[175,214,215], possibly due to high incidences of infectious diseases.

There are multiple risk factors that contribute to the aetiology of most GIT cancers. Non-modifiable risk factors include the advancement of age, gender, genetics and heredity. Modifiable

factors are typically lifestyle-related: smoking, alcohol consumption, physical inactivity and dietary habits (i.e. minimum consumption or low intake of fibre-foods, fruit and vegetables)[217–219].

It has been strongly implied that exposure to environmental metals that emerge from soil may be risk factor for GIT cancer in humans. However, the few prior studies have directly addressed this possibility. Most of the contemporary studies exploring this aspect of research were conducted in countries such as Turkey,[220] Iran,[221,222] Taiwan[49] and China[223,224], where areas of severe contamination are widespread than elsewhere in the world. The method for quantifying exposure to soil elements were restricted to only surrogate measures of levels in soils through using either biomarkers levels in blood[49,220] (or urine) or levels of in locally grown foods.[220,223,224] In the UK, studies exploring this issue were carried out between the late 1950s and 1980s (and may not, therefore, reflect current risks), but were reliant on ecological study designs, limiting their value in assessing the likelihood of causation.[99–103,225–229]

Therefore, the purpose of this chapter is to establish a plausible explanation for the attribution of soil metals and how they may lead towards GIT cancer. A contemporary approach carried out in three stages will use (1) data mining, (2) population-based cohort design and (3) multivariate analysis using competing risk modelling to assess the associations between GIT cancer and residential levels of soil metals in UK.

6.2 Potential mechanism for gastrointestinal tract cancers in relation to soil elements

In previous chapters, we have shown a conceptual framework for how soil metals may be related to BCC and lung cancer, and shown the important pathways for exposure were through dermal absorption and inhalation via respiratory tract, respectively. In terms of GIT cancer, the largest pathway of concern is through the digestive tract. There are multiple means by which ingestion of soil particulates containing contaminants may occur. In particular, ingestion of airborne soil particulates in a form of household dust, ingestion of soil particulates that are attached to consumable plants, and the consumption of contaminated foods grown or raised on soil polluted with toxic metals. These are examples that can be integrated into our food chain indirectly.

6.2.1 Upper gastrointestinal tract

The biological mechanism for metallic elements that may trigger a carcinogenic effect differs depending on the type of organ associated with digestion. For instance, the effect of metallic elements on upper GIT organs responsible for ingestion [i.e. oral cavity (mouth) and oesophagus] may differ entirely from those of the stomach and colorectal tract involved with digestion and absorption, respectively. For example, recent studies conducted in Taiwan have shown that elevated concentrations of soil arsenic and nickel in farm soils may be associated with increased incidence of oral cancer,[230,231] with

associations being attributed to the fact that residents were consuming locally grown produce from soils with high concentrations of arsenic and nickel.[87,220,232] A study from Eastern Turkey,[220] where GIT cancers are endemic, have shown that oesophageal cancers were common among residents that consumed foods grown on soil highly contaminated with cadmium, lead, copper and cobalt.[220] It indicated that study population were consuming foods grown on soils with concentrations of cadmium, lead, copper and cobalt being 50, 6, 4 and 2-fold, respectively, greater than the standard soil guideline levels outlined in Turkey.[220] It has been hypothesised that the potential mechanism for soil metals to have carcinogenic effects on such tissues would require the lodging of contaminated material in the crevices of the oral cavity and oesophagus during ingestion.[142-144]

6.2.2 Stomach cancer

The carcinogenic effects of metallic elements are not limited to tissues of the upper GIT. For instance, past studies using tissues specimens extracted from the autopsied patients who had stomach or other cancers have demonstrated that long-term trace amount of exposure to environmental metals from air, water and soil leads to accumulation of elements inside tissues.[77-82] Studies based on analysing trace metals build-up in tissues have support the notion of accumulation of elements in stomach tissues may potentially induce cancer.[77-82] Health studies conducted in Iran and China have added to some degree of a possible link between elevated concentration levels

of metals in farm soil and stomach cancer by showing areas endemic to stomach cancer are typically situated at locations with farm land containing excessive concentrations of arsenic, chromium, cadmium, lead and selenium.[221,223] The authors acknowledged that higher concentrations of soil elements in farm soil play a key role in gastric cancer development, and that this problem has led to certain areas being endemic with stomach cancer.

6.2.3 **Bowel cancer**

The potential mechanism for bowel (or colorectal) cancer in relation to soil elements may involve one of the following processes: (1) the long-term lodging of a carcinogen in the fissures of the intestinal wall during absorption in the small or large intestine,[233-235] or (2) direct absorption of a carcinogen into the tissues of the intestine where upon they will remain and accumulate over time.[77-82] This was proven by few clinical studies that have observed trace levels of harmful elements present in intestinal tissue samples from autopsied patients affected with colorectal polyps or tumour growth.[233-235] It has been hypothesised that such mechanisms, if indeed, are occurring over a long period of time may be sufficient enough to trigger the development of tumour at the target organ. In general direct epidemiological research concerning the relationship between specific soil elements and bowel cancers are limited to studies (notably in China, Iran and Turkey) reporting general GIT cancers in rural areas where farming and soil cultivation are common practice. They usually

report GIT cancers area endemic to farm lands with soil containing excessive amount of toxic metals;[220,222,224,236] however, they do not report whether there is an association. The authors acknowledge this due to the limitation and lack of data, and therefore mention that further research must be warranted to establish a clear association.

6.3 Soil metallic elements and potential risk of gastrointestinal tract cancer in the United Kingdom

GIT cancers are among the most common malignancies reported in the UK.[237-239] In particular, bowel cancers are the fourth most diagnosed cancer following after breast, lung and prostate cancer.[240-242] According to Cancer Research UK, bowel cancer account for ~13% (41,581) of all incident cases reported in UK (2011) with an overall incidence rate of ~66 cases per 100,000 person year. Oesophageal, oral and stomach cancers are less common than large bowel cancers, being ranked 13th, 15th and 16th among UKs top 20 cancers by site, respectively.[237-239]

Like most industrialised countries - the incidence of GIT cancers in the UK is heavily influenced by modifiable factors such as physical inactivity, smoking, dietary regime and alcohol consumption.[214,215,217-219] The current literature pertaining to GIT cancers is therefore strongly focused on these risk factors as they are deemed to be the most readily addressable in public health terms. While there is also a considerable body of research exploring the relationship of GIT cancers and environmental carcinogens, most studies are focused on

occupational populations. The UK studies published to date on exposure to metallic elements from soil in relation to GIT cancers are limited, and are predominately over thirty years old.[99-103,225-229]

Studies assessing links between soil metals and GIT cancers were mostly carried out during the early 1950s to late-1980s, with an emphasis on finding the health impacts of trace elements – especially for cadmium, chromium, cobalt, nickel and lead[102,243-247]. At the time, investigations were performed in the North West (Cheshire) and Southwest of England (Somerset & Devonshire), as well as Northern Wales (Anglesey & Montgomeryshire) - due to rapid increase in cancer mortality and stomach cancers observed in these areas, as well as the known contamination present in soils of those areas. The increased incidence was a matter of huge public health concern that attracted considerable attention from the research community, and the subsequent search for potential environmental risk factors included several studies investigating the health impacts of soil elements[102,243-247].

A series of ecological studies were conducted, whereby investigators hypothesised that the nature of soil had a potential health impact, as well as the possibility of residents living in areas reported to have the greatest incidence of stomach cancer were more likely to be exposed to excessive concentrations of trace elements at their soil location. These studies indicated that the high incidence of stomach cancers were mainly observed in areas with soils that were peaty in texture[243],

which contained large excessive amounts of chromium, cobalt, copper, and zinc[102,244].

From 1978, until the late-80s, cadmium became a focus of public health concern in England which received much attention due to high concentrations of cadmium detected in soils throughout the Shipham Parish (South West England)[104,248-250]. The possibility of a health impact could not be ignored and so an array of exposure, risk assessment and epidemiologic studies similar to those performed in Northern Wales and parts of England during the early 50s were carried out. Substantial evidence of cadmium toxicity was found in a pilot study which that residents of Shipham whose garden soils contained between 60.0 and 998.0 mg/kg of cadmium had elevated blood-cadmium levels. The observed health consequences included hypertension, renal tubular damage and cancer of the colon in some patients[248]; however, this association was refuted in subsequent studies which only showed a small but nonetheless statistically significant risk for ovarian carcinoma only and not with GITs cancers – and thus concluded that risks of other health outcomes were unlikely to be explained by soil cadmium exposure[104,249-251].

The latest UK study aimed at elucidating local relationships between various metallic elements and twelve different cancer types was carried out in Northern Ireland. McKinley et al. (2013) used a spatial and ecological study design which demonstrated that the incidence of stomach cancer (at a ward-level) was correlated with arsenic levels in

soil. In addition, the strongest association for stomach cancer in the study was observed for wards linked to soil data having arsenic soil levels greater than 43.7 mg/kg (above UK the national safety value of 32 mg/kg).[189] However, the one of the major limitations of this study were that soil samples were aggregated at a geographical ward-level in Northern Ireland. This means that is ecological bias and any risk quantified cannot be related to a single individual due to exposure being measured on a geographical scale.

The linkage of THIN-GBASE removes this limitation by providing soil data on a postcode-level which can be utilised to quantify cancer risk at an individual-level. Therefore, the purpose of this chapter is to address this research gap. By using the THIN-GBASE database, we used series of epidemiological analysis that adopts a three-stage approach to assess the impact of soil metallic elements on GIT cancer risk.

6.4 Methods

A three stage process was used to determine which of the 15 soil elements in the linked database were associated with GIT cancers. The initial process involved the usage of data mining techniques as an exploratory exercise to find which of the element(s) (or best subset of elements) were predictive of GIT cancer. The next stage used a population-based cohort design for the purpose of quantifying the effect size for overall GIT cancer risk for individuals living in a residential area with high concentration levels for the selected soil elements. The third stage, multivariate analysis using competing risks

survival models were applied to determine the association between exposure to the subset of elements and the risk of developing cancer at certain region of the GIT [upper gastrointestinal cancer (includes oral & oesophagus), stomach cancer, and small (ileum) & large bowel cancer (colon, rectum & anal canal)].

6.4.1 **Study population**

6.4.1.1 **Case definition for gastrointestinal tract cancer**

The case definition for GIT cancers were patients typically recorded with any type of malignant neoplasm found in accessory organs actively involved with food digestion (mouth, oesophagus, stomach, ileum, colon, rectum and anal canal) - in THIN, GIT-related neoplasms are coded as site-specific cancers. Clinical gastrointestinal experts at the University of Nottingham were consulted with to determine which of the GIT cancer Read Codes were the most appropriate to be used for this analysis.

Patients with malignant neoplasms were identified using Read Codes under the following hierarchies: malignant neoplasms of lip, oral cavity and pharynx *B0…00*; malignant neoplasm of oesophagus *B10…00*; malignant neoplasm of stomach *B11…00*; malignant neoplasm of small intestines (ileum) *B12…00*; malignant neoplasm of colon *B13…00*; and malignant neoplasm of rectum, rectosigmoid junction & anus *B14…00*. The group of Read Codes with the above hierarchies

were extracted from the THIN-GBASE databases' patient medical and AHD records.

Patients found with any rare genetic-related GIT illness, or neoplasms located on the walls of the peritonum [or in any section of the intra-abdominal cavity rather than on the digestive organs themselves] were excluded from the analysis. Only malignant neoplasms appearing on organs that are actively involved with food digestion were considered; although the liver, pancreas, and biliary tract play an important role in digestion, their role is to store or secrete of fluids that break down food particles for absorption in the small and large intestine. Since they do not normally come into direct contact with ingested matter, they were excluded from the analysis.

In stages 1 and 2 the main outcome variable was treated as a binary indicator; as the presence or absence any GIT cancer. However, in stage 3 the study adopted a multivariate framework, whereby outcomes were generated separately for each site-specific GIT cancer (upper GIT, stomach and bowel cancer) as a three-category (polychotomous) indicator in which a category coded with 0 means complete absence of cancer outcome. The value 1 represented the main outcome of interest, while the third category (coded as 2) was classed as the group competing with main outcome. For example, the multivariate outcome or indictor for upper GIT cancers - the absence of any cancer would be coded as 0. The main outcome, diagnosis with upper GIT (mouth or oesophageal) cancer during the study period was

coded as 1. Lastly, the third category which was the competing risk group representing any GITs (i.e. stomach or bowel) other than upper GIT cancers were coded as 2. The definitions for upper GITs, stomach and bowel cancers are as follows: (1) Upper GIT cancers are any oral and oesophageal malignant neoplasms, (2) stomach cancers including any form of gastric-related neoplasms, and (3) bowel (or colorectal) cancers having Read Codes that refer to malignant neoplasms of the duodenum, ileum, colon, rectum and anal canal.

6.4.1.2 Inclusion criteria

The specification for inclusion was similar to that outlined in Chapter 5 (see section 5.3.2.2). Eligible participants were those aged 18+ years, which were registered at their GP practice and contributing data for at least one year before the study start date (1-January-2004). Participants were excluded if they had a GIT cancer diagnosis before the start date of the study or did not have complete exposure data on the 15 soil elements. Patients were only included if the practice at which they were registered had an acceptable mortality recording (AMR) date before the start of the study. The end date for the study was 31-December-2013.

Participants diagnosed with a GIT cancer during the course of the study (i.e. between 1-January-2004 and 31-December-2013) were right-censored at the date of their first GIT cancer recording. Secondly, participants with a death record within the duration of the study were right-censored at the date they were recorded to have

died. Lastly, participants with no event throughout the duration of the study were automatically right-censored on 31-December-2013.

6.4.1.3 Exposure and confounding variables

The methodologies used for categorising all 15 soil elements were similar to those applied in Chapter 5 (see section 5.3.2.3). Non-modifable risk factors treated as potential confounding variables included age, gender, and socioeconomic deprivation. Age was categorised as below 40, 41-50, 51-60, 61-70, 71-80 and 81+ years. Quintiles of Townsend indices of deprivation were used to measure socioeconomic deprivation.

Modifiable or lifestyle-related confounding variables were smoking status, body mass index (BMI) and alcohol consumption. An individual's smoking status and the amount of alcohol consumed per week before the start date of study was extracted from the AHD files using validated Read Codes. Smoking status of participants were categorised as never smoked, non-smoker, ever-smoked or unknown if information about the patient's smoking status was unavailable.[190] A participant's alcohol consumption level (per week) was categorised in accordance with the UK sensible drinking guideline as moderate (men: <21 units per week; or women: <14 units per week), hazardous (men: 21-50 units per week; or women: 14-35 units per week), harmful (men: 50+ unit per week; or women: 35+ units per week) or unknown if information about regarding the patient's drinking patterns was unavailable. BMI of participants were directly extracted from the AHD

files in THIN-GBASE. The BMI was directly estimated using the height (in m) and mass (in kg) of a participant as an alternative approach when he/she had missing information. BMI was categories in accordance with WHO guideline for adults – underweight: <18.5 BMI; acceptable: 18.5-24.9 BMI; pre-obesity: 25.0-29.9 BMI; Obesity (class I-III): 30+ BMI.

The type of residential setting (or living environment the patient resided in) was extracted and classified as urban (>10,000 buildings), suburban (town or fringe) and rural (village, hamlet or isolated village). In stage 2 – the categories were treated as strata (for subsequent stratification analysis) rather than confounders, so as to test the effects of each soil metal on GIT cancers in each geographic setting.

6.4.2 Statistical analysis

6.4.2.1 Filter method for feature selection (Stage 1)

The CFS method was used to determine which restricted group of soil metals were highly correlated with overall GIT cancer outcome. The default search algorithm used for CFS was the forward greedy stepwise technique to select elements in a successive order of importance (from highest importance to lowest) to form larger and larger subsets.

The CFS algorithm was applied using the greedy stepwise technique whereby exposures were procedurally generated in successive order of

importance to form a larger subset. The search process ends once it reaches the optimal subset of exposures. The algorithm produces a summary statistic called the merit score which determines the overall strength of error in adding attributes into the subset – a summary score of 15% and below is usually preferred. Note that the CFS algorithm does not guarantee model estimates for exposure categories in selected attributes to be statistically significant; rather, it ensures that the model constructed is fully optimised by returning the best log-likelihood score. All data mining analyses in stage one were performed in Weka 3.7.12 (University of Waikato, Hamilton & Tauranga, New Zealand). All elements derived from this stage are used throughout subsequent stages.

6.4.2.2 Multivariable Cox regression modelling (Stage 2)

Multivariable Cox regression models were used to determine the association between the selected group of soil elements derived from stage 1 and GIT cancers. Our initial model comprised of a mutual adjusted model containing only the selected set of soil metals. Subsequently, a corrected model was fitted with both soil elements and confounding variables. A stratified analysis was also carried out to assess the impacts within a type of residential setting (urban, suburban and rural). Results from the mutually adjusted and fully corrected models were presented in tables as Hazard Ratios (HR) with 95% confidence intervals (95% CI), where statistical significance was accepted if 95% CI excluded the null value of 1. Results from stratified

analyses were presented graphically through the use of modified scatter plots to display HRs added with range plot with capped spikes to show 95% CI.

The diagnostics involved a trend test using orthogonal polynomial contrasts to assess whether the hazards for GIT cancer increased linearly with exposure categories going higher for each soil metal. P-values were generated whereby a value < 0.05 indicates that the increase in hazard ratios across increasing exposure categories adequately (or crudely) follows a linear pattern[194]. In addition, we assessed whether the hazards of exposures and confounding variables were proportional using a test based on the Schoenfeld's residuals, which provides p-values to determine non-proportionality overall or per exposure category[195,196]. Where violation of the assumption was identified (p-values less than 0.05) - Aalen plots in a form spline were generated to identify the time points where the hazard function changed significantly[168]. Indicator variables were then created on the cut points where there is an obvious change in the pattern of the hazard function, and re-fitted into the multivariable regression model. This procedure eliminates any time-varying effect that is present in the exposure.

For stage 2, all statistical analyses were performed in Stata 12.0 MP for Windows 7.0 (Stata Corporation, Station College, Texas, USA).

6.4.2.3 Sensitivity analysis using multivariate competing risk models (Stage 3)

We carried out a sensitivity analysis to see if the patterns of risk detected in our previous models differed according to the type of cancer (upper GIT vs. stomach vs. colorectal cancer). In order to obtain site-specific estimates, we used competing risk survival models (as described by Fine and Gray[252,253]). As noted previously, our case definition required a first ever cancer as it is difficult to distinguish between subsequent primary diagnoses, metastases, and follow up visits for the first cancer. When considering specific sites, it is necessary to censor patients who develop a cancer at another site on the date of this diagnosis, as they can no longer develop a first ever cancer in the specific site of interest. This competing risk may artificially diminish the observed hazard ratio at the site of interest. Competing risks models attempt to remove this bias by estimating the sub-hazard ratio in the site of interest alone.

For each site, the outcome was categorised as described in section 6.4.1.1 (no outcome, outcome at site of interest, outcome at other site – treated as competing risk). The results presented for this analysis were sub-hazard ratios (SHR) with 95% confidence intervals and an overall p-value for trend for each soil metal. All models were adjusted for the same confounding variables as shown in the previous corrected Cox models. Note that SHRs are not directly comparable to the HRs from stage 2 as they are derived from a subgroup of

outcomes. Furthermore, SHRs are not directly comparable between models (e.g. SHRs for upper GIT cancers will inhabit a different scale from those for stomach or bowel cancers); they can only be compared within a single model. The main estimates of interest were therefore the p-values for trend, which can be compared with those derived from the corrected model. For stage 3, all statistical analyses were performed in Stata 12.0 MP for Windows 7.0 (Stata Corporation, Station College, Texas, USA).

6.5 Results

6.5.1 Demographic characteristics

The study was based on a prospective UK cohort of 1,812,477 individuals in the English and Welsh areas. The total time for follow-up in our study population between 2004 and 2013 was 17.3 million years. Among the participants, 17,477 developed a GIT cancer – these have an average duration of follow-up time of 4.89 years (SE: 0.021, IQR: 2.45–7.31). Out of 17,477 individuals who developed GIT cancer, malignant neoplasms of the large intestine (colon) were commonly recorded in THIN (7,461, 42.8%); followed by cancer of the rectum and anal canal (3,862, 22.2%); oesophageal cancer (2,867, 16.4%); stomach cancer (1,675, 9.6%); oral cancer (1,406, 8.1%); and cancer of the small intestine (ileum) (166, 0.95%).

The overall demographic characteristics of the study population in relation GIT cancer are presented in (Table 6.1). It indicates that

proportion of those diagnosed with GIT cancers was greater in men (58.7%) than women (41.3%). It also shows that participants developing GITs were more likely to be in the older age groups (≥61 years), from the least deprived groups (group I: 28.1%; group II: 23.7%), whose a current drinker (moderate & hazardous: 52.8%) and some whose ever-smoked (current & ex-smoker: 45.0%).

Table 6.1: Baseline demographic characteristics of participants for GIT cancer study, using THIN-GBASE database from 01-Jan-2004 to 31-Dec-2013

Characteristics	Without GIT cancer (1,800,040)		With GIT cancer (17,477)		Total (1,817,477)	
	N	%	N	%	N	%
Sex						
Male	885,737	49.2	10,227	58.7	895,964	49.3
Female	914,303	50.8	7,210	41.3	921,513	50.7
Agegroup						
≤40	670,099	37.2	406	2.3	670,505	36.9
41-50	336,678	18.7	1,366	7.8	338,044	18.6
51-60	310,924	17.3	3,471	19.9	314,395	17.3
61-70	222,079	12.3	5,130	29.4	227,209	12.5
71-80	163,577	9.1	5,040	28.9	168,617	9.3
81+	96,683	5.4	2,024	11.6	98,707	5.4
BMI						
<18.5	46,205	2.6	334	1.9	46,539	2.6
18.5-24.9	605,418	33.6	5,167	29.6	610,585	33.6
25-29.9	463,488	25.7	5,907	33.9	469,395	25.8
30+	261,709	14.5	3,217	18.4	264,926	14.6
Unknown	423,220	23.5	2,812	16.1	426,032	23.4
Alcohol drinking patterns						
Never	224,903	12.5	2,372	13.6	227,275	12.5
Ex-drinker	20,428	1.1	293	1.7	20,721	1.1

Moderate	766,496	42.6	7,661	43.9	774,157	42.6
Hazardous	123,493	6.9	1,552	8.9	125,045	6.9
Harmful	18,081	1	339	1.9	18,420	1
Unknown	646,639	35.9	5,220	29.9	651,859	35.9
Smoking status						
Never	562,733	31.3	4,983	28.6	567,716	31.2
Non	281,787	15.7	2,726	15.6	284,513	15.7
Ex-smoker	247,682	13.8	4,408	25.3	252,090	13.9
Current	385,027	21.4	3,429	19.7	388,456	21.4
Unknown	322,811	17.9	1,891	10.8	324,702	17.9
Socioeconomic deprivation						
1st	514,246	28.6	4,903	28.1	519,149	28.6
2nd	398,736	22.2	4,141	23.7	402,877	22.2
3rd	363,350	20.2	3,466	19.9	366,816	20.2
4th	309,415	17.2	3,047	17.5	312,462	17.2
5th	199,567	11.1	1,788	10.3	201,355	11.1
Unknown	14,726	0.8	92	0.5	14,818	0.8
Residential setting						
Urban	1,429,290	79.4	13,515	77.5	1,442,805	79.4
Suburban	215,818	12	2,406	13.8	218,224	12
Rural	140,981	7.8	1,449	8.3	142,430	7.8
Unknown	13,951	0.8	67	0.4	14,018	0.8
Health authorities						
London	231,889	12.9	1,693	9.7	233,582	12.9

East Midlands	58,761	3.3	600	3.4	59,361	3.3
East of England	137,463	7.6	1,312	7.5	138,775	7.6
West Midlands	203,723	11.3	1,933	11.1	205,656	11.3
North East	61,121	3.4	677	3.9	61,798	3.4
North West	219,452	12.2	2,470	14.2	221,922	12.2
Yorkshire & Humber	40,051	2.2	438	2.5	40,489	2.2
South Central	292,507	16.3	2,677	15.4	295,184	16.2
South East Coast	226,854	12.6	2,079	11.9	228,933	12.6
South West	196,679	10.9	2,144	12.3	198,823	10.9
Wales	131,540	7.3	1,414	8.1	132,954	7.3

6.5.2 **Exploratory analysis for soil elements**

The greedy stepwise search algorithm for CFS found the following metals to be the most appropriate subset for modelling the risk of GIT cancer: aluminium, calcium, lead, manganese, phosphorus, uranium and zinc. Being that the elements were selected in successive order of importance – the result indicated that aluminium was the most important attribute present in the subset, followed by phosphorus, and zinc as the third. The attribute with the lowest importance was lead. The overall merit score was 0.0234 (or 2.34%) which indicated the error rate for selecting, and including attributes in a subset for model construction was low. The remaining elements not included were arsenic, chromium, copper, iron, nickel, selenium, silicon and vanadium (Table 6.2).

The median concentrations for the selected soil metals among cases with GIT cancer (*n* = 17,477) were as follows: aluminium (median: 52,300 mg/kg, IQR: 41,700-60,000 mg/kg); calcium (median: 5,600 mg/kg, IQR: 3,200-13,200 mg/kg); lead (median: 69.0 mg/kg, IQR: 46.0-137.0 mg/kg); manganese (median: 499.0 mg/kg, IQR: 373.0-790.0 mg/kg); phosphorus (median: 960.0 mg/kg, IQR: 724.0-1,301.0 mg/kg); uranium (median: 2.04 mg/kg, IQR: 1.63-2.45 mg/kg); and zinc (median: 94.0 mg/kg, IQR: 65.0-148.0 mg/kg) (Table 6.3). When comparing the soil concentration levels for elements that may potentially be harmful (or carcinogenic) – it was observed that the

median soil concentration levels for aluminium and uranium only were higher in cases than in controls.

Table 6.2: Show the soils elements selected for model construction and the order of sequence in which the subset were generated using the Correlation-based Filter Selection method

Sequence	Order of selected attribute	Subset generated	
	Start set: no attribute	-	[Start]
1	Aluminium	{Aluminium}	
2	Phosphorus	{Aluminium, Phosphorus}	
3	Zinc	{Aluminium, Phosphorus, Zinc}	
4	Uranium	{Aluminium, Phosphorus, Zinc, Uranium}	
5	Calcium	{Aluminium, Phosphorus, Zinc, Uranium, Calcium}	
6	Manganese	{Aluminium, Phosphorus, Zinc, Uranium, Calcium, Manganese}	
7	Lead	{Aluminium, Phosphorus, Zinc, Uranium, Calcium, Manganese, Lead}	[End]
		Overall merit score (i.e. error rate) 0.0234	
	Excluded attributes:	Arsenic, Chromium, Copper, Iron, Nickel, Selenium, Silicon, Vanadium	

Table 6.3: Showing the residential soil's average (arithmetic mean) and median (with interquartile ranges) concentration levels for selected metallic elements among study population from THIN-GBASE data

Soil metallic element	Without GIT cancer		With GIT cancer	
	Mean (SE)	Median (IQR)	Mean (SE)	Median (IQR)
Aluminium [mg/kg]	49,981.5 (12.1)	51,400.0 (40,500-59,300)	50,959.3 (121.2)	52,300.0 (41,700-60,000)
Calcium [mg/kg]	15,369.4 (20.9)	6,600.0 (3,300-14,300)	14,987.4 (214.2)	5,600.0 (3,200-13,200)
Lead [mg/kg]	126.2 (0.11)	72.0 (47-158)	114.4 (0.93)	69.0 (46-137)
Manganese [mg/kg]	613.4 (0.29)	492.0 (380-765)	623.8 (3.11)	499.0 (373-790)
Phosphorus [mg/kg]	1084.5 (0.40)	986.0 (729-1352)	1057.8 (3.98)	960.0 (724-1301)
Uranium [mg/kg]	2.10 (0.00048)	1.98 (1.58-2.40)	2.06 (0.00493)	2.04 (1.63-2.45)
Zinc [mg/kg]	132.7 (0.099)	99.0 (67-164)	124.4 (0.927)	94.0 (65-148)

6.5.3 Results for overall gastrointestinal cancers

6.5.3.1 Mutually adjusted Cox multivariable regression model

Before deriving the hazard ratios for the soil exposure categories for each of the selected elements, we carried out a diagnostic test based on the Schoenfeld's residuals to assess the proportional-hazards assumption. It shows no evidence that the model specification for the subset of selected elements (aluminium, calcium, lead, manganese, phosphorus and zinc) violates the proportional-hazards assumption (global test: p-value = 0.40). In addition, all p-values for each category for the individual soil elements were non-significant ($p > 0.05$) (Table 6.4).

Our mutually adjusted model containing the subset of soil metals only suggested that there was no evidence of increased, nor a decreased risk of GIT cancer for participants living in areas with soil containing calcium, lead and manganese (Table 6.5). However, it was observed that participants living on residential soil with elevated concentrations of uranium were likely to have an increased risk of developing GIT cancer (Group II: HR 1.10, 95% CI: 1.05-1.17; Group III: HR 1.11, 95% CI: 1.06-1.18; Group IV: HR 1.09, 95% CI: 1.03-1.16; Group V: HR 1.15, 95% CI: 1.08-1.22). Especially, those with the highest exposure (Group V: 2.5-61.2 uranium mg/kg of topsoil) were at a 15% increased risk of developing the malignancy compared to those in the lowest exposure (Group I: < 1.49 uranium mg/kg of

topsoil). The trends test indicates that the pattern of increase in risk of GIT cancer. Although, the patterns were increasing non-linear, the risk patterns appeared to be significant for elevated exposure groups for soil uranium ($p < 0.001$) (Table 6.5).

There was an increased risk of GIT cancer for participants living in areas with medium to highest soil concentration levels of aluminium (Group III: HR 1.07, 95% CI: 1.01-1.14; Group IV: HR 1.09, 95 CI%: 1.03-1.17; Group V: HR 1.07, 95% CI: 1.01-1.15). The inspection of HRs shows that the increased risk of GIT cancer in groups III to V remains stationary between 7.0-9.0%. The overall pattern in terms of increased risk was significant ($p < 0.001$); such pattern was probably captured in the apparent increase that occurs between exposure group II and IV (Table 6.5).

Evidence of an increased risk of GIT cancer was seen for soil phosphorus, whereby the apparent risks appeared to be confined to residents living on soil with medium concentration levels (Group II: HR 1.09, 95% CI: 1.04-1.15; Group III: HR 1.12, 95% CI: 1.07-1.19; Group III: HR 1.08, 95% CI: 1.02-1.15). However, due the inverted U-shape seen for the exposure categories for phosphorus and such trends in elevated risk of GIT cancer were rendered non-significant ($p = 0.36$) (Table 6.5).

Participants in areas with elevated soil concentrations of zinc were at a significantly reduced risk of developing GIT cancer (Group II: HR 0.91, 95% CI: 0.85-0.96; Group III: HR 0.83, 95% CI: 0.78-0.85; Group

IV: HR 0.85, 95% CI: 0.79-0.92; Group V: HR 0.83, 95% CI: 0.76-0.90). With the exception of group II exposures (60.0-85.0 zinc mg/kg of topsoil), it is apparent that the reduced effect of soil zinc on GIT cancer risk remain stationary between 15.0-17.0% across exposure groups III to V. Although the patterns in HRs seem to appear non-linear between groups I-III, and plateaued after zinc concentrations of above 85.0 mg/kg; the overall reduction in risk patterns was significant ($p < 0.001$) (Table 6.5).

Table 6.4: Test of proportional-hazards assumption for mutually adjusted model for assessing risk of gastrointestinal tract cancer with the selected group of soil metals using the Schoenfeld's residual test

Soil element	Al[1]	Ca[2]	Pb[3]	Mn[4]	P[5]	U[6]	Zn[7]
	p-value	p-value	p-value	p-value	p-value	p-value	p-value
Exposure groups							
Group I	-	-	-	-	-	-	-
Group II	0.44	0.33	0.41	0.22	0.16	0.51	0.97
Group III	0.18	0.39	0.93	0.32	0.23	0.32	0.47
Group IV	0.75	0.70	0.71	0.80	0.66	0.79	0.94
Group V	0.39	0.71	0.42	0.77	0.41	0.18	0.23

[1] Soil aluminium **[Al]** (mg/kg) was categorised as quintiles: group I (Al < 37,100), group II (37,100 ≤ Al < 47,200), group III (47,200 ≤ Al < 54,700), group IV (54,700 ≤ Al < 61,600) and group V (Al ≥ 61,600); [2] Soil calcium **[Ca]** (mg/kg) was categorised as quintiles: group I (Ca < 3,000), group II (3,000 ≤ Ca < 4,700), group III (4,700 ≤ Ca < 8,800), group IV (8,800 ≤ Ca < 17,100) and group V (Ca ≥ 17,100); [3] Soil lead **[Pb]** (mg/kg) was categorised as quintiles: group I (Pb < 44.0), group II (44.0 ≤ Pb < 60.0), group III (60.0 ≤ Pb < 95.0), group IV (95.0 ≤ Pb < 184.0) and group V (Pb ≥ 184.0); [4] Soil manganese **[Mn]** (mg/kg) was categorised as quintiles: group I (Mn < 345), group II (345 ≤ Mn < 444), group III (444 ≤ Mn < 572), group IV (572 ≤ Mn < 867) and group V (Mn ≥ 867); [5] Soil phosphorus **[P]** (mg/kg) was categorised as quintiles: group I (P < 680), group II (680 ≤ P < 873), group III (873 ≤ P < 1,127), group IV (1,127 ≤ P < 1,456) and group V (P ≥ 1,456); [6] Soil uranium **[U]** (mg/kg) was categorised as quintiles: group I (U < 1.49), group II (1.49 ≤ U < 1.83), group III (1.83 ≤ U < 2.16), group IV (2.16 ≤ U < 2.50) and group V (U ≥ 2.50); [7] Soil zinc [**Zn**] (mg/kg) was categorised as quintiles: group I (Zn < 60.0), group II (60.0 ≤ Zn < 85.0), group III (85.0 ≤ Zn < 115.0), group IV (115.0 ≤ Zn < 186.0) and group V (Zn ≥ 186.0) – for completeness – its left in

Table 6.5: Using mutually adjusted multivariable Cox regression model to estimate hazard ratios (HR) for gastrointestinal tract (GIT) cancer in association with aluminium, calcium, lead, manganese, phosphorus, uranium and zinc, using THIN-GBASE database from 01-Jan-2004 to 31-Dec-2013

Soil element	Al[1]	Ca[2]	Pb[3]	Mn[4]	P[5]	U[6]	Zn[7]
	HR (95% CI)	HR (95% CI)	HR (95% CI)	HR (95% CI)	HR (95% CI)	HR (95% CI)	HR (95% CI)
Exposure groups							
Group I	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Group II	0.93 (0.88-1.00)	1.02 (0.97-1.08)	0.95 (0.91-1.00)	0.92 (0.87-0.97)	1.09 (1.04-1.15)	1.10 (1.05-1.17)	0.91 (0.85-0.96)
Group III	1.07 (1.01-1.14)	0.98 (0.94-1.04)	1.04 (0.98-1.09)	0.95 (0.90-1.01)	1.12 (1.07-1.19)	1.11 (1.06-1.18)	0.83 (0.78-0.89)
Group IV	1.09 (1.03-1.17)	0.94 (0.88-1.00)	1.01 (0.95-1.07)	0.99 (0.93-1.04)	1.08 (1.02-1.15)	1.09 (1.03-1.16)	0.85 (0.79-0.92)
Group V	1.07 (1.01-1.15)	0.99 (0.94-1.05)	0.94 (0.87-1.01)	1.01 (0.95-1.07)	1.04 (0.97-1.11)	1.15 (1.08-1.22)	0.83 (0.76-0.90)
Trend test	p < 0.001	p = 0.14	p = 0.39	p = 0.18	p = 0.36	p < 0.001	p < 0.001

Hazard ratio (HR); 95% Confidence Interval (95% CI)

[1] Soil aluminium **[Al]** (mg/kg) was categorised as quintiles: group I (Al < 37,100), group II (37,100 ≤ Al < 47,200), group III (47,200 ≤ Al < 54,700), group IV (54,700 ≤ Al < 61,600) and group V (Al ≥ 61,600); [2] Soil calcium **[Ca]** (mg/kg) was categorised as quintiles: group I (Ca < 3,000), group II (3,000 ≤ Ca < 4,700), group III (4,700 ≤ Ca < 8,800), group IV (8,800 ≤ Ca < 17,100) and group V (Ca ≥ 17,100); [3] Soil lead **[Pb]** (mg/kg) was categorised as quintiles: group I (Pb < 44.0), group II (44.0 ≤ Pb < 60.0), group III (60.0 ≤ Pb < 95.0), group IV (95.0 ≤ Pb < 184.0) and group V (Pb ≥ 184.0); [4] Soil manganese **[Mn]** (mg/kg) was categorised as quintiles: group I (Mn < 345), group II (345 ≤ Mn < 444), group III (444 ≤ Mn < 572), group IV (572 ≤ Mn < 867) and group V (Mn ≥ 867); [5] Soil phosphorus **[P]** (mg/kg) was categorised as quintiles: group I (P < 680), group II (680 ≤ P < 873), group III (873 ≤ P < 1,127), group IV (1,127 ≤ P < 1,456) and group V (P ≥ 1,456); [6] Soil uranium **[U]** (mg/kg) was categorised as quintiles: group I (U < 1.49), group II (1.49 ≤ U < 1.83), group III (1.83 ≤ U < 2.16), group IV (2.16 ≤ U < 2.50) and group V (U ≥ 2.50); [7] Soil zinc [**Zn**] (mg/kg) was categorised as quintiles: group I (Zn < 60.0), group II (60.0 ≤ Zn < 85.0), group III (85.0 ≤ Zn < 115.0), group IV (115.0 ≤ Zn < 186.0) and group V (Zn ≥ 186.0)

6.5.3.2 Corrected Cox multivariable regression model

Testing of the proportional-hazards assumption was performed to ensure that the proportional-hazards assumption remained valid when assessing the hazard ratios for the selected soil elements with the inclusion of confounding factors - including age, gender, BMI, drinking status, smoking status and socioeconomic deprivation.

