BMJ Open

Improving Predictive Asthma Algorithms with Modelled Environment Data for Scotland

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2018-023289
Article Type:	Protocol
Date Submitted by the Author:	02-Apr-2018

Complete List of Authors:	Soyiri, Ireneous; University of Edinburgh, Usher Institute of Population Health Sciences and Informatics Sheikh, Aziz; University of Edinburgh, Usher Institute of Population Health Sciences and Informatics Reis, Stefan; Natural Environment Research Council, NERC Centre for Ecology & Hydrology, Biosphere Atmosphere Interactions Kavanagh, Kimberly; University of Strathclyde, Department of Mathematics and Statistics Vieno, Massimo; Natural Environment Research Council, NERC Centre for Ecology & Hydrology, Biosphere Atmosphere Interactions Clemens, Tom; University of Edinburgh, School of Geography & Geosciences, Department of Geography & Sustainable Development Carnell, Edward; Natural Environment Research Council, NERC Centre for Ecology & Hydrology, Biosphere Atmosphere Interactions Pan, Jiafeng; University of Strathclyde, Department of Mathematics and Statistics King, Abby; University of Edinburgh, Usher Institute of Population Health Sciences and Informatics Beck, Rachel; Natural Environment Research Council, NERC Centre for Ecology & Hydrology, Biosphere Atmosphere Interactions Ward, Hester; NHS Scotland National Services Division, Information Services Division (ISD) & Health Protection Scotland (HPS) Dibben, Chris; University of Edinburgh, School of Geography & Geosciences, Department of Geography & Sustainable Development Robertson, Chris; University of Strathclyde, Department of Mathematics and Statistics Simpson, Colin; University of Edinburgh, Usher Institute of Population Health Sciences and Informatics

Keywords:	Asthma < THORACIC MEDICINE, PRIMARY CARE, STATISTICS & RESEARCH METHODS, Learning Health System, Environmental Epidemiology, Pollution Effects

Improving Predictive Asthma Algorithms with Modelled Environment Data for Scotland

Ireneous N Soyiri[1*], Aziz Sheikh[1], Stefan Reis[2], Kimberley Kavanagh[3], Massimo Vieno[2], Tom Clemens[4], Edward J Carnell[2], Jiafeng Pan[3], Abby King[1], Rachel C Beck[2], Hester Ward[5], Chris Dibben[4], Chris Robertson[3], Colin R Simpson[1,6]

[1]The University of Edinburgh, Usher Institute of Population Health Sciences and Informatics, Centre for Medical Informatics, Edinburgh, EH8 9AG, UK

[2]NERC Centre for Ecology & Hydrology, Atmospheric Chemistry and Effects, Edinburgh, EH26 0QB, UK

[3]The University of Strathclyde, Department of Mathematics and Statistics, Livingstone Tower, Glasgow, G1 1XH, UK

[4]The University of Edinburgh, School of Geosciences, Institute of Geography, Edinburgh, EH8 9XP, UK

[5]NHS National Services Scotland, Information Services Division & Health Protection Scotland, Edinburgh, EH12 9EB, UK

[6]Victoria University of Wellington, Faculty of Health, Wellington, New Zealand

***Correspondence:**
Ireneous Soyiri, University of Edinburgh, Usher Institute of Population Health Sciences and Informatics, Centre for Medical Informatics, Medical School, Teviot Place, Edinburgh, EH8 9AG, UK. Email: I.Soyiri@ed.ac.uk

Funding information: This project is funded by the Natural Environment Research Council (NERC), the Medical Research Council (MRC) and the Scottish Government Chief Scientist Office (CSO); Grant Ref: R8/H12/83/NE/P011012/1 under the Innovation grant scheme on Improving Health Care with Environment Data with support from the Farr Institute and the Asthma UK Centre for Applied Research.

Word count: 2,101

Abstract

Introduction: Asthma has a considerable, but potentially, avoidable burden on many populations globally. Scotland has some of the poorest health outcomes from asthma. Although ambient pollution, weather changes and sociodemographic factors have been associated with asthma attacks, it remains unclear whether modelled environment data and geospatial information can improve population-based asthma predictive algorithms. We aim to create the afferent loop of a national learning health system for asthma in Scotland. We will investigate the associations between ambient pollution, meteorological, geospatial and sociodemographic factors, and asthma attacks.