There was an overall violation of the model assumptions (global test: $p < 0.001$) caused by the following categories in BMI (i.e. category I: <18.5; category III: 25-29.9; and category IV: 30+) and ages (i.e. group IV: 61-70; group V: 71-80; and group VI: 80+ years) (Table 6.6; Table 6.7). With the exception of BMI and age groups, the model specification for the main soil exposure, and remaining confounding factors did not violate the proportional-hazard assumptions, and therefore, further diagnostic test using the Aalen plots were carried out on the identified variables to remove the time-varying effects.

As performed in previous sections (see 5.4.3.2), the time-varying effects for each category affected by non-proportional hazards were identified through Aalen plots to examine the patterns of the estimated cumulative hazards regression coefficients. Visual inspection of the Aalen plots enables parameterisation and removal of time-vary effects through the creation of time-based interval-specific covariates based on the affected categories (Figure 6.1-6.6).

Table 6.6: Test of proportional-hazards assumption for the corrected model which includes confounding factors for assessing the risk of gastrointestinal tract cancer with the selected group soil elements using Schoenfeld's residuals (part one)

Soil element	Al[1]	Ca[2]	Pb[3]	Mn[4]	P[5]	U[6]	Zn[7]
	p-value	p-value	p-value	p-value	p-value	p-value	p-value
Exposure groups							
Group I	-	-	-	-	-	-	-
Group II	0.40	0.23	0.52	0.22	0.10	0.35	0.73
Group III	0.15	0.36	0.91	0.27	0.17	0.39	0.73
Group IV	0.66	0.53	0.83	0.77	0.57	0.90	0.86
Group V	0.30	0.81	0.52	0.73	0.47	0.20	0.27

[1] Soil aluminium **[Al]** (mg/kg) was categorised as quintiles: group I (Al < 37,100), group II (37,100 ≤ Al < 47,200), group III (47,200 ≤ Al < 54,700), group IV (54,700 ≤ Al < 61,600) and group V (Al ≥ 61,600); [2] Soil calcium **[Ca]** (mg/kg) was categorised as quintiles: group I (Ca < 3,000), group II (3,000 ≤ Ca < 4,700), group III (4,700 ≤ Ca < 8,800), group IV (8,800 ≤ Ca < 17,100) and group V (Ca ≥ 17,100); [3] Soil lead **[Pb]** (mg/kg) was categorised as quintiles: group I (Pb < 44.0), group II (44.0 ≤ Pb < 60.0), group III (60.0 ≤ Pb < 95.0), group IV (95.0 ≤ Pb < 184.0) and group V (Pb ≥ 184.0); [4] Soil manganese **[Mn]** (mg/kg) was categorised as quintiles: group I (Mn < 345), group II (345 ≤ Mn < 444), group III (444 ≤ Mn < 572), group IV (572 ≤ Mn < 867) and group V (Mn ≥ 867); [5] Soil phosphorus **[P]** (mg/kg) was categorised as quintiles: group I (P < 680), group II (680 ≤ P < 873), group III (873 ≤ P < 1,127), group IV (1,127 ≤ P < 1,456) and group V (P ≥ 1,456); [6] Soil uranium **[U]** (mg/kg) was categorised as quintiles: group I (U < 1.49), group II (1.49 ≤ U < 1.83), group III (1.83 ≤ U < 2.16), group IV (2.16 ≤ U < 2.50) and group V (U ≥ 2.50); [7] Soil zinc [**Zn**] (mg/kg) was categorised as quintiles: group I (Zn < 60.0), group II (60.0 ≤ Zn < 85.0), group III (85.0 ≤ Zn < 115.0), group IV (115.0 ≤ Zn < 186.0) and group V (Zn ≥ 186.0); [8] **Global test; p-value < 0.001**

Table 6.7: Test of proportional-hazards assumption for the corrected model which includes confounding factors for assessing the risk of GIT cancer with the selected group soil elements using Schoenfeld's residuals (part two)

Confounding variables	Schoenfeld test
	p-value
Sex	
Male (referent)	-
Female	0.06
Age group (years)	
≤ 40 (referent)	-
41-50	0.73
51-60	0.06
61-70	< 0.001
71-80	< 0.001
+81	< 0.001
Body mass index (BMI)	
<18.5	0.002
18.5-24.9 (referent)	-
25.0-29.9	0.001
+30	< 0.001
Unknown	0.12
Smoking status	
Never smoked (referent)	-
Non-smoker	0.69
Ex-smoker	0.83
Current smoker	0.26
Unknown	0.57
Drinking status	
Never (referent)	-
Ex-drinker	0.44
Moderate	0.23
Hazardous	0.21
Harmful	0.28
Unknown	0.59
Socioeconomic deprivation	
Group I (referent)	-
Group II	0.85
Group III	0.30
Group IV	0.15
Group V	0.11
Unknown	0.72

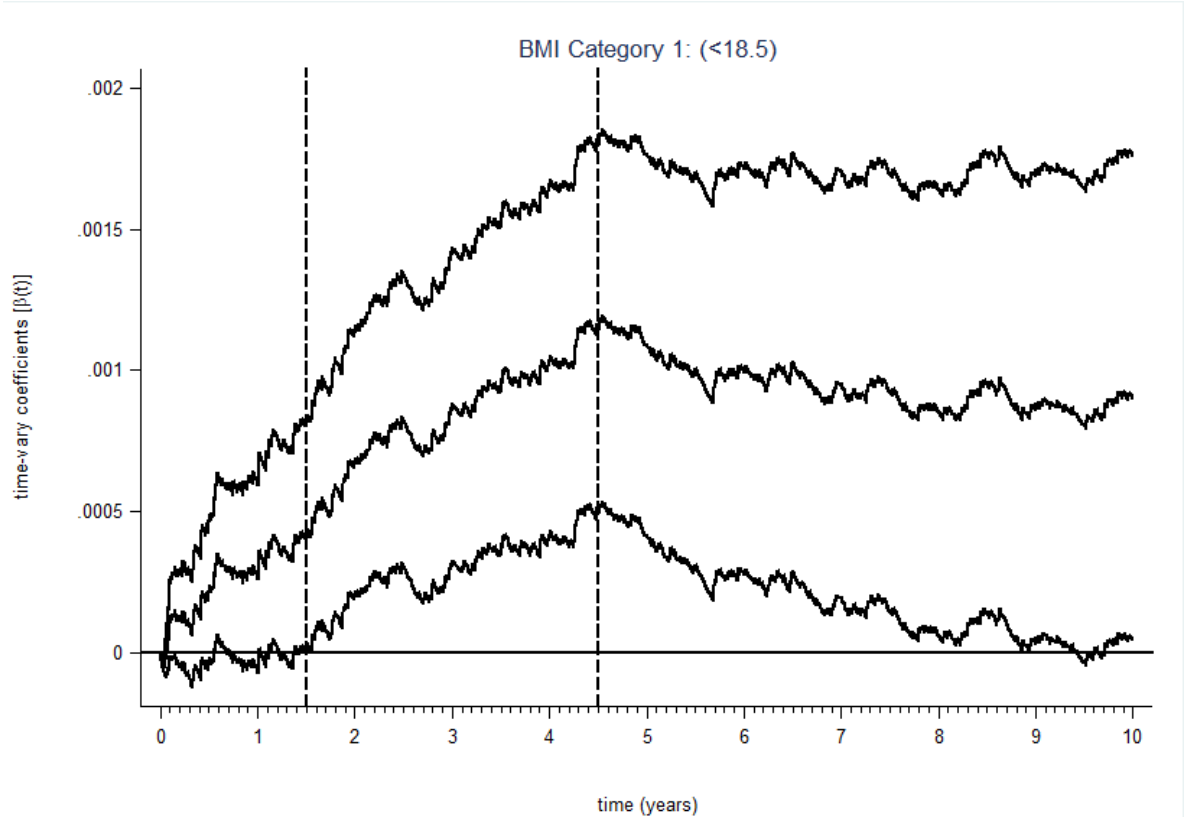

Figure 6.1: Aalen plot showing the estimated cumulative regression coefficients for GIT cancer patients (with 95% confidence interval) with BMI value < 18.5 (versus those with a BMI between 18.5-24.9). The vertical dashed lines are cut-points at 1.5 & 4.5 which show the change in slope of the cumulative hazard function. The following three time-based interval-specific effects for the BMI category were generated using the above cut-points: Early effects (t ≤ 1.5), Middle effects (1.5 < t ≤ 4.5) and late effects (t > 4.5)

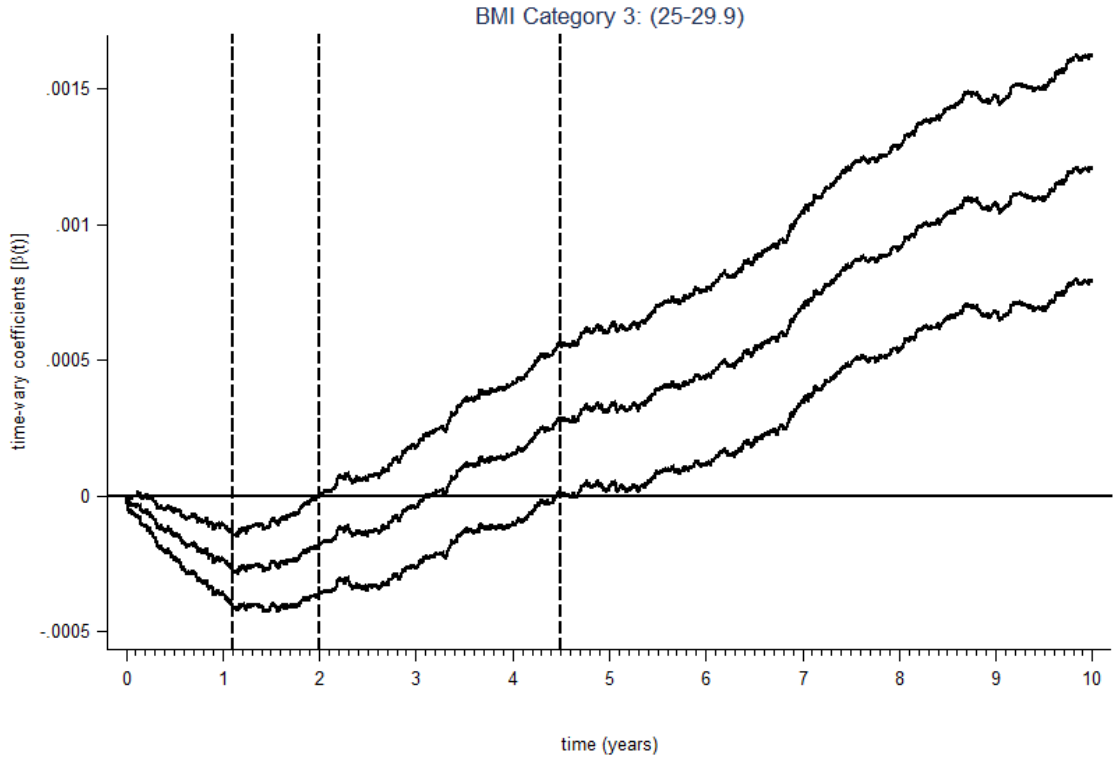

Figure 6.2: Aalen plot showing the estimated cumulative regression coefficients for GIT cancer patients (with 95% confidence interval) with BMI value between 25-29.9 (versus those with a BMI between 18.5-24.9). The vertical dashed lines are cut-points at 1.2, 2.0 & 4.5 which show the change in slope of the cumulative hazard function. The following four time-based interval-specific effects for the BMI category were generated using the above cut-points: Early effects (t ≤ 1.2), early-to-middle effects (1.2 < t ≤ 2.0), middle-late effects (2.0 < t ≤ 4.5) and late effects (>4.5)

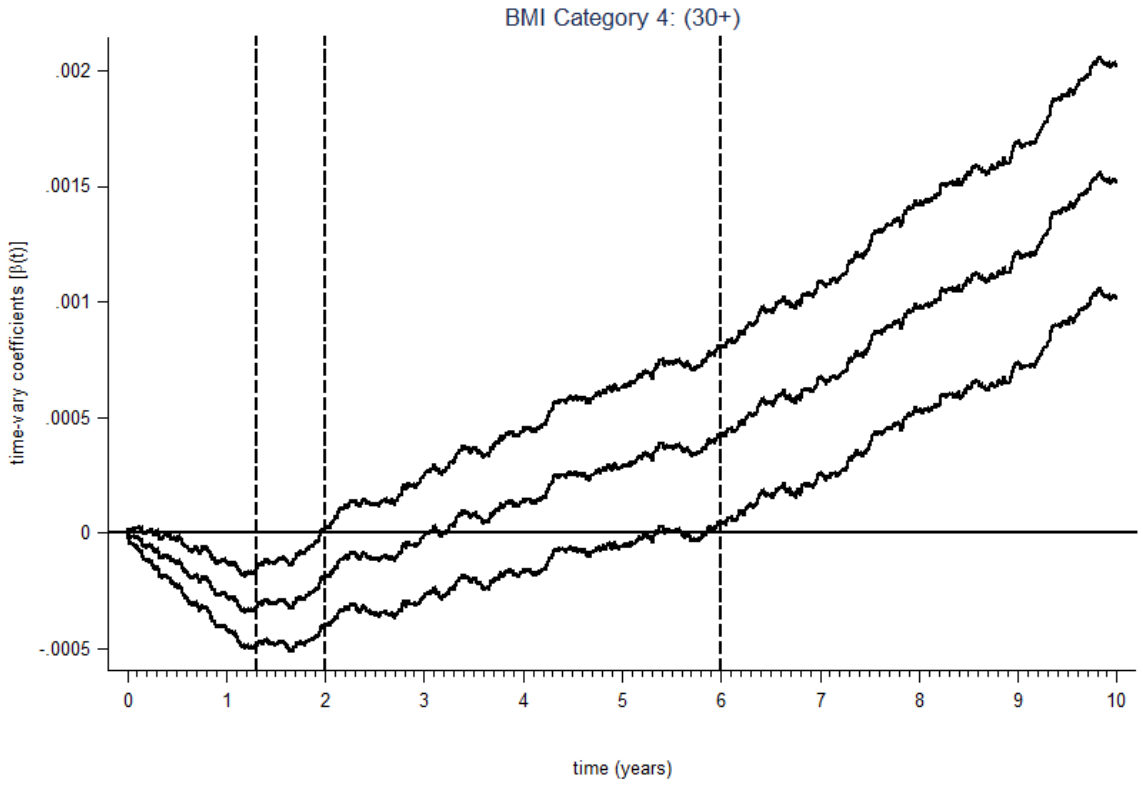

Figure 6.3: Aalen plot showing the estimated cumulative regression coefficients for GIT cancer patients (with 95% confidence interval) with BMI value 30+ (versus those with a BMI between 18.5-24.9). The vertical dashed lines are cut-points at 1.2, 2.0 & 4.5 which show the change in slope of the cumulative hazard function. The following four time-based interval-specific effects for the BMI category were generated using the above cut-points: Early effects (t ≤ 1.2), early-to-middle effects (1.2 < t ≤ 2.0), middle-late effects (2.0 < t ≤ 6.0) and late effects (>6.0)

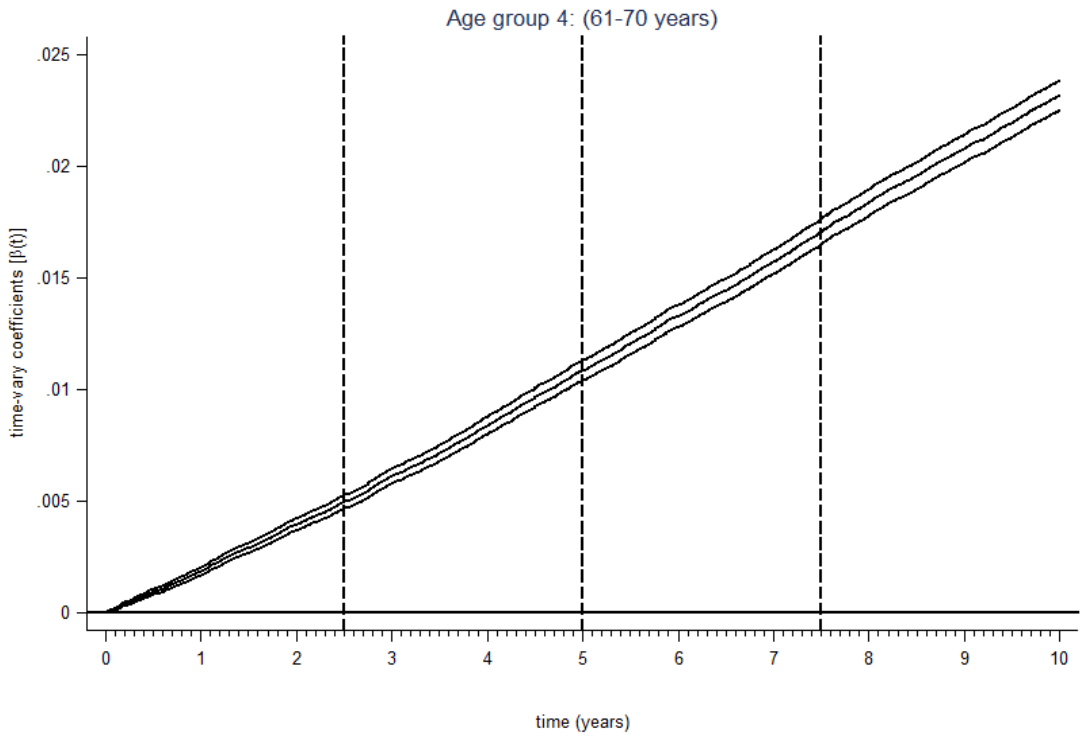

Figure 6.4: Aalen plot showing the estimated cumulative regression coefficients for GIT cancer patients (with 95% confidence interval) who were 61-70 years of age (versus those with ages ≤ 40 years). The changes in the hazard function in the above output were inconspicuous therefore cut-points were specified at a set interval of 2.5 years to eliminate the time-varying effects. The following four time-based interval-specific effects for the age group were generated using the proposed cut-points: Early effects (t ≤ 2.5), early-middle effects (2.5 < t ≤ 5.0), middle-late effects (5.0 < t ≤ 7.5) and late effects (> 7.5)

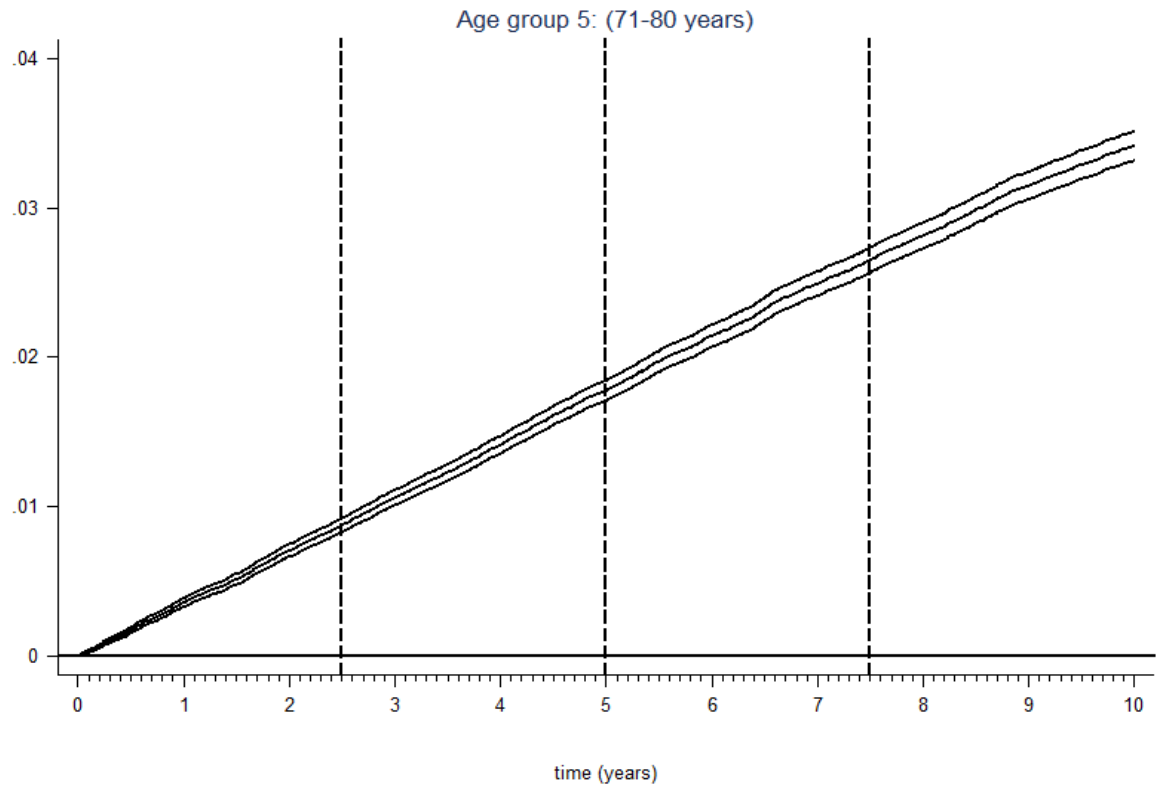

Figure 6.5: Aalen plot showing the estimated cumulative regression coefficients for GIT cancer patients (with 95% confidence interval) who were 71-80 years of age (versus those with ages ≤ 40 years). The changes in the hazard function in the above output were inconspicuous therefore cut-points were specified at a set interval of 2.5 years to eliminate the time-varying effects. The following four time-based interval-specific effects for the age group were generated using the proposed cut-points: Early effects (t ≤ 2.5), early-middle effects (2.5 < t ≤ 5.0), middle-late effects (5.0 < t ≤ 7.5) and late effects (> 7.5)

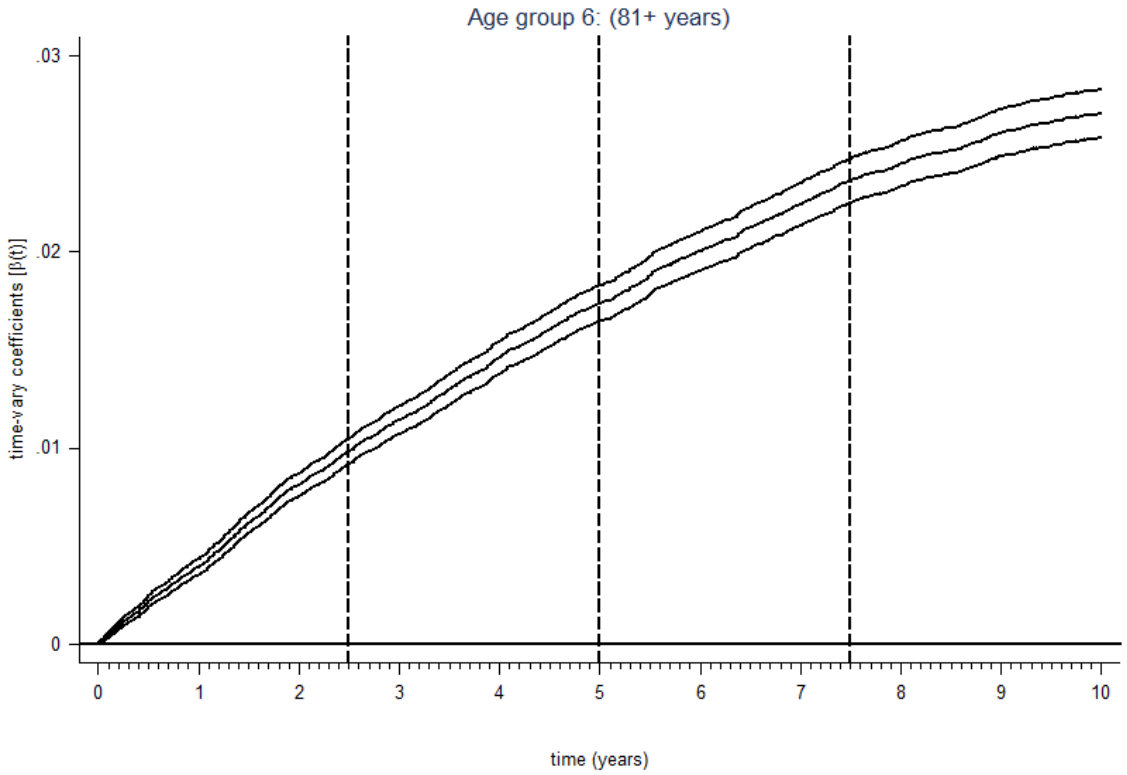

Figure 6.6: Aalen plot showing the estimated cumulative regression coefficients for GIT cancer patients (with 95% confidence interval) who were 81+ years of age (versus those with ages ≤ 40 years). The changes in the hazard function in the above output were inconspicuous therefore cut-points were specified at a set interval of 2.5 years to eliminate the time-varying effects. The following four time-based interval-specific effects for the age group were generated using the proposed cut-points: Early effects (t ≤ 2.5), early-middle effects (2.5 < t ≤ 5.0), middle-late effects (5.0 < t ≤ 7.5) and late effects (> 7.5)

The original categories were replaced and refitted in the corrected models using their time-based interval-specific counterparts to ensure successful removal of any time-varying effects (all variables with $p > 0.05$) (Table 6.8). After adjusting for confounding factors in the corrected model, this lead to marginal reductions in the estimated hazard ratios of the selection of soil elements (Table 6.9). With the exception of soil phosphorus which retained it statistical significance for its exposure groups (Group III: HR 1.08, 95% CI: 1.02-1.14; Group IV: HR 1.07, 95% CI: 1.01-1.13; Group V: HR 1.07, 95% CI: 1.01-1.13); soil aluminium, uranium and zinc estimates were rendered non-significant after inclusion of confounding variables. All other exposures for the remaining elements were not statistically significant. In terms of trends, the important patterns of association that we examined for aluminium and phosphorus did not differ from previous results found in our mutually adjusted model ($p < 0.05$). In contrast, previous patterns that were found to be significant for soil uranium and zinc were negated after including confounding factors ($p > 0.05$) (Table 6.9).

Table 6.8: Test of proportional-hazards assumption for confounding factors in the corrected multivariable Cox regression model using Schoenfeld's residuals after using Aalen plots to remove time-varying effects for age groups and BMI

Confounding variables		P-value
Sex		
Male (referent)		-
Female		0.06
Age group (years)		
≤ 40 (referent)		-
41-50		0.73
51-60		0.06
61-70		< 0.001
(1) early effect	t = [0.0, 2.5)	0.08
(2) early-to-middle effect	t = [2.5, 5.0)	0.24
(3) mid to late effect	t = [5.0, 7.5)	0.25
(4) late effect	t = [7.5, 10)	0.34
71-80		< 0.001
(1) early effect	t = [0.0, 2.5)	0.20
(2) early to mid effect	t = [2.5, 5.0)	0.19
(3) mid to late effect	t = [5.0, 7.5)	0.21
(4) late effect	t = [7.5, 10)	0.24
+81		< 0.001
(1) early effect	t = [0.0, 2.5)	0.17
(2) early to mid effect	t = [2.5, 5.0)	0.18
(3) mid to late effect	t = [5.0, 7.5)	0.29
(4) late effect	t = [7.5, 10)	0.28
Body mass index (BMI)		
<18.5 (category 1)		
(1) early effect	t = [0.0, 1.5)	0.98
(2) mid effect	t = [1.5, 4.5)	0.99
(3) late effect	t = [4.5, 10)	0.61
18.5-24.9 (referent)		-
25.0-29.9 (category 3)		
(1) early effect	t = [0.0, 1.5)	0.45
(2) early to mid effect	t = [1.5, 2.0)	0.79
(3) mid to late effect	t = [2.0, 4.5)	0.56
(4) late effect	t = [4.5, 10)	0.41
+30 (category 4)		
(1) early effect	t = [0.0, 1.5)	0.58
(2) early to mid effect	t = [1.5, 2.0)	0.78
(3) mid to late effect	t = [2.0, 6.0)	0.71
(4) late effect	t = [6.0, 10)	0.28
Unknown (category 5)		0.12
Smoking status		
Never smoked (referent)		-
Non-smoker		0.80
Ex-smoker		0.86
Current smoker		0.25
Unknown		0.60
Drinking status		
Never (referent)		-
Ex-drinker		0.41
Moderate		0.21
Hazardous		0.19
Harmful		0.27
Unknown		0.53
Socioeconomic deprivation		
Group I (referent)		-
Group II		0.85
Group III		0.34
Group IV		0.12
Group V		0.06
Unknown		0.86

Global test: p-value = 0.89

Table 6.9: Using a corrected multivariable Cox regression model to estimate hazard ratios (HR) for GIT cancer in association with aluminium, calcium, lead, manganese, phosphorus, uranium and zinc, using THIN-GBASE database from 01-Jan-2004 to 31-Dec-2013

Soil element	Al[1]	Ca[2]	Pb[3]	Mn[4]	P[5]	U[6]	Zn[7]
	HR (95% CI)	HR (95% CI)	HR (95% CI)	HR (95% CI)	HR (95% CI)	HR (95% CI)	HR (95% CI)
Exposure groups							
Group I	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Group II	0.94 (0.89-1.00)	0.96 (0.91-1.01)	0.99 (0.95-1.04)	0.97 (0.92-1.03)	1.03 (0.98-1.09)	0.99 (0.95-1.05)	0.98 (0.93-1.05)
Group III	1.04 (0.98-1.10)	0.95 (0.91-1.01)	1.08 (1.03-1.14)	0.99 (0.93-1.05)	1.08 (1.03-1.14)	1.00 (0.95-1.06)	0.92 (0.86-0.98)
Group IV	1.07 (1.02-1.14)	0.91 (0.86-0.97)	1.02 (0.96-1.08)	0.96 (0.91-1.02)	1.06 (1.01-1.13)	0.98 (0.92-1.04)	0.94 (0.87-1.01)
Group V	1.06 (0.99-1.13)	0.94 (0.89-1.00)	0.97 (0.90-1.05)	1.01 (0.95-1.06)	1.07 (1.01-1.14)	1.02 (0.97-1.09)	0.93 (0.86-1.02)
Trend test	p = 0.001	p = 0.01	p = 0.81	p = 0.99	p = 0.04	p = 0.63	p = 0.08

Hazard ratio (HR); 95% Confidence Interval (95% CI)
Adjustments include age groups (post-Aalen diagnostics), gender, smoking status, BMI (post-Aalen diagnostics), drinking behaviour and socioeconomic deprivation
[1] Soil aluminium **[Al]** (mg/kg) was categorised as quintiles: group I (Al < 37,100), group II (37,100 ≤ Al < 47,200), group III (47,200 ≤ Al < 54,700), group IV (54,700 ≤ Al < 61,600) and group V (Al ≥ 61,600); [2] Soil calcium **[Ca]** (mg/kg) was categorised as quintiles: group I (Ca < 3,000), group II (3,000 ≤ Ca < 4,700), group III (4,700 ≤ Ca < 8,800), group IV (8,800 ≤ Ca < 17,100) and group V (Ca ≥ 17,100); [3] Soil lead **[Pb]** (mg/kg) was categorised as quintiles: group I (Pb < 44.0), group II (44.0 ≤ Pb < 60.0), group III (60.0 ≤ Pb < 95.0), group IV (95.0 ≤ Pb < 184.0) and group V (Pb ≥ 184.0); [4] Soil manganese **[Mn]** (mg/kg) was categorised as quintiles: group I (Mn < 345), group II (345 ≤ Mn < 444), group III (444 ≤ Mn < 572), group IV (572 ≤ Mn < 867) and group V (Mn ≥ 867); [5] Soil phosphorus **[P]** (mg/kg) was categorised as quintiles: group I (P < 680), group II (680 ≤ P < 873), group III (873 ≤ P < 1,127), group IV (1,127 ≤ P < 1,456) and group V (P ≥ 1,456); [6] Soil uranium **[U]** (mg/kg) was categorised as quintiles: group I (U < 1.49), group II (1.49 ≤ U < 1.83), group III (1.83 ≤ U < 2.16), group IV (2.16 ≤ U < 2.50) and group V (U ≥ 2.50); [7] Soil zinc [**Zn**] (mg/kg) was categorised as quintiles: group I (Zn < 60.0), group II (60.0 ≤ Zn < 85.0), group III (85.0 ≤ Zn < 115.0), group IV (115.0 ≤ Zn < 186.0) and group V (Zn ≥ 186.0)

6.5.3.3 Stratified analysis based on residential settings

The patterns of GIT cancer risk in relation to soil aluminium, calcium, phosphorus and uranium differed depending on the types of residential environment. Aluminium appeared to be the only element which significantly increased the risk of GIT cancer among suburban residents only - whereby the increased risks of GIT in suburban areas appeared to be between 21.0 and 24.0% for exposure groups III, IV & V ($p = 0.011$) (Figure 6.7). For calcium, there was a significant dose-response pattern of a linear nature within the urban residential areas, whereby the risks decreased significantly with increasing concentration levels of soil calcium between exposure groups II-IV ($p = 0.014$) (Figure 6.8). For soil phosphorus, the risks were confined to urban residents where the risk GIT cancer increases as elevated concentration levels for phosphorus reaches up to 1,127 mg/kg, and the risks of GIT cancer appears to plateau ranging between 9.0-12.0% (p-value = 0.034) (Figure 6.11). Finally, there was a significant increase in risk for those living in urban areas with the highest soil concentrations of soil uranium (group V: HR 1.08 95% CI: 1.01-1.17); however, there are no clear risk patterns shown for uranium in other residential settings ($p = 0.07$) (Figure 6.12).

The remaining exposure groups for all other elements – lead (Figure 6.9), manganese (Figure 6.10) and zinc (Figure 6.13) - were non-significant in all different environmental settings, with unclear patterns ($p > 0.05$).

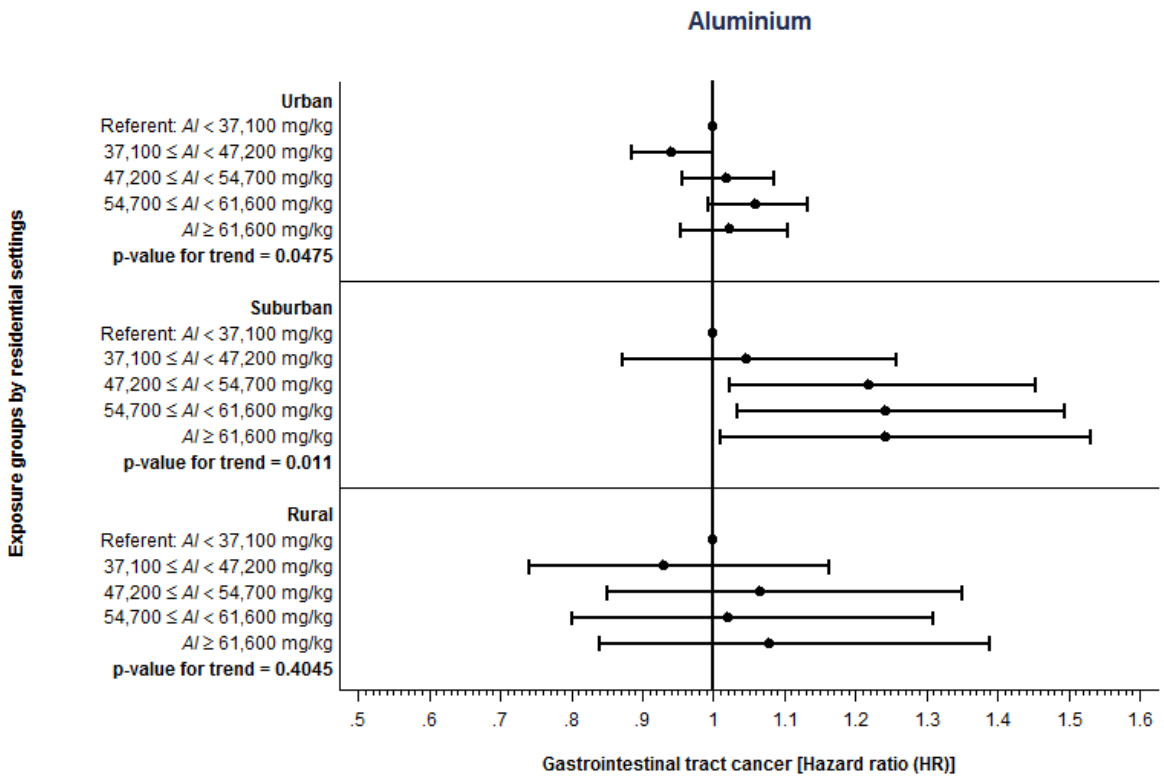

Figure 6.7: Modified scatter plot with range capped spikes were used to represent HR estimates derived from a stratified Cox regression model based on residential classification to show the association between GIT cancer risk and soil aluminium, using THIN-GBASE database 01-01-2004 to 31-12-2013. The stratified model included other primary exposures - soil calcium, lead, manganese, phosphorus, uranium and zinc, and were adjusted for sex, age group, BMI, drinking alcohol status, smoking status, socioeconomic deprivation and categories fitted as time-based interval-specified covariates derived from earlier Aalen plots

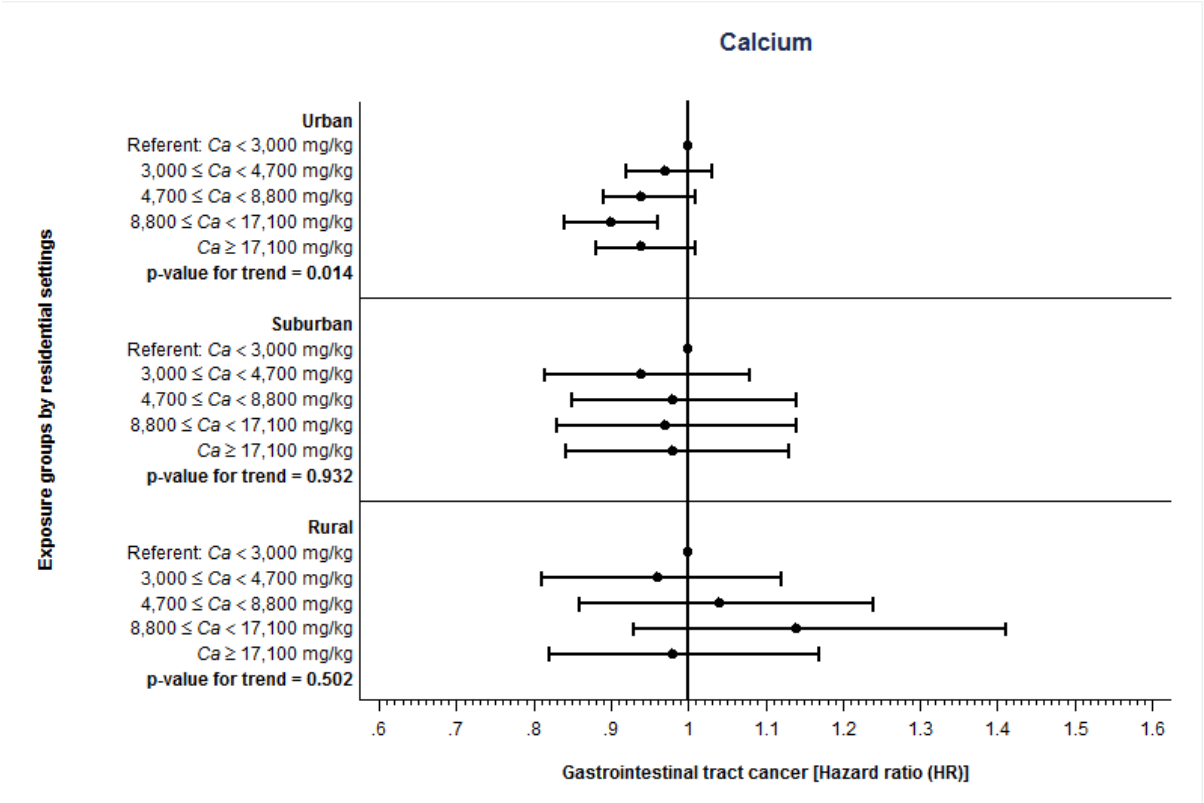

Figure 6.8: Modified scatter plot with range capped spikes were used to represent HR estimates derived from a stratified Cox regression model based on residential classification to show the association between GIT cancer risk and soil calcium, using THIN-GBASE database 01-01-2004 to 31-12-2013. The stratified model included other primary exposures - soil aluminium, lead, manganese, phosphorus, uranium and zinc, and were adjusted for sex, age group, BMI, drinking alcohol status, smoking status, socioeconomic deprivation and categories fitted as time-based interval-specified covariates derived from earlier Aalen plots

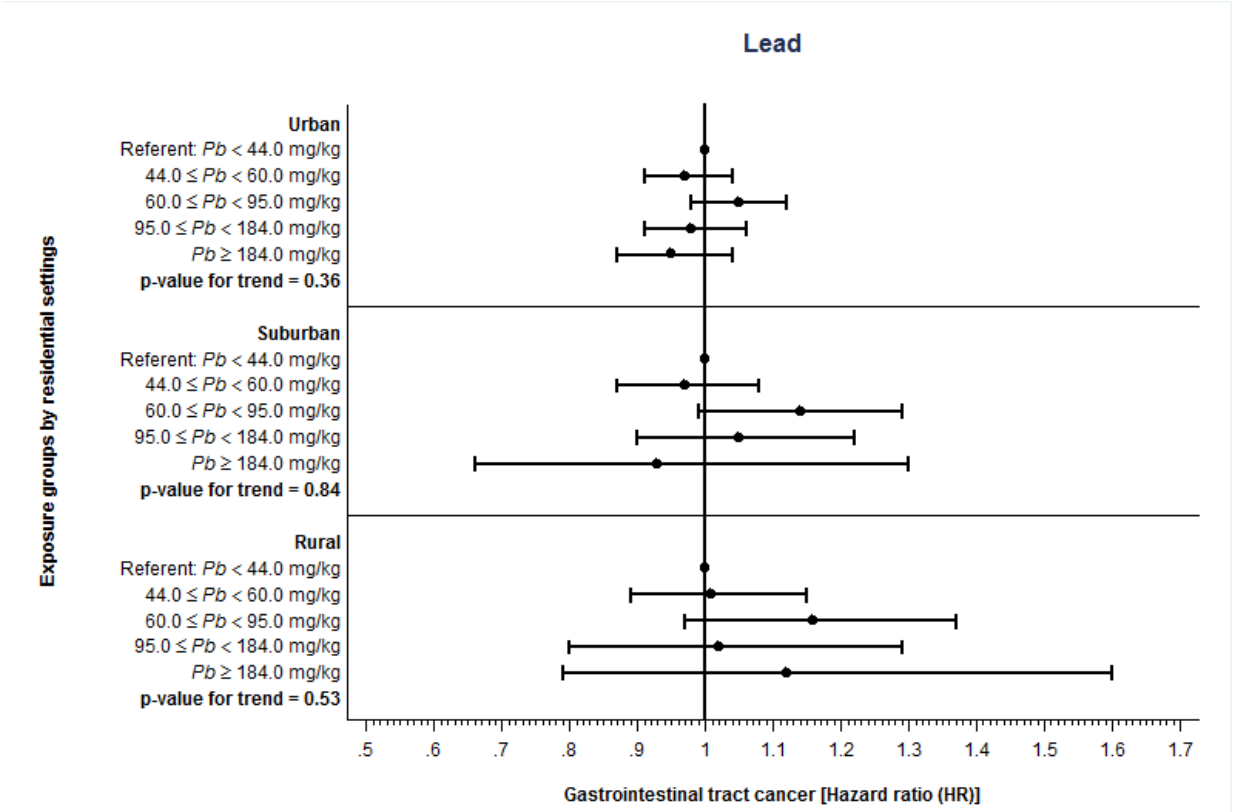

Figure 6.9: Modified scatter plot with range capped spikes were used to represent HR estimates derived from a stratified Cox regression model based on residential classification to show the association between GIT cancer risk and soil lead, using THIN-GBASE database 01-01-2004 to 31-12-2013. The stratified model included other primary exposures - soil aluminium, calcium, manganese, phosphorus, uranium and zinc, and were adjusted for sex, age group, BMI, drinking alcohol status, smoking status, socioeconomic deprivation and categories fitted as time-based interval-specified covariates derived from earlier Aalen plots

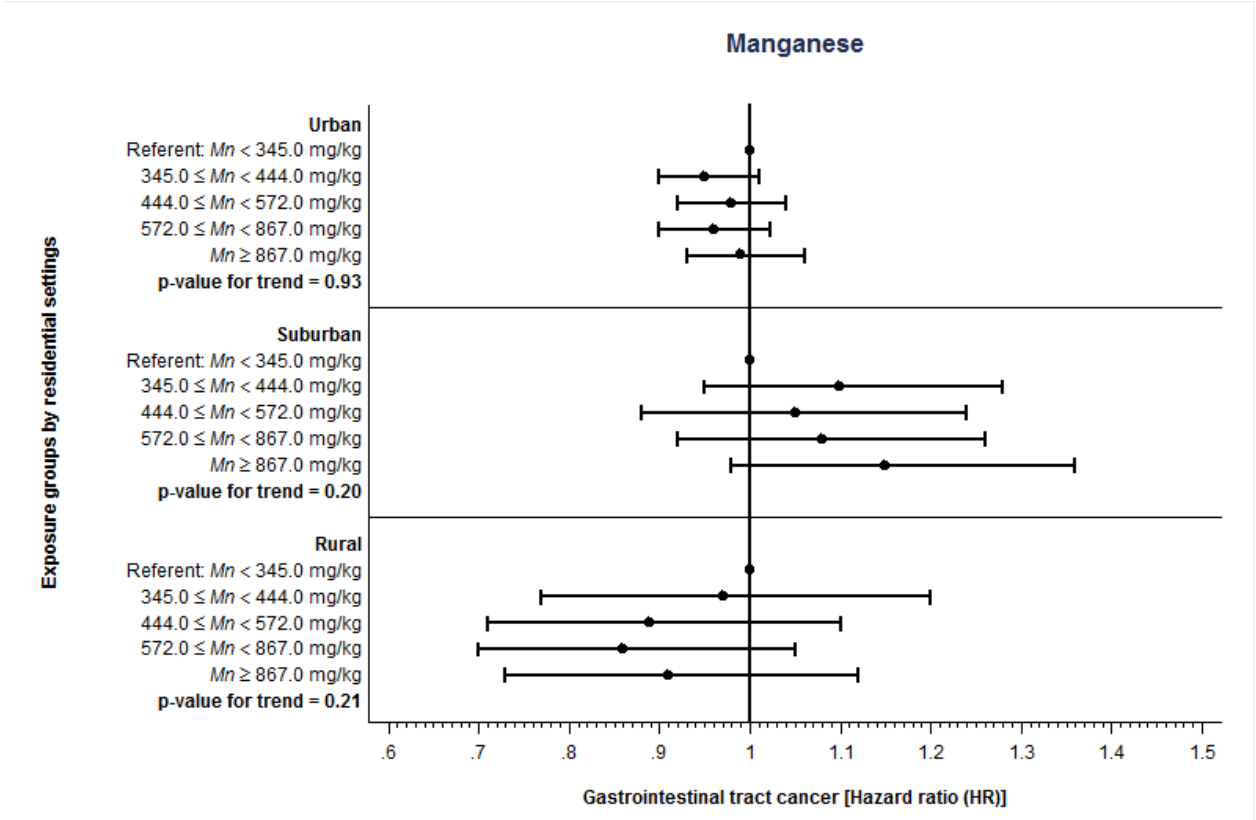

Figure 6.10: Modified scatter plot with range capped spikes were used to represent HR estimates derived from a stratified Cox regression model based on residential classification to show the association between GIT cancer risk and soil manganese, using THIN-GBASE database 01-01-2004 to 31-12-2013. The stratified model included other primary exposures - soil aluminium, calcium, lead, phosphorus, uranium and zinc, and were adjusted for sex, age group, BMI, drinking alcohol status, smoking status, socioeconomic deprivation and categories fitted as time-based interval-specified covariates derived from earlier Aalen plots

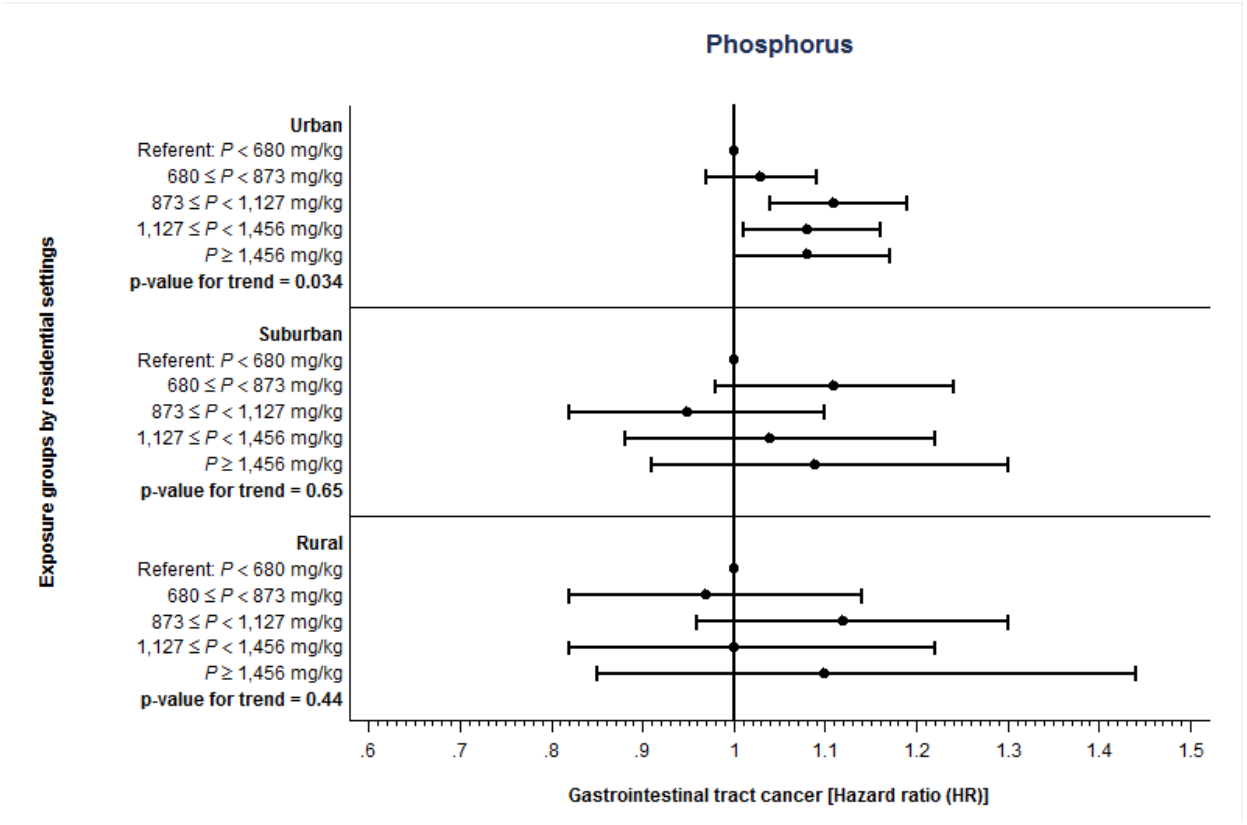

Figure 6.11: Modified scatter plot with range capped spikes were used to represent HR estimates derived from a stratified Cox regression model based on residential classification to show the association between GIT cancer risk and soil phosphorus, using THIN-GBASE database 01-01-2004 to 31-12-2013. The stratified model included other primary exposures - soil aluminium, calcium, lead, manganese, uranium and zinc, and were adjusted for sex, age group, BMI, drinking alcohol status, smoking status, socioeconomic deprivation and categories fitted as time-based interval-specified covariates derived from earlier Aalen plots

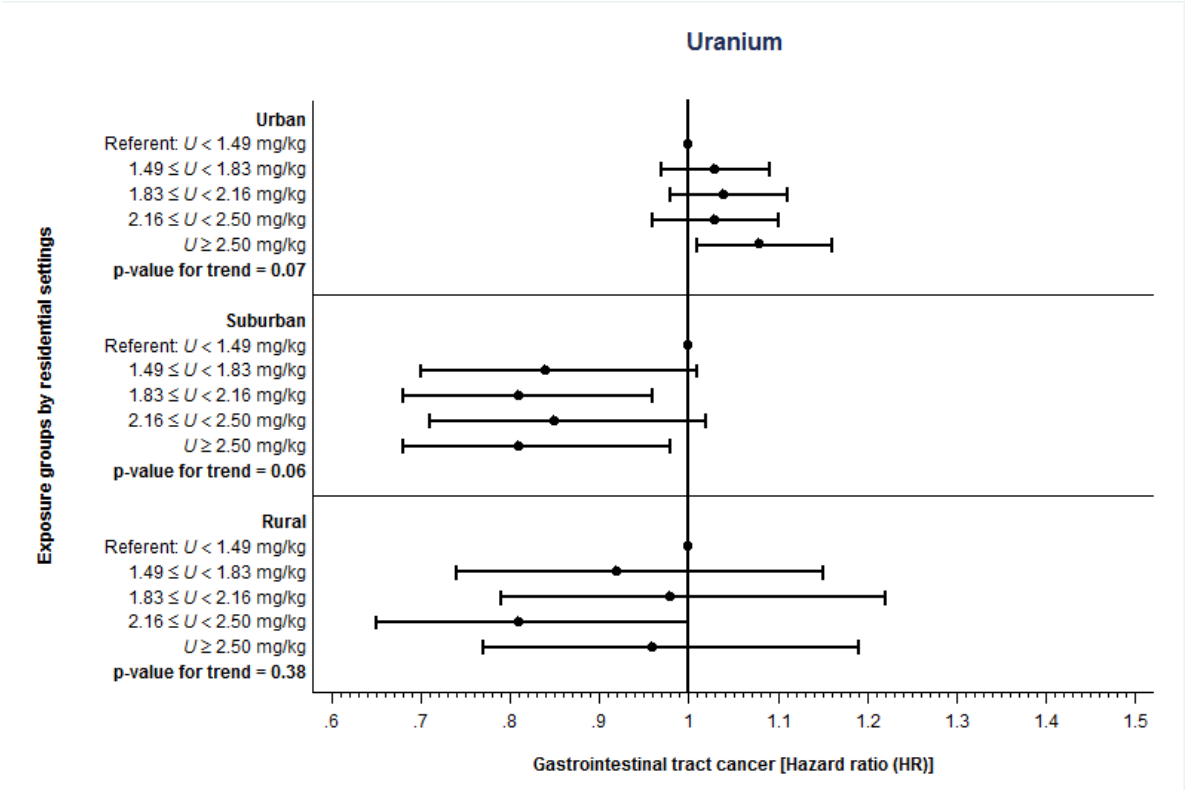

Figure 6.12: Modified scatter plot with range capped spikes were used to represent HR estimates derived from a stratified Cox regression model based on residential classification to show the association between GIT cancer risk and soil uranium, using THIN-GBASE database 01-01-2004 to 31-12-2013. The stratified model included other primary exposures - soil aluminium, calcium, lead, manganese, phosphorus and zinc, and were adjusted for sex, age group, BMI, drinking alcohol status, smoking status, socioeconomic deprivation and categories fitted as time-based interval-specified covariates derived from earlier Aalen plots

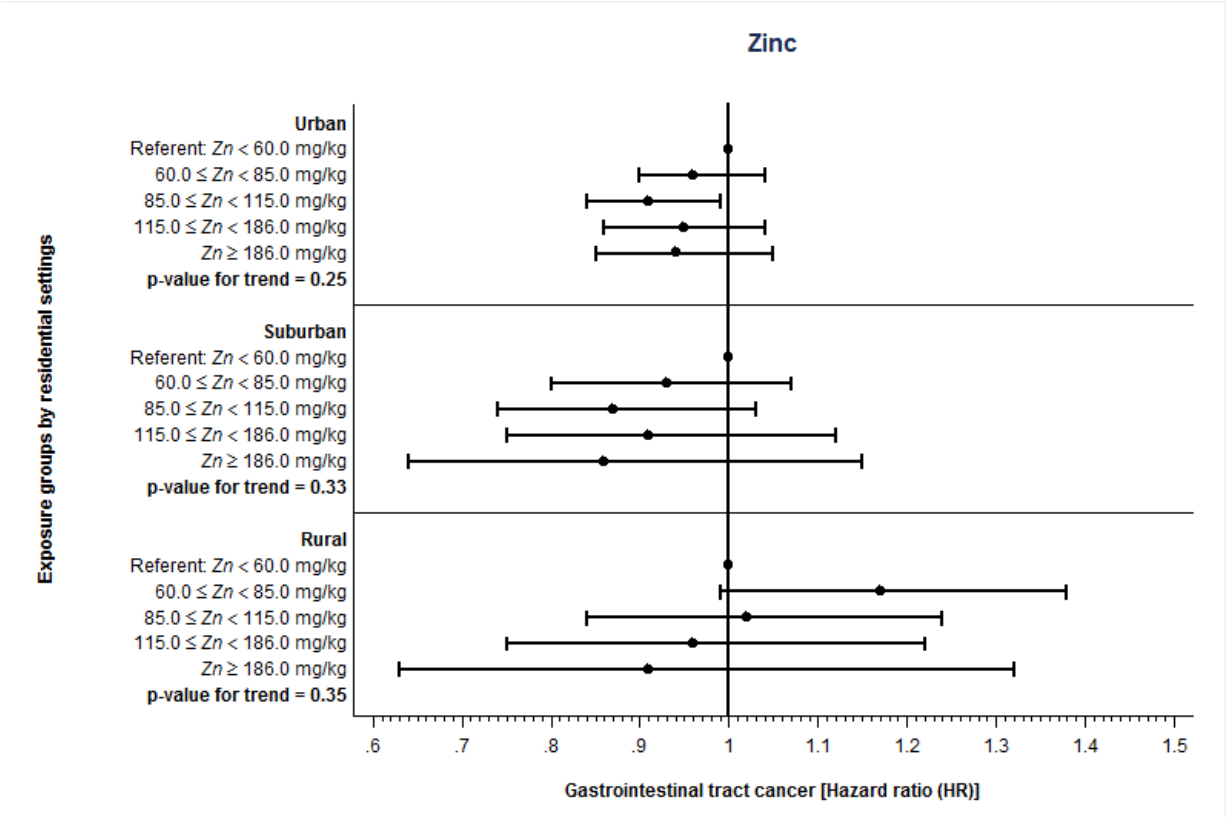

Figure 6.13: Modified scatter plot with range capped spikes were used to represent HR estimates derived from a stratified Cox regression model based on residential classification to show the association between GIT cancer risk and soil zinc, using THIN-GBASE database 01-01-2004 to 31-12-2013. The stratified model included other primary exposures - soil aluminium, calcium, lead, manganese and uranium, and were adjusted for sex, age group, BMI, drinking alcohol status, smoking status, socioeconomic deprivation and categories fitted as time-based interval-specified covariates derived from earlier Aalen plots

6.5.3.4 Sensitivity analyses using competing risk models

When comparing trend p-values derived from competing risk models we found that the direction and patterns of risk for certain elements were more pronounced for particular sub outcomes. For upper GIT cancers, the relationship for soil calcium shows an overall downwards pattern indicating a significant reduction in risk of upper GIT cancer with increased concentrations of soil calcium ($p = 0.03$). This pattern appeared to be consistent with those found in the mutually adjusted model. However, it appears the trends' p-value for aluminium and phosphorus was sensitive to the stratification of outcomes, their previous trend patterns of an increased risk were rendered non-significant for upper GIT cancer ($p= 0.05$ & 0.08, respectively) when negating the competing effects of stomach and bowel cancers.

However, the patterns of risk for aluminium appeared to be non-sensitive when stomach cancer was the primary outcome (vs. upper GIT & bowel cancer). Similar to the correct model, it retains its significant positive association, indicating that increased exposure to soil aluminium may specifically be linked to stomach cancer ($p=0.02$). The trend patterns for the remaining elements were not significant in all models despite the attempt to negate the effects of competing outcomes.

Table 6.10: Using an adjusted multivariate competing risk model to estimate sub-hazard ratios (SHR) for upper GIT cancers in association with selected soil elements, using THIN-GBASE database from 01-Jan-2004 to 31-Dec-2013

Soil element	Al[1]	Ca[2]	Pb[3]	Mn[4]	P[5]	U[6]	Zn[7]
	SHR (95% CI)	SHR (95% CI)	SHR (95% CI)	SHR (95% CI)	SHR (95% CI)	SHR (95% CI)	SHR (95% CI)
Exposure groups							
Group I	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Group II	1.02 (0.91-1.14)	0.98 (0.89-1.09)	1.00 (0.91-1.12)	0.89 (0.80-1.00)	1.02 (0.92-1.13)	1.05 (0.94-1.16)	1.01 (0.89-1.12)
Group III	1.14 (1.02-1.28)	0.93 (0.83-1.03)	1.08 (0.97-1.20)	0.95 (0.85-1.06)	1.04 (0.94-1.17)	1.00 (0.89-1.13)	0.95 (0.84-1.08)
Group IV	1.15 (1.02-1.30)	0.81 (0.72-0.92)	1.06 (0.94-1.19)	0.88 (0.79-0.98)	1.05 (0.93-1.18)	1.01 (0.89-1.13)	0.93 (0.79-1.07)
Group V	1.08 (0.94-1.23)	0.95 (0.85-1.06)	1.03 (0.88-1.20)	0.91 (0.81-1.01)	1.07 (0.93-1.22)	1.05 (0.92-1.18)	0.95 (0.80-1.13)
Trend test	**p = 0.08**	**p = 0.03***	**p = 0.55**	**p = 0.12**	**p = 0.34**	**p = 0.74**	**p = 0.38**

Sub-hazard ratio (SHRs); 95% Confidence Interval (95% CI)

[1] Soil aluminium **[Al]** (mg/kg) was categorised as quintiles: group I (Al < 37,100), group II (37,100 ≤ Al < 47,200), group III (47,200 ≤ Al < 54,700), group IV (54,700 ≤ Al < 61,600) and group V (Al ≥ 61,600); [2] Soil calcium **[Ca]** (mg/kg) was categorised as quintiles: group I (Ca < 3,000), group II (3,000 ≤ Ca < 4,700), group III (4,700 ≤ Ca < 8,800), group IV (8,800 ≤ Ca < 17,100) and group V (Ca ≥ 17,100); [3] Soil lead **[Pb]** (mg/kg) was categorised as quintiles: group I (Pb < 44.0), group II (44.0 ≤ Pb < 60.0), group III (60.0 ≤ Pb < 95.0), group IV (95.0 ≤ Pb < 184.0) and group V (Pb ≥ 184.0); [4] Soil manganese **[Mn]** (mg/kg) was categorised as quintiles: group I (Mn < 345), group II (345 ≤ Mn < 444), group III (444 ≤ Mn < 572), group IV (572 ≤ Mn < 867) and group V (Mn ≥ 867); [5] Soil phosphorus **[P]** (mg/kg) was categorised as quintiles: group I (P < 680), group II (680 ≤ P < 873), group III (873 ≤ P < 1,127), group IV (1,127 ≤ P < 1,456) and group V (P ≥ 1,456); [6] Soil uranium **[U]** (mg/kg) was categorised as quintiles: group I (U < 1.49), group II (1.49 ≤ U < 1.83), group III (1.83 ≤ U < 2.16), group IV (2.16 ≤ U < 2.50) and group V (U ≥ 2.50); [7] Soil zinc [**Zn**] (mg/kg) was categorised as quintiles: group I (Zn < 60.0), group II (60.0 ≤ Zn < 85.0), group III (85.0 ≤ Zn < 115.0), group IV (115.0 ≤ Zn < 186.0) and group V (Zn ≥ 186.0)

* p-value for trend < 0.05

Table 6.11: Using an adjusted multivariate competing risk model to estimate sub-hazard ratios (SHR) for stomach cancers in association with selected soil elements, using THIN-GBASE database from 01-Jan-2004 to 31-Dec-2013

Soil element	Al[1]	Ca[2]	Pb[3]	Mn[4]	P[5]	U[6]	Zn[7]
	SHR (95% CI)	SHR (95% CI)	SHR (95% CI)	SHR (95% CI)	SHR (95% CI)	SHR (95% CI)	SHR (95% CI)
Exposure groups							
Group I	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Group II	0.84 (0.72-1.01)	0.98 (0.84-1.16)	1.04 (0.86-1.23)	0.95 (0.81-1.14)	0.99 (0.99-1.14)	0.98 (0.91-1.04)	0.89 (0.73-1.08)
Group III	0.98 (0.84-1.20)	0.98 (0.82-1.17)	1.27 (1.07-1.52)	0.90 (0.76-1.10)	1.05 (1.05-1.18)	0.99 (0.93-1.08)	0.99 (0.80-1.22)
Group IV	1.04 (0.87-1.28)	0.90 (0.76-1.10)	1.20 (0.99-1.45)	0.98 (0.83-1.18)	0.99 (0.99-1.16)	0.96 (0.89-1.03)	0.97 (0.76-1.24)
Group V	1.17 (0.98-1.47)	0.95 (0.80-1.14)	1.22 (0.95-1.56)	0.90 (0.75-1.10)	0.99 (0.99-1.27)	0.99 (0.93-1.08)	0.86 (0.64-1.13)
Trend test	p < 0.001	p = 0.14	p = 0.39	p = 0.18	p = 0.36	p < 0.001	p < 0.001

Sub-Hazard ratio (HR); 95% Confidence Interval (95% CI)

[1] Soil aluminium **[Al]** (mg/kg) was categorised as quintiles: group I (Al < 37,100), group II (37,100 ≤ Al < 47,200), group III (47,200 ≤ Al < 54,700), group IV (54,700 ≤ Al < 61,600) and group V (Al ≥ 61,600); [2] Soil calcium **[Ca]** (mg/kg) was categorised as quintiles: group I (Ca < 3,000), group II (3,000 ≤ Ca < 4,700), group III (4,700 ≤ Ca < 8,800), group IV (8,800 ≤ Ca < 17,100) and group V (Ca ≥ 17,100); [3] Soil lead **[Pb]** (mg/kg) was categorised as quintiles: group I (Pb < 44.0), group II (44.0 ≤ Pb < 60.0), group III (60.0 ≤ Pb < 95.0), group IV (95.0 ≤ Pb < 184.0) and group V (Pb ≥ 184.0); [4] Soil manganese **[Mn]** (mg/kg) was categorised as quintiles: group I (Mn < 345), group II (345 ≤ Mn < 444), group III (444 ≤ Mn < 572), group IV (572 ≤ Mn < 867) and group V (Mn ≥ 867); [5] Soil phosphorus **[P]** (mg/kg) was categorised as quintiles: group I (P < 680), group II (680 ≤ P < 873), group III (873 ≤ P < 1,127), group IV (1,127 ≤ P < 1,456) and group V (P ≥ 1,456); [6] Soil uranium **[U]** (mg/kg) was categorised as quintiles: group I (U < 1.49), group II (1.49 ≤ U < 1.83), group III (1.83 ≤ U < 2.16), group IV (2.16 ≤ U < 2.50) and group V (U ≥ 2.50); [7] Soil zinc [**Zn**] (mg/kg) was categorised as quintiles: group I (Zn < 60.0), group II (60.0 ≤ Zn < 85.0), group III (85.0 ≤ Zn < 115.0), group IV (115.0 ≤ Zn < 186.0) and group V (Zn ≥ 186.0)

Table 6.12: Using an adjusted multivariate competing risk model to estimate sub-hazard ratios (SHR) for colorectal cancers in association with selected soil elements, using THIN-GBASE database from 01-Jan-2004 to 31-Dec-2013

Soil element	Al[1]	Ca[2]	Pb[3]	Mn[4]	P[5]	U[6]	Zn[7]
	SHR (95% CI)	SHR (95% CI)	SHR (95% CI)	SHR (95% CI)	SHR (95% CI)	SHR (95% CI)	SHR (95% CI)
Exposure groups							
Group I	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Group II	0.93 (0.88-0.99)	0.96 (0.89-1.02)	0.99 (0.93-1.05)	1.02 (0.85-1.08)	1.05 (0.99-1.12)	0.98 (0.91-1.04)	0.99 (0.92-1.06)
Group III	1.01 (0.95-1.08)	0.98 (0.90-1.03)	1.16 (0.99-1.13)	1.03 (0.96-1.10)	1.11 (1.05-1.18)	0.99 (0.93-1.07)	0.89 (0.82-0.96)
Group IV	1.06 (0.98-1.14)	0.96 (0.88-1.03)	0.99 (0.91-1.06)	1.00 (0.94-1.07)	1.07 (0.99-1.15)	0.96 (0.89-1.03)	0.94 (0.85-1.02)
Group V	1.03 (0.96-1.12)	0.95 (0.88-1.00)	0.92 (0.85-1.02)	1.07 (0.99-1.04)	1.08 (0.99-1.16)	0.99 (0.93-1.07)	0.94 (0.84-1.05)
Trend test	p < 0.001	p = 0.14	p = 0.39	p = 0.18	p = 0.36	p < 0.001	p < 0.001

Sub-Hazard ratio (SHR); 95% Confidence Interval (95% CI)

[1] Soil aluminium **[Al]** (mg/kg) was categorised as quintiles: group I (Al < 37,100), group II (37,100 ≤ Al < 47,200), group III (47,200 ≤ Al < 54,700), group IV (54,700 ≤ Al < 61,600) and group V (Al ≥ 61,600); [2] Soil calcium **[Ca]** (mg/kg) was categorised as quintiles: group I (Ca < 3,000), group II (3,000 ≤ Ca < 4,700), group III (4,700 ≤ Ca < 8,800), group IV (8,800 ≤ Ca < 17,100) and group V (Ca ≥ 17,100); [3] Soil lead **[Pb]** (mg/kg) was categorised as quintiles: group I (Pb < 44.0), group II (44.0 ≤ Pb < 60.0), group III (60.0 ≤ Pb < 95.0), group IV (95.0 ≤ Pb < 184.0) and group V (Pb ≥ 184.0); [4] Soil manganese **[Mn]** (mg/kg) was categorised as quintiles: group I (Mn < 345), group II (345 ≤ Mn < 444), group III (444 ≤ Mn < 572), group IV (572 ≤ Mn < 867) and group V (Mn ≥ 867); [5] Soil phosphorus **[P]** (mg/kg) was categorised as quintiles: group I (P < 680), group II (680 ≤ P < 873), group III (873 ≤ P < 1,127), group IV (1,127 ≤ P < 1,456) and group V (P ≥ 1,456); [6] Soil uranium **[U]** (mg/kg) was categorised as quintiles: group I (U < 1.49), group II (1.49 ≤ U < 1.83), group III (1.83 ≤ U < 2.16), group IV (2.16 ≤ U < 2.50) and group V (U ≥ 2.50); [7] Soil zinc [**Zn**] (mg/kg) was categorised as quintiles: group I (Zn < 60.0), group II (60.0 ≤ Zn < 85.0), group III (85.0 ≤ Zn < 115.0), group IV (115.0 ≤ Zn < 186.0) and group V (Zn ≥ 186.0)

6.6 Discussion

This study uses a unique three-stage approach in an attempt to show the association between the risk of developing various GIT cancers and with elevated concentrations of metallic elements present in residential soil.

In stage one; aluminium, calcium, lead, manganese, phosphorus, uranium and zinc were identified as the set of important exposures for modelling GIT cancer risk. In stage two; a mutually adjusted model showed evidence of an increased risk of GIT cancer for participants residing in areas with higher concentrations for soil aluminium, phosphorus and uranium. It also showed a reduction in risk of GIT cancer for participants living on residential soil containing elevated levels soil zinc. However, these findings from the initial model disappeared after taking into account of age, gender, alcohol consumption, smoking status, BMI and levels of socioeconomic deprivation. The only exposure group that retained its statistical significance after adjustments for confounding was soil phosphorus. Further analysis in stage 2 included the usage of stratified models. It showed that the marginal risks appeared to be confined to residents living in urban and suburban areas, whereby they had the highest exposure for soil phosphorus (>1,127 mg/kg) and uranium (> 2.5 mg/kg); and a weak positive dose-relationship for soil aluminium, respectively. Finally, in stage three, our competing risk models had shown a possible association for the following: (1) a trends reduction

in risk of upper GIT cancer in relation to soil calcium; and (2) an increased risk of stomach cancer in association with soil aluminium. There were no meaningful relationships for any of the remaining elements with the sub outcomes when attempting to negate the effects of competing events.