Methods: We will develop and implement a secured data governance and linkage framework to incorporate primary care health data, modelled environment data, geospatial population and sociodemographic data. Data from 75 recruited primary care practices (n=500,000 patients) in Scotland will be used. Modelled environment data on key air pollutants at a horizontal resolution of 5 km x 5 km at hourly time steps will be generated using the EMEP4UK atmospheric chemistry transport modelling system for the datazones of the primary care practices' populations. Scottish population census and education databases will be incorporated into the linkage framework for analysis. We will then undertake a longitudinal retrospective observational analysis. Asthma outcomes include asthma hospitalizations and oral steroid prescriptions. Using a nested-case control study design, associations between all covariates will be measured using conditional logistic regression to account for the matched design, and to identify suitable predictors and potential candidate algorithms for an asthma learning health system in Scotland.

Conclusions: Findings from this study will contribute to the development of predictive algorithms for asthma outcomes and be used to form the basis for our learning health system prototype.

Ethics and Dissemination: The study received National Health Service Research Ethics Committee (NHS REC) approval (16/SS/0130), and also obtained permissions via the Public and Patient Privacy Panel (PBPP) in Scotland to access, collate and use the following datasets: population and housing census for Scotland; Scottish education data via the Scottish Exchange of Data (ScotXed) and primary care data from general practice Data Custodians. Study protocol and analytic codes will be published in open source.

Key words: asthma, environmental epidemiology, pollution, predictive, modelled environment data, learning health system, algorithms

Article Summary

Strengths and limitations of this study

- The study will provide important evidence on how modelled environment data and geospatial information can improve population-based asthma predictive algorithms.

- We develop and implement a novel secured data governance and linkage framework to incorporate primary care health data, modelled environment data, geospatial population and sociodemographic data.

- Data from 75 recruited primary care practices (n=500,000 patients) in Scotland will be used to provide a highly powered statistical analysis.

- Like many data sources, the accuracy of modelled environment data (including ambient pollution and meteorological measures) that was generated from an atmospheric chemistry transport modelling system may affect the predictive asthma algorithm. Nonetheless with new data, the modelling process improves over time.

1. INTRODUCTION

Scotland has amongst the highest prevalence of asthma in the world and some of the poorest health outcomes from asthma [1]. Asthma is, as a result, responsible for considerable – potentially avoidable – morbidity, healthcare utilization and mortality [2], with an estimated cost to Scotland of at least £92m per annum, over £10m of which is spent on unscheduled care [3]. Serial confidential enquiries into UK asthma deaths have consistently highlighted major deficiencies in care and have come to the striking conclusion that around half of asthma deaths per year in the UK are potentially avoidable [4]. Furthermore, our detailed analysis of Scottish data has highlighted that there has been no decrease in rate of asthma hospitalisations or deaths in Scotland over the last decade [3]. To help address these poor health outcomes, we are working to develop a Learning Health System (LHS) for asthma in Scotland, which will allow real-time access to, and interrogation of, a wide range of data sources to provide tailored feedback to health care professionals in a format that they can readily action.

Pollution, meteorological, lifestyle and sociodemographic factors are known to contribute to asthma outcomes including morbidity and mortality. These factors by themselves have complex interactions that may mediate their overall impact on asthma outcomes such as asthma attacks. However, there are no comprehensive studies, which have examined the effects of ambient pollution, sociodemographic and lifestyle factors on asthma in large populations. Epidemiological studies have demonstrated the short- and long-term adverse respiratory/asthma health effects resulting from exposure to pollution and weather conditions such as extremes in temperature and humidity [5-7]. For instance, school proximity to urban infrastructure (e.g. highways and industry) has been linked to poor asthma outcomes [8-10]. Indoor air pollution, (e.g. from tobacco smoke or volatile organic compounds), as well as exposure to ambient air pollutants (including fine particles and nitrogen oxides) and pollen have all been shown to be associated with asthma attacks. [6 11-15] Key asthma care gaps can also lead to asthma attacks, e.g. poor clinical guideline adherence [16] and a lack of guided self-management/ provision of personal asthma action plans [17]. Nonetheless, there has been little work to demonstrate the combined effects of air pollutants, meteorological factors, sociodemographic factors and healthcare gaps on asthma attacks.

The current approaches to risk prediction have been weakened by an inability to include reliable modelled ambient air pollution and meteorological data [7 18 19]. Previous studies have estimated pollution exposures e.g. particulate matter from monitoring stations, attributing the value at the monitoring station to all locations around it (frequently up to 5-10 kms away). This lack of spatial representativeness of existing monitoring network data presents a substantial barrier for the identification of associations between pollution events and health effects [20]. In addition, the use of only residential location has meant that the spatio-temporal representativeness of measured air pollutant concentrations for the estimation of population exposure has been sub-optimal [21 22].