One of the key advantage for using the CFS method was that it provided an analytical, and yet, an objective approach for determining which set of soil metallic elements were appropriate for model construction that would optimally predict GIT cancer risk. The algorithm for this analysis had accurately selected meaningful elements that, in some way, have a plausible relationship with various GIT cancers. For instance, recent studies have shown evidence of the beneficial effects for dietary calcium, manganese and zinc; they have indicated that higher levels and intake of these trace elements in food (or in supplements) were associated with a reduced risk of developing colon cancers.[254-259] Also, the IARC have ranked lead and uranium as carcinogenic agents with the potential to cause a wide spectrum of cancers in humans.[46,197,198] In contrast, aluminium and phosphorous are yet to be recognised by the IARC as carcinogenic agents[46].

For aluminium, an initial positive association was found with GIT cancer after mutually adjusted for other elements. However, this positive relationship disappeared after correcting for confounding factors. The results derived from the corrected model takes precedence over the mutually adjusted model since it takes into

account confounding; therefore, we found no consistent evidence of any links between soil aluminium and overall risk of developing GIT cancer. In addition, our model detected an increasing trend for risk of stomach cancers only. Due to the limitations of this epidemiological research (i.e. residual confounding and bias), further studies are needed to confirm or refute our findings. According to some research, the amounts of aluminium derived from soil and absorbed via GIT tract are considered negligible to cause toxicity.[202,205,260] This assertion is based on empirical studies that have derived oral bioavailability indexes for aluminium based on data that has been theoretically assumed for humans, which have shown that aluminium is derived largely from particulate matter and contaminated drinking water. These study have shown that the oral bioavailability for aluminium from foods were estimated to be between 0.05-0.1%; whereas from ambient air and drinking water, they were >0.1-0.2% and ~0.3%, respectively.[202,205,260] At present, there is no current data showing the uptake of soil aluminium in animals.

A similar observation was found for uranium. The initial model had shown a positive association with GIT cancer with increased exposure of uranium. However, the association disappeared after correcting for other risk factors. The results derived from the corrected model takes precedence over the mutually adjusted model; therefore, we were unable to identify an overall association for GIT cancers with soil uranium. In addition, our sensitivity analyses were unable to pick up any significant trends despite negating the effects of competing

outcomes. The carcinogenic nature of uranium has been established in previous studies;[197,261] however, most studies relating uranium exposures are focused of drinking water. For instance, a study used an ecological design and observed that elevated concentrations of uranium in wells use local residents in rural areas of South Carolina (US) had an increased risk of colorectal, breast, kidney and prostate cancer.[197] Furthermore, past studies in Pennsylvania (US) and New Mexico (US) have shown that the incidence of stomach cancers were in correlation with elevated levels of uranium and radon in both drinking water and soil. [251] This was the first study to explore the potential relationship between soil uranium and risk of GIT cancers in the UK. Overall, this study suggests that there is no consistent or strong association which is a reassuring result because currently the presence of uranium in major areas of South West England and Southern Wales is of a huge concern, where elevated levels and decay of top soil uranium that gives significant rise to radon. It is emitted as radon in a form of radioactive gas typically at low-levels. The nature of this gas is that it has the properties of permeating through built environments in which it accumulates leading to increased levels of indoor radon in air.[199,200] Long-term exposure through inhalation or ingestion of indoor radon contributes significantly to the risk of developing lung and GIT cancer.[181] [197,261] This may possibly explain the observed risk found among urban residents with the highest exposure of soil uranium ranging from levels of 2.5 mg/kg and above who had a modest increase in risk of GI cancer (Figure 6.12); however, this result

should be interpreted with caution due to the limitations of other residual confounding in this study.

The observed relationship between elevated concentrations of soil phosphorus and GIT cancer (overall) remained consistent throughout both models. It showed that there was general increased risk of GIT cancers for residents living on soil phosphorus that have concentrations above 873.0 mg/kg, and that, such increased risk ranged between 6-8%. Our results are similar to findings of a previous study which quantified cancer risks using serum levels of inorganic phosphate as a proxy for environmental exposure.[262] Laboratory-based studies using rodents have shown elevated serum levels of inorganic phosphate causes a modification in their gene expression and protein translation, which, in turn, affects the rate of cell proliferation *in vitro*. In addition, the entry of large amounts of inorganic phosphate in diet has been indicated to increase the development of lung and cutaneous cancers in rodents. It is through these observations that the potential link with carcinogenesis was extrapolated on to humans. In addition, soil phosphorous concentrations are usually high in residential areas mostly due to urban development; the frequent contact with environmental phosphorus in urban areas could be a possible explanation for finding a weak positive dose-response relationship between elevated levels of soil phosphorus and GIT cancer.

For zinc, our initial model had shown that elevated soil levels were associated with a risk reduction of developing GIT cancer. Although the direction in the patterns of risk remained in the corrected model; however, the findings were no longer significant after including confounding variables. In addition, we could not find any plausible link with the sub outcomes. Zinc is among the elements considered to be essential to humans.[5] Zinc deficiency commonly leads to growth retardation, cell-mediated immune dysfunctions and cognitive impairment.[257,259] The most important property of zinc is that it functions as an antioxidant and anti-inflammatory agent which are two key mechanisms that prevent oxidative stress and chronic inflammation (i.e. two mechanisms that triggers the development of cancer), respectively.[257,259] Experimental studies have suggested that zinc supplements may be efficacious in the prevention (and treatment) of cancers of the head, neck, upper GIT, pancreas and colon,[257,259] to which, in some way are agreement with reduced risk patterns observed in our results. The authors of this areas stated that further studies in the preventative properties of zinc are needed to confirm it usage in management and chemoprevention of cancer. [257,259] If the efficacious nature of zinc is true, then the likely explanation of the patterns observed in our study may be due to the fact that it is an abundant metal in soil, whereby the general population commonly derives its zinc supplements through the food chain (i.e. soil-plant-human or soil-plant-animal-human). However,

this result should be interpreted with caution due to the limitations of other residual confounding in this study.

For this research, we believe that we managed to minimise any potential selection bias in terms of selecting GIT cases and non-cases. Like chapters 4 and 5, this study uses incident cases of GIT cancer that were recorded in THIN before the time this study was conducted. Selection bias is not a major concern as we attempted to use the entire cohort in the THIN-GBASE. We have provided a clear definition in our inclusion and exclusion criteria that limits our population of interests to adults (aged 18 years and above) registered at a GP practice at least 1-year before start date of study (i.e. January 1st, 2004) with linked soil data. One source of selection bias may arise from the dates we have chosen to initiate our study - by setting the start date from 2004, we are only capturing incident cases of GIT cancer after this year and assessing the risks at a much shorter time window (i.e. 2004-2014). However, we chose this date because of it was the year that the QOF scheme was introduced to encourage GPs in report new cases of GIT cancers when they are observed.

The recording of GIT cancers in THIN appear to be relatively complete, a recent study has shown that specific recording of these solid cancers has increased over time after 2004,[213] and indicated that with increasing expertise with Vision software these records are now close to that expected based on cancer registry data.[213] There are differences in the case ascertainment rates for GIT cancers at a GP-

level across England and Wales, if ascertainment rates across GP practices are associated with the geographical variation in exposures for the selected group soil elements - it would mean that our risk estimates derived for GIT cancer are biased.[169] We attempted to making corrections for these by including SHA in our models; however, the study would have benefitted if we were able to include a better indicator such as practice-level performance variable which accounts for how well practices tend to record clinical outcomes.

Finally, a robust statistical approach was adopted to objectively determine which selection soil elements are appropriate predictors of GIT cancer through data mining, before deriving detailed hazard estimates using Cox survival regression analysis, where we removed any time-varying effects through using a recognised diagnostic tool (the Aalen additive survival model).[168,193]

One of the main limitations was our inability to incorporate into our analyses the exact location of where a patient lived, and the spatial point of where soil samples were measured. The inability to display spatially by means of usage of high resolution maps could have potentially provided a convincing picture as to whether these soil elements (where they are highly concentrated at) are linked to large clusters of incident GIT cancer cases. Furthermore, by incorporating that spatial component into the analysis would have reduced any error (spatial variability) that exists between samples measured. Another limitation was that although we were able to make adjustments for

meaningful confounding variables, we unable to include other potentially important confounders; for example: the dietary records; [217-219] biomarkers for exposure includes finger nail (or toenail)[27,30,31] and household source of drinking water and levels of elements measured.[159] At present, this study is the only nationwide UK study to analyse the potential relationship between soil elements and GIT cancer at an individual-level; therefore, we recommend further research in this aspect and propose that much of limitations should be incorporated in future studies.

Similar to previous chapters, we are unable to provide a descriptive account on the spatial characteristics related to: 1.) a patient's address; 2.) the location of general practices he/she attends; and 3.) whether exposures fall within G-BASE rural, urban or NSI(XRFS) settings. These limitations were due to ethical implications outlined by THIN. The geospatial details of GIT cancer patients were anonymised, and therefore, we could not utilise this resource to directly ascertain the distribution of those that fall within (or close to) a sampling location; let alone, quantify the distribution of those that lived on soils classified as either a G-BASE urban or rural terrain, or NSI(XRFS) areas. If this information was made available in the linked database, we could have stratified the population at risk of GIT cancer in accordance to these three types of sampling locations. However, we used the THIN-derived residential setting indicators as a proxy to differentiate participants that lived on an urban or rural terrain, and incorporated them in our stratified Cox regression model

in attempting to minimise potential information biases that may affect our risk estimates for GIT cancer due to the systematic differences in exposure caused by the locations of sampling points, as well as the different densities through which the soil samples were collected.

In conclusion, this study suggests that there may be evidence of an increased risk of general GIT cancers for residents living areas with high levels of phosphorus. This study also indicates the possibility of a modest risk of GIT cancers in general, where there are elevated levels of soil aluminium and uranium in urban and suburban environments. However, when we account for confounding, the increased risk for GIT cancer disappears, which is reassuring. Our competing risk models indicated that the trend in risk reduction for upper GIT cancers were more pronounced for calcium, while the trends in an increased risk of stomach cancer were linked to elevated soil levels of aluminium only. Overall, it is difficult to conclude whether these metals can genuinely be carcinogenic due to the ambiguous results observed from this study. Further investigation is warranted to establish a clear association.

Conclusion and Implications of the work

Chapter 7

7.1 Commentary on findings

7.1.1 For basal cell carcinoma

In chapter two, the primary aims were to produce contemporary incidence rates of BCC throughout the UK, and to determine whether elevated concentrations levels of soil arsenic in residential areas was an aetiological factor for the development BCC. We were able to achieve our first aim through a large population-based ecological study using the THIN-GBASE database to obtain regional breakdowns of the incidence rates of BCC in the UK, and an updated coverage for country and regional incidence of BCC. We have also presented novel estimates for stratified incidence of BCC for levels of socioeconomic deprivation in the UK.

The work pertaining to this section has shown that at the country-level, Wales has the largest incidence of BCC than England, Scotland and Northern Ireland. At a regional-level (i.e. the 10 SHAs in England, Wales, Northern Ireland and Scotland), the South East Coast has England's highest incidence of BCC. Levels of socioeconomic deprivation was major contributing factor to the increased incidence rates of BCC whereby those in the least deprived group were at a higher risk of developing BCC. This study provided novel findings in terms of informing which regions of UK had higher rates of BCC, as well as which socioeconomic groups were at a higher risk of BCC. The second objective which sought to use the UK soil guideline values to assess whether arsenic levels in residential soils were linked to the

development of BCC were also achieved through a population-based cohort study. The work presented in this thesis has shown that soil arsenic present at low-levels have the potential to contribute towards the BCC burden in UK. We quantified residential exposure levels for soil arsenic based on the C4SL value of 35.0 mg/kg. By definition, the C4SL system deems that the level of health risk soil arsenic concentrations below the value of 35.0 mg/kg to low; while, areas with soil arsenic above this threshold are typically considered to be contaminated lands with the possibility of causing harmful effects to humans. Our study population for this study had approximately 5.0% of participants contributing to the THIN-GBASE database that lived in areas with residential soils contaminated with arsenic (i.e. > 35.0 mg/kg). Compared with the remaining 95.0% of participants that resided on lands with acceptable levels of arsenic that were below 35.0 mg/kg; those living on residential soil with levels exceeding this criterion were at 8.0-17.0% increased risk of developing BCC [(1) 35.0-70.0 mg/kg: HR 1.08, 95% CI: 1.02-1.14; (2) ≥70.0 mg/kg: HR 1.17, 95% CI: 1.09-1.28], and urban residents with the highest exposure were significantly at the greatest risk of developing BCC (≥ 70.0 mg/kg: HR 1.18, 95% CI: 1.06-1.36).

7.1.2 For lung cancer

In chapter five, we sought to use a two-stage process to first determine which group of soil elements were best suited as predictors for modelling lung cancer risk. Before using a population-based cohort

study to incorporate the selected group of elements into a statistical model in order to quantify the effect size for lung cancer risk for individuals living in a residential area with high concentration levels for the selected group of soil elements.

The first objective was achieved by using data mining techniques i.e. correlation-based filter selection method which identified aluminium, lead and uranium as the most appropriate subset of soil elements to be modelled as risk factors for lung cancer. The selection of these elements are plausible variables modelling risk of lung cancer since they are among the group agents ranked as carcinogenic due to their toxic properties for inducing cell-mediated immune dysfunctions, oxidative stress, cell proliferation and genetic alterations, which are known mechanisms for triggering cancer. We were able to determine such associated risks between the selected group of metals and lung cancer. Our mutually adjusted model have shown that the risk of developing lung cancer was significantly higher among individuals living in areas with elevated soil concentrations for aluminium, lead and uranium; however, these risk patterns were only retained for aluminium after correction of confounding factors. We also observed that the risk of lung cancer was only isolated to urban residents on residential soil with aluminium concentrations above 47,200 mg/kg.

7.1.3 For GI tract cancers

In chapter 6, we sought to use a three-stage process to determine the following: (1) using data mining to find which restricted group of soil

elements were appropriate modelling GIT cancer risk, (2) using a population-based cohort design for quantifying the effect size for GIT cancer risk with the elements found in the first stage, and (3) further analysis of multiple GIT cancer outcomes using competing risk models.

The data mining model identified the following seven soil elements as appropriate predictors for GIT cancer: aluminium, calcium, lead, phosphorus, manganese, uranium and zinc. In stage 2, once the model was mutually fitted for all the selected elements, we found that residents in areas with elevated levels of aluminium, phosphorus and uranium had a significant increase in risk of GIT cancer. Further adjusts showed zinc to have a protective effect against GIT cancer, while soil phosphorus was the only element to have retained its significance throughout the analysis. In stage 3, once the model was adjusted for different GIT cancers as competing events, no consistent relationships were identified between any of the selected groups of elements and the GIT-specific cancer outcomes.

7.2 Implications

The THIN-GBASE database is an invaluable resource for studying the health impacts of geochemical soil contaminants in the UK. It is a huge resource that contains complete coverage on the soil contamination levels of fifteen soil elements in England and Wales. One of the main advantages of the resource is that the concentration levels of each contaminant has been individually linked to person contributing to THIN by postcode, and so exposure-levels of

individuals in the linkage are quantified at a fine resolution. Another of the main advantages of THIN-GBASE has made possible for this research to conduct a series of epidemiological studies on a national scale, and so the results presented in these studies are reliable and representative of the population residing on English and Welsh soil. Since several outcomes in THIN have been validated; however, the results cannot be generalised to neighbouring countries – Northern Ireland and Scotland, until full geochemical coverage is achieved in Northern Ireland and Scotland. Furthermore, this resource has enabled us to test a series of hypotheses to examine the health impact of low-level concentrations of toxic soil elements on BCC, lung and GI tract cancer incidence in the England and Wales.

7.2.1 **Causality**

For BCC, the study has shown an indication of an increased hazard of BCC for residents living on soil with arsenic concentration above the UK national safety limits. We have incorporated UK arsenic C4SL as a tool for categorising residential exposure levels of soil arsenic, which have shown that the risk of BCC has a modest positive dose-relationship with elevated concentrations of arsenic. The patterns of risk for BCC in association with soil arsenic concentration were unambiguous and remained consistent throughout the multiple stages of risk estimation in our analyses. Furthermore, our results have corroborated other studies with similar findings in India, Bangladesh, Taiwan and in the mainland of Europe.[87,169,232,263,264] These studies

have provided a theoretical and plausible basis for environmental arsenic to be potential carcinogen for BCC. In light of the findings for soil arsenic and BCC, the viewpoint of this research stands in support that there is a possibility of cause and effect relation between potential exposures to soil arsenic and BCC.

The same can be inferred for aluminium and phosphorus, these were the only soil metals that appeared to be associated with lung and GI tract cancer, respectively. The viewpoint of this research stands in favour of these metals may have a carcinogenic causal effect for lung and GI tract cancer. Uranium and the remaining 14 soil elements (i.e. lead, chromium, copper, iron…) had a very weak, or no effect for these outcomes. Although most of these metals have a plausible and theoretical basis for causing cancer, our result failed to establish any forms of a significant statistical association and dose-response pattern. The patterns of risk were ambiguous and not consistent throughout the modelling stages. By taking in to account of these results, the viewpoint of this PhD thesis stands in favour that these groups of metals have a non-causal effect which is a reassuring result.

7.2.2 **Public health implications**

The implications of these findings indicate that more studies are required to fully understand the causal nature between such potential exposures and cancer outcomes. As a public health implication in terms of minimising exposure to soil aluminium, arsenic and phosphorus; the following simple measures can be implemented: 1)

Garden owners growing their own produces should test their soils frequently for signs of any contamination using soil testing kits; 2) Protective clothing (i.e. face masks, aprons and gloves) should be worn by frequent garden users to avoid soil particulates entering the main exposure pathways (oral, inhalation and dermal absorption); 3) Builders and contractors establishing residential settlements should ensure that the safely soil levels are met before constructing a building; and 4) Environmental agencies should utilise electronic geochemical atlases that are updated frequently to monitor soil levels for contaminations, so as to take pre-emptive measures in preventing local areas from becoming heavily contaminated.

7.2.3 **Lessons**

A population-based prospective cohort study is the best approach for analysing the effects of environmental contaminants on cancer outcomes. For instance, we initially adopted a population-based case-control study design to analyse the effects of soil arsenic and BCC whereby the cases were matched by age, sex and GP at a ratio of 1:5. The intentions were to have the cases and controls have similar demographic characteristics only. Unfortunately, the soil arsenic estimates between cases and controls were also similar – this was probably caused by a data artefact in G-BASE [i.e. soil estimates for elements in GBASE were spatially interpolated, and one of the key assumptions for interpolating the actual soil metal samples over continuous surface is spatial autocorrelation, which assumes that the

pixel point(s) of surface closest to the index sampling sites in GBASE must be approximately similar to one another, while those farthest away from an index sampling location may have a different value]. This data artefact was realised through matching patients registered to similar practices (i.e. within catchment of a radius of 250 meters) had yielded sub-groups having the same arsenic value. By revising the study design to a prospective cohort, we were able to negate the effects of spatial autocorrelation.

We discovered that survival analyses using the Cox proportional hazard regression model was a much flexible approach for analysing geochemical data. By adopting this model, we were able to account for time dependency between soil exposure (at the start of the study) and outcome. Most importantly, we were able to incorporate advance diagnostic techniques such as the Aalen additive survival model to ensure a non-violation of the proportional hazards assumption. The Aalen additive survival model is an example of a spline-based diagnostic tool for detecting points of occurrences for time-varying effects. This robust approach enabled the removal of such time-varying artefacts ensuring validity of our risk estimates.

Finally, to optimise the predictive accuracy of our Cox models, we adopted a multi-disciplinary approach utilising data mining techniques such as correlation-based filter selection as a tool for generating a restricted group of soil exposures for modelling lung and GI tract cancer risk. Although, such technique does not guarantee the model

to churn out hazard ratios that will be statistically significant, its algorithm uses a robust approach in forwardly (or backwardly) selecting the significant attributes to help the investigator to further generate hypothesis in terms of biological plausibility, so that they can be modelled appropriately.

7.3 Avenues for further research

Despite the limitations outlined in prior chapters, the work presented has shown huge potential, and the usefulness of this unique THIN-GBASE database for future medical geological and epidemiological research in cancer. Although, several potential confounders were measured and accounted for in our studies; however, in addition to the information that was provided in this work, we strongly urge to obtain additional variables to remove the possibility of residual confounding that was inherently present in our studies.

The most important variables to include the following for future studies: (1) toxicological information such as biomarkers as an indicator for metal toxicity – this would include the collection of hair, toenail, urine and blood samples. More specifically, lung cancer induced by long-term metallic toxicity can be measured in exhaled breath and expelled mucus of the lung (i.e. sputum), and for GIT cancer induced by metal toxicity can also be measured through stool examination; (2) the inclusion of dietary information and nutritional status; (3) the GPS coordinates of the participant so as to map areas

using GIS showing high risk patients residing on soil with elevated levels of soil contaminants.

7.3.1 Protocol for developing an exposure model for subsequent environmental epidemiologic analysis

7.3.1.1 Introduction

The methods presented in the chapters for this PhD provides a data-driven and viable approach for making direct associations for a specific group of cancers in light of potential exposures to environmental levels of certain soil contaminants in England and Wales. The interesting aspect of this research was the adoption of a multi-disciplinary approach, using both statistical and data mining methodologies, to quantifying cancer risk in relation to a group of soil metals. To extend this research, we take the opportunity to explore these databases further.

One of the advantages of this new and unique resource is that it provides an unprecedented population size with enough statistical power to research on the effects of geochemical soil metals on human health.[119] The linkage of soil data with details documented in EMRs in THIN provides us with the ability to adjust for range of potential confounding factors – these typically include a range of additional health and lifestyle details (such as smoking, alcohol consumption and BMI), as well as area-level measurements (such as quintile estimates for socioeconomic deprivation and air pollution for certain

compounds).[105,106,109,119] Finally, our recent validation study, which has just been accepted to a peer-review publication, has established that the observed soil exposure levels for each contaminant among THIN patients are similar to the wider population, and thus, studies utilising the linked resource are likely to produce generalizable results.[119]

However, the greatest limitation of this resource – exposure estimates are measured at an environmental level and they by no means reflect *in-vivo* or internal exposures that occur within an individual, and so, studies using this resource may be subject to exposure bias. Due to geographic variations in concentrations for the soil metals, as well as paucity of GPS readings on THIN patients and soil sampling sites in the database - studies using this resource will be subject to geographic bias – unless, such information is made available and adjusted for in the analysis.

One of the challenging aspects of assessing the environmental impacts of soil elements on cancer risk in the UK are due to some of the outlined limitations of this database, as well as the paucity of personal exposure data which is most often determined from biological specimens (e.g. fingernail and hair samples) and lack of spatial data. In this protocol, we proposal an exposure-based assessment study and provide a brief outline for how we can utilise the THIN-GBASE (and collaborating directly with GP practices using THIN) to develop an exposure model for estimating the amounts of soil contaminant ingested by an individual, which, in turn, can be used in

subsequent epidemiological cancer studies; and also, as an attempt to address some of the limitations outline in this thesis.

7.3.1.2 Study area and population

The intention is to carry out an exposure-based assessment study on a small-scale targeting family members in one of the major Boroughs in London. The determination of the area from which we will draw our sample will be dependent on the characteristics such as the population, surface area and total number of people within the Borough that have access to a GP practice. For instance, Ealing is the third largest in terms of population in London, and it also ranks as the eleventh borough with the largest surface area covering the north-western parts of London.[265] It has an estimated total land surface area of 55.53km^2. The borough itself is sub-divided into 23 wards which has an estimated total population size of 345,038 (males ~ 174,423 (50.6%)).[265] There are currently 79 GP practices responsible for slightly over 400,000 people – the discrepancy that exist between the Ealing population and patients registered to Ealing GPs is not unusual because people other neighbouring Boroughs register to a practice. According to the maps we derived from GBASE (Figure 3.1), Ealing is a GBASE urban zone which might have full coverage of sampling points at a resolution of 1 sampling site per every 0.25km^2. This is an example of the criteria selecting a potential study area.

7.3.1.3 Data collection

A feasible approach for collecting biological samples from individuals will be through GP practices using THIN (and practices linked to GBASE). GP practices agreeing to participate will be the channel through which registered individuals at the participating practice will be contacted and sent a consent form. These consent forms will be distributed to the addresses of individuals registered at the practices that typically fall within 250m of their practice's catchment area. Individuals who agree to provide a written consent will receive a package through the post which will contain a number of sealable plastic bags for submitting hair strands and fingernail clippings.

For each participant, a questionnaire will be distributed to collect information about the food and dietary habits, residential histories (which includes the number of years of stay at the current address or previous addresses in the UK), type of household (i.e. a house, apartment, store building etc.) and ownership of a garden. Participants will be urged to also provide other additional health information such as their ages, gender, body weight, smoking status and frequency of hair product usage. The collected fingernail and hair samples from participants will be sent for laboratory analysis whereby total concentrations for elements such as arsenic, copper, lead, nickel and other elements present in the GBASE dataset will be determined using the appropriate chemical analysis.

7.3.1.4 Statistical analysis

The estimated exposure doses via the ingestion route for an individual is the primary outcome of interest. For each soil metal, we will able to develop an exposure model to estimate the amount of soil (with the contaminants) ingested using a set of dose-exposure mathematical equations outlined by the ATSDR guidelines.[266] These set of equations are highly modifiable and can be tailored to estimate dose exposures at an individual-level. The predicted outcome is typically dependent on the elemental concentrations found in the soil at the residence of the individual, as well as their duration at their residence, age group (i.e. child or adulthood) and his/her body weight.[266] After the derivation of these estimates, they can be implemented in regression models in subsequent epidemiologic studies for predicting cancer risk.

7.3.2 Other suggestions for future studies and recommendation

The suggestions for future research are below:

1. **Geostatistical prediction of the intensity of basal cell carcinoma risk in association with elevated arsenic concentrations in UK topsoil;**

   Health information of patients contributing to the THIN is georeferenced at a SHA-level, which renders the spatial analysis between geochemical risk factors and cancer outcomes to be

limited. In order to use a geostatistical approach in predicting the intensity of BCC risk in association with elevated concentrations of soil arsenic over a continuous surface would require the investigator to be engaged with collecting data directly from consenting participants, as well as obtaining their health information directly from the GPs they are registered with (i.e. GPs whom have agreed to be involved with the project).

Data collection would include a participant's current address recorded as a postcode or geographical coordinate (i.e. latitude and longitude). Data on incident BCC occurrence can also be collected directly from participating practices in which the patients have registered with. This data, in turn, can be postcode linked to soil sample records in GBASE to facilitate the use of geostatistical modelling of BCC risk in relation soil arsenic.

2. **Implementing environmental, toxicological and dietary risk factors through linking UK Biobank with GBASE, to assess the effects of topsoil metal exposure and general cancers in the UK;**

The UK Biobank is a unique resource containing large range of biological samples obtained from over 500,000 individual volunteers in the UK. These include blood, saliva, urine, faecal, hair and finger and toenail samples submitted by participating volunteers contributing to the UK Biobank. The advantages of using UK Biobank database allow researchers to follow-up samples of volunteers to perform further biochemical analysis - e.g. detecting

levels of toxic elements stored in blood, urine, nail and hair samples to determine the extent of exposure. Furthermore, dietary information of volunteers contributing to the database is available in the UK Biobank.

One of the major limitations of the investigations carried out in this PhD was our inability to include dietary information and biomarkers for exposures as confounding risk factors. By linking UK Biobank to GBASE, we will have environmental, toxicological and dietary information to increase the accuracy of predicting the risks of BCC, lung and GI tract cancers associated with soil metal exposure.

3. **Using the THIN-GBASE linkage and adopting risk-based approach in deriving UK exposure (or exceedance) limits for each the of the soil metal;**

   The UK C4SLs are national safety limits or trigger values used to monitor and ensure that specific elements (or compounds) do not exceed certain concentration levels in soils that are typical considered to be in a residential, agricultural and industrial settings. One of the major problems encountered for this research were that the guideline values were limited to only arsenic, chromium, lead and selenium. There are currently no C4SLs for the remaining 12 elements in the THIN-GBASE database. Using mathematical modelling approach, the linkage can be exploited to determine for each metal the minimum environmental exposure

needed to have an incident recording of a cancer outcome in THIN-GBASE; where C4SLs are currently limited to only arsenic, chromium and lead, this resource can also be utilised to derive other C4SLs for toxic metals such as aluminium, copper, nickel, phosphorous and uranium.

7.4 Overall conclusions

In conclusion, this thesis showed a modest relationship between BCC, lung and GIT cancers, and potential exposures to elevated concentrations of soil arsenic, aluminium and phosphorus, respectively. The series of investigations conducted for this research are one of the first, if not, contemporary UK-based study to present novel estimates for a group of ill-defined pollutants. This research demonstrates that linking geochemical data with electronic primary care medical records can be a valuable for the investigation of long-term potential exposures to low-level soil contaminants may have a health consequence in the population.

Bibliography

1. Bunnell JE, Finkelman RB, Centeno JA, Selinus O. Medical Geology: a globally emerging discipline. Geologica Acta: an international earth science journal. 2007;5(3):273–281.

2. Selinus O. Medical Geology: An Opportunity for the Future. AMBIO: A Journal of the Human Environment. 2007;36(1):114–116.

3. Stüwe K. Geodynamics of the Lithosphere: An Introduction. Springer; 2007. 497 p.

4. Alloway BJ. Sources of heavy metals and metalloids in soils. In: Chapter 2 Alloway BJ Heavy Metals in Soils: Trace Metals and Metalloids in Soils and their Bioavailability. New York (US) & London (UK): Springer Science & Business Media; 2012. p. 11–50.

5. World Health Organisation (WHO). A. Essential trace elements. In: Trace elements in human nutrition and health [Internet]. 1996 [cited 2016 Aug 3]. p. 47–160. Available from: http://www.who.int/nutrition/publications/micronutrients/9241561734/en/

6. World Health Organisation (WHO). B. Trace elements that are probably essential. In: Trace elements in human nutrition and health [Internet]. 1996 [cited 2016 Aug 3]. Available from: http://www.who.int/nutrition/publications/micronutrients/9241561734/en/

7. Kabata-Pendias A, Mukherjee AB. Trace Elements from Soil to Human. Springer Science & Business Media; 2007. 561 p.

8. World Health Organisation (WHO). C. Potential toxic elements, some possibly with essential functions. In: Trace elements in human nutrition and health [Internet]. 1996 [cited 2016 Aug 3]. p. 185–230. Available from: http://www.who.int/nutrition/publications/micronutrients/9241561734/en/

9. Wilbur S, Abadin H, Fay M, Yu D, Tencza B, Ingerman L, et al. Toxicological profile for chromium. 2012 [cited 2016 Aug 5]; Available from: http://europepmc.org/abstract/med/24049864

10. Faroon O, Ashizawa A, Wright S, Tucker P, Jenkins K, Ingerman L, et al. Toxicological profile for cadmium. 2012 [cited 2016 Aug 5]; Available from: http://europepmc.org/abstract/med/24049863

11. Abadin H, Ashizawa A, Stevens Y-W, Llados F, Diamond G, Sage G, et al. Toxicological profile for lead. 2007 [cited 2016 Aug 5]; Available from: http://europepmc.org/abstract/med/24049859

12. Toxicological Profile for Mercury. In: ATSDR's Toxicological Profiles [Internet]. CRC Press; 2002 [cited 2016 Aug 5]. Available from: http://www.crcnetbase.com/doi/abs/10.1201/9781420061888_ch109

13. Skinner HCW. The Earth, Source of Health and Hazards: An Introduction to Medical Geology. Annual Review of Earth and Planetary Sciences. 2007;35(1):177–213.

14. Abrahams PW. Soils: their implications to human health. Science of the Total Environment. 2002;291(1):1–32.

15. Oliver MA. Soil and human health: a review. European Journal of Soil Science. 1997;48(4):573–592.

16. Macklin Y, Foxall K, Pollitt F, Cush M, Gadeberg B. Contaminated land and public health. In: Chapter 6 Bradley et al Essentials of Environmental Public Health Science: A Handbook for Field Professionals. OUP Oxford; 2014. p. 115–45.

17. Kamanyire R, Urquhart G, Stewart L. Emerging issues. In: Chapter 8 Bradley et al Essentials of Environmental Public Health Science: A Handbook for Field Professionals. OUP Oxford; 2014. p. 175–97.

18. Cave M. Bioaccessibility of potentially harmful soil elements. Environmental Scientist. 2012;21(3):26–29.

19. Prüss-Ustün A, Vickers C, Haefliger P, Bertollini R. Knowns and unknowns on burden of disease due to chemicals: a systematic review. Environ Health [Internet]. 2011 [cited 2013 Jul 11];10(9). Available from: http://www.biomedcentral.com/content/pdf/1476-069X-10-9.pdf

20. Nieuwenhuijsen MJ, Brunekreef B. Environmental exposure assessment. In: Chapter 3 Nieuwenhuijsen et al Environmental epidemiology: Study methods and application. Oxford University Press; 2008. p. 41–72.

21. Appleton JD, Cave MR, Wragg J. Anthropogenic and geogenic impacts on arsenic bioaccessibility in UK topsoils. Science of The Total Environment. 2012 Oct;435–436:21–9.

22. Denys S, Tack K, Caboche J, Delalain P. Assessing metals bioaccessibility to man in human health risk assessment of

contaminated site. In 2006 [cited 2016 Feb 13]. p. NC. Available from: http://hal-ineris.ccsd.cnrs.fr/ineris-00973245/document

23. Klinck B, Palumbo-Roe B, Cave MR, Wragg J. Arsenic dispersal and bioaccessibility in mine contaminated soils : a case study from an abandoned arsenic mine in Devon, UK [Internet]. 2005 [cited 2015 Oct 8]. Available from: http://www.bgs.ac.uk

24. Lu Y, Yin W, Huang L, Zhang G, Zhao Y. Assessment of bioaccessibility and exposure risk of arsenic and lead in urban soils of Guangzhou City, China. Environ Geochem Health. 2011 Apr;33(2):93–102.

25. Luo X-S, Ding J, Xu B, Wang Y-J, Li H-B, Yu S. Incorporating bioaccessibility into human health risk assessments of heavy metals in urban park soils. Sci Total Environ. 2012 May 1;424:88–96.

26. Wragg J, Cave MR. In-vitro methods for the measurement of the oral bioaccessibility of selected metals and metalloids in soils: a critical review [Internet]. Environment Agency; 2003 [cited 2016 Aug 4]. Available from: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/290321/sp5-062-tr-1-e-e.pdf

27. Button M, Jenkin GRT, Harrington CF, Watts MJ. Human toenails as a biomarker of exposure to elevated environmental arsenic. J Environ Monit. 2009;11(3):610–617.

28. Hwang J, Na S, Lee H, Lee D. Correlation between preoperative serum levels of five biomarkers and relationships between these biomarkers and cancer stage in epithelial overian cancer. J Gynecol Oncol. 2009 Sep;20(3):169–75.

29. Vandevijvere S, Geelen A, Gonzalez-Gross M, van't Veer P, Dallongeville J, Mouratidou T, et al. Evaluation of food and nutrient intake assessment using concentration biomarkers in European adolescents from the Healthy Lifestyle in Europe by Nutrition in Adolescence study. Br J Nutr. 2013 Feb 28;109(4):736–47.

30. Freeman LEB, Dennis LK, Lynch CF, Thorne PS, Just CL. Toenail Arsenic Content and Cutaneous Melanoma in Iowa. Am J Epidemiol. 2004 Oct 1;160(7):679–87.

31. Hinwood AL, Sim MR, Jolley D, De Klerk N, Bastone EB, Gerostamoulos J, et al. Hair and toenail arsenic concentrations of residents living in areas with high environmental arsenic concentrations. Environmental health perspectives. 2003;111(2):187.

32. Järup L. Hazards of heavy metal contamination. Br Med Bull. 2003 Dec 1;68(1):167–82.

33. Hayes RB. The carcinogenicity of metals in humans. Cancer Causes Control. 1997 May 1;8(3):371–85.

34. British Geological Survey. Geochemical Baseline Survey of the Environment (G-BASE) [Internet]. [cited 2013 Sep 12]. Available from: http://www.bgs.ac.uk/gbase/

35. Hawkes SJ. What Is a "Heavy Metal"? Journal of Chemical Education. 1997 Nov;74(11):1374.

36. Duruibe JO, Ogwuegbu MOC, Egwurugwu JN, others. Heavy metal pollution and human biotoxic effects. Int J Phys Sci. 2007;2(5):112–118.

37. Wuana RA, Okieimen FE, Wuana RA, Okieimen FE. Heavy Metals in Contaminated Soils: A Review of Sources, Chemistry, Risks and Best Available Strategies for Remediation, Heavy Metals in Contaminated Soils: A Review of Sources, Chemistry, Risks and Best Available Strategies for Remediation. International Scholarly Research Notices, International Scholarly Research Notices. 2011 Oct 24;2011, 2011:e402647.

38. Tchounwou PB, Yedjou CG, Patlolla AK, Sutton DJ. Heavy Metal Toxicity and the Environment. In: Luch A, editor. Molecular, Clinical and Environmental Toxicology [Internet]. Springer Basel; 2012 [cited 2014 Feb 17]. p. 133–64. (Experientia Supplementum). Available from: http://link.springer.com/chapter/10.1007/978-3-7643-8340-4_6

39. Polmear I, John DS. Light alloys: from traditional alloys to nanocrystals [Internet]. Butterworth-Heinemann; 2005 [cited 2015 Sep 28]. Available from: https://books.google.co.uk/books?hl=en&lr=&id=td0jD4it63cC&oi=fnd&pg=PP2&dq=Light+Alloys:+From+Traditional+Alloys+to+Nanocrystals&ots=pap8ta3Bsm&sig=gI_XI21t4B3hbtD10d8N6CF77Uo

40. Brook GB. Light metals handbook [Internet]. Butterworth-Heinemann; 1998 [cited 2015 Sep 28]. Available from: https://books.google.co.uk/books?hl=en&lr=&id=RNQpGonP3CgC&oi=fnd&pg=PP1&dq=Light+Metals+Handbook&ots=Ol6O65bmYk&sig=OtXKA4e5JBZum4b_jeaGt0gY3MI

41. Neuendorf KKE. Glossary of Geology. Springer Science & Business Media; 2005. 802 p.

42. Pani B. Textbook of Environmental Chemistry. I. K. International Pvt Ltd; 2007.

43. ATSDR - Toxicological Profile: Aluminum [Internet]. [cited 2016 Aug 5]. Available from: http://www.atsdr.cdc.gov/ToxProfiles/tp.asp?id=191&tid=34

44. ATSDR - Toxicological Profile: Copper [Internet]. [cited 2016 Aug 5]. Available from: http://www.atsdr.cdc.gov/toxprofiles/TP.asp?id=206&tid=37

45. ATSDR - Toxicological Profile: Zinc [Internet]. [cited 2016 Aug 5]. Available from: http://www.atsdr.cdc.gov/toxprofiles/TP.asp?id=302&tid=54

46. IARC. IARC Monographs [Internet]. Agents Classified by the IARC Monographs, Volumes 1–116. 2016 [cited 2016 Aug 5]. Available from: http://monographs.iarc.fr/ENG/Classification/index.php

47. Hughes MF, Beck BD, Chen Y, Lewis AS, Thomas DJ. Arsenic Exposure and Toxicology: A Historical Perspective. Toxicol Sci. 2011 Oct 1;123(2):305–32.

48. Hopenhayn-Rich C, Biggs ML, Smith AH. Lung and kidney cancer mortality associated with arsenic in drinking water in Cordoba, Argentina. International Journal of Epidemiology. 1998;27(4):561–569.

49. Chiang C-T, Chang T-K, Hwang Y-H, Su C-C, Tsai K-Y, Yuan T-H, et al. A critical exploration of blood and environmental chromium concentration among oral cancer patients in an oral cancer prevalent area of Taiwan. Environ Geochem Health. 2011 Oct;33(5):469–76.

50. Landrigan PJ. Occupational and community exposures to toxic metals: lead, cadmium, mercury and arsenic. West J Med. 1982;137(6):531–9.

51. Cempel M, Nikel G. Nickel: a review of its sources and environmental toxicology. Polish Journal of Environmental Studies. 2006;15(3):375–382.

52. ATSDR - Toxicological Profile: Cobalt [Internet]. [cited 2016 Aug 6]. Available from: http://www.atsdr.cdc.gov/toxprofiles/TP.asp?id=373&tid=64

53. Rawlins BG, McGrath SP, Scheib A, Breward N, Cave MR, Lister B, et al. The advanced soil geochemical atlas of England and Wales [Internet]. Nottingham, UK: British Geological Survey; 2012 [cited 2015 May 21]. 227 p. Available from: http://www.bgs.ac.uk/GBASE/advSoilAtlasEW.html

54. Rawlins BG, Webster R, Lister TR. The influence of parent material on topsoil geochemistry in eastern England. Earth Surface Processes and Landforms. 2003;28(13):1389–1409.

55. Stüwe K. Geodynamics of the lithosphere: an introduction [Internet]. Springer Science & Business Media; 2007 [cited 2015 Sep 29]. Available from: https://books.google.co.uk/books?hl=en&lr=&id=peagD0TnM4cC&oi=fnd&pg=PA1&dq=geodynamics+of+the+lithosphere+an+introduction&ots=tmgkpo6vtN&sig=B-ye49eEL2a4dzKhQMswfExTjWs

56. Mandal BK, Suzuki KT. Arsenic round the world: a review. Talanta. 2002 Aug 16;58(1):201–35.

57. Clapp BW. An Environmental History of Britain Since the Industrial Revolution. Routledge; 2014. 283 p.

58. Ashton TS, others. The industrial revolution 1760-1830. OUP Catalogue [Internet]. 1997 [cited 2016 Aug 6]; Available from: https://ideas.repec.org/b/oxp/obooks/9780192892898.html

59. Mannion AM. The Environmental Impact of War & Terrorism [Internet]. Department of Geography, University of Reading; 2003 [cited 2016 Aug 6]. Available from: https://www.reading.ac.uk/web/FILES/geographyandenvironmentalscience/GP169.pdf

60. Jensen MC. The modern industrial revolution, exit, and the failure of internal control systems. the Journal of Finance. 1993;48(3):831–880.

61. Turnbull G. Canals, coal and regional growth during the industrial revolution. The Economic History Review. 1987 Nov 1;40(4):537–60.

62. Pless-Mulloli T, Edwards R, Paepke O, Schilling B. Report on the analysis of PCCD/PCDF and heavy metals in footpaths and soil samples related to the Byker incinerator. Newcastle upon Tyne, UK7 University of Newcastle. 2000;

63. Pless-Mulloli T, Paepke O, Schilling B. PCDD/PCDF and heavy metals in vegetable samples from Newcastle allotments: assessment of the role of ash from the Byker incinerator. Newcastle upon Tyne, UK7 University of Newcastle. 2001;

64. Cresswell PA, Scott JES, Pattenden S, Vrijheid M. Risk of congenital anomalies near the Byker waste combustion plant. Journal of Public Health. 2003;25(3):237–242.

65. Participation E. Public Health Act 1936 [Internet]. [cited 2016 Aug 6]. Available from:

http://www.legislation.gov.uk/ukpga/Geo5and1Edw8/26/49/contents

66. Nathanail P. A brief history of land contamination in the UK. In Nottingham, UK; 2011 [cited 2016 Aug 6]. p. 2 pages. Available from: http://www.commonforum.eu/Documents/Meetings/2011/Nottingham/S1_brief_history_of_contaminated_land_in_the_UK.pdf

67. Using soil guideline values - SGV Introduction. Enivronmetal Agency (UK). 2009;Science Report SC050021/SGV Introduction:1–26.

68. Defra. SP1010 Development of category 4 screening levels for assessment of affected by contamination [Internet]. 2014 [cited 2014 Apr 11]. Available from: http://randd.defra.gov.uk/Default.aspx?Module=More&Location=None&ProjectID=18341

69. Defra. SP1010 Appendix H - Lead [Internet]. 2014 [cited 2014 Apr 11]. (Category 4 screening levels (C4SLs)). Available from: http://randd.defra.gov.uk/Default.aspx?Module=More&Location=None&ProjectID=18341

70. Defra. SP1010 Appendix F - Cadmium [Internet]. 2014 [cited 2014 Apr 11]. (Category 4 screening levels (C4SLs)). Available from: http://randd.defra.gov.uk/Default.aspx?Module=More&Location=None&ProjectID=18341

71. Defra. SP1010 Appendix C Provisional C4SLS for Arsenic [Internet]. 2014 [cited 2014 Apr 11]. Available from: http://randd.defra.gov.uk/Default.aspx?Module=More&Location=None&ProjectID=18341

72. Defra. SP1010 Appendix G Provisional C4SLs for Chromium [Internet]. 2014 [cited 2014 Apr 11]. Available from: http://randd.defra.gov.uk/Default.aspx?Module=More&Location=None&ProjectID=18341

73. Bradley N, Harrison H, Hodgson G, Kamanyire R, Kibble A, Murray V. Essentials of Environmental Public Health Science: A Handbook for Field Professionals. OUP Oxford; 2014. 225 p.

74. Science Communication Unitnit, University of the West of England, Bristol. Science for Environment Policy In-depth Report: Soil Contamination: Impacts on Human Health. European Commission DG Environment; 2013.

75. Baker D, Nieuwenhuijsen MJ. Environmental Epidemiology: Study methods and application. OUP Oxford; 2008. 414 p.

76. Abrahams PW. Soils: their implications to human health. Science of the Total Environment. 2002;291(1):1–32.

77. Diez M, Arroyo M, Cerdan FJ, Munoz M, Martin MA, Balibrea JL. Serum and tissue trace metal levels in lung cancer. Oncology. 1989;46(4):230–234.

78. Guidotti TL, McNamara J, Moses MS. The interpretation of trace element analysis in body fluids. Indian J Med Res. 2008 Oct;128(4):524–32.

79. Mulay IL, Roy R, Knox BE, Suhr NH, Delaney WE. Trace-metal analysis of cancerous and non-cancerous human tissues. Journal of the National Cancer Institute. 1971;47(1):1–13.

80. Pasha Q, Malik SA, Shaheen N, Shah MH. Investigation of trace metals in the blood plasma and scalp hair of gastrointestinal cancer patients in comparison with controls. Clinica Chimica Acta. 2010 Apr 2;411(7–8):531–9.

81. Silvera SAN, Rohan TE. Trace elements and cancer risk: a review of the epidemiologic evidence. Cancer Causes Control. 2007 Feb 1;18(1):7–27.

82. Versieck J. Trace elements in human body fluids and tissues. Crit Rev Clin Lab Sci. 1985;22(2):97–184.

83. Reilly C, Henry J. Geophagia: why do humans consume soil? Nutrition Bulletin. 2000;25(2):141–144.

84. Barnes RM. Childhood soil ingestion: How much dirt do kids eat? Analytical Chemistry. 1990;62(19):1023A–1033A.

85. Abrahams PW. Geophagy and the involuntary ingestion of soil. In: Essentials of medical geology [Internet]. Springer; 2013 [cited 2016 Aug 6]. p. 433–454. Available from: http://link.springer.com/chapter/10.1007/978-94-007-4375-5_18

86. Charlesworth S, Everett M, McCarthy R, Ordonez A, De Miguel E. A comparative study of heavy metal concentration and distribution in deposited street dusts in a large and a small urban area: Birmingham and Coventry, West Midlands, UK. Environment International. 2003;29(5):563–573.

87. Guo HR, Yu HS, Hu H, Monson RR. Arsenic in drinking water and skin cancers: cell-type specificity (Taiwan, ROC). Cancer Causes Control. 2001 Dec;12(10):909–16.

88. Guo HR, Lipsitz SR, Hu H, Monson RR. Using ecological data to estimate a regression model for individual data: the association

between arsenic in drinking water and incidence of skin cancer. Environ Res. 1998;79(2):82–93.

89. Alam MGM, Allinson G, Stagnitti F, Tanaka A, Westbrooke M. Arsenic contamination in Bangladesh groundwater: a major environmental and social disaster. Int J Environ Health Res. 2002;12(3):235–53.

90. Kile ML, Hoffman E, Rodrigues EG, Breton CV, Quamruzzaman Q, Rahman M, et al. A Pathway-based Analysis of Urinary Arsenic Metabolites and Skin Lesions. Am J Epidemiol. 2011 Apr 1;173(7):778–86.

91. Tseng C-H. Arsenic methylation, urinary arsenic metabolites and human diseases: current perspective. J ENVIRON SCI HEALTH PART C ENVIRON CARCINOG ECOTOXICOL REV. 2007;25(1):1–22.

92. Kavanagh P, Farago ME, Thornton I, Goessler W, Kuehnelt D, Schlagenhaufen C, et al. Urinary arsenic species in Devon and Cornwall residents, UK. A pilot study†. Analyst. 1998;123(1):27–29.

93. Tinwell H, Stephens SC, Ashby J. Arsenite as the probable active species in the human carcinogenicity of arsenic: mouse micronucleus assays on Na and K arsenite, orpiment, and Fowler's solution. Environ Health Perspect. 1991;95(ei0, 0330411):205–10.

94. Liou G-Y, Storz P. Reactive oxygen species in cancer. Free Radic Res [Internet]. 2010 May [cited 2016 Aug 6];44(5). Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3880197/

95. Nishigori C, Hattori Y, Toyokuni S. Role of reactive oxygen species in skin carcinogenesis. Antioxid Redox Signal. 2004;6(3):561–70.

96. Waris G, Ahsan H. Reactive oxygen species: role in the development of cancer and various chronic conditions. J Carcinog. 2006 May 11;5:14.

97. Bergamini C, Gambetti S, Dondi A, Cervellati C. Oxygen, Reactive Oxygen Species and Tissue Damage. Current Pharmaceutical Design. 2004 May 1;10(14):1611–26.

98. Loft S, Poulsen HE. Cancer risk and oxidative DNA damage in man. Journal of Molecular Medicine. 1996 Jun 12;74(6):297–312.

99. Davies RI, Griffith GW. Cancer and soils in the county of Anglesey. British journal of cancer. 1954;8(1):56.

100. Tromp SW, Diehl JC. A Statistical Study of the Possible Relationship between Cancer of the Stomach and Soil. Br J Cancer. 1955 Sep;9(3):349–57.

101. Millar IB. Gastro-intestinal cancer and geochemistry in north Montgomeryshire. British journal of cancer. 1961;15(2):175.

102. Stocks P, Davies RI. Zinc and Copper Content of Soils Associated with the Incidence of Cancer of the Stomach and other Organs. Br J Cancer. 1964 Mar;18(1):14–24.

103. Ashley DJ, Davies HD. Gastric cancer in Wales. Gut. 1966;7(5):542–548.

104. Philipp R, Hughes AO. Health effects of cadmium. British medical journal (Clinical research ed). 1981;282(6281):2054.

105. Blak BT, Thompson M, Dattani H, Bourke A. Generalisability of The Health Improvement Network (THIN) database: demographics, chronic disease prevalence and mortality rates. Inform Prim Care. 2011;19(4):251–5.

106. Cegedim Strategic Data UK (CSD). THIN (2013) statistics [Internet]. [cited 2014 Dec 26]. Available from: http://csdmruk.cegedim.com/our-data/statistics.shtml

107. Meal A, Leonardi-Bee J, Smith C, Hubbard R, Bath-Hextall F. Validation of THIN data for non-melanoma skin cancer. Qual Prim Care. 2008;16(1):49–52.

108. Hall GC. Validation of death and suicide recording on the THIN UK primary care database. Pharmacoepidemiology and Drug Safety. 2009 Feb;18(2):120–31.

109. Langley TE, Szatkowski L, Gibson J, Huang Y, McNeill A, Coleman T, et al. Validation of The Health Improvement Network (THIN) primary care database for monitoring prescriptions for smoking cessation medications. Pharmacoepidemiol Drug Saf. 2010 Jun;19(6):586–90.

110. Lewis JD, Schinnar R, Bilker WB, Wang X, Strom BL. Validation studies of the health improvement network (THIN) database for pharmacoepidemiology research. Pharmacoepidemiol Drug Saf. 2007 Apr;16(4):393–401.

111. British Geological Survey. History of the G-BASE project (version 1.1) [Internet]. 2006. Available from: http://www.bgs.ac.uk/downloads/start.cfm?id=860

112. British Geological Survey. Aims and Objectives G-BASE [Internet]. [cited 2013 Sep 12]. Available from: http://www.bgs.ac.uk/gbase/objective.html

113. Johnson CC, Breward N, Ander EL, Ault L. G-BASE: baseline geochemical mapping of Great Britain and Northern Ireland. Geochemistry: Exploration, Environment, Analysis. 2005;5(4):347–357.

114. British Geological Survey. Summary of geochemical atlas information (version 1.1) [Internet]. 2006 [cited 2013 Sep 12]. Available from: http://www.bgs.ac.uk/downloads/start.cfm?id=861

115. Ander EL, Cave MR, Palumbo-Roe B. Normal background concentrations of contaminants in the soil of England. Available data and data exploration [Internet]. Keyworth, Nottingham: British Geological Survey; 2011 [cited 2017 Mar 4] p. 124pp. (British Geological Survey Commission Report). Report No.: CR/11/145. Available from: http://nora.nerc.ac.uk/19958/

116. British Geological Survey (BGS). Geochemical baselines [Internet]. [cited 2017 Mar 4]. Available from: http://www.bgs.ac.uk/gbase/sampleindexmaps/soilnsi.html

117. British Geological Survey (BGS). Normal background concentrations (NBCs) of contaminants in English and Welsh soils [Internet]. [cited 2017 Mar 4]. Available from: http://www.bgs.ac.uk/gbase/NBCDefraProject.html

118. Johnson CC. G-BASE Field procedures manual (2005). British Geological Survey (BGS). 2005;Internal report no. IR/04/134.

119. Gibson J, Ander EL, Pinder L, Cave MR, Bath-Hextall F, Musah A, et al. Linkage of national soil quality measurements in England and Wales to primary care medical records: a new resource for investigating environmental impacts on human health. Population Health Metrics [Accepted] (In Press). 2016;

120. NHS choices - Your health, your choices. Skin cancer (non-melanoma) [Internet]. 2012 [cited 2013 Sep 18]. Available from: http://www.nhs.uk/Conditions/Cancer-of-the-skin/Pages/Introduction.aspx

121. Cancer Research UK. About skin cancer (non melanoma): a quick guide [Internet]. 2011 [cited 2013 Jan 9]. Available from: http://www.cancerresearchuk.org/cancer-help/prod_consump/groups/cr_common/@cah/@gen/documents/generalcontent/about-skin-cancer.pdf

122. World Health Organisation (WHO). Nonmealanoma skin cancer [Internet]. How common is skin cancer? 2013 [cited 2013 Oct 17]. Available from: http://www.who.int/uv/faq/skincancer/en/index1.html

123. Madan V, Lear JT, Szeimies R-M. Non-melanoma skin cancer. Lancet. 2010;375(9715):673–85.

124. Lomas A, Leonardi-Bee J, Bath-Hextall F. A systematic review of worldwide incidence of nonmelanoma skin cancer. Brit J Dermatol. 2012;166(5):1069–1080.

125. Demers AA, Nugent Z, Mihalcioiu C, Wiseman MC, Kliewer EV. Trends of nonmelanoma skin cancer from 1960 through 2000 in a Canadian population. J Am Acad Dermatol. 2005;53(2):320–328.

126. Jung GW, Metelitsa AI, Dover DC, Salopek TG. Trends in incidence of nonmelanoma skin cancers in Alberta, Canada, 1988–2007. Brit J Dermatol. 2010;163(1):146–154.

127. Staples MP, Elwood M, Burton RC, Williams JL, Marks R, Giles GG. Non-melanoma skin cancer in Australia: the 2002 national survey and trends since 1985. Med J Aust. 2006;184(1):6–10.

128. Rogers H, Weinstock M. Incidence estimate of nonmelanoma skin cancer in the united states, 2006. Arch Dermatol. 2010;146(3):283–7.

129. Kricker A, Armstrong BK, English DR, Heenan PJ. Does intermittent sun exposure cause basal cell carcinoma? a case-control study in Western Australia. Int J Cancer. 1995;60(4):489–494.

130. Kricker A, Armstrong BK, English DR, Heenan PJ. A dose-response curve for sun exposure and basal cell carcinoma. Int J Cancer. 2006;60(4):482–488.

131. Karagas MR, Stannard VA, Mott LA, Slattery MJ, Spencer SK, Weinstock MA. Use of tanning devices and risk of basal cell and squamous cell skin cancers. J Natl Cancer I. 2002;94(3):224–226.

132. Tran H, Chen K, Shumack S. Epidemiology and aetiology of basal cell carcinoma. Brit J Dermatol. 2003;149:50–52.

133. Roewert-Huber J, Lange-Asschenfeldt B, Stockfleth E, Kerl H. Epidemiology and aetiology of basal cell carcinoma. Brit J Dermatol. 2007;157:47–51.

134. Marks R, Staples M, Giles GG. Trends in non-melanocytic skin cancer treated in Australia: The second national survey. Int J Cancer. 1993;53(4):585–590.

135. Green A, Battistutta D, Hart V, Leslie D, Weedon D. Skin cancer in a subtropical Australian population: incidence and lack of association with occupation. Am J Epidemiol. 1996;144(11):1034–1040.

136. Lear JT, Tan BB, Smith AG, Bowers W, Jones P, Heagerty A, et al. Risk factors for basal cell carcinoma in the UK: case-control study in 806 patients. J Roy Soc Med. 1997;90(7):371.

137. Bath-Hextall F, Leonardi-Bee J, Smith C, Meal A, Hubbard R. Trends in incidence of skin basal cell carcinoma. Additional evidence from a UK primary care database study. Int J Cancer. 2007;121(9):2105–2108.

138. Holme SA, Malinovszky K, Roberts D. Changing trends in non-melanoma skin cancer in South Wales, 1988–98. Brit J Dermatol. 2000;143(6):1224–1229.

139. Ko CB, Walton S, Keczkes K, Bury HPR, Nicholson C. The emerging epidemic of skin cancer. Brit J Dermatol. 1994;130(3):269–272.

140. Brewster DH, Bhatti LA, Inglis JHC, Nairn ER, Doherty VR. Recent trends in incidence of nonmelanoma skin cancers in the East of Scotland, 1992–2003. Brit J Dermatol. 2007;156(6):1295–1300.

141. Hoey S, Devereux C, Murray L, Catney D, Gavin A, Kumar S, et al. Skin cancer trends in Northern Ireland and consequences for provision of dermatology services. Brit J Dermatol. 2007;156(6):1301–1307.

142. Skellett AM, Hafiji J, Greenberg DC, Wright KA, Levell NJ. The incidence of basal cell carcinoma in the under-30s in the UK. Clin Exp Dermatol. 2011;37(3):227–9.

143. Doherty V, Brewster D, Jensen S, Gorman D. Trends in skin cancer incidence by socioeconomic position in Scotland, 1978–2004. Brit J Cancer. 2010;102(11):1661–1664.

144. Van Hattem S, Aarts MJ, Louwman WJ, Neumann HAM, Coebergh JWW, Looman CWN, et al. Increase in basal cell carcinoma incidence steepest in individuals with high socioeconomic status: results of a cancer registry study in the Netherlands. Brit J Dermatol. 2009;161(4):840–845.

145. Bourke A, Dattani H, Robinson M. Feasibility study and methodology to create a quality-evaluated database of primary care data. Inform Prim Care. 2004;12(3):171–177.

146. Townsend P. Deprivation. J Soc Policy. 1987;16(02):125–146.

147. Boyle P, Parkin DM. Statistical methods for registries. In: Jensen O.M., Parkin D.M., MacLennan R., Muir C.S., Skeet R.G. (eds.). Lyon: IARC; 1991.

148. Office of National Statistics. National Population Projections [Internet]. Office of National Statistics (ONS). 2011 [cited 2012 Nov 15]. Available from: http://www.ons.gov.uk/ons/rel/npp/national-population-projections/2010-based-projections/index.html

149. Corona R DE. Risk factors for basal cell carcinoma in a mediterranean population: Role of recreational sun exposure early in life. Arch Dermatol. 2001;137(9):1162–8.

150. Dessinioti C, Tzannis K, Sypsa V, Nikolaou V, Kypreou K, Antoniou C, et al. Epidemiologic risk factors of basal cell carcinoma development and age at onset in a Southern European population from Greece. Exp Dermatol. 2011;20(8):622–626.

151. Saraiya M, Glanz K, Briss PA, Nichols P, White C, Das D, et al. Interventions to prevent skin cancer by reducing exposure to ultraviolet radiation. Am J Prev Med. 2004;27(5):422–466.

152. Freedman DM, Dosemeci M, McGlynn K. Sunlight and mortality from breast, ovarian, colon, prostate, and non-melanoma skin cancer: a composite death certificate based case-control study. Occup Environ Med. 2002;59(4):257–62.

153. Swerdlow AJ. Incidence of malignant melanoma of the skin in England and Wales and its relationship to sunshine. Brit Med J. 1979;2(6201):1324–7.

154. Freedman DM, Sigurdson A, Doody MM, Mabuchi K, Linet MS. Risk of Basal Cell Carcinoma in Relation to Alcohol Intake and Smoking. Cancer Epidemiol Biomarkers Prev. 2003;12(12):1540–3.

155. Briggs D. Environmental pollution and the global burden of disease. Br Med Bull. 2003 Dec 1;68(1):1–24.

156. Gawkrodger DJ. Occupational skin cancers. Occup Med. 2004;54(7):458–63.

157. Huang H-H, Huang J-Y, Lung C-C, Wu C-L, Ho C-C, Sun Y-H, et al. Cell-type specificity of lung cancer associated with low-dose soil heavy metal contamination in Taiwan: An ecological study. BMC Public Health. 2013;13(1):330.

158. Gilbert-Diamond D, Li Z, Perry AE, Spencer SK, Gandolfi AJ, Karagas MR. A Population-based Case-Control Study of Urinary Arsenic Species and Squamous Cell Carcinoma in New

Hampshire, USA. Environmental Health Perspectives [Internet]. 2013 Jul 19 [cited 2014 Jul 16]; Available from: http://ehp.niehs.nih.gov/1206178

159. IARC. Some drinking-water disinfectants and contaminants, including arsenic. IARC Monogr Eval Carcinog Risks Hum. 2004;84:1–477.

160. Leonardi G, Vahter M, Clemens F, Goessler W, Gurzau E, Hemminki K, et al. Inorganic Arsenic and Basal Cell Carcinoma in Areas of Hungary, Romania, and Slovakia: A Case–Control Study. Environmental health perspectives. 2012;120(5):721.

161. Ander EL, Johnson CC, Cave MR, Palumbo-Roe B, Nathanail CP, Lark RM. Methodology for the determination of normal background concentrations of contaminants in English soil. Science of The Total Environment. 2013;454:604–618.

162. Defra. Technical Guidance Sheet on normal levels of contaminants in English soils: Arsenic - supplementary information. Department of Environment Food and Rural Affairs (Defra), London; 2012.

163. Johnson CC, Ander EL, Cave MR. Technical guidance on normal levels of contaminants in Welsh soil: Arsenic (As): January 2013. 2013 [cited 2013 Sep 10]; Available from: http://nora.nerc.ac.uk/501674/

164. Martin I, De Burca R, Morgan H. Soil Guideline Values for inorganic arsenic in soil. Enivronmetal Agency (UK). 2009;Science Report SC050021:1–11.

165. Fordyce FM, Brown SE, Ander EL, Rawlins BG, O'Donnell KE, Lister TR, et al. GSUE: urban geochemical mapping in Great Britain. Geochemistry: Exploration, Environment, Analysis. 2005;5(4):325–336.

166. Oliver MA, Loveland PJ, Frogbrook ZL, Webster R, McGrath SP. Statistical and geostatistical analysis of the national soil inventory of England and Wales. Defra (formerly MAFF) Soil Programme Technical Report Project SP0124. 2002;

167. Met Office. UK climate - Historic station data [Internet]. UK Meteorological Office (Met Office). 2015 [cited 2015 May 22]. Available from: http://www.metoffice.gov.uk/public/weather/climate-historic/#?tab=climateHistoric

168. Hosmer DW, Royston P. Using Aalen's linear hazards model to investigate time-varying effects in the proportional hazards regression model. Stata J. 2002;2:331–350.

169. Leonardi G, Vahter M, Clemens F, Goessler W, Gurzau E, Hemminki K, et al. Inorganic arsenic and basal cell carcinoma in areas of Hungary, Romania, and Slovakia: a case-control study. Environ Health Perspect. 2012;120(5):721–6.

170. Leiter U, Garbe C. Epidemiology of Melanoma and Nonmelanoma Skin Cancer—The Role of Sunlight. In: Reichrath J, editor. Sunlight, Vitamin D and Skin Cancer [Internet]. Springer New York; 2008 [cited 2017 Mar 24]. p. 89–103. (Advances in Experimental Medicine and Biology). Available from: http://link.springer.com/chapter/10.1007/978-0-387-77574-6_8

171. Leary A. Lung Cancer: A Multidisciplinary Approach. John Wiley & Sons; 2012.

172. Witschi H. A Short History of Lung Cancer. Toxicol Sci. 2001;64(1):4–6.

173. Knaapen AM, Borm PJA, Albrecht C, Schins RPF. Inhaled particles and lung cancer. Part A: Mechanisms. Int J Cancer. 2004 May 10;109(6):799–809.

174. International Ageny for Research on Cancer (IARC), World Health Organisation (WHO). Cancer fact sheets - Lung cancer [Internet]. GLOBOCAN 2012: Estimated cancer incidence, mortality and prevalence worldwide in 2012. [cited 2014 Nov 24]. Available from: http://globocan.iarc.fr/Pages/fact_sheets_cancer.aspx

175. Jemal A, Bray F, Center MM, Ferlay J, Ward E, Forman D. Global cancer statistics. CA: A Cancer Journal for Clinicians. 2011;61(2):69–90.

176. Ferlay J, Soerjomataram I, Dikshit R, Eser S, Mathers C, Rebelo M, et al. Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012. Int J Cancer. 2014;[early view version].

177. Darby S, Hill D, Auvinen A, Barros-Dios JM, Baysson H, Bochicchio F, et al. Radon in homes and risk of lung cancer: collaborative analysis of individual data from 13 European case-control studies. Bmj. 2005;330(7485):223.

178. Krewski D, Lubin JH, Zielinski JM, Alavanja M, Catalan VS, Field RW, et al. Residential radon and risk of lung cancer: a combined analysis of 7 North American case-control studies. Epidemiology. 2005;16(2):137–145.

179. Pershagen G, Akerblom G, Axelson O, Clavensjo B, Damber L, Desai G, et al. Residential radon exposure and lung cancer in Sweden. New England journal of medicine. 1994;330(3):159–164.

180. Lubin JH, Boice JD. Lung cancer risk from residential radon: meta-analysis of eight epidemiologic studies. Journal of the National Cancer Institute. 1997;89(1):49–57.

181. WHO | Radon and cancer [Internet]. WHO. [cited 2013 Sep 9]. Available from: http://www.who.int/mediacentre/factsheets/fs291/en/

182. Yin J, Harrison RM, Chen Q, Rutter A, Schauer JJ. Source apportionment of fine particles at urban background and rural sites in the UK atmosphere. Atmospheric Environment. 2010;44(6):841–851.

183. Viana M, Kuhlbusch TAJ, Querol X, Alastuey A, Harrison RM, Hopke PK, et al. Source apportionment of particulate matter in Europe: a review of methods and results. Journal of Aerosol Science. 2008;39(10):827–849.

184. Harrison RM, Deacon AR, Jones MR, Appleby RS. Sources and processes affecting concentrations of PM 10 and PM 2.5 particulate matter in Birmingham (UK). Atmospheric Environment. 1997;31(24):4103–4117.

185. Cohen AJ, Arden Pope Iii C, Speizer FE. Ambient air pollution as a risk factor for lung cancer. Salud Pública de México. 1997 Jul;39(4):346–55.

186. Beeson WL, Abbey DE, Knutsen SF. Long-term concentrations of ambient air pollutants and incident lung cancer in California adults: results from the AHSMOG study.Adventist Health Study on Smog. Environ Health Perspect. 1998 Dec;106(12):813–22.

187. Lung cancer incidence statistics [Internet]. Cancer Research UK. 2015 [cited 2016 Aug 4]. Available from: http://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/lung-cancer/incidence

188. Iyen-Omofoman B, Hubbard RB, Smith CJ, Sparks E, Bradley E, Bourke A, et al. The distribution of lung cancer across sectors of society in the United Kingdom: a study using national primary care data. BMC public health. 2011;11(1):857–66.

189. McKinley JM, Ofterdinger U, Young M, Barsby A, Gavin A. Investigating local relationships between trace elements in soils and cancer data. Spatial Statistics. 2013 Aug;5:25–41.

190. Szatkowski LC. Can primary care data be used to evaluate the effectiveness of tobacco control policies? Data quality, method development and assessment of the impact of smokefree legislation using data from the Health Improvement Network [Internet] [Doctoral thesis]. [UK]: University of Nottingham;

2011. Available from: http://eprints.nottingham.ac.uk/11902/1/Final_thesis_Lisa_Szatkowski.pdf

191. Hall MA. Correlation-based Feature Selection for Discrete and Numeric Class Machine Learning. In: Proceedings of the Seventeenth International Conference on Machine Learning [Internet]. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.; 2000 [cited 2017 Mar 22]. p. 359–366. (ICML '00). Available from: http://dl.acm.org/citation.cfm?id=645529.657793

192. Guyon I, Elisseeff A. An Introduction to Variable and Feature Selection. J Mach Learn Res. 2003;3:1157–1182.

193. Witten IH, Frank E, Hall MA. Data Mining: Practical Machine Learning Tools and Techniques: Practical Machine Learning Tools and Techniques. Elsevier; 2011. 1285 p.

194. Vittinghoff E, Glidden DV, Shiboski SC, McCulloch CE. Regression Methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models. Springer Science & Business Media; 2012. 526 p.

195. Cleves M. An Introduction to Survival Analysis Using Stata, Second Edition. Stata Press; 2008. 398 p.

196. Kleinbaum DG, Klein M. Survival Analysis: A Self-Learning Text. Springer Science & Business Media; 2005. 616 p.

197. Wagner SE, Burch JB, Bottai M, Puett R, Porter D, Bolick-Aldrich S, et al. Groundwater uranium and cancer incidence in South Carolina. Cancer Causes Control. 2010;22(1):41–50.

198. Do MT. Ionizing Radiation Exposure and Risk of Gastrointestinal Cancer: A Study of the Ontario Uranium Miners [Internet] [Thesis]. 2010 [cited 2015 Dec 7]. Available from: https://tspace.library.utoronto.ca/handle/1807/24304

199. Appleton JD, Miles JCH. Soil uranium, soil gas radon and indoor radon empirical relationships in the UK and other European countries [Internet]. Conference presented at: 10th International Workshop on the Geological Aspects of Radon Risk Mapping; 2010 [cited 2015 Dec 7]; Prague, Czech Republic. Available from: http://www.radon.eu/workshop2010/index.html

200. Schmid K, Kuwert T, Drexler H. Radon in Indoor Spaces. Dtsch Arztebl Int. 2010 Mar;107(11):181–6.

201. Ball TK, Cameron DG, Colman TB, Roberts PD. Behaviour of radon in the geological environment: a review. Quarterly Journal of Engineering Geology and Hydrogeology. 1991;24(2):169–182.

202. Krewski D, Yokel RA, Nieboer E, Borchelt D, Cohen J, Harry J, et al. Human health risk assessment for aluminium, aluminium oxide, and aluminium hydroxide. J Toxicol Environ Health B Crit Rev. 2007;10(Suppl 1):1–269.

203. Gibbs GW, Labrèche F. Cancer Risks in Aluminum Reduction Plant Workers. J Occup Environ Med. 2014 May;56(5 Suppl):S40–59.

204. Cancer risk among workers of a secondary aluminium smelter [Internet]. [cited 2016 Aug 5]. Available from: http://occmed.oxfordjournals.org/content/66/5/412.abstract

205. Agency for Toxic Substances and Disease Registry (ATSDR). Toxicological profile for Aluminum [Internet]. Atlanta: GA: U.S. Department of Health and Human Services; 2008 [cited 2015 Sep 30]. Available from: http://www.atsdr.cdc.gov/toxprofiles/tp22.pdf

206. Kraus T, Schaller KH, Angerer J, Letzel S. Aluminium dust-induced lung disease in the pyro-powder-producing industry: detection by high-resolution computed tomography. Int Arch Occup Environ Health. 73(1):61–4.

207. Mazzoli-Rocha F, Dos Santos AN, Fernandes S, Ferreira Normando VM, Malm O, Nascimento Saldiva PH, et al. Pulmonary function and histological impairment in mice after acute exposure to aluminum dust. Inhal Toxicol. 2010 Aug;22(10):861–7.

208. Anttila A, Heikkilä P, Pukkala E, Nykyri E, Kauppinen T, Hernberg S, et al. Excess lung cancer among workers exposed to lead. Scand J Work Environ Health. 1995 Dec;21(6):460–9.