The specific challenge of predicting individuals likely to have an asthma attack who require clinical care can be addressed through the interrogation and analysis of general practice and other healthcare data linked to data sources such as meteorological, pollution, school and census data. Asthma represents an appropriate exemplar condition to investigate the value of bringing these data assets together because it is very common, affects people of all ages, both genders, and urban and rural populations, and because pollution and weather are both important risk factors for asthma attacks. We aim to investigate the associations between ambient pollution, meteorological, geospatial and sociodemographic factors, which will aid the future development of predictive algorithms for asthma outcomes.

Our objectives are to:

1. Gain the necessary permissions required to compile and use longitudinal disparate datasets from clinical and health services databases, environment, geospatial, population and sociodemographic and education databases;
2. Model potential meteorological and pollution predictors of asthma at higher resolutions;
3. Investigate geospatial locations and mobility, exposure at work place, schools and residence;
4. Understand the associations between environmental and health covariates and clinically relevant asthma outcomes;
5. Investigate the feasibility of creating a daily pollution and meteorological forecast for use in an operational learning health system.

2. METHODS

2.1. Ethics, data governance and management and patient involvement

The study received National Health Service Research Ethics Committee (NHS REC) approval (16/SS/0130). It has also obtained permissions via the Public and Patient Privacy Panel (PBPP) in Scotland to access, collate and use the following datasets: population and housing census for Scotland; Scottish education data via the Scottish Exchange of Data (ScotXed) and primary care data from general practice Data Custodians. We will conduct all data linkage, management and analysis using the secured environment provided by the Scottish Government's electronic Data Research and Innovation Service (eDRIS) hosted at the Farr Institute in Scotland. The standard protocols provided by eDRIS will be adopted for our data management.

Patients and or the public were not involved in the development of the research question and outcome measures, nor the study design. Patients were not involved in the recruitment to and conduct of the study. We plan to use the Asthma UK Centre for Applied Research Patient and Public Involvement Platform in the dissemination of our results (www.aukcar.ac.uk/public-involvement).

2.2. Study design, data sources and study period

A longitudinal retrospective observational study, involving at least two nested case control studies, using a national primary care linked database will be carried out. These studies will inform our further studies. The period of analysis will range from 1/1/2013 to 31/12/ 2015. The respective health (including key outcomes), environment, education and census datasets that will act as the exposure variables and confounding factors/effect modifiers are described subsequently and the process of data linkage illustrated in Figure 1.

2.3. Health data

Almost all individuals resident in Scotland are registered with a general practice, which provides and co-ordinates healthcare services, including prescriptions, free of charge. We will use data from about 500,000 patients from 75 primary care practices from across Scotland. The main health outcomes of interest are daily asthma hospitalisations and daily oral steroid prescriptions. These variables, together with other patient characteristics such as sex, body mass index (BMI) and previous attacks, which are linked to the encrypted unique patient identifier - the Community Health Index number (CHI), will be extracted. The anonymised primary care data will be linked (using CHI) to the national unscheduled care DataMart, which includes Accident and Emergency (Emergency Room known as A&E in the UK) and ambulance data, and hospital/mortality datasets (held by the Information Services Division (ISD) for the purpose of linkage and research).

2.4. Environmental data

We will utilise the EMEP4UK atmospheric chemistry transport modelling system (http://www.emep4uk.ceh.ac.uk/), which comprises the state-of-the-art Weather Research and Forecasting (WRF) meteorological model, and the EMEP4UK Chemistry Transport Model. EMEP4UK generates geospatially and temporally resolved data on key air pollutants at a horizontal resolution of 5 km x 5 km at hourly time steps for nitrogen oxides (NO, NO_2), ground level ozone (O_3), sulphur dioxide (SO_2), and the main components of fine particulate matter ($PM_{2.5}$ and PM_{10}) as well as meteorological factors including temperature, wind speed/direction, and humidity.

2.5. Census and school data

To enhance our exposure estimation for people with asthma we will link patient data from primary care practices to the 2011 Scottish national census and ScotXed, an education database with school attendance at the individual patient level. We will extract information relating to school and work location, hours spent at work, mode of transport taken when travelling to work as well as other individual level socio-economic and demographic data. This will generate greater contextual information about the patients [23-25]. We will use Data Zones as our area unit of analysis and calculate exposures at the locations provided by school and census data.