209. Lundström NG, Nordberg G, Englyst V, Gerhardsson L, Hagmar L, Jin T, et al. Cumulative lead exposure in relation to mortality and lung cancer morbidity in a cohort of primary smelter workers. Scand J Work Environ Health. 1997 Feb;23(1):24–30.

210. Schumann RR, Owen DE. Relationships between geology, equivalent uranium concentration, and radon in soil gas, Fairfax County, Virginia [Internet]. US Geological Survey,; 1988 [cited 2016 Aug 5]. Available from: https://pubs.er.usgs.gov/publication/ofr8818

211. Nazaroff WW. Radon transport from soil to air. Reviews of Geophysics. 1992;30(2):137–160.

212. Sevc J, Kunz E, Placek V. Lung cancer in uranium miners and long-term exposure to radon daughter products. Health Physics. 1976;30(6):433–437.

213. Haynes K, Forde KA, Schinnar R, Wong P, Strom BL, Lewis JD. Cancer incidence in The Health Improvement Network. Pharmacoepidem Drug Safe. 2009 Aug 1;18(8):730–6.

214. Kampman E, Bueno-De-Mesquita HB, Boeing H, Gonzalez CA, Stam B, Veer PV, et al. Gastrointestinal cancer: epidemiology. In: Chapter 1 Kelson DP, Daly JM, Kern SE, Levin B, Tepper JE & Cutsem EV (editors) Principles and Practice of Gastrointestinal Oncology. Second Edition. Philadelphia, USA: Lippincott Williams & Wilkins; 2008. p. 3–14.

215. Dikshit R. Epidemiology of gastrointestinal cancers. In: Chapter 1 Sirohi B & Nundy S (editors) Adjuvant and neoadjuvant therapy in gastrointestinal cancer. New Delhi, India: Elsevier; 2014. p. 1–15.

216. Bishehsari F, Mahdavinia M, Vacca M, Malekzadeh R, Mariani-Costantini R. Epidemiological transition of colorectal cancer in developing countries: Environmental factors, molecular pathways, and opportunities for prevention. World J Gastroenterol. 2014 May 28;20(20):6055–72.

217. Bingham SA, Day NE, Luben R, Ferrari P, Slimani N, Norat T, et al. Dietary fibre in food and protection against colorectal cancer in the European Prospective Investigation into Cancer and Nutrition (EPIC): an observational study. The Lancet. 2003 May 3;361(9368):1496–501.

218. Howe GR, Benito E, Castelleto R, Cornée J, Estève J, Gallagher RP, et al. Dietary Intake of Fiber and Decreased Risk of Cancers of the Colon and Rectum: Evidence From the Combined Analysis of 13 Case-Control Studies. JNCI J Natl Cancer Inst. 1992 Dec 16;84(24):1887–96.

219. Terry P, Giovannucci E, Michels KB, Bergkvist L, Hansen H, Holmberg L, et al. Fruit, Vegetables, Dietary Fiber, and Risk of Colorectal Cancer. JNCI J Natl Cancer Inst. 2001 Apr 4;93(7):525–33.

220. Türkdoğan MK, Kilicel F, Kara K, Tuncer I, Uygan I. Heavy metals in soil, vegetables and fruits in the endemic upper gastrointestinal cancer region of Turkey. Environmental Toxicology and Pharmacology. 2003 Apr;13(3):175–9.

221. Mohajer R, Salehi MH, Mohammadi J, Emami MH, Azarm T. The status of lead and cadmium in soils of high prevalenct

gastrointestinal cancer region of Isfahan. J Res Med Sci. 2013 Mar;18(3):210–4.

222. Keshavarzi B, Moore F, Najmeddin A, Rahmani F. The role of selenium and selected trace elements in the etiology of esophageal cancer in high risk Golestan province of Iran. Science of the Total Environment. 2012;433:89–97.

223. Zhao Q, Wang Y, Cao Y, Chen A, Ren M, Ge Y, et al. Potential health risks of heavy metals in cultivated topsoil and grain, including correlations with human primary liver, lung and gastric cancer, in Anhui province, Eastern China. Science of The Total Environment. 2014 Feb 1;470–471:340–7.

224. Khan S, Cao Q, Zheng YM, Huang YZ, Zhu YG. Health risks of heavy metals in contaminated soils and food crops irrigated with wastewater in Beijing, China. Environmental pollution. 2008;152(3):686–692.

225. Stocks P, Davies RI. Epidemiological evidence from chemical and spectrographic analyses that soil is concerned in the causation of cancer. British journal of cancer. 1960;14(1):8.

226. Carruthers M, Smith B. Evidence of cadmium toxicity in a population living in a zinc-mining area: pilot survey of Shipham residents. The Lancet. 1979;313(8121):845–847.

227. Philipp R, Hughes AO, Robertson MC. Stomach cancer and soil metal content. British journal of cancer. 1982;45(3):482.

228. Philipp R, Hughes AO. Morbidity and soil levels of cadmium. International journal of epidemiology. 1982;11(3):257–260.

229. Philipp R. Cadmium–risk assessment of an exposed residential population: a review. Journal of the Royal Society of Medicine. 1985;78(4):328.

230. Su C-C, Lin Y-Y, Chang T-K, Chiang C-T, Chung J-A, Hsu Y-Y, et al. Incidence of oral cancer in relation to nickel and arsenic concentrations in farm soils of patients' residential areas in Taiwan. BMC Public Health. 2010;10:67.

231. Su C-C, Tsai K-Y, Hsu Y-Y, Lin Y-Y, Lian I-B. Chronic exposure to heavy metals and risk of oral cancer in Taiwanese males. Oral Oncology. 2010 Aug;46(8):586–90.

232. Chen Y-C, Guo Y-LL, Su H-JJ, Hsueh Y-M, Smith TJ, Ryan LM, et al. Arsenic methylation and skin cancer risk in southwestern Taiwan. J Occup Environ Med. 2003;45(3):241–8.

233. Kucharzewski M, Braziewicz J, Majewska U, Gózdz S. Iron concentrations in intestinal cancer tissue and in colon and rectum polyps. Biol Trace Elem Res. 2003 Oct;95(1):19–28.

234. Alimonti A, Bocca B, Lamazza A, Forte G, Rahimi S, Mattei D, et al. A study on metals content in patients with colorectal polyps. J Toxicol Environ Health Part A. 2008;71(5):342–7.

235. Klimczak M, Dziki A, Kilanowicz A, Sapota A, Duda-Szymańska J, Daragó A. Concentrations of cadmium and selected essential elements in malignant large intestine tissue. Prz Gastroenterol. 2016;11(1):24–9.

236. Baykara O, Dogru M. Measurements of radon and uranium concentration in water and soil samples from East Anatolian Active Fault Systems (Turkey). Radiation Measurements. 2006;41(3):362–367.

237. Bowel cancer incidence statistics | Cancer Research UK [Internet]. [cited 2016 Aug 5]. Available from: http://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/bowel-cancer/incidence

238. Oral cancer statistics | Cancer Research UK [Internet]. [cited 2016 Aug 5]. Available from: http://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/oral-cancer

239. Stomach cancer statistics | Cancer Research UK [Internet]. [cited 2016 Aug 5]. Available from: http://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/stomach-cancer

240. Breast cancer statistics | Cancer Research UK [Internet]. [cited 2016 Aug 5]. Available from: http://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/breast-cancer

241. Lung cancer incidence statistics [Internet]. Cancer Research UK. 2015 [cited 2016 Aug 5]. Available from: http://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/lung-cancer/incidence

242. Prostate cancer statistics [Internet]. Cancer Research UK. 2015 [cited 2016 Aug 5]. Available from: http://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/prostate-cancer

243. Griffith GW, Davies RI. Cancer and Soils in the County of Anglesey.—A Revised Method of Comparison. Br J Cancer. 1954 Dec;8(4):594–8.

244. Stocks P, Davies RI. Epidemiological Evidence from Chemical and Spectrographic Analyses that Soil is Concerned in the Causation of Cancer. Br J Cancer. 1960 Mar;14(1):8–22.

245. Millar IB. Gastro-intestinal Cancer and Geochemistry in North Montgomeryshire. Br J Cancer. 1961 Jun;15(2):175–99.

246. Ashley DJ, Davies HD. Gastric cancer in Wales. Gut. 1966;7(5):542–548.

247. Philipp R, Hughes AO, Robertson MC. Stomach cancer and soil metal content. Br J Cancer. 1982 Mar;45(3):482.

248. Carruthers M, Smith B. Evidence of cadmium toxicity in a population living in a zinc-mining area: Pilot survey of Shipham residents. The Lancet. 1979 Apr 21;313(8121):845–7.

249. Philipp R, Hughes AO. Morbidity and Soil Levels of Cadmium. Int J Epidemiol. 1982 Sep 1;11(3):257–60.

250. Philipp R. Cadmium--risk assessment of an exposed residential population: a review. J R Soc Med. 1985 Apr;78(4):328–33.

251. Philipp R, Hughes A. Health risks from exposure to cadmium in soil. Occup Environ Med. 2000 Sep 1;57(9):647–8.

252. Fine JP, Gray RJ. A Proportional Hazards Model for the Subdistribution of a Competing Risk. Journal of the American Statistical Association. 1999;94(446):496–509.

253. Andersen PK, Geskus RB, de Witte T, Putter H. Competing risks in epidemiology: possibilities and pitfalls. Int J Epidemiol. 2012 Jun;41(3):861–70.

254. Wu K, Willett WC, Fuchs CS, Colditz GA, Giovannucci EL. Calcium Intake and Risk of Colon Cancer in Women and Men. JNCI J Natl Cancer Inst. 2002 Mar 20;94(6):437–46.

255. Kasum CM, Jacobs DR, Nicodemus K, Folsom AR. Dietary risk factors for upper aerodigestive tract cancers. Int J Cancer. 2002 May 10;99(2):267–72.

256. Lamprecht SA, Lipkin M. Chemoprevention of colon cancer by calcium, vitamin D and folate: molecular mechanisms. Nat Rev Cancer. 2003 Aug;3(8):601–14.

257. Dhawan DK, Chadha VD. Zinc: A promising agent in dietary chemoprevention of cancer. Indian J Med Res. 2010 Dec;132(6):676–82.

258. Li P, Xu J, Shi Y, Ye Y, Chen K, Yang J, et al. Association between zinc intake and risk of digestive tract cancers: a

systematic review and meta-analysis. Clin Nutr. 2014 Jun;33(3):415–20.

259. Skrovanek S, DiGuilio K, Bailey R, Huntington W, Urbas R, Mayilvaganan B, et al. Zinc and gastrointestinal disease. World J Gastrointest Pathophysiol. 2014 Nov 15;5(4):496–513.

260. Yokel RA, Rhineheimer SS, Brauer RD, Sharma P, Elmore D, McNamara PJ. Aluminum bioavailability from drinking water is very low and is not appreciably influenced by stomach contents or water hardness. Toxicology. 2001;161(1):93–101.

261. Auvinen A, Salonen L, Pekkanen J, Pukkala E, Ilus T, Kurttio P. Radon and other natural radionuclides in drinking water and risk of stomach cancer: A case-cohort study in Finland. Int J Cancer. 2005 Mar 10;114(1):109–13.

262. Wulaningsih W, Michaelsson K, Garmo H, Hammar N, Jungner I, Walldius G, et al. Inorganic phosphate and the risk of cancer in the Swedish AMORIS study. BMC Cancer. 2013 May 24;13(1):1–8.

263. Argos M, Rahman M, Parvez F, Dignam J, Islam T, Quasem I, et al. Baseline comorbidities in a skin cancer prevention trial in Bangladesh. Eur J Clin Invest. 2013;43(6):579–88.

264. Surdu S, Fitzgerald EF, Bloom MS, Boscoe FP, Carpenter DO, Haase RF, et al. Occupational exposure to arsenic and risk of nonmelanoma skin cancer in a multinational European study. Int J Cancer. 2013;133(9):2182–91.

265. Ealing Council. Ealing facts and figures - Census and population density [Internet]. [cited 2017 Mar 24]. Available from: https://www.ealing.gov.uk/downloads/201051/census

266. Agency for Toxic Substances and Disease Registry (ATSDR). Appendix G: Calculating Exposure Doses | PHA Guidance Manual [Internet]. [cited 2017 Mar 24]. Available from: https://www.atsdr.cdc.gov/hac/phamanual/appg.html

8 List of Appendices

8.1 BCC incidence in the UK (publication)

EPIDEMIOLOGY AND HEALTH SERVICES RESEARCH

BJD
British Journal of Dermatology

Regional variations of basal cell carcinoma incidence in the U.K. using The Health Improvement Network database (2004–10)

A. Musah,[1] J.E. Gibson,[1] J. Leonardi-Bee,[1] M.R. Cave,[2] E.L. Ander[2] and F. Bath-Hextall[3]

[1]Division of Epidemiology and Public Health, University of Nottingham, Clinical Sciences Building (Phase 2), Nottingham NG5 1PB, U.K.
[2]British Geological Survey, Environmental Science Centre, Nottingham NG12 5GG, U.K.
[3]Centre of Evidence Based Dermatology, University of Nottingham, King's Meadow Campus, Nottingham NG7 2NR, U.K.

Correspondence
Anwar Musah.
E-mail: moxm23@nottingham.ac.uk

Accepted for publication
14 May 2013

Funding sources
The authors acknowledge the following institutions for their support: Biomedical Research Committee, The University of Nottingham, Nottingham, U.K., British Geological Survey, Nottingham, U.K., and Cegedim Strategic Data – Medical Research, U.K.

Conflicts of interest
None declared.

DOI 10.1111/bjd.12446

Summary

Background Basal cell carcinoma (BCC) is one of the most common types of nonmelanoma skin cancer affecting the white population; however, little is known about how the incidence varies across the U.K.

Objectives To determine the variation in BCC throughout the U.K.

Methods Data from 2004 to 2010 were obtained from The Health Improvement Network database. European and world age-standardized incidence rates (EASRs and WASRs, respectively) were obtained for country-level estimates and levels of socioeconomic deprivation, while strategic health-authority-level estimates were directly age and sex standardized to the U.K. standard population. Incidence-rate ratios were estimated using multivariable Poisson regression models.

Results The overall EASR and WASR of BCC in the U.K. were 98·6 per 100 000 person-years and 66·9 per 100 000 person-years, respectively. Regional-level incidence rates indicated a significant geographical variation in the distribution of BCC, which was more pronounced in the southern parts of the country. The South East Coast had the highest BCC rate followed by South Central, Wales and the South West. Incidence rates were substantially higher in the least deprived groups and we observed a trend of decreasing incidence with increasing levels of deprivation ($P < 0·001$). Finally, in terms of age groups, the largest annual increase was observed among those aged 30–49 years.

Conclusions Basal cell carcinoma is an increasing health problem in the U.K.; the southern regions of the U.K. and those in the least deprived groups had a higher incidence of BCC. Our findings indicate an increased incidence of BCC for younger age groups below 49 years.

What's already known about this topic?

- The incidence rate of basal cell carcinoma in the U.K. is high compared with other types of nonmelanoma skin cancers.
- Risk factors include exposure to ultraviolet radiation through sunlight or tanning beds, and age and ethnicity.
- Incidence in the younger population is rising, although incidence increases after 40 years of age.

What does this study add?

- These findings provide novel estimates for regional incidence rates across the U.K.
- They also provide novel estimates for levels of socioeconomic deprivation in the U.K.

Basal cell carcinoma (BCC), a form of nonmelanoma skin cancer (NMSC), is the most common malignant neoplasm found in humans, and the incidence is increasing in the U.S.A., Canada, Australia and most European countries.[1–5] Risk factors include exposure to ultraviolet (UV) radiation through sunlight or tanning beds,[6–8] advancement of age,[9] sex,[9,10] skin type (fair, white or freckled skin),[9,11–13] history of skin cancer[14] and some environmental and occupational factors.[15]

The incidence of BCC in the U.K. is increasing at an unprecedented rate. The overall incidence of NMSC is estimated to be well over 100 000 cases per year, with BCC accounting for 75% of cases, squamous cell carcinoma for 20% and other rare skin cancer types for 5% (e.g. Merkel cell carcinoma, Kaposi sarcoma and T-cell lymphoma of the skin).[16] Recent studies have shown that while the incidence of BCC varies geographically in the U.K.,[1,17] rapidly increasing incidence has been observed in many areas. For instance, in West Glamorgan (Wales), the incidence rate (IR) increased by 60% between 1988 and 1998.[18] Similarly, in England, the incidence of BCC in North Humberside tripled over the 13-year period from 1978 to 1991.[1,19] Scotland and Northern Ireland have lower incidence of BCC relative to England and Wales; however, within the past two decades the incidence of BCC among men has risen by approximately 16% in Scotland and 18% in Northern Ireland.[20,21] The elderly population contributes substantially to the disease burden, with the risk of BCC increasing after 40 years of age; however, we are now seeing an increased incidence in younger people, and in particular those aged ≤ 30 years.[22]

Socioeconomic status and deprivation are also known to modify the risk of BCC. Some studies suggest that BCC appears to be more common in those belonging to higher social class.[13,23,24] However, such associations are not well understood, and the distribution of BCC in terms of levels of socioeconomic deprivation in the U.K. population is unknown.

We therefore used data from a U.K.-wide database of primary care medical records to derive contemporary regional breakdowns of the incidence of BCC in the U.K. We present novel incidence-rate estimates stratified by the level of socioeconomic deprivation in the U.K., and additional analyses examining whether BCC incidence has continued to increase in recent years, particularly in the younger age groups.

Materials and methods

Study design and data source

We conducted a large, population-based study using data from The Health Improvement Network (THIN). THIN is a large database comprised of anonymized primary-care electronic medical records of more than 10 million patients from across all regions of the U.K. The information contained within THIN includes details on all diagnoses made by or reported to general practitioners, as well as other additional health information relevant to primary care. THIN is recognized for its completeness and accuracy of data recording, and has been validated for its suitability for use in medical research,[25] including specific validation of diagnoses of BCC.[26] In addition, several sociodemographic indicators are available in THIN, including quintiles of the Townsend Deprivation Index[27] in each patient's postcode of residence.

Study population

The medical histories and deprivation indicators of all adults with a first recorded diagnosis of BCC between 1 January 2004 and 31 December 2010 were extracted from THIN. Subjects diagnosed with basal cell naevus syndrome (Gorlin syndrome), organoid naevi or other genetic syndromes were excluded from the study. Patient ages were categorized into 10–15-year bands (18–29, 30–39, 40–49, 50–64, 65–79 and ≥ 80 years). Patients were categorized into 13 regions based on the Strategic Health Authority (SHA; in England) or devolved government administration (Wales, Scotland and Northern Ireland) to which each patient's primary care practice belongs. The regions are as follows: North East, North West, Yorkshire and Humber, East Midlands, West Midlands, East of England, London, South East Coast, South Central, South West, Wales, Scotland and Northern Ireland. These are the only spatially referenced data available from the anonymized patient records.

Statistical analysis

The primary outcome measures were IRs of BCC in the whole of the U.K., in its constituent countries and principalities, and in each English SHA region. We also estimated the incidence of BCC across quintiles of socioeconomic deprivation groups. IRs were calculated as the number of patients receiving their first diagnosis of BCC divided by the total number of person-years at risk. Diagnoses within the first year of registration with a participating primary care practice were excluded, as such recordings can relate to prevalent, rather than incident, events (being an artefact of back entry of records from a previous practitioner). Second or subsequent diagnoses were also excluded as these are difficult to differentiate from recurrences and follow-up consultations in primary care records. Population denominators were midyear (1 July) total numbers of persons registered for at least 1 year at a primary care practice enrolled in THIN. IRs are presented as rates per 100 000 person-years. We derived estimates for European and world age-standardized incidence rates (EASRs and WASRs, respectively) using a direct standardization method to allow direct comparisons between country-level incidence rates (i.e. U.K., England, Scotland, Northern Ireland and Wales) with other populations.[28] IRs of BCC at a regional level were directly age and sex standardized to the U.K. standard population.[29]

A Poisson multivariable regression model was used to examine the effects of all factors (i.e. calendar year of diagnosis, socioeconomic deprivation and regions) on the incidence of BCC adjusted for sex and age groups. Stratified Poisson multivariable analyses were used to determine whether

associations between all factors and the incidence of BCC were modified by sex, while controlling for age groups. For secondary analyses, we further used stratified models by age groups to assess calendar years as a continuous variable in order to determine the average change (per year) in incidence of BCC. IR ratios (IRRs) were estimated with 95% confidence intervals (CIs). All statistical analyses were carried out using Stata version 12 (StataCorp, College Station, TX, U.S.A.).

Results

There were 38 121 incident cases of BCC identified from 546 general practices in the THIN database. The mean age was 64 ± 13 years, with slightly more men (52·4%) diagnosed with a BCC.

Incidence rates of basal cell carcinoma at country and regional level

The crude IR of BCC between 2004 and 2010 in our THIN database was 171·9 per 100 000 person-years (95% CI 170·1–173·6). The crude IR of BCC was higher among men (183·1, 95% CI 180·5–185·6) than women (161·0, 95% CI 158·7–163·4) (Table 1). When comparing the overall figures between 2004 and 2010, we found that there was an increase from 154·0 to 182·0 per 100 000 person-years. Our Poisson multivariable regression model showed an overall, significant 16% increase in incidence in 2010 compared with 2004 (IRR 1·16, 95% CI 1·11–1·20), which equates to an average increase of 2·5% per year (95% CI 1·9–3·0; P for trend < 0·001) (Table 2).

Comparatively, at a country level, Wales has significantly the highest overall crude rate of BCC (IR 196·4, 95% CI 189·2–203·8), followed by England (IR 178·5, 95% CI 176·5–180·5). We observed that the incidences were low in Scotland (IR 128·7, 95% CI 124·6–133·0) and Northern Ireland (IR 131·6, 95% CI 123·3–140·3) (Table 1), and from the 95% CIs there is no significant difference in the incidence of BCC between Scotland and Northern Ireland.

We observed important geographical variations in the distribution of BCC (Fig. 1). The age- and sex-standardized IRs of BCC in the South East Coast, South Central, Wales and South West are significantly higher than in other regions of the U.K. (Table 3). Our models show that the incidence of BCC was significantly lower in the West Midlands (IRR 0·92, 95% CI 0·88–0·97), Northern Ireland (IRR 0·92, 95% CI 0·85–0·98) and Scotland (IRR 0·87, 95% CI 0·83–0·91) than in London (the referent). Conversely, we found that the incidence of BCC was significantly higher in the South East Coast (IRR 1·28, 95% CI 1·22–1·34), Wales (IRR 1·23, 95% CI 1·16–1·29), South Central (IRR 1·21, 95% CI 1·15–1·27) and South West (IRR 1·15, 95% CI 1·09–1·21) than in London. Our results show no substantial sex-specific difference in the incidence of BCC in any region.

Table 1 Crude and sex-specific age-standardized incidence rates of basal cell carcinoma in the U.K. and by country, The Health Improvement Network database (2004–10)

	Men (n)	Women (n)	Overall (N)
U.K.			
Crude	183·1 (19 960)	161·0 (18 161)	171·9 (38 121)
EASR	112·2	88·1	98·6
WASR	74·8	60·7	66·9
England			
Crude	189·9 (16 079)	167·5 (14 671)	178·5 (30 750)
EASR	114·9	91·4	101·5
WASR	76·6	63·1	69·0
Northern Ireland			
Crude	144·7 (502)	119·3 (439)	131·6 (941)
EASR	99·6	67·5	81·6
WASR	66·2	45·2	54·6
Scotland			
Crude	137·9 (1904)	119·8 (1704)	128·7 (3608)
EASR	89·3	65·6	75·9
WASR	59·1	44·7	51·1
Wales			
Crude	208·1 (1475)	185·0 (1347)	196·4 (2822)
EASR	128·7	103·1	114·4
WASR	86·4	71·3	78·1

EASR, European age-standardized rate; WASR, world age-standardized rate.

Trends in basal cell carcinoma over time by age group

Models were stratified by age groups to determine the effects of calendar years on the incidence of BCC. Our results show a small average increase of 0·4% per year in the age group of 18–29 years; however, this failed to reach statistical significance (95% CI −8·0 to 9·3%, P = 0·91). In particular, the largest average increases in incidence were observed for those aged 30–39 years (3·9% per year, 95% CI 0·2–7·7, P = 0·04) and 40–49 years (4·0% per year, 95% CI 2·0–6·1, P < 0·001) (Fig. 2).

Incidence rates by socioeconomic deprivation

The crude incidence of BCC was significantly highest for those living in areas with lower levels of deprivation, with estimates of 222·5 per 100 000 person-years (95% CI 218·5–226·5) and 203·2 (95% CI 199·1–207·3) for those in the 1st quintile (least deprived) and 2nd quintile, respectively (Table 4). We observed that the incidence of BCC was lowest for those living in the most deprived areas (IR 110·7, 95% CI 106·8–114·7).

Using our models, there appeared to be a linear effect of decreasing incidence of BCC with increasing levels of deprivation (P for trend < 0·001). We found that those living in the least deprived areas were significantly more likely (by 50%) to have a BCC than those with the highest levels of deprivation (IRR 1·50, 95% CI 1·44–1·56). Our models also show a substantial difference in the magnitude of BCC incidence between men and women, where the IR was higher in men than in women.

Table 2 Overall and sex-specific incidence-rate ratio (IRR) estimates showing associations between incidence of basal cell carcinoma and risk factors

	Men[a] IRR (95% CI)	Women[a] IRR (95% CI)	Overall[b] IRR (95% CI)
Year			
2004	1	1	1
2005	1·07 (1·02–1·14)	1·00 (0·94–1·06)	1·04 (0·99–1·08)
2006	1·10 (1·04–1·16)	1·07 (1·01–1·13)	1·09 (1·04–1·13)
2007	1·14 (1·08–1·21)	1·16 (1·10–1·23)	1·15 (1·10–1·20)
2008	1·16 (1·10–1·22)	1·16 (1·10–1·23)	1·16 (1·12–1·20)
2009	1·15 (1·09–1·22)	1·13 (1·07–1·20)	1·15 (1·10–1·19)
2010	1·12 (1·06–1·18)	1·21 (1·14–1·27)	1·16 (1·12–1·21)
Annual increase, %	1·8 (1·1–2·5)	3·2 (2·5–4·0)	2·5 (1·9–3·0)
P for trend	0·003	0·008	< 0·001
Socioeconomic deprivation[c]			
5th (most deprived)	1	1	1
4th	1·13 (1·06–1·20)	1·01 (0·95–1·08)	1·07 (1·02–1·12)
3rd	1·28 (1·21–1·36)	1·13 (1·07–1·20)	1·21 (1·16–1·26)
2nd	1·47 (1·39–1·56)	1·26 (1·19–1·33)	1·37 (1·31–1·43)
1st (least deprived)	1·62 (1·53–1·72)	1·36 (1·28–1·44)	1·50 (1·44–1·56)
Unknown	1·22 (1·11–1·35)	1·12 (1·01–1·23)	1·17 (1·09–1·25)
P for trend	< 0·001	< 0·001	< 0·001
Regions			
London	1	1	1
Scotland	0·91 (0·85–0·98)	0·82 (0·76–0·88)	0·87 (0·83–0·91)
Northern Ireland	0·99 (0·89–1·10)	0·84 (0·76–0·94)	0·92 (0·85–0·98)
West Midlands	0·93 (0·87–1·00)	0·92 (0·85–0·99)	0·92 (0·88–0·97)
North West	0·98 (0·91–1·05)	0·93 (0·87–1·00)	0·96 (0·91–1·01)
Yorkshire & Humber	1·04 (0·96–1·15)	0·98 (0·89–1·08)	1·01 (0·95–1·08)
East Midlands	1·02 (0·93–1·11)	1·02 (0·93–1·12)	1·02 (0·96–1·09)
North East	1·08 (0·97–1·19)	1·03 (0·93–1·14)	1·05 (0·98–1·13)
East of England	1·05 (0·98–1·13)	1·02 (0·95–1·10)	1·04 (0·98–1·09)
Wales	1·25 (1·17–1·35)	1·19 (1·11–1·29)	1·23 (1·16–1·29)
South Central	1·24 (1·16–1·32)	1·18 (1·10–1·26)	1·21 (1·15–1·27)
South West	1·19 (1·12–1·28)	1·10 (1·03–1·18)	1·15 (1·09–1·21)
South East Coast	1·30 (1·21–1·39)	1·27 (1·18–1·36)	1·28 (1·22–1·34)

CI, confidence interval. [a]Models were stratified by sex, include all covariates and are adjusted for age groups: 18–29, 30–39, 40–49, 50–64, 65–79 and ≥ 80 years. [b]Overall model includes all covariates and is adjusted for sex and age bands, as above. [c]Quintiles of Townsend Deprivation Index.

Discussion

Our results indicate that the IR of BCC is increasing in the general population, in particular among those aged 30–69 years. They show that Wales and the southern parts of England have the highest recorded rates of BCC. For socioeconomic deprivation, incidence of BCC was consistently higher in the least deprived groups. Based on our estimates (i.e. EASRs), they show that approximately 61 500 new cases of BCC are diagnosed annually in the U.K. population. Previous reports using EASRs have estimated that 53 000 cases of BCC were reported yearly, using a cohort between 1996 and 2003.[17] Comparatively, this represents an overall increase of 16% in diagnosis rates in the past decade.

This study has several strengths; to our knowledge it uses the largest sample size of incident cases of BCC compared with previous research conducted in the U.K.[13,17–21,23] Due to our large sample size, our findings are unlikely to be by chance. Also, the data were obtained from a national database and prospectively recorded by general practitioners, thus excluding the possibility of recording or recall bias in either exposure or outcome. The major limitation is our inability to account for important factors such as history of sun exposure during childhood and adolescence (frequency of sunburns and overseas holidays),[30,31] latitudinal position (proximity to the equator),[9,32] settings of occupation (indoor, mixed or outdoor)[33] and skin type (fair, white or freckled skin).[13] In addition, we were unable to classify adults according to subtypes of BCC (superficial, nodular or infiltrative).

Our results for country-level IRs are consistent with previous studies showing escalating rates in England, Northern Ireland, Scotland and Wales.[18–21] The most likely explanation for the rise in incidence may be linked to previous behaviour with regards to sun exposure during childhood or adolescence. Exposure to UV radiation during this stage plays a significant role in the future development of BCC. Previous studies have shown that subjects who have reported travelling

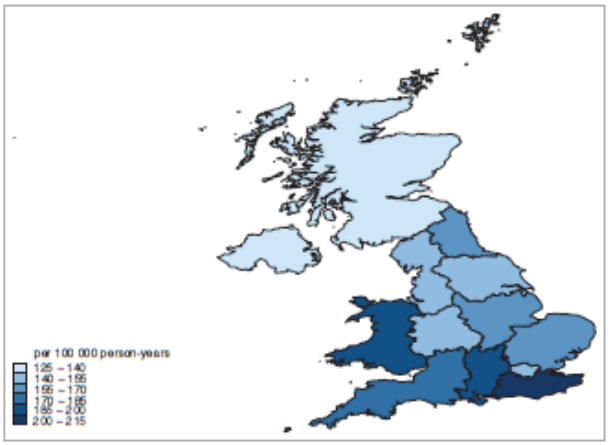

Fig 1. Thematic map showing direct age and sex-standardized incidence rates of basal cell carcinoma in the U.K. standard population (The Health Improvement Network database 2004–10).

Table 3 Regional-level estimates for sex-specific and age and sex-standardized incidence rates of basal cell carcinoma in the U.K., The Health Improvement Network database (2004–10)

Regions	Age- and sex-standardized rate[a] Overall	Sex-specific age-standardized rate[b] Men (n)	Women (n)
Scotland	127·9	139·5 (1904)	116·8 (1704)
Northern Ireland	138·4	155·4 (502)	122·2 (439)
London	144·0	148·6 (1432)	139·6 (1415)
West Midlands	144·1	152·2 (1537)	136·3 (1422)
North West	146·5	156·6 (1785)	136·8 (1627)
Yorkshire & Humber	151·4	163·3 (686)	140·0 (618)
North East	156·0	165·2 (539)	147·2 (503)
East Midlands	158·6	166·4 (776)	151·2 (718)
East of England	161·1	170·7 (1457)	151·8 (1325)
South West	180·2	196·2 (2438)	165·0 (2123)
Wales	185·7	197·6 (1475)	174·4 (1347)
South Central	193·5	208·7 (2998)	178·9 (2645)
South East Coast	202·7	215·0 (2431)	191·0 (2275)

[a]Estimates are directly age and sex standardized using the U.K. as the standard population. [b]Sex-specific estimates are directly age standardized using the U.K. as the standard population.

frequently and spending more than 4–5 weeks per year at the beach before the age of 20 years were more likely to have developed the skin malignancy in adulthood.[30,31] Although we were unable to account for this factor, history of sun exposure through frequent holidays to sunnier places has been a strong predictor for BCC. Another likely explanation may be the U.K.'s ageing population. BCC is highly prevalent in the older age groups; in our cohort, the number of cases diagnosed with the skin malignancy was consistently high among those aged ≥ 50 years.

We found a significant increase in the incidence of BCC among those aged 30–39 and 40–49 years. A previous study has shown similar findings, where the annual increases in incidence were estimated to be approximately 3·9% and 5·2% for those aged 30–39 and 40–49 years, respectively, although these estimates did not reach statistical significance.[17] Interestingly, we found an increase among those aged 18–29 years, although our models showed no statistical significance. The incidence of BCC in this particular age group rose to approximately 5·2 per 100 000 person-year in 2009 (Fig. 2), which is consistent with the escalating rates observed by others.[22]

We observed especially high IRs for the South East Coast, South Central, South West and Wales. Compared with London, we found there was a significant increase in the risk of developing BCC in these areas. This observation may be linked to several environmental factors. The most prominent is the latitudinal position of a location.[9,32] Areas in proximity to the equator, but situated in the temperate zone, usually experience a prolonged duration of sunlight in the summer season. In the U.K., the hours of sunshine are normally greater in the southern than the north-

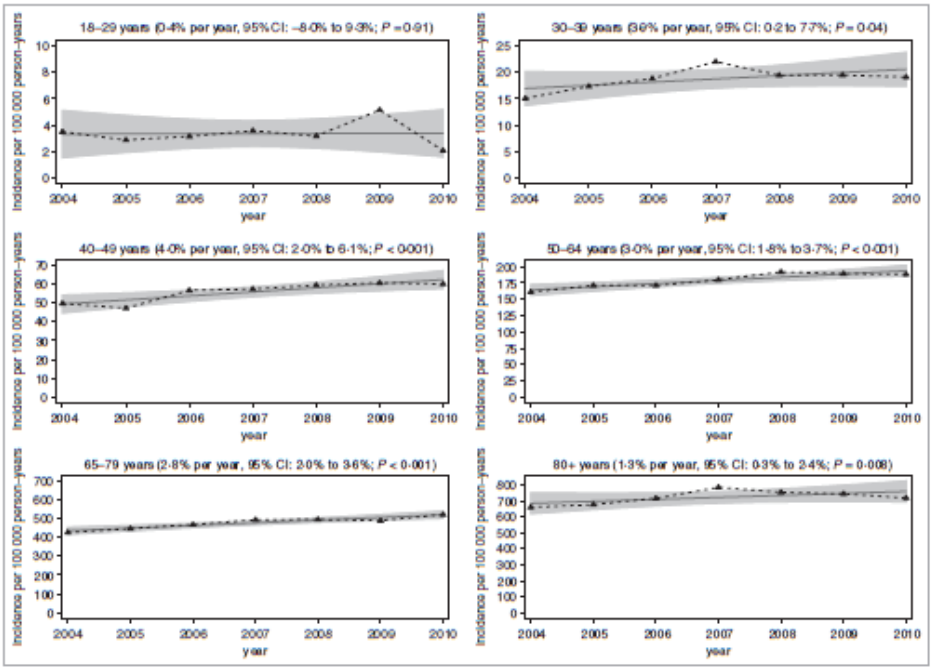

Fig 2. Average change in incidence of basal cell carcinoma in the U.K. stratified by age groups (The Health Improvement Network database 2004–10). CI, confidence interval.

Table 4 Crude and sex-specific age-standardized incidence rates of basal cell carcinoma in the U.K., by quintiles of Townsend Deprivation Index, The Health Improvement Network database (2004–10)

Deprivation index	1	2	3	4	5	Unknown
Overall						
Crude	222·5 (12 070)	203·2 (9575)	162·1 (7175)	131·7 (5173)	110·7 (3013)	115·6 (1115)
EASR	120·2	106·9	92·2	79·4	70·6	–
WASR	82·1	72·7	62·4	53·5	47·2	–
Men						
Crude	246·7 (6602)	220·8 (5097)	169·7 (3677)	134·3 (2583)	106·0 (1441)	119·0 (560)
EASR	137·5	122·0	104·7	90·4	77·4	–
WASR	92·3	81·5	69·6	59·7	50·9	–
Women						
Crude	198·9 (5468)	186·2 (4478)	154·7 (3498)	129·2 (2590)	115·3 (1572)	112·4 (555)
EASR	105·7	94·9	83·1	71·9	66·2	–
WASR	73·2	65·7	57·1	49·2	44·7	–

Number of population given in brackets. EASR, European age-standardized rate; WASR, world age-standardized rate.

ern regions of the U.K. Especially during the summer season, the southern parts of England and Wales are usually known to receive the greatest hours of annual sunshine.[34]

Our findings for socioeconomic deprivation show that the incidence of BCC is highest among the least deprived groups, and that the risk of BCC tends to decrease as the level of deprivation increases. Our results are consistent with previous studies conducted in the U.K. and the Netherlands.[13,26] It is interesting to note the wide difference in the incidence of BCC between the least and most deprived groups, which may be an indication that socioeconomic status or deprivation is a risk factor for BCC. This observation may be linked to higher levels of income for frequent holidays overseas to sunnier places, thereby exposing the skin to sunlight, or having avail-

able funds for pursuing other lifestyle habits that are risk factors for BCC, for instance the frequent use of tanning beds[8–10] or consumption of alcohol.[35] Interestingly, we also observed that the incidence differs substantially between sexes; this may be due to differences in behaviour in terms of sun exposure, clothing habits and tanning.[13,24]

Basal cell carcinoma is an increasingly important health problem in the U.K., with extremely high levels observed in the least deprived groups and in the southern parts of the U.K. Due to the multifactorial nature of BCC, further work is warranted to identify causes and to investigate the detailed reasons for these findings. Our results demonstrate that the incidence of BCC will continue to rise much higher in all age bands if it remains unchecked, which will have a significant impact on the workload and costs for health services. Better strategies are required to inform the public of the risk factors associated with the skin malignancy, and to implement measures to avoid future development of BCC.

References

1 Lomas A, Leonardi-Bee J, Bath-Hextall F. A systematic review of worldwide incidence of nonmelanoma skin cancer. *Br J Dermatol* 2012; **166**:1069–80.
2 Demers AA, Nugent Z, Mihalcioiu C *et al.* Trends of nonmelanoma skin cancer from 1960 through 2000 in a Canadian population. *J Am Acad Dermatol* 2005; **53**:320–8.
3 Jung GW, Metelitsa AI, Dover DC, Salopek TG. Trends in incidence of nonmelanoma skin cancers in Alberta, Canada, 1988–2007. *Br J Dermatol* 2010; **163**:146–54.
4 Staples MP, Elwood M, Burton RC *et al.* Non-melanoma skin cancer in Australia: the 2002 national survey and trends since 1985. *Med J Aust* 2006; **184**:6–10.
5 Rogers H, Weinstock M. Incidence estimate of nonmelanoma skin cancer in the United States, 2006. *Arch Dermatol* 2010; **146**:283–7.
6 Kricker A, Armstrong BK, English DR, Heenan PJ. Does intermittent sun exposure cause basal cell carcinoma? A case–control study in Western Australia. *Int J Cancer* 1995a; **60**:489–94.
7 Kricker A, Armstrong BK, English DR, Heenan PJ. A dose–response curve for sun exposure and basal cell carcinoma. *Int J Cancer* 1995b; **60**:482–8.
8 Karagas MR, Stannard VA, Mott LA *et al.* Use of tanning devices and risk of basal cell and squamous cell skin cancers. *J Natl Cancer Inst* 2002; **94**:224–6.
9 Tran H, Chen K, Shumack S. Epidemiology and aetiology of basal cell carcinoma. *Br J Dermatol* 2003; **149**:50–2.
10 Roewert-Huber J, Lange-Asschenfeldt B, Stockfleth E, Kerl H. Epidemiology and aetiology of basal cell carcinoma. *Br J Dermatol* 2007; **157**:47–51.
11 Marks R, Staples M, Giles GG. Trends in non-melanocytic skin cancer treated in Australia: the second national survey. *Int J Cancer* 1993; **53**:585–90.
12 Green A, Battistutta D, Hart V *et al.* Skin cancer in a subtropical Australian population: incidence and lack of association with occupation. *Am J Epidemiol* 1996; **144**:1034–40.
13 Lear JT, Tan BB, Smith AG *et al.* Risk factors for basal cell carcinoma in the U.K.: case–control study in 806 patients. *J R Soc Med* 1997; **90**:371–4.
14 Madan V, Lear JT, Szeimies R-M. Non-melanoma skin cancer. *Lancet* 2010; **375**:673–85.
15 Gawkrodger DJ. Occupational skin cancers. *Occup Med* 2004; **54**:458–63.
16 Cancer Research U.K. About skin cancer (non melanoma): a quick guide. Available at: http://www.cancerresearchuk.org/cancer-help/prod_consump/groups/cr_common/@cah/@gen/documents/generalcontent/about-skin-cancer.pdf (last accessed 9 July 2013).
17 Bath-Hextall F, Leonardi-Bee J, Smith C *et al.* Trends in incidence of skin basal cell carcinoma. Additional evidence from a U.K. primary care database study. *Int J Cancer* 2007; **121**:2105–8.
18 Holme SA, Malinovszky K, Roberts D. Changing trends in non-melanoma skin cancer in South Wales, 1988–98. *Br J Dermatol* 2000; **143**:1224–9.
19 Ko CB, Walton S, Kecskes K *et al.* The emerging epidemic of skin cancer. *Br J Dermatol* 1994; **130**:269–72.
20 Brewster DH, Bhatti LA, Inglis JHC *et al.* Recent trends in incidence of nonmelanoma skin cancers in the East of Scotland, 1992–2003. *Br J Dermatol* 2007; **156**:1295–300.
21 Hoey S, Devereux C, Murray L *et al.* Skin cancer trends in Northern Ireland and consequences for provision of dermatology services. *Br J Dermatol* 2007; **156**:1301–7.
22 Skellett AM, Hafiji J, Greenberg DC *et al.* The incidence of basal cell carcinoma in the under-30s in the U.K. *Clin Exp Dermatol* 2011; **37**:227–9.
23 Doherty V, Brewster D, Jensen S, Gorman D. Trends in skin cancer incidence by socioeconomic position in Scotland, 1978–2004. *Br J Cancer* 2010; **102**:1661–4.
24 van Hattem S, Aarts MJ, Louwman WJ *et al.* Increase in basal cell carcinoma incidence steepest in individuals with high socioeconomic status: results of a cancer registry study in the Netherlands. *Br J Dermatol* 2009; **161**:840–5.
25 Meal A, Leonardi-Bee J, Smith C *et al.* Validation of THIN data for non-melanoma skin cancer. *Qual Prim Care* 2008; **16**:49–52.
26 Bourke A, Dattani H, Robinson M. Feasibility study and methodology to create a quality-evaluated database of primary care data. *Inform Prim Care* 2004; **12**:171–7.
27 Townsend P. Deprivation. *J Soc Policy* 1987; **16**:125–46.
28 Boyle P, Parkin DM. Statistical methods for registries. In: *Cancer Registration: Principles and Methods* (Jensen OM, Parkin DM, MacLennan R, Muir CS, Skeet RG, eds). Lyon: International Agency for Research on Cancer Scientific Publications, 1991.
29 Office of National Statistics. National population projections, 2010-based projections. Available at: http://www.ons.gov.uk/ons/rel/npp/national-population-projections/2010-based-projections/index.html (last accessed 9 July 2013).
30 Corona R, Dogliotti E, D'Errico M *et al.* Risk factors for basal cell carcinoma in a Mediterranean population: role of recreational sun exposure early in life. *Arch Dermatol* 2001; **137**:1162–8.
31 Dessinioti C, Tzannis K, Sypsa V *et al.* Epidemiologic risk factors of basal cell carcinoma development and age at onset in a Southern European population from Greece. *Exp Dermatol* 2011; **20**:622–6.
32 Saraiya M, Glanz K, Briss PA *et al.* Interventions to prevent skin cancer by reducing exposure to ultraviolet radiation. *Am J Prev Med* 2004; **27**:422–66.
33 Freedman DM, Dosemeci M, McGlynn K. Sunlight and mortality from breast, ovarian, colon, prostate, and non-melanoma skin cancer: a composite death certificate based case–control study. *Occup Environ Med* 2002; **59**:257–62.
34 Swerdlow AJ. Incidence of malignant melanoma of the skin in England and Wales and its relationship to sunshine. *BMJ* 1979; **2**:1324–7.
35 Freedman DM, Sigurdson A, Doody MM *et al.* Risk of basal cell carcinoma in relation to alcohol intake and smoking. *Cancer Epidemiol Biomarkers Prev* 2003; **12**:1540–3.

8.2 Approved Protocol for data mining analysis and cohort studies

1. Title of the study

The influence of typical and widespread concentrations of potentially harmful elements in soil, on cancer risk in the general population in England and Wales

2. Background

Human exposure to potentially harmful elements from soil is common due to their ubiquity as naturally occurring elements, wide use in industry, and environmental persistence [1]. Historically, the greatest exposures occurred in the workplace, or in populations in close proximity to industrial sources. Most epidemiological research into the influence of exposure to the potentially harmful elements on the risk of cancer has been focused solely on occupational groups [2], because their levels of exposure are substantially greater than that of the general population, potentially limiting the generalizability of the findings.

Most potentially harmful elements that are commonly found in highly exposed occupational groups are in fact naturally occurring geochemical elements that are widely distributed in the environment [3]. These geochemical contaminants include arsenic, chromium, copper, lead, mercury, nickel and uranium [4]; and many of these elements have been shown to increase the risk of bladder, cutaneous, kidney, lung and digestive tract cancer [4]–[6]. For instance, the mass poisoning and numerous cancer outbreaks in the Indian and Bangladeshi population due to long-term consumption of geochemically contaminated ground water [7], [8]. Their major lithological and hydrological sources includes soil and groundwater, respectively [3], [9]. The primary pathways for environmental exposure are inhalation of particulate matter that were transported from the earth (or soil) to the atmosphere as a result of anthropogenic activities or wind erosion, ingestion of contaminated foods or drinking water obtained from contaminated areas, and through direct contact with the skin [1], [4], [10]

The potential mechanisms for cancer development as a result of geochemical exposure are somewhat similar to those derived from occupational exposures, except that the processes are on-going and long-term. The individual is exposed to a particular contaminant from the environment (soils) via the primary pathways; continued exposure eventually leads to accumulation of such toxicants which are broken down to metabolites causing toxicity to the body which, in turn, have genotoxic effects on various tissues such as the lungs, kidney or skin thereby resulting in significant DNA damage which leads to cancer [11]. There has been little or no focus on the potential health impacts of potentially harmful elements in soil emerging from lithological sources, in particular soils, due to the belief that exposures are not sufficiently direct, and levels of exposure are far lower than those found in occupational settings. For this reason, there is little direct evidence showing that potentially harmful elements in soil are risk factors for cancer [4], [12].

In the UK, there are widespread areas with elevated concentrations of potentially harmful elements, with specific geological formations yielding especially high concentrations in some areas of England and Wales [9], [13], [14]. These increased concentrations resulted mainly from the extensive mining activities that took place before the 1970s, and also due to natural lithological processes occurring in

soils such as mineralisation and rock formation [9]. Other parts of England and Wales have equally variable, but typically lower concentrations for specific elements. Of those elements considered potentially harmful, e.g. arsenic and lead, different processes are important in each case, and thus the elements are not necessarily positively correlated in their spatial distribution; this makes study of each of these important. In the UK, the Department for Environmental Food and Rural Affairs (DEFRA) have a set of screening values to assess soil safety in relation to human exposure to soil. For example, the SGV recommends that arsenic and nickel concentrations for residential soils should not exceed 32 and 130 mg/kg, respectively [15], [16]; however, there are residential soils in many areas of England and Wales with levels exceeding 100 mg/kg and 300 mg/kg, respectively [15]–[17]; whether this kind of long-term exposure for people living on soils with elevated arsenic concentration is contributory to the cancer incidence in UK remains unclear.

A small number of studies have found concentrations of geochemical elements, notably arsenic and lead, to correlate with biomarkers of exposure [4], [18], and that these biomarkers are, in turn, associated with an increased risk of various cancers [4][4], [6]. These possible causal pathways indicate the possibility of residents living in environments where potentially harmful element soil concentration is elevated, are prone to exposure via air, food and/or drinking water with increasing soil concentration. To our knowledge, there has been no research directly examining the relationship between geochemical exposure, at typical and widespread environmental concentrations, cancer risk in the UK. We therefore propose a series epidemiological studies using data mining and cohort studies to assess such influence of concentrations of environmental soil constituents on nonmelanoma skin, lung, gastrointestinal cancer risk in the England and Wales.

3. Purpose

The purpose of this research is to determine which of the 15 soil constituents[1] are associated with each of the following outcomes:

i. Nonmelanoma skin cancer (NMSC)
ii. Lung cancer
iii. Digestive tract cancer

A series of studies will be performed for each of the diseases; they will go through two stages with the following objectives:

Stage 1 Exploration for the possible soil constituents that will most likely be associated with cancer. In stage 1, we will establish a list of appropriate elements before using them in an epidemiologic study in stage 2.

Stage 2 A population based epidemiological study will be used to determine whether individuals living on residential soils with high soil elemental concentrations are more likely to

[1] Chemical elements include: Arsenic, Selenium, Lead, Zinc, Copper, Nickel, Chromium, Vanadium, Uranium, Aluminium, Iron, Calcium, Manganese, Phosphorus and Silicon

develop cancer compared to those on residential soils with the lower concentrations of potentially harmful elements.

4. Data source

THIN is an appropriate resource with which to carry out these studies because:

i. THIN is recognised for its accuracy and completeness of recording of medical data, has been validated for its suitability for use in medical research for several clinical outcomes and health indicators.

ii. THIN provides a range of additional information on potential confounding factors such as patients' demographics, rural vs. urban residence and smoking history.

iii. THIN has recently been linked, using patients' postcodes of residence, to data from the Geochemical Baseline Survey of the Environment (G-BASE) project – a systematic survey of chemical elements in soil samples (see above footnote) are collected throughout England and Wales only, which provides (for the majority of THIN patients in England and Wales) an estimate for 15 geochemical element concentration within 1-2km of individuals' homes.

5. Methods

5.1. Study design

Stage 1 Data mining will be used to determine which of the 15 constituents are significantly associated with each cancer outcome in our linked database.

Stage 2 A population based cohort study design to quantify the risks associated between the development of each cancer, with increasing levels of exposure to specific elements in soil at patient level.

5.2. Study population

For inclusion, all eligible patients registered in THIN (at a practice with G-BASE coverage) must be at the age of 18 years or above without any previous history of any cancer diagnosis before 1st January 2000. They must be registered with a general practice where G-BASE coverage is available. In addition, they must be registered with a general practice for least one year before the start date of study (i.e. 1st January 2000). Furthermore, they must be registered from a general practice which has an AMR date before the 1st January 2000.

Identification of cases and extraction of individual medical histories from the main THIN database will be carried out using PostgreSQL 9.3 for OSX 10.10. Stage 1 of the data mining analyses will be performed in R version 2.14.0 (R foundation for statistical computing, Vienna, Austria). Stage 2 of the statistical analyses will be carried out using Stata MP4 version 12 (Stata Corporation, College Station, Texas, US) for Windows x64.

5.2.1. Nonmelanoma skin cancer (Basal cell carcinoma only)

Patients with a first recorded diagnosis of basal cell carcinoma (BCC) between January 1st 2000 and December 31st 2011 will be identified using our previously validated code list [19]. Patients found with Basal cell nevus syndrome (or Gorlin's syndrome), organoid naevi or other BCC-related genetic diseases will be excluded from the study. Patients with cutaneous squamous cell carcinoma (SCC) will be excluded from the analysis. We have taken this approach because the way in which the code lists for cutaneous SCC are presented in THIN poses a challenge for us to authentically identify patients that were diagnosed with cutaneous SCC.

5.2.2. Lung cancer

Patients with a first recorded diagnosis of lung cancers between the study period January 1st 2000 and December 31st 2011 with included into the study. Subjects identified with the following malignancies: small or large cell carcinoma, SCC of the lung, adenocarcinoma or any lung–related neoplasms, will be extracted using the medical read codes under the following hierarchies: malignant neoplasm of the trachea, bronchus and lung *B22..00*; Carcinoma in situ of the respiratory system *B81..00;* and lung-related neoplasms otherwise specified *By...00*. Patients found with any restrictive fibrotic or rare lung-related genetic diseases will be excluded from the study. Additional lung cancer data from the Higher Episode Statistics (HES) and the national Lung Cancer Audit (LUCADA) will be added to this study, minimising the risk that cases diagnosed in hospital may be missed if details are not returned to the primary care practitioner or properly coded when added to the electronic medical record. Data from secondary care will also provide useful additional details such as disease stage at diagnosis.

5.2.3. Digestive tract cancer

Patients will be identified with a first recorded diagnosis of a digestive tract cancer from the 1st of January 1st 2000 to the 31st December 2011. Subjects identified with malignancies of the oesophagus, stomach, ileum, colon, rectum or any gastrointestinal tract related cancer will be extracted using the medical read codes under the following hierarchies: malignant neoplasm of the digestive organs and peritoneum *B1...00*; and neoplasms of the digestive organs otherwise specified *By...00*. Additional cancer data from the Higher Episode Statistics (HES) will be included to this study.

5.3. Study variables

Main outcome variable(s):

- The presence or absence for the following disease outcome(s): BCC, lung cancers (see section 5.2.2), and gastrointestinal tract cancers (see section 5.2.3).

Main exposures:

- The soil elements are aluminium, arsenic, calcium, chromium, copper, iron, lead, manganese, nickel, phosphorous, selenium, silicon, uranium, vanadium and zinc. Those which can be soil contaminant will be categorised according to the UK's current screening values (2014) for residential soils [20]; other elements (such as manganese) will be categorised in rank classification according to the overall distribution of abundance in England and Wales.

Adjustments for potential confounding variables:

- Important variables associated with main exposure(s) (geochemical soil concentrations) and outcomes (nonmelanoma, lung and digestive tract cancer) are: Townsend deprivation index categorised as quintiles, where a higher score equals a higher level of deprivation; type of living environment the patient resides in are categorised as urban (> 10,000 buildings), suburban (town or fringe) and rural (village, hamlet or isolated dwellings); the 10 strategic health authority units in England, as well as Wales; and demographic factors (i.e. age and sex).

5.4. Statistical analyses

5.4.1. Stage One

The association rule mining technique will be used to generate a set of items or exposure groups (rules) on whether patients diagnosed with, or without cancer (i.e. skin, lung or digestive tract cancer), lived in areas with potentially harmful element concentrations measured below or above the current residential screening levels.

Soil measurements are solely used as categorical variables (see section 5.3) for this type of analysis to generate a set of items. These items may be generated as a single, a pair or a set with more than two different soil metal exposures together. For each item, we will derive an estimate called the *confidence* (%), which is defined as the probability that a patient diagnosed with cancer was due to the effects of one, or the combine effects of more than one of the measured soil constituents.

The *confidence* estimates for each rule will be ranked in a descending order, which will show the most common exposure (or groups of exposure) in patients diagnosed with cancer.

5.4.2. Stage Two

Cox proportional-hazards models will be used, with years as the time scale, to determine the risk of developing cancer according to exposure levels for each soil contaminant. Soil elements that were significantly identified in stage one will be used in stage 2's analysis. Two models will be created using different methods in quantifying the exposure:

i. Soil elements modelled as a continuous measure in mg/kg

ii. Soil elements categorised as below or above the UK's current Category 4 Screening Levels (C4SLs) value for residential soils [21]–[23]

- The soil elements are aluminium, arsenic, calcium, chromium, copper, iron, lead, manganese, nickel, phosphorous, selenium, silicon, uranium, vanadium and zinc. Those which can be soil contaminant will be categorised according to the UK's current screening values (2014) for residential soils [20]; other elements (such as manganese) will be categorised in rank classification according to the overall distribution of abundance in England and Wales.

Adjustments for potential confounding variables:

- Important variables associated with main exposure(s) (geochemical soil concentrations) and outcomes (nonmelanoma, lung and digestive tract cancer) are: Townsend deprivation index categorised as quintiles, where a higher score equals a higher level of deprivation; type of living environment the patient resides in are categorised as urban (> 10,000 buildings), suburban (town or fringe) and rural (village, hamlet or isolated dwellings); the 10 strategic health authority units in England, as well as Wales; and demographic factors (i.e. age and sex).

5.4. Statistical analyses

5.4.1. Stage One

The association rule mining technique will be used to generate a set of items or exposure groups (rules) on whether patients diagnosed with, or without cancer (i.e. skin, lung or digestive tract cancer), lived in areas with potentially harmful element concentrations measured below or above the current residential screening levels.

Soil measurements are solely used as categorical variables (see section 5.3) for this type of analysis to generate a set of items. These items may be generated as a single, a pair or a set with more than two different soil metal exposures together. For each item, we will derive an estimate called the *confidence* (%), which is defined as the probability that a patient diagnosed with cancer was due to the effects of one, or the combine effects of more than one of the measured soil constituents.

The *confidence* estimates for each rule will be ranked in a descending order, which will show the most common exposure (or groups of exposure) in patients diagnosed with cancer.

5.4.2. Stage Two

Cox proportional-hazards models will be used, with years as the time scale, to determine the risk of developing cancer according to exposure levels for each soil contaminant. Soil elements that were significantly identified in stage one will be used in stage 2's analysis. Two models will be created using different methods in quantifying the exposure:

i. Soil elements modelled as a continuous measure in mg/kg

ii. Soil elements categorised as below or above the UK's current Category 4 Screening Levels (C4SLs) value for residential soils [21]–[23]

Our results will be presented as Hazard Ratios (HR) with 95% CI, where statistical significance is determined by p-values of 0.05 or less, and the exclusion of the null value 1 in the 95% confidence interval estimates. The proportional-hazards assumption for each model will be examined by comparing the cumulative hazards plots grouped on the exposure, which is dichotomised according to their screening levels. Where assumptions are violated due to a time interaction with the soil constituent, we will use alternative models as appropriate [24].

We will adjust for potential confounding variables, whereby any adjustments for variables that produces at least 10% change in the strength of association between the exposure and outcome will be identified as a confounder, and we will use likelihood ratio tests (LRT) as a statistical test for assessing effect modification and comparing goodness-of-fit between null and adjusted models.

5.4.3. Sensitivity analysis

In addition, we will use a sensitivity analysis in both stages to test whether changing the conditions for specific types of lung and digestive tract cancers will make a difference to the result.

For lung cancer, we will focus our attention on the four major lung cancers in the UK – small, large cell carcinoma, adenocarcinoma and SCC of the lung. For simplicity, the sensitivity analysis will be based on two outcome groups in which patients are categorised:

i. Small cell lung cancer (i.e. small cell carcinoma)
ii. Non-small cell lung cancer (i.e. SCC of the lung, large cell carcinoma and adenocarcinoma)

Similar for digestive tract cancers, the sensitivity analysis will be based on major groups:

i. Upper gastrointestinal tract cancer (i.e. oral, oesophageal, stomach and duodenal cancers)
ii. Lower gastrointestinal tract cancer (i.e. ileum and colorectal cancers)

6. Limitations

i. Exposure to geochemical soil elements are measured at an environmental level rather than *in vivo* at individual level. If an association is found in our proposed studies, then this will demonstrate the need for further research using biological samples such as biomarker measurements in nails, urine or hair samples.

ii. Our analysis is limited to measurements of soil elemental concentrations at a single time point. We are unable to quantify or assess the effects of variations in these levels over time, although a project is currently underway at the British Geological Survey to contribute further to understanding this issue, and we will consider the findings when preparing our results for publication. However, by adopting a cohort study design, it will allow us to take into account the implicit time factor that exists between the effect of a soil contaminant and cancer (i.e. nonmelanoma, lung and digestive tract).

7. References

[1] A. Prüss-Ustün, C. Vickers, P. Haefliger, and R. Bertollini, "Knowns and unknowns on burden of disease due to chemicals: a systematic review," *Environ Health*, vol. 10, no. 9, 2011.
[2] R. B. Hayes, "The carcinogenicity of metals in humans," *Cancer Causes Control*, vol. 8, no. 3, pp. 371–385, May 1997.
[3] B. J. Alloway, *Heavy Metals in Soils*. Springer, 1995.
[4] D. J. Gawkrodger, "Occupational skin cancers," *Occup Med*, vol. 54, no. 7, pp. 458 –463, 2004.
[5] Y. Yuan, G. Marshall, C. Ferreccio, C. Steinmaus, J. Liaw, M. Bates, and A. H. Smith, "Kidney cancer mortality: fifty-year latency patterns related to arsenic exposure," *Epidemiology*, vol. 21, no. 1, pp. 103–108, Jan. 2010.
[6] H.-H. Huang, J.-Y. Huang, C.-C. Lung, C.-L. Wu, C.-C. Ho, Y.-H. Sun, P.-C. Ko, S.-Y. Su, S.-C. Chen, and Y.-P. Liaw, "Cell-type specificity of lung cancer associated with low-dose soil heavy metal contamination in Taiwan: An ecological study," *BMC Public Health*, vol. 13, no. 1, p. 330, Apr. 2013.
[7] B. K. Mandal, T. R. Chowdhury, G. Samanta, G. K. Basu, P. P. Chowdhury, C. R. Chanda, D. Lodh, N. K. Karan, R. K. Dhar, and D. K. Tamili, "Arsenic in groundwater in seven districts of West Bengal, India- the biggest arsenic calamity in the world," *Current science. Bangalore*, vol. 70, no. 11, pp. 976–985, 1996.
[8] M. G. M. Alam, G. Allinson, F. Stagnitti, A. Tanaka, and M. Westbrooke, "Arsenic contamination in Bangladesh groundwater: a major environmental and social disaster.," *Int J Environ Health Res*, vol. 12, no. 3, pp. 235–53, 2002.
[9] E. L. Ander, C. C. Johnson, M. R. Cave, B. Palumbo-Roe, C. P. Nathanail, and R. M. Lark, "Methodology for the determination of normal background concentrations of contaminants in English soil," *Science of The Total Environment*, vol. 454, pp. 604–618, 2013.
[10] D. Briggs, "Environmental pollution and the global burden of disease," *Br Med Bull*, vol. 68, no. 1, pp. 1–24, Dec. 2003.
[11] P. B. Tchounwou, C. G. Yedjou, A. K. Patlolla, and D. J. Sutton, "Heavy Metal Toxicity and the Environment," in *Molecular, Clinical and Environmental Toxicology*, A. Luch, Ed. Springer Basel, 2012, pp. 133–164.
[12] M. A. Oliver, "Soil and human health: a review," *European Journal of Soil Science*, vol. 48, no. 4, pp. 573–592, 1997.
[13] Defra, *Technical Guidance Sheet on normal levels of contaminants in English soils: Arsenic - supplementary information*. Department of Environment Food and Rural Affairs (Defra), London: , 2012.
[14] C. C. Johnson, E. L. Ander, and M. R. Cave, "Technical guidance on normal levels of contaminants in Welsh soil: Arsenic (As): January 2013," 2013.
[15] I. Martin, R. De Burca, and H. Morgan, "Soil Guideline Values for inorganic arsenic in soil," *Enivronmetal Agency (UK)*, vol. Science Report SC050021, pp. 1–11, 2009.
[16] I. Martin, H. Morgan, C. Jones, E. Waterfall, and J. Jeffries, "Soil Guideline Values for nickel in soil," *Enivronmetal Agency (UK)*, vol. Science Report SC050021/ Nickel SGV, pp. 1–10, 2009.
[17] B. G. Rawlins, S. P. McGrath, A. J. Scheib, N. Breward, M. Cave, T. R. Lister, M. Ingham, C. Gowing, and S. Carter, *The advanced soil geochemical atlas of England and Wales*. Keyworth, Nottingham: British Geological Survey, 2012.
[18] A. L. Hinwood, M. R. Sim, D. Jolley, N. De Klerk, E. B. Bastone, J. Gerostamoulos, and O. H. Drummer, "Hair and toenail arsenic concentrations of residents living in areas with high environmental arsenic concentrations.," *Environmental health perspectives*, vol. 111, no. 2, p. 187, 2003.
[19] A. Meal, J. Leonardi-Bee, C. Smith, R. Hubbard, and F. Bath-Hextall, "Validation of THIN data for non-melanoma skin cancer," *Qual Prim Care*, vol. 16, no. 1, pp. 49–52, 2008.
[20] Defra, "SP1010 Development of category 4 screening levels for assessment of affected by contamination," 2014. [Online]. Available: http://randd.defra.gov.uk/Document.aspx?Document=11964_SP1010DevelopmentofCategory4ScreeningLevelsMainReport.pdf.
[21] Defra, "SP1010 Appendix C Provisional C4SLS for Arsenic," 2014. [Online]. Available: http://randd.defra.gov.uk/Document.aspx?Document=11966_SP1010AppendixC-Arsenic.pdf.
[22] Defra, "SP1010 Appendix G Provisional C4SLs for Chromium," 2014. [Online]. Available: http://randd.defra.gov.uk/Document.aspx?Document=11970_SP1010AppendixG-ChromiumVI.pdf.
[23] Defra, "SP1010 Appendix G Provisional C4SLs for Lead," 2014. [Online]. Available: http://randd.defra.gov.uk/Document.aspx?Document=11971_SP1010AppendixH-Lead.pdf.
[24] D. W. Hosmer and P. Royston, "Using Aalen's linear hazards model to investigate time-varying effects in the proportional hazards regression model," *Stata J*, vol. 2, pp. 331–350, 2002.

8.3 BCC THIN Read codes

Description	Code
Basal cell carcinoma	B33..11
Epithelioma	B33..12
Rodent Ulcer	B33..13
Epithelioma Basal cell	B33..16
Naevoid basal cell carcinoma	B33z100
[M]Basal cell neoplasms	BB3..00
[M]Basal cell tumour	BB30.00
[M]Basal cell carcinoma NOS	BB31.00
[M]Multicentric basal cell carcinoma	BB32.00
[M]Basal cell carcinoma (morphoea type)	BB33.00
[M]Basal cell carcinoma (fibroepithelial)	BB34.00
[M]Basal cell neoplasm NOS	BB3z.00

8.4 Lung cancer THIN Read codes

Description	Code
Malignant neoplasm of trachea; bronchus and lung	B22..00
Malignant neoplasm of trachea	B220.00
Malignant neoplasm of mucosa of trachea	B220100
Malignant neoplasm of trachea NOS	B220z00
Malignant neoplasm of main bronchus	B221.00
Malignant neoplasm of carina of bronchus	B221000
Malignant neoplasm of hilus of lung	B221100
Malignant neoplasm of main bronchus NOS	B221z00
Malignant neoplasm of upper lobe; bronchus or lung	B222.00
Malignant neoplasm of upper lobe bronchus	B222000
Malignant neoplasm of upper lobe of lung	B222100
Pancoast's syndrome	B222.11
Malignant neoplasm of upper lobe; bronchus or lung NOS	B222z00
Malignant neoplasm of middle lobe; bronchus or lung	B223.00
Malignant neoplasm of middle lobe bronchus	B223000
Malignant neoplasm of middle lobe of lung	B223100
Malignant neoplasm of middle lobe; bronchus or lung NOS	B223z00
Malignant neoplasm of lower lobe; bronchus or lung	B224.00
Malignant neoplasm of lower lobe bronchus	B224000
Malignant neoplasm of lower lobe of lung	B224100
Malignant neoplasm of lower lobe; bronchus or lung NOS	B224z00
Malignant neoplasm of overlapping lesion of bronchus & lung	B225.00
Malignant neoplasm of other sites of bronchus or lung	B22y.00
Malignant neoplasm of bronchus or lung NOS	B22z.00

Lung cancer	B22z.11
Carcinoma in situ of the respiratory tract	B81..00
respiratory tract adenomas and adenocarcinomas	BB5S.00
any lung-related malignant neoplasms otherwise stated	By…00

8.5 GIT cancer THIN Read codes

Description	Code
Malignant neoplasm of lip; oral cavity and pharynx	B0...00
Malignant neoplasm of lip	B00..00
Malignant neoplasm of upper lip; vermilion border	B000.00
Malignant neoplasm of upper lip; external	B000000
Malignant neoplasm of upper lip; lipstick area	B000100
Malignant neoplasm of upper lip; vermilion border NOS	B000z00
Malignant neoplasm of lower lip; vermilion border	B001.00
Malignant neoplasm of lower lip; external	B001000
Carcinoma of lip	B00..11
Malignant neoplasm of lower lip; lipstick area	B001100
Malignant neoplasm of lower lip; vermilion border NOS	B001z00
Malignant neoplasm of upper lip; inner aspect	B002.00
Malignant neoplasm of upper lip; frenulum	B002100
Malignant neoplasm of upper lip; mucosa	B002200
Malignant neoplasm of upper lip; oral aspect	B002300
Malignant neoplasm of upper lip; inner aspect NOS	B002z00
Malignant neoplasm of lower lip; inner aspect	B003.00
Malignant neoplasm of lower lip; buccal aspect	B003000
Malignant neoplasm of lower lip; frenulum	B003100
Malignant neoplasm of lower lip; mucosa	B003200
Malignant neoplasm of lower lip; oral aspect	B003300
Malignant neoplasm of lower lip; inner aspect NOS	B003z00
Malignant neoplasm of lip unspecified; inner aspect	B004.00
Malignant neoplasm of lip unspecified; buccal aspect	B004000
Malignant neoplasm of lip unspecified; mucosa	B004200
Malignant neoplasm of lip; oral aspect	B004300
Malignant neoplasm of commissure of lip	B005.00
Malignant neoplasm of overlapping lesion of lip	B006.00
Malignant neoplasm of lip; unspecified	B007.00
Malignant neoplasm of lip; unspecified; external	B00z000
Malignant neoplasm of lip; unspecified; lipstick area	B00z100
Malignant neoplasm of lip; vermilion border NOS	B00zz00
Malignant neoplasm of tongue	B01..00
Malignant neoplasm of base of tongue	B010.00
Malignant neoplasm of base of tongue dorsal surface	B010000
Malignant neoplasm of posterior third of tongue	B010.11

Malignant neoplasm of fixed part of tongue NOS	B010z00
Carcinoma of lip; oral cavity and pharynx	B0...11
Malignant neoplasm of dorsal surface of tongue	B011.00
Malignant neoplasm of dorsum of tongue NOS	B011z00
Malignant neoplasm of tongue; tip and lateral border	B012.00
Malignant neoplasm of ventral surface of tongue	B013.00
Malignant neoplasm of anterior 2/3 of tongue ventral surface	B013000
Malignant neoplasm of frenulum linguae	B013100
Malignant neoplasm of ventral tongue surface NOS	B013z00
Malignant neoplasm of anterior 2/3 of tongue unspecified	B014.00
Malignant neoplasm of tongue; junctional zone	B015.00
Malignant neoplasm of lingual tonsil	B016.00
Malignant overlapping lesion of tongue	B017.00
Malignant neoplasm of other sites of tongue	B01y.00
Malignant neoplasm of tongue NOS	B01z.00
Malignant neoplasm of major salivary glands	B02..00
Malignant neoplasm of parotid gland	B020.00
Malignant neoplasm of submandibular gland	B021.00
Malignant neoplasm of sublingual gland	B022.00
Malignant neoplasm of other major salivary glands	B02y.00
Malignant neoplasm of major salivary gland NOS	B02z.00
Malignant neoplasm of gum	B03..00
Malignant neoplasm of upper gum	B030.00
Malignant neoplasm of lower gum	B031.00
Malignant neoplasm of other sites of gum	B03y.00
Malignant neoplasm of gum NOS	B03z.00
Malignant neoplasm of floor of mouth	B04..00
Malignant neoplasm of anterior portion of floor of mouth	B040.00
Malignant neoplasm of lateral portion of floor of mouth	B041.00
Malignant neoplasm; overlapping lesion of floor of mouth	B042.00
Malignant neoplasm of other sites of floor of mouth	B04y.00
Malignant neoplasm of floor of mouth NOS	B04z.00
Malignant neoplasm of other and unspecified parts of mouth	B05..00
Malignant neoplasm of cheek mucosa	B050.00
Malignant neoplasm of buccal mucosa	B050.11
Malignant neoplasm of vestibule of mouth	B051.00
Malignant neoplasm of upper buccal sulcus	B051000
Malignant neoplasm of lower buccal sulcus	B051100
Malignant neoplasm of hard palate	B052.00
Malignant neoplasm of soft palate	B053.00
Malignant neoplasm of uvula	B054.00
Malignant neoplasm of palate unspecified	B055.00
Malignant neoplasm of junction of hard and soft palate	B055000
Malignant neoplasm of roof of mouth	B055100
Malignant neoplasm of palate NOS	B055z00
Malignant neoplasm of retromolar area	B056.00