2.6. Data linkage

We will perform data linkage involving cross-sectoral administrative data and environmental and NHS health data. Available information from pollution, census and education databases will be linked to the individual patient health data (using the individual CHI number and geocodes) to allow us to implement more detailed exposure models by linking to concentration estimates pertaining to residential, work and commuting locations (Figure 1).

2.7. Modelling environmental factors

We will use the EMEP4UK atmospheric chemistry transport modelling system (http://www.emep4uk.ceh.ac.uk/), and provide both historic model outputs for a multi-year period to input into a statistical analysis, and to develop and evaluate an operational Scotland-wide high resolution (~1 km x 1 km) atmospheric composition forecast for priority gaseous air pollutants (nitrogen oxides - NO, NO_2 and ozone - O_3) and the main components of $PM_{2.5}$ and PM_{10}. The following tasks will be undertaken:

I. Undertake a scoping study for setting up the model framework and input parameters for both the meteorology (Weather Research Forecast model, WRF) and the atmospheric chemistry (atmospheric chemistry transport model, EMEP4UK);
II. Compare model forecasts with hourly observations of air pollutant concentrations from existing regulatory monitoring networks to validate model output at the appropriate scale and resolution;
III. Develop a robust statistical method to test the EMEP4UK model forecast skills and use this information to quantify uncertainties in the model results, which feeds into the final statistical analysis.

2.8. Statistical analysis

For the two outcomes specified – asthma hospitalisations and oral steroid prescriptions, we will construct from the cohort, two nested-case control studies whereby the cases are those with the relevant outcome. As not everyone with asthma may be identifiable (in terms of a risk group coding) from the data prior to an attack, we will include all individuals in the extraction process for cases and the subsequent matching of controls. All cases will be identified with corresponding date of exacerbation dependent on the outcome - either date of hospitalisation or date of prescription dispense. Controls will be assigned to cases matching on age (within two years) and gender. We will aim to assign up to six controls for each case. For inclusion in the nested-case control study, we will restrict to individuals who have been registered with their GP for at least one year prior to the case date (with controls assigned a pseudo-date matching the date of outcome for their case) to allow appropriate GP history to construct the health related variables. For all cases and controls, all health related and comorbidity variables will be calculated relative to the case date. Environmental exposure and metrological variables will be linked based on the appropriate datazones (place of residence, place of work, place of education) for the three days prior to the case date, to investigate the time dependent effects of short-term exposure to air pollution and temperature. We will also use a longer-term look back four to seven days previously, as well as eight to 35 days. Average exposure will be calculated for the longer durations. Average as well as maximum exposure will be calculated for the short-term durations. Associations between all covariates (health and environmental) will be measured using conditional logistic regression to account for the matched design.

3. REPORTING AND DISSEMINATION

The study protocol will be registered with the European Union electronic Register of Post-Authorisation Studies (EU PAS Register) as a non-interventional post-authorisation study (PAS) before the commencement of data analysis. We shall also aim to publish our analytic codes in the open source GitHub website (https://github.com/asthmalhs).

To culminate the project work, a dissemination workshop will be held in which key stakeholders will be invited to participate. The meeting session will aim to leverage support from affiliated organisations and from individuals who are experts in this area of research to help chart a new research lines that will extend our pilot study.

Data Statement

The technical appendix and the dataset used for this study will be made accessible via the eDRIS secured platform under the project number: 1516-0489.

4. CONCLUSIONS

Findings from this project will aid the future development of predictive algorithms for asthma outcomes and be used to form the basis for our learning health system.

ACKNOWLEDGEMENTS

This project is funded by the Natural Environment Research Council (NERC), the Medical Research Council (MRC) and the Scottish Government Chief Scientist Office (CSO); Grant Ref: R8/H12/83/NE/P011012/1 under the Innovation grant scheme on Improving Health Care with Environment Data with support from the Farr Institute and the Asthma UK Centre for Applied Research.

CONFLICTS OF INTERESTS

The authors have no competing interests.

Author Contributions

Conceptualization: CRS, AS, INS, SR, MV, KK
Methodology: CRS, AS, INS, SR, MV, KK, CR
Supervision: CRS, SR, CR
Writing original draft: INS, CRS
Writing review & editing: CRS, AS, INS, SR, KK, MV, TC, EJC, JP, AK, RCB, HW, CD, CR