Overlapping lesion of other and unspecified parts of mouth	B057.00
Malignant neoplasm of other specified mouth parts	B05y.00
Malignant neoplasm of mouth NOS	B05z.00
Kaposi's sarcoma of palate	B05z000
Malignant neoplasm of oropharynx	B06..00
Malignant neoplasm of tonsil	B060.00
Malignant neoplasm of faucial tonsil	B060000
Malignant neoplasm of palatine tonsil	B060100
Malignant neoplasm of overlapping lesion of tonsil	B060200
Malignant neoplasm tonsil NOS	B060z00
Malignant neoplasm of tonsillar fossa	B061.00
Malignant neoplasm of tonsillar pillar	B062.00
Malignant neoplasm of faucial pillar	B062000
Malignant neoplasm of glossopalatine fold	B062100
Malignant neoplasm of palatoglossal arch	B062200
Malignant neoplasm of palatopharyngeal arch	B062300
Malignant neoplasm of tonsillar fossa NOS	B062z00
Malignant neoplasm of vallecula	B063.00
Malignant neoplasm of anterior epiglottis	B064.00
Malignant neoplasm of epiglottis; free border	B064000
Malignant neoplasm of glossoepiglottic fold	B064100
Malignant neoplasm of anterior epiglottis NOS	B064z00
Malignant neoplasm of junctional region of epiglottis	B065.00
Malignant neoplasm of lateral wall of oropharynx	B066.00
Malignant neoplasm of posterior wall of oropharynx	B067.00
Malignant neoplasm of oropharynx; other specified sites	B06y.00
Malignant neoplasm of other specified site of oropharynx NOS	B06yz00
Malignant neoplasm of oropharynx NOS	B06z.00
Malignant neoplasm of nasopharynx	B07..00
Malignant neoplasm of roof of nasopharynx	B070.00
Malignant neoplasm of posterior wall of nasopharynx	B071.00
Malignant neoplasm of adenoid	B071000
Malignant neoplasm of pharyngeal tonsil	B071100
Malignant neoplasm of posterior wall of nasopharynx NOS	B071z00
Malignant neoplasm of lateral wall of nasopharynx	B072.00
Malignant neoplasm of pharyngeal recess	B072000
Malignant neoplasm of lateral wall of nasopharynx NOS	B072z00
Malignant neoplasm of anterior wall of nasopharynx	B073.00
Malignant neoplasm posterior margin nasal septum and choanae	B073200
Malignant neoplasm of anterior wall of nasopharynx NOS	B073z00
Malignant neoplasm; overlapping lesion of nasopharynx	B074.00
Malignant neoplasm of other specified site of nasopharynx	B07y.00
Malignant neoplasm of nasopharynx NOS	B07z.00
Malignant neoplasm of hypopharynx	B08..00
Malignant neoplasm of postcricoid region	B080.00
Malignant neoplasm of pyriform sinus	B081.00

Malignant neoplasm aryepiglottic fold; hypopharyngeal aspect	B082.00
Malignant neoplasm of posterior pharynx	B083.00
Malignant neoplasm of other specified hypopharyngeal site	B08y.00
Malignant neoplasm of hypopharynx NOS	B08z.00
Malig neop other/ill-defined sites lip; oral cavity; pharynx	B0z..00
Malignant neoplasm of pharynx unspecified	B0z0.00
Malignant neoplasm of Waldeyer's ring	B0z1.00
Malignant neoplasm of laryngopharynx	B0z2.00
Malignant neoplasm of other sites lip; oral cavity; pharynx	B0zy.00
Malignant neoplasm of lip; oral cavity and pharynx NOS	B0zz.00
Malignant neoplasm of oesophagus	B10..00
Malignant neoplasm of cervical oesophagus	B100.00
Malignant neoplasm of thoracic oesophagus	B101.00
Malignant neoplasm of abdominal oesophagus	B102.00
Malignant neoplasm of upper third of oesophagus	B103.00
Malignant neoplasm of middle third of oesophagus	B104.00
Malignant neoplasm of lower third of oesophagus	B105.00
Malignant neoplasm; overlapping lesion of oesophagus	B106.00
Siewert type I adenocarcinoma	B107.00
Malignant neoplasm of other specified part of oesophagus	B10y.00
Malignant neoplasm of oesophagus NOS	B10z.00
Oesophageal cancer	B10z.11
Malignant neoplasm of stomach	B11..00
Malignant neoplasm of cardia of stomach	B110.00
Malignant neoplasm of cardiac orifice of stomach	B110000
Malignant neoplasm of cardio-oesophageal junction of stomach	B110100
Malignant neoplasm of gastro-oesophageal junction	B110111
Malignant neoplasm of cardia of stomach NOS	B110z00
Malignant neoplasm of pylorus of stomach	B111.00
Malignant neoplasm of prepylorus of stomach	B111000
Gastric neoplasm	B11..11
Malignant neoplasm of pyloric canal of stomach	B111100
Malignant neoplasm of pylorus of stomach NOS	B111z00
Malignant neoplasm of pyloric antrum of stomach	B112.00
Malignant neoplasm of fundus of stomach	B113.00
Malignant neoplasm of body of stomach	B114.00
Malignant neoplasm of lesser curve of stomach unspecified	B115.00
Malignant neoplasm of greater curve of stomach unspecified	B116.00
Malignant neoplasm; overlapping lesion of stomach	B117.00
Siewert type II adenocarcinoma	B118.00
Siewert type III adenocarcinoma	B119.00
Malignant neoplasm of other specified site of stomach	B11y.00
Malignant neoplasm of anterior wall of stomach NEC	B11y000
Malignant neoplasm of posterior wall of stomach NEC	B11y100
Malignant neoplasm of other specified site of stomach NOS	B11yz00
Malignant neoplasm of stomach NOS	B11z.00

Malignant neoplasm of colon	B13..00
Malignant neoplasm of hepatic flexure of colon	B130.00
Malignant neoplasm of transverse colon	B131.00
Malignant neoplasm of descending colon	B132.00
Malignant neoplasm of sigmoid colon	B133.00
Malignant neoplasm of caecum	B134.00
Carcinoma of caecum	B134.11
Malignant neoplasm of appendix	B135.00
Malignant neoplasm of ascending colon	B136.00
Malignant neoplasm of splenic flexure of colon	B137.00
Malignant neoplasm; overlapping lesion of colon	B138.00
Hereditary nonpolyposis colon cancer	B139.00
Malignant neoplasm of other specified sites of colon	B13y.00
Malignant neoplasm of colon NOS	B13z.00
Colonic cancer	B13z.11
Malignant neoplasm of rectum	B141.00

8.6 Examiner's feedback and list of amendments

8.6.1 Comments from Internal examiner

	Internal examiner's comments	Responses and amendments	Location of changes
1.	Abstract, Conclusion: Given the modest overall effect size and total number of soil elements evaluated I believe the statement that there is "strong evidence" overstates the findings	Thank you for your comments. It has been change to: *"Conclusion: There appears to be slight evidence of BCC, respiratory and GIT cancer risk with elevated exposure to soil arsenic, aluminium and phosphorus, respectively."*	Abstract, page iv, lines 8-10
2.	p. 32: At the start of section 2 at least one paragraph is needed describing exactly how the linkage was carried out as understanding this linkage is crucial to the understanding of the entire work. Saying this was carried out by "experts from University of Nottingham and BGS" is probably not sufficient.	Thank you for your comments. A paragraph has been inserted to give a brief description of how the linkage was carried out. The added paragraph was: *"G-BASE database contains geochemical information on the normal background concentrations for different trace elements in UK topsoil. Soil samples were collected throughout England and Wales from urban and rural soils within 1-2km of an individual's home. Soil samples were analysed using the X-ray fluorescence spectroscopy to detect geochemical composition and concentrations levels for each element. The soil concentrations for fifteen elements were spatially interpolated over a continuous raster shapefile for England and Wales to produce point estimates at a pixel-level. Spatially referenced point estimates that overlapped the street postcode of patient registered to a general practice using the THIN were merged to produce the THIN-GBASE database."*	Chapter 2, section 2, page 32, lines 9-20

3.	p. 36: The last sentence could be improved grammatically, also "at least 85,803,247 person-years" is rather a strange statement. This seems quite an exact number	Thank you for your comments. The last sentence has been formatted to: *"The overall number of patients contributing computerised data in THIN is 85,803,247 person-years."*	Chapter 3, section 3.1, page 36, lines 20-21
4.	p. 44: The thesis would really benefit from a table detailing the number and percentage of the overall cohort size of 6,825,382 patients who had missing data for each of the 15 elements	Thank you for your comments, this additional information would have been beneficial for this PhD. However, this change is not feasible. To provide results based on 6,825,382 patients, I would need to have complete access to the entire THIN database, which is not possible. The dataset provided for this PhD was limited to a cohort of up-to 2.3 million patients who were alive, active at their recent GP practice which is within GBASE coverage.	No changes made
5.	p. 44: The was signposted to ref no. 116 for important comparing the linked population with the general population. However, if this paper has not appeared in print, by the time a revised version of this thesis is submitted would it be possible to include the manuscript as an appendix?	Thank you for your comments. The first author and I discussed copyright concerns and therefore I have refrained from adding the full manuscript to the appendix. However, a revised version of the article has recently been accepted to the Population Health Metric Journal. I have updated its citation in the bibliography, and added its information to my list of publications since I have co-authored it. You will find that I have provided the contact details of the first author who would be happy to distribute a copy of the manuscript for your perusal.	Page v, lines 7-12
6.	p. 45: The median level of silicon was 299,000 in the text (in line 12) but 29,900 in Table 3.2. The IQRs are also discrepant by the same order of magnitude.	The errors for the median levels of silicon have been corrected. The median and IQR estimates for silicon in the text have been updated and are now consistent with the values shown in table 3.2. The sentence has been changed to the following: *"Overall, the elements with highest median concentrations were aluminium (median: 51,000 mg/kg, IQR: 40,300-59,300 mg/kg)*	Chapter 3, section 3.3.4, page 48, lines 4-7

		followed by silicon (median: 29,900 mg/kg, IQR: 26,500-32,900 mg/kg)".	
7.	p. 45: Whilst description of the distribution of soil elements by sampling point is informative what will be even more helpful is a description of the number of linked THIN participants who have been assigned to each sampling point (i.e. exposure level). The minimum, maximum, median and IQR for this distribution would be informative. This could be done separately for G-BASE and NSI-XRFS sampling points	Thank you for your comment, this piece of information would have been beneficial for this PhD. However, this change is not feasible. This is because of the limitations imposed by partners at THIN/EPIC. The linkage between databases were carried out in manner that censored any spatial details (postcode, address and geographic coordinates) of a soil sampling locations. In addition, sensitive information pertained to a patient's home address, residential history and location of the GP were stripped from their medical records after the linkage. Therefore, the dataset provided for this research contained cohort of 2.3 million patient records that were anonymised.	No changes made
8.	p. 46 (Table 3.2): Please clarify whether the total of 1,742,205 relates to the total number of sampling points rather than participants in the THIN dataset	Thank you for your comment. The value of 1,742,205 corresponds to the total number of patients in the database who have soil data across all fifteen elements. The following information has been added to provide clarification: "*Descriptive analyses were performed accordingly on 1,742,205 patients in the THIN-GBASE who have soil data across all 15 elements. Table 3.2 provides the statistical summaries in a form of median, interquartile ranges (IQRs) and maximum value observed among the cohort of participants who are contributing data to the THIN-GBASE database. For instance, 1,664,155 (out of 1,742,205) (95.52%) participants have a concentration value for aluminium – among the participants the estimates typically ranged between 2,000 to 116,700 mg/kg with a median of 51,000 mg/kg (IQR: 40,300-59,300 mg/kg). The remaining participants (4.48%) either had concentrations for aluminium that were deemed as estimates below detection limits*	Chapter 3, section 3.3.4, page 44, lines 10-21

		(i.e. less than 2,000 mg/kg (coded as -1)) or not available (coded -2).”	
9.	Figure 3.3 to 3.16: For some elements (e.g. aluminium, calcium, silicon), the x-axis presents values as a percent and in mg/kg units we would expect. Is there a reason for this?	Thank you for your comment. The concentrations (in mg/kg) for these elements were too high, and so, I have reported them as weight percentage (%) in the histograms: *Weight percent (%) = soil metal concentration (in mg/kg) ÷ 10,000*. Whereby, a 1.0% equivalent to 10,000 parts-per million. Also, I have included the information below to the legends for Figure 3.3, 3.5 and 3.14: *“The concentrations for [aluminium, calcium or silicon] were converted to a weight percentage (mg/kg÷10,000), whereby 1.0% = 10,000 (of [aluminium, calcium or silicon]) parts-per million”*	Chapter 3, section 3.3.4, figure 3.3 (page 52, lines 5-6); figure 3.5 (page 54, lines 5-6); and figure 3.14 (page 63, lines 5-6)
10.	p. 74: I am not clear how the confidence intervals for each country were used to establish statistical significance. Statistical significance between the IRs of 2 countries can only be established from CI around the difference between the two IRs and not by whether or not they overlap	Thank you for this comment. This statement has been removed and formatted to: *“We observed the incidences are low, and similar for Scotland and Northern Ireland”*	Chapter 4, section 4.2.3.1, page 77, lines 5-6
11.	p. 80: The final sentence is quite difficult to follow. I assume you mean the “IRRs for deprivation were higher in men than in women” not “the incidence rate was higher…”	Thank you for your comments. This error has been corrected to: *“…the IRRs for socioeconomic deprivation were higher in men than in women”*	Chapter 4, section 4.2.3.3, page 84, lines 1-2

12.	p. 88: The justification for choosing arsenic for chapter 4 and not other soil elements needs to be stronger	A paragraph has been inserted at the opening of the summary section of chapter four to reinforce the justification for conducting both incidence and soil arsenic-BCC study: *"Basal cell carcinoma (BCC) is one of the most common types of non-melanoma skin cancer in the UK. There is a well-established link between environmental arsenic exposure and the development of BCC, but at considerably higher levels of exposure than those likely to be observed in the UK. We therefore carried out a study to determine whether there is evidence that more modest levels of arsenic in soil increase the risk of BCC, as an example of testing a specific, evidence-informed hypothesis using the new data source (other elements were therefore not considered, except where there was concern they might modify the effect of arsenic). As little is known about how the incidence of BCC varies across the UK, we first took the opportunity to quantify the variation. Therefore, this chapter describes two studies:"*	Chapter 4, section 4, page 68, lines 2-14
13.	p. 95: Given you have presented column percentages in table 4.5 (row percentages may be more intuitive) it would be correct to say "The proportion of men was greater among those with BCC than for those without". Also, the % for group I with BCC is 35.5 in the text but 35.2 in the table.	The sentence has been corrected, and the percentage for group I with BCC has been updated (in accordance to what's in table 4.5) to following: "*The proportion of men was greater among those with BCC than for those without. Participants developing BCC were more likely to be in the older age groups and from the least deprived group (Group I: 35.2%; Group II: 25.6%).*"	Chapter 4, section 4.3.3.1, page 98, lines 14-17
14.	p. 115-124: Please correct the page numbering and formatting errors (text in landscape rather than portrait)	The corrections and formatting has been made.	

15.	p. 120: The filter method for feature selection algorithm needs a greater explanation and a reference which can be easily traced (journal article rather than a book section). How this corrects for the potential of a type I error needs to be explained with particular care.	Thank you for your comments. The filter methods have been explained in greater details and references have been provided. The following information have been added: *"We have applied feature selection data mining techniques to our database to generate new hypothesis that may aid in determining the relationship between soil elements from GBASE and clinical outcomes in THIN. Feature selection is a very useful data mining tool which has a suite of wrapper, filter and embedded methods for searching potential exposures used for optimising risk predictions of certain outcome variables in a large database.*[191,192] *The filter methods are especially useful in determining which exposure, or subset of exposures that are relevant for building a predictive model. When applying filter methods to a large database, they act as filters for identifying relevant exposures needed for building and optimising risk models, whilst, at the same time removing any exposures that are redundant and do not contribute to the accuracy of the predictive model.*[191,192] *This technique is certainly helpful in building our own risk models because there is paucity in literature that establish any direct relationships between specific soil elements and lung cancer. Contemporary studies that have shown associations between soil elements and lung cancer have conflicting results,*[157,189] *and so it would be inappropriate to rely on their results to build our predictive models for risk of lung cancer. We therefore relied on these filter-based methods to select the relevant soil elements. The technique optimises risk prediction of lung cancer based on the selected group of soil element and thus do not guarantee any statistical significance thereby limiting the potential of a type-I error to occur in our results."*	Chapter 5, section 5.3.3.1, pages 126 (lines 17-23) and 127 (lines 1-17)

16.	p. 123: “sample population”: I assume you mean either sample or population.	The error has been corrected to: “*The final extract contained a sample of 1,823,312 participants (Figure 5.2).*”	Chapter 5, section 5.4.1, page 130, lines 9-11
17.	p. 123: The statement that 35.1% of people developing lung cancer live in the South Eastern or Western region of England does not tally with the numbers in table 5.1	The sentence has been corrected to: “*Participants who developed lung cancer were more likely to be older (51-60 years: 19.5%; 61-70 years: 31.5% and 71-80 years: 31.8%), a male (57.3%) current smoker (48.0%) from the south of England (36.2%), and from the deprived groups (group IV: 21.2%).*”	Chapter 5, section 5.4.1, page 130, lines 14-18
18.	p. 133: Interpretation of figure 5.6: A test for “linear trend” is not the converse of a “test for non-linearity”. The test for linear trends states the probability of obtaining your results assuming that any true relationship is linear (and that the null hypothesis of no relationship is true). A plateau in effect sizes with increasing exposure can occur simply due to statistical error. A test for non-linearity compares nested models where in one instance group number (I to V) is fitted as a linear term and another where it is fitted as a category.	The interpretation for the trends test for aluminium has been revised to: “*Our test seems to indicate a significant linear trends relationship for increased exposure of aluminium and risk of lung cancer ($p < 0.001$) (Figure 5.6). However, the patterns of risk for lung cancer in relation to aluminium show a plateau effect which are unclear.*”	Chapter 5, section 5.4.3.1, page 139, lines 14-18
19.	p. 138: Please rephrase “the proportional hazards assumption was highly significant”. I think you mean there is significant evidence that the assumptions of PH was violated	The sentence has been rephrased to: “*Overall, we found significant evidence that the assumption of the proportional-hazards was violated (i.e. global test: p-value < 0.0001).*”	Chapter 5, section 5.4.3.2, page 144, Lines 6-7

20.	Figures 5.7 to 5.10: These plots show the "cumulative regression coefficients" not the "cumulative hazard functions" as stated in the figure titles. Also, as written you have implied that the assumption of PH has been violated for females, 71-80 years, etc., but this actually relates to the comparison with the reference group (i.e. coefficient comparing women with men). The title of the figures should therefore be changed to reflect this.	All changes to the figure titles have been made to reflect the following: 1.) the estimates plotted are cumulative regression coefficients, 2.) the reference groups for each of the variables used in the Aalen exercise are added to the figure titles.	Chapter 5, section 5.4.3.2, pages 147-150
21.	p. 148: Some differences were highlighted between urban, suburban and rural areas, but were tests for interaction carried out? These could be chance findings and therefore the fairly strong conclusion on p. 157 that such risks are limited to residential areas may be overstated.	Thank you for this comment. No tests for interactions between the soil exposure groups and type residential environment were conducted for this research. The analysis was only limited to stratifications based on the type of residential environmental because: 1.) I wanted to know magnitude of risk for the cancers across the different, and 2.) this indicator was used as a proxy for G-BASE rural, G-BASE urban, and NSI(XRFS) areas, in attempt to account for differences caused by the sampling areas. The conclusions made have been modified to: *"In conclusion, the current study suggests that those living in areas with soil aluminium levels above 47,200 mg/kg may have a greater risk of developing lung cancer. The result suggests that aluminium exposure among urban residents may be the cause of lung cancer for this group. While, the results indicate statistical significance, they need to be interpreted with caution due to the limitations that are present for this study. Further studies will be needed to validate the findings made in this investigation."*	Chapter 5, section 5.5, page 165 (lines 20-24) and 166 (lines 1-3)

22.	p. 177: The approach to how competing risks were taken account needs to be explained more fully so that someone is able to clearly understand and appraise the approach used. As sub-hazards ratios are present, I assume the method of Fine and Gray was used. Could you therefore cite their paper if this is the case?	Thank you for this comment. I have provided additional information regarding the competing risk models used in chapter 6: *"In order to obtain site-specific estimates, we used competing risk survival models (as described by Fine and Gray[252,253]). As noted previously, our case definition required a first ever cancer as it is difficult to distinguish between subsequent primary diagnoses, metastases, and follow up visits for the first cancer. When considering specific sites, it is necessary to censor patients who develop a cancer at another site on the date of this diagnosis, as they can no longer develop a first ever cancer in the specific site of interest. This competing risk may artificially diminish the observed hazard ratio at the site of interest. Competing risks models attempt to remove this bias by estimating the sub-hazard ratio in the site of interest alone."* The methods used were based on the Fine and Gray – I have cited their paper which appears as reference number 250 in the bibliography.	Chapter 6, section 6.4.2.3, page 186, lines 5-16
23.	p. 189: Table 6.4 is not especially informative. The main information that is contained in the footnote and this has been copied and pasted directly from other tables. I believe it will be simply to say okay in the text that tests for proportional hazards were non-significant for all seven elements (p > 0,05)	Thank you for your comments. However, I wish to keep this table because it has the added value of contributing to the consistency within this section and across chapters, as in previous sections I included similar tables.	No changes made
24.	p. 191 to 200: Please incorporate all the above suggestions re: presentation of results	Thank you for your comments. All changes to the figure titles in figure 6.1 to 6.6 have been applied to reflect the following changes: 1.) the estimates plotted are cumulative regression coefficients, 2.) the	Chapter 6, section 6.5.3.2, pages 204-209

	from the Aalen plots in the same way you have done for chapter 5.	reference groups for each of the variables used in the Aalen exercise are added to the figure titles	
	Presentational aspects:		
25.	Where typographical and grammatical errors occurred these tend to cluster in the same sections of the thesis, in particular the abstract, e.g. “much” rather than “many” (line 1), “utilise a new resource…” (line 15), “…the findings for the ecological study…” (p. ii, last line 4). Please could you correct these and check the thesis carefully for any further errors of this type before resubmitting.	Thank you for your comments. All the errors that have been highlighted in the abstract have been corrected. The entire thesis has been carefully checked for any further errors of this sort.	Abstract, page i (lines 2 and 15) and ii (line 22)
26.	References: Several web links provided were broken (ref. 43-45, 52, 68, 71 and 72). Also, please make sure the date last accessed is provided for all internet references in some instances only the year is provided (ref. 69 and 70)	Thank you for your comments. I have checked all internet references, and most appear to be okay. The references that were broken are those numbered 68 – 72. These references are from the same source (the Department for Environment, Food and Rural Affairs (DEFRA)) and so I have amended their URLs by providing DEFRA’s parent webpage where the PDF documents for the soil elements can be downloaded. The dates at which an internet reference provided in this thesis was last accessed has also been provided.	

8.6.2 **Comments from External examiner**

	Comments and responses
1.	**1.1. External examiner**: The report does not make entirely clear the extent to which the different sources of the "interpolated smooth surface over the map of England and Wales" mentioned on page 42, provided the main information on potential exposure in different geographic areas within the study area. ***Response**: Thank you for your comments. To address the above statement, I have added a new section to chapter 3 (sections 3.3.1 page 38 (lines 17-22) and page 39 (lines 1-15)) which provides a descriptive account about the 3 primary soil datasets used in the G-BASE project (G-BASE rural, G-BASE urban and NSI(XRFS)), and provided a breakdown on the number of sampling sites that were derived for G-BASE rural, G-BASE-urban and NSI(XRFS). Also, in section 3.3.3 (Page 44, lines 2-16) of chapter 3, I have formatted the text to make it a bit clear on how the spatial interpolation for each substance was carried out between sampling points.* **1.2. External examiner**: Although section 3.2 on page 37 and following, explains the G-BASE is joint project with BGS re-analysis of the National Soil Inventory x-ray fluorescence spectrometry NSI(XRFS) samples, there was no consideration given to the differential error in epidemiology that could derive from systematic differences across the geochemistry from G-BASE data were completely absent from many parts of England and Wales (see map on page 40), and therefore the interpolation would have been supplemented on NSI(XRFS) data where G-BASE data were not available, but a THIN database derived postcode was available and required the corresponding geochemical area value. The validity of the interpolated smoothed surface would be correspondingly variable according to the different density of sampling points in predominantly G-BASE derived areas compared to predominantly NSI(XRFS) derived areas (average 1 every 25km^2). In any case, the geochemical information would have been referring to actual geochemical samples representing geographical areas larger than most population-containing areas (postcodes) used for the linking of THIN information. It would be desirable to provide an overall tabulation of the percentages of population for whom the exposure model was derived mainly from G-BASE and those for whom it was derived mainly from NSI(XRFS). If this is not feasible, it would seem essential at least to add to the discussion a consideration of the potential information bias that might affect all cancer risk estimates due to systematic differences in exposure method between population groups across study areas, in Page 104 (i.e. discussion of BCC associations), Page 156, (discussions of lung cancer associations), Page 221, discussion of gastrointestinal cancer associations. ***Response**: Thank you for your comments. To address the above comments regarding the potential for information bias that might have occurred due to the inability to control for the sampling areas defined as G-BASE urban, G-BASE rural and NSI(XRFS) - the following information have been added to the discussions for each chapter:*

	- *In chapter 4, i.e. BCC (section 4.3.4, pages 110 (lines 19-25) and 111 (lines: 1-13))* - *In chapter 5, i.e. lung cancer (section 5.5, page 164 (lines 7-25) and 165 (lines 1-3))* - *In chapter 6, i.e. GIT cancer (section 6.6, pages 232 (lines 10-25) and 233 (lines 1-5))* *Also, I agree that the fact that this thesis would have benefitted with the inclusion of a tabulation showing the percentages of the population (i.e. cases and non-cases) that were classified as G-BASE urban, rural or NSI(XRFS). However, this is addition is not feasible due the limitations of the dataset. The linkage between THIN and the G-BASE database was carried out in manner that censored any spatial details (i.e. postcode, address and geographic coordinates) relating to the soil sampling locations. In addition, any sensitive information pertained to a patient's home address, residential history and location of the GP were stripped from their medical records after the linkage.* **1.3. External examiner:** In addition, the title of overall dissertation would seem more appropriate as referring to BGS data, as in "Environmental Exposure to Metallic Soil Elements and Risk of Cancer in the UK Population, Using a Unique Linkage Between THIN and BGS Databases. ***Response**: Thank you for your comments, the title has been changed to the above suggestion (see title page).* **1.4. External examiner:** Another design aspects that would have been interesting to see discussed in more detail in same discussion sections, is regarding for differential ascertainment rates for cancers, completeness of cancer diagnosis reports based on histology available at GP level, within SHA areas, and the potential implications for selection bias in this study, given the pattern of spatial distribution of the main exposure of interest. ***Response**: Thank you for your comments, in the discussion sections for each chapters 4, 5 and 6, I have included text discussing the completeness of the data, as well as potential for selection bias to occur due to certain aspects pertained to the selection of our case and non-case population. I have also discussed potential biases that my might arise if GP-level cancer ascertainment rates are associated with the geographical variations in the exposure for soil elements in THIN-GBASE. Please see the following sections:* - *In chapter 4, i.e. BCC (section 4.3.4, pages 107 (lines 11-24) and 108 (lines: 1-25))* - *In chapter 5, i.e. lung cancer (section 5.5, page 161 (lines 17-25), 162 (lines: all) and 163 (lines 1-4))* - *In chapter 6, i.e. GIT cancer (section 6.6, pages 230 (lines 3-24) and 231 (lines 1-8))*
2.	**<u>Comments on the exposure model</u>** **2.1. External examiner:** The building of a valid exposure model for the population studied is one of the most crucial tasks for an environmental epidemiology analysis. The challenges involved in addressing this task are such that a specific report on the population exposure model building and

	results is often considered a requirement or certainly very helpful, before proceeding to estimation of disease risk based on exposure model. Although the exposure model developed and used by Anwar represents perhaps the main result included in his thesis, numerous aspects of such model did not benefit from as full a description as one would wish. ***Response**: Thank you for your comments. However, I believe I have provided a full description and an account of the models used to quantify the risk for each type of cancer. In Chapter 5 and 6 includes a detailed description of the data mining methods used for determining which subset of soil metals before implementing them into my risk model (in sections under 5.3.3, and sections under 6.4.2).* *The comments raised in 2.1 are similar those in 2.4 - I have provided an in-depth response as to why this research did not include an exposure model for calculating lifetime dose (or intake).* **2.2. External examiner:** The limitations of an exposure model to chemical elements that does not include a reference to validity of such exposure as could be confirmed by estimation of biomarkers was acknowledged in the thesis. In view of this, it would have helped to refer to "potential exposure" rather than simple "exposure" especially when discussing conclusions on causality of any associations identified ***Response**: Thank you for your comments, where applicable in the discussions sections of chapter 4 (section 4.3.4), chapter 5 (section 5.5), chapter 6 (section 6.6) and chapter 7, the term "exposure" as has been replace with the terms "potential exposure(s)".* **2.3. External examiner:** It would have been desirable to include a descriptive account and tables on the number and variation in the number of individuals contained in the geographical areas that were used as units of observation would have helped, and so would have also a general coherent description of such spatial units in terms of their size and other characteristics potentially meaningful for the analysis, and the description for the overall cohort population, and also perhaps separately for the two areas of geochemical sampling underlying the smooth surface, of each element median and range by quintile of its distribution. ***Response**: Thank you for your comments. I agree with the above statement and the fact that this thesis would have benefitted with the inclusion of a descriptive table and a visual map showing the number or density of individuals (i.e. cases and non-cases), respectively, contained in the geographical study areas. The work would have also benefitted with a full description of the cohort population from THIN within these three sampling areas (i.e. G-BASE urban, rural or NSI(XRFS)). However, this addition is not feasible because the records were anonymised. These limitations have been acknowledged in the discussions in chapters 4, 5 and 6. But, in chapter 3, I do provide a description on the overall numbers of patients with soil measurements across all elements, as well as the distribution of patients with specific concentrations for each of the 15 elements provided in the linkage (see chapter 3, 3.3.4, table 3.2 (page 49) and figure 3.3-3.17 (pages 52-66)).*

2.4. External examiner: A further element of distribution of population exposure, as opposed to environmental concentration of an element, that would have helped, would be the accounting for estimated intake into the body by year of life, based on several assumptions required to be made mode explicit. This could have led to some approximate estimate of the lifetime of the study population represented by values in this exposure model, in terms of percentage, and to description of such exposure indicator and its variation by space and time.

***Response**: Thank you for your comments. I agree with the above statement, the implementation of an exposure model to estimate the cumulative (i.e. lifetime) dose (or intake) of a contaminant via ingestion of soil would have been a useful exposure variable for this research. However, I refrained from using this approach for the following reasons: 1.) An exposure model is typically a function of the length of exposure an individual may experience at a source (i.e. postcode). Usually, the time spent living at a residence is a proxy for measuring length of exposure; however, I was unable to derive these dose exposure estimates because of the paucity of records in THIN pertained to a person's residential history and the time spent at the address; 2.) The exposure model is also dependent on other environmental and dietary factors such as source of drinking water (i.e. name of water suppliers, boreholes, wells etc.) and type of food consumed (and their trace levels of elements from such food sources etc.) for which such information is not present in THIN; 3.) In absence of such data in THIN, the parameterisation of an exposure model will be problematic. One would have to rely on scientific and grey literature elsewhere to abstract values (which may not be generalizable or representative of the UK population) in order to parameterise the exposure model for calculating dose exposure (or intake) – which I believe is inappropriate as this will introduce certain bias in my risk estimates for BCC, lung and GIT cancer.*

2.5. External examiner: Similarly, several other population exposure indicators could have been defined more clearly, such as exposure indicators for period within the life of the exposed population, either the first 20 years or other period. In the absence of this information, the discussion sections in each of the cancer risk chapters may be supplemented by a paragraph commenting on the possible types of error in risk estimates due to measurement errors in population indices of potential exposure.

***Response**: Thank you for your comments. Please see the responses made for comments numbered as 1.4.*

2.6. External examiner: Also, I would like to comment on the selection of the chemical elements to be analysed, described on pages 42-44. The circumstance by which Anwar used a dataset linked by collaborators who had agreed this before he started his PhD work appears to indicate that he was not involved in choices made regarding the selection of chemical elements (15) analysed in his dissertation, compared to the total number actually available (48). Within his PhD work, Anwar adopted a "screening" procedure, described as "Stage 1", for further selection of the elements to be analysed, which was essentially a data-driven approach. Alternative approaches could have been discussed on page 42 or in the relevant discussion section, such as a selection justified entirely based on literature evidence of recognised or potential harm to human health.

	***Response**: Thank you for your comments. 15 of the elements (out of the 48) were selected for the linkage with the medical records in THIN for the following reasons: 1.) they are major elements with the greatest abundance in soil (aluminium, calcium and silicon); 2.) their influence in terms of mobility of trace elements in soil (e.g. iron's influence on the mobility of arsenic); and 3.) interests due to known or suspected impacts (i.e. beneficial or adverse) on health human (arsenic, chromium, copper, lead, manganese, nickel, phosphorus, selenium, uranium, vanadium and zinc). The remaining element that were not included in the linkage were excluded because 75% of their samples were typically below detection limits.* *In response to comments made regarding the use of data mining – I have included a justification for using filter-based methods for the selection of potential exposure variables in chapter 5 (section 5.3.3.1, page 127, lines 7-14)*
3.	**<u>Comment on confounders</u>** **3.1. External examiner:** It was good to see an attempt to control for the major known cause of BCC by inclusion of a proxy indicator for UV exposure, as well as inclusion of a smoking indicator for analyses of other cancers. It would have been interesting to see a discussion of the value and potential limitations of such variables tailored to each cancer analysis. ***Response**: Thank you for your comments, the acquisition of sunlight data from the UK Meteorological office, as well as, the calculation of the lifetime UV exposure measure was a challenge. I have included in the discussions the merits and limitations of the inclusion of such proxy indicator as adjustments in our BCC models. See chapter 4 (section 4.3.4, page 109, lines 8-20)* **3.2. External examiner:** The presence of numerous possible confounders in an ecological design makes the task of focusing on arear-level potential confounders particularly challenging. Although adjustment for Strategic Health Authority (SHA) was included in the models, the presence of multiple possible confounders at smaller area level suggests that some further discussion of the potential role of factors operating at area level, such as variability in the practice of pathologists proving histological report for BCC in particular but also other cancers, variability in the ascertainment rate of such cancers by different primary care/hospitals arrangements, and their overlapping with spatial variation in the exposure model. ***Response**: Thank you for your comments. Please see the responses made for comments numbered as 1.4.*
4.	**<u>Comment on choice of health endpoints</u>**

	4.1. External examiner: I consider that cancers were appropriate endpoints for this analysis, as they are of grave concern to society and cancer registry ascertainment is among the most complete for any chronic disease. This is not the case for BCC, and although its inclusion may be justified by the literature on arsenic effects, the possible error in BCC risk estimates arising from differential ascertainment within SHA (due among others to variation in practice by GPs and /or pathologists providing histology diagnosis), and methods for minimising this error could be more explicitly recognised in the discussion. Among cancers, a more toxicologically-based discussion of choices for selection of relevant endpoints might have been conducted, but as illustrated of a new method, those selected appear relevant. ***Response**: Thank you for your comments. Please see the responses made for comments numbered as 1.4.*
5.	**Comments on causality discussion** **5.1. External examiner:** Dose-response and plausibility were mentioned, consistency with other evidence was alluded to. Although the conclusions arrived at would have appeared stronger if a more full discussion of possible bias due to selection or confounding had been provided, on the whole they appear reasonable. ***Response**: No changes made*
6.	**Comments on suggested future work** **External examiner:** All the activities mentioned in section 7.3 (page 232) appear desirable, but in my view before proceeding with any of them, it would be important to document the achievements of the present thesis with a detailed account of two related topics (which might be combined) to produce possible future peer-reviewed publications: (a) exposure results for this population, entirely based on the work already completed, and providing tables with information accounting for the methods considered from the population point of view, and quantitative estimates of potential exposure estimated based on these methods. Such results could be described by explicitly adopting the concepts and language of environmental epidemiology, in other words defining population based exposure indicators derived from the information available. These indicators could be computed by making explicit all the assumptions required in applying soil concentrations of metals considered implicitly as proxies for doses ingested by human beings over relevant time periods; (b) suggested protocol for the ideal epidemiological analysis, that would include new elements for the design (such as the access to individual level THIN data as mentioned by the candidate), an exposure model (such as based on exposure as reported in (a) but adding a validation study in a subsample of the population using biomarkers of exposure), and several choices for health endpoints and confounders. ***Response**: Thank you for your comments, a protocol has been added to chapter 7, section 7.3.1, pages 244-247*