REFERENCES

1. Patel SP, Jarvelin M-R, Little MP. Systematic review of worldwide variations of the prevalence of wheezing symptoms in children. *Environ Health* 2008;7(1):57.
2. Anandan C, Gupta R, Simpson C, et al. Epidemiology and disease burden from allergic disease in Scotland: analyses of national databases. *Journal of the Royal Society of Medicine* 2009;102(10):431-42.
3. Mukherjee M, Stoddart A, Gupta RP, et al. The epidemiology, healthcare and societal burden and costs of asthma in the UK and its member nations: analyses of standalone and linked national databases. *BMC Medicine* 2016;14(1):1-15. doi: 10.1186/s12916-016-0657-8
4. Levy ML. The national review of asthma deaths: what did we learn and what needs to change? *Breathe* 2015;11(1):14.
5. Gauderman WJ, Urman R, Avol E, et al. Association of Improved Air Quality with Lung Development in Children. *New England Journal of Medicine* 2015;372(10):905-13. doi: doi:10.1056/NEJMoa1414123
6. Guarnieri M, Balmes JR. Outdoor air pollution and asthma. *The Lancet* 2014;383(9928):1581-92. doi: http://dx.doi.org/10.1016/S0140-6736(14)60617-6
7. Lim SS, Vos T, Flaxman AD, et al. A comparative risk assessment of burden of disease and injury attributable to 67 risk factors and risk factor clusters in 21 regions, 1990–2010: a systematic analysis for the Global Burden of Disease Study 2010. *The Lancet* 2012;380(9859):2224-60. doi: http://dx.doi.org/10.1016/S0140-6736(12)61766-8
8. Wu Y-C, Batterman SA. Proximity of schools in Detroit, Michigan to automobile and truck traffic. *Journal of exposure science and environmental epidemiology* 2006;16(5):457-70.
9. Kopnina H. asthma and air pollution: connecting the Dots. *A Companion to the Anthropology of Environmental Health* 2016:142.
10. Son J-Y, Kim H, Bell ML. Does urban land-use increase risk of asthma symptoms? *Environmental Research* 2015;142:309-18. doi: http://dx.doi.org/10.1016/j.envres.2015.06.042
11. Dick S, Friend A, Dynes K, et al. A systematic review of associations between environmental exposures and development of asthma in children aged up to 9 years. *Bmj Open* 2014;4(11) doi: 10.1136/bmjopen-2014-006554
12. Saulyte J, Regueira C, Montes-Martinez A, et al. Active or passive exposure to tobacco smoking and allergic rhinitis, allergic dermatitis, and food allergy in adults and children: a systematic review and meta-analysis. *PLoS medicine* 2014;11(3):e1001611-e11. doi: 10.1371/journal.pmed.1001611
13. Rufo CJ, Madureira J, Fernandes OE, et al. Volatile organic compounds in asthma diagnosis: a systematic review and meta-analysis Corrigendum (vol 71, 175, 2016). *Allergy* 2016;71(5):733-33.
14. Nurmatov UB, Tagiyeva N, Semple S, et al. Volatile organic compounds and risk of asthma and allergy: a systematic review. *European respiratory review : an official journal of the European Respiratory Society* 2015;24(135):92-101. doi: 10.1183/09059180.00000714
15. Bousquet J, O'Hehir RE, Anto JM, et al. Assessment of thunderstorm-induced asthma using Google Trends. *Journal of Allergy and Clinical Immunology* 2017;140(3):891-93.e7. doi: https://doi.org/10.1016/j.jaci.2017.04.042

16. Bush A, Kleinert S, Pavord ID. The asthmas in 2015 and beyond: a Lancet Commission. *The Lancet* 2015;385(9975):1273-75.
17. Pinnock H. Supported self-management for asthma. *Breathe* 2015;11(2):98-109. doi: 10.1183/20734735.015614
18. Jerrett M, Arain A, Kanaroglou P, et al. A review and evaluation of intraurban air pollution exposure models. *J Expo Anal Environ Epidemiol* 2004;15(2):185-204.
19. Reis EJ, Guzmán RM. An econometric model of Amazon deforestation. 2015
20. Willocks LJ, Bhaskar A, Ramsay CN, et al. Cardiovascular disease and air pollution in Scotland: no association or insufficient data and study design? *BMC Public Health* 2012;12(1):1-6. doi: 10.1186/1471-2458-12-227
21. Dibben C, Clemens T. Place of work and residential exposure to ambient air pollution and birth outcomes in Scotland, using geographically fine pollution climate mapping estimates. *Environmental Research* 2015;140:535-41. doi: http://dx.doi.org/10.1016/j.envres.2015.05.010
22. Butland BK, Armstrong B, Atkinson RW, et al. Measurement error in time-series analysis: a simulation study comparing modelled and monitored data. *BMC Medical Research Methodology* 2013;13(1):136. doi: 10.1186/1471-2288-13-136
23. Strickland MJ, Klein M, Darrow LA, et al. The issue of confounding in epidemiological studies of ambient air pollution and pregnancy outcomes. *Journal of Epidemiology and Community Health* 2009;63(6):500-04. doi: 10.1136/jech.2008.080499
24. Woodruff TJ, Parker JD, Darrow LA, et al. Methodological issues in studies of air pollution and reproductive health. *Environmental Research* 2009;109(3):311-20. doi: http://dx.doi.org/10.1016/j.envres.2008.12.012
25. Clemens T, Dibben C. A method for estimating wage, using standardised occupational classifications, for use in medical research in the place of self-reported income. *BMC Medical Research Methodology* 2014;14(1):1-8. doi: 10.1186/1471-2288-14-59

Figures

Figure 1: Data linkage framework in secured environment

(See attached)

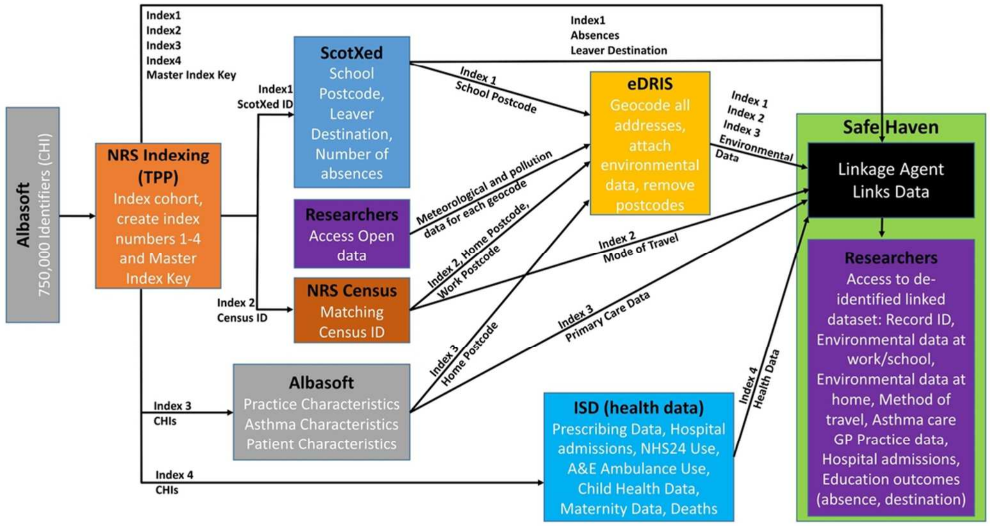

Figure 1: Data linkage framework in secured environment

49x27mm (600 x 600 DPI)

BMJ Open

Improving Predictive Asthma Algorithms with Modelled Environment Data for Scotland: an observational cohort study protocol

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2018-023289.R1
Article Type:	Protocol
Date Submitted by the Author:	13-Apr-2018
Complete List of Authors:	Soyiri, Ireneous; University of Edinburgh, Usher Institute of Population Health Sciences and Informatics Sheikh, Aziz; University of Edinburgh, Usher Institute of Population Health Sciences and Informatics Reis, Stefan; Natural Environment Research Council, NERC Centre for Ecology & Hydrology, Biosphere Atmosphere Interactions Kavanagh, Kimberly; University of Strathclyde, Department of Mathematics and Statistics Vieno, Massimo; Natural Environment Research Council, NERC Centre for Ecology & Hydrology, Biosphere Atmosphere Interactions Clemens, Tom; University of Edinburgh, School of Geography & Geosciences, Department of Geography & Sustainable Development Carnell, Edward; Natural Environment Research Council, NERC Centre for Ecology & Hydrology, Biosphere Atmosphere Interactions Pan, Jiafeng; University of Strathclyde, Department of Mathematics and Statistics King, Abby; University of Edinburgh, Usher Institute of Population Health Sciences and Informatics Beck, Rachel; Natural Environment Research Council, NERC Centre for Ecology & Hydrology, Biosphere Atmosphere Interactions Ward, Hester; NHS Scotland National Services Division, Information Services Division (ISD) & Health Protection Scotland (HPS) Dibben, Chris; University of Edinburgh, School of Geography & Geosciences, Department of Geography & Sustainable Development Robertson, Chris; University of Strathclyde, Department of Mathematics and Statistics Simpson, Colin; University of Edinburgh, Usher Institute of Population Health Sciences and Informatics
Primary Subject Heading:	Respiratory medicine
Secondary Subject Heading:	Epidemiology
Keywords:	Asthma < THORACIC MEDICINE, PRIMARY CARE, STATISTICS & RESEARCH METHODS, Learning Health System, Environmental Epidemiology, Pollution Effects

Improving Predictive Asthma Algorithms with Modelled Environment Data for Scotland: an observational cohort study protocol

Ireneous N Soyiri[1*], Aziz Sheikh[1], Stefan Reis[2,3], Kimberley Kavanagh[4], Massimo Vieno[2], Tom Clemens[5], Edward J Carnell[2], Jiafeng Pan[4], Abby King[1], Rachel C Beck[2], Hester JT Ward[6], Chris Dibben[5], Chris Robertson[4], Colin R Simpson[1,7]

[1]The University of Edinburgh, Usher Institute of Population Health Sciences and Informatics, Centre for Medical Informatics, Edinburgh, EH8 9AG, UK

[2]NERC Centre for Ecology & Hydrology, Atmospheric Chemistry and Effects, Edinburgh, EH26 0QB, UK

[3]University of Exeter Medical School, Knowledge Spa, Truro, TR1 3HD, UK

[4]The University of Strathclyde, Department of Mathematics and Statistics, Livingstone Tower, Glasgow, G1 1XH, UK

[5]The University of Edinburgh, School of Geosciences, Institute of Geography, Edinburgh, EH8 9XP, UK

[6]NHS National Services Scotland, Information Services Division & Health Protection Scotland, Edinburgh, EH12 9EB, UK

[7]Victoria University of Wellington, Faculty of Health, Wellington, New Zealand

***Correspondence:**
Ireneous Soyiri, University of Edinburgh, Usher Institute of Population Health Sciences and Informatics, Centre for Medical Informatics, Medical School, Teviot Place, Edinburgh, EH8 9AG, UK. Email: I.Soyiri@ed.ac.uk

Funding information: This project is funded by the Natural Environment Research Council (NERC), the Medical Research Council (MRC) and the Scottish Government Chief Scientist Office (CSO); Grant Ref: R8/H12/83/NE/P011012/1 under the Innovation grant scheme on Improving Health Care with Environment Data with support from the Farr Institute and the Asthma UK Centre for Applied Research.

Word count: 2,244

Abstract

Introduction: Asthma has a considerable, but potentially, avoidable burden on many populations globally. Scotland has some of the poorest health outcomes from asthma. Although ambient pollution, weather changes and sociodemographic factors have been associated with asthma attacks, it remains unclear whether modelled environment data and geospatial information can improve population-based asthma predictive algorithms. We aim to create the afferent loop of a national learning health system for asthma in Scotland. We will investigate the associations between ambient pollution, meteorological, geospatial and sociodemographic factors, and asthma attacks.

Methods: We will develop and implement a secured data governance and linkage framework to incorporate primary care health data, modelled environment data, geospatial population and sociodemographic data. Data from 75 recruited primary care practices (n=500,000 patients) in Scotland will be used. Modelled environment data on key air pollutants at a horizontal resolution of 5 km x 5 km at hourly time steps will be generated using the EMEP4UK atmospheric chemistry transport modelling system for the datazones of the primary care practices' populations. Scottish population census and education databases will be incorporated into the linkage framework for analysis. We will then undertake a longitudinal retrospective observational analysis. Asthma outcomes include asthma hospitalisations and oral steroid prescriptions. Using a nested-case control study design, associations between all covariates will be measured using conditional logistic regression to account for the matched design, and to identify suitable predictors and potential candidate algorithms for an asthma learning health system in Scotland.

Conclusions: Findings from this study will contribute to the development of predictive algorithms for asthma outcomes and be used to form the basis for our learning health system prototype.

Ethics and Dissemination: The study received National Health Service Research Ethics Committee approval (16/SS/0130), and also obtained permissions via the Public Benefit and Privacy Panel for Health and Social Care in Scotland to access, collate and use the following datasets: population and housing census for Scotland; Scottish education data via the Scottish Exchange of Data and primary care data from general practice Data Custodians. Analytic code will be made available in the open

source GitHub website. The results of this study will be published in international peer reviewed journals.

Key words: asthma, environmental epidemiology, pollution, predictive, modelled environment data, learning health system, algorithms

Article Summary

Strengths and limitations of this study

- Development and implementation of a novel data governance and linkage framework to incorporate primary care health, modelled environmental, geospatial population and sociodemographic data.
- Data from 75 recruited primary care practices (n=500,000 patients) in Scotland will be used, 10% Scottish population
- We will create a daily model forecast of ambient air pollution and meteorology for 72 hours for the study area at a high geospatial resolution and for specific pollutants.
- Like many data sources, the accuracy of modelled environment data (including ambient pollution and meteorological measures) that was generated from an atmospheric chemistry transport modelling system may affect the predictive asthma algorithm.
- Will need to evaluate whether the new model affects care and improves outcomes for patients in future studies.

1. INTRODUCTION

Scotland has amongst the highest prevalence of asthma in the world and some of the poorest health outcomes from asthma [1]. Asthma is, as a result, responsible for considerable – potentially avoidable – morbidity, healthcare utilisation and mortality [2], with an estimated cost to Scotland of at least £92m per annum, over £10m of which is spent on unscheduled care [3]. Serial confidential enquiries into UK asthma deaths have consistently highlighted major deficiencies in care and have come to the striking conclusion that around half of asthma deaths per year in the UK are potentially avoidable [4]. Furthermore, our detailed analysis of Scottish data has highlighted that there has been no decrease in rate of asthma hospitalisations or deaths in Scotland over the last decade [3]. To help address these poor health outcomes, we are working to develop a Learning Health System (LHS) for asthma in Scotland, which will allow real-time access to, and interrogation of, a wide range of data sources to provide tailored feedback to health care professionals in a format that they can readily action.

Pollution, meteorological, lifestyle and sociodemographic factors are known to contribute to asthma outcomes including morbidity and mortality. These factors by themselves have complex interactions that may mediate their overall impact on asthma outcomes such as asthma attacks. However, there are no comprehensive studies, which have examined the effects of ambient pollution, sociodemographic and lifestyle factors on asthma in large populations. Epidemiological studies have demonstrated the short- and long-term adverse respiratory/asthma health effects resulting from exposure to pollution and weather conditions such as extremes in temperature and humidity [5-7]. For instance, school proximity to urban infrastructure (e.g. highways and industry) has been linked to poor asthma outcomes [8-10]. Indoor air pollution, (e.g. from tobacco smoke or volatile organic compounds), as well as exposure to ambient air pollutants (including fine particles and nitrogen oxides) and pollen have all been shown to be associated with asthma attacks. [6 11-15] Key asthma care gaps can also lead to asthma attacks, e.g. poor clinical guideline adherence [16] and a lack of guided self-management/ provision of personal asthma action plans [17]. Nonetheless, there has been little work to demonstrate the combined effects of air pollutants, meteorological factors, sociodemographic factors and healthcare gaps on asthma attacks.

The current approaches to risk prediction have been weakened by an inability to include reliable modelled ambient air pollution and meteorological data [7 18 19]. Previous studies have estimated pollution exposures e.g. particulate matter from monitoring stations, attributing the value at the monitoring station to all locations around it (frequently up to 5-10 kms away). This lack of spatial representativeness of existing monitoring network data presents a substantial barrier for the identification of associations between pollution events and health effects [20]. In addition, the use of only residential location has meant that the spatio-temporal representativeness of measured air pollutant concentrations for the estimation of population exposure has been sub-optimal [21 22].

The specific challenge of predicting individuals likely to have an asthma attack who require clinical care can be addressed through the interrogation and analysis of general practice and other healthcare data linked to data sources such as meteorological, pollution, school and census data. Asthma represents an appropriate exemplar condition to investigate the value of bringing these data assets together because it is very common, affects people of all ages, both genders, and urban and rural populations, and because pollution and weather are both important risk factors for asthma attacks. We aim to investigate the associations between ambient pollution, meteorological, geospatial and sociodemographic factors, which will aid the future development of predictive algorithms for asthma outcomes.

Our objectives are to:

1. Gain the necessary permissions required to compile and use longitudinal disparate datasets from clinical and health services databases, environment, geospatial, population and sociodemographic and education databases;
2. Model potential meteorological and pollution predictors of asthma at higher resolutions;
3. Investigate geospatial locations and mobility, exposure at work place, schools and residence;
4. Understand the associations between environmental and health covariates and clinically relevant asthma outcomes;
5. Investigate the feasibility of creating a daily pollution and meteorological forecast for use in an operational learning health system.

2. METHODS

2.1. Patient and Public Involvement (PPI)

Patients and or the public were not involved in the development of the research question and outcome measures of this study, however the PPI group of the Asthma UK Centre for Applied Research (AUKCAR) commented on the early design of the LHS for asthma in Scotland project. Patients were not involved in the recruitment to and conduct of the study. We plan to use the AUKCAR's Patient and Public Involvement Platform in the dissemination of our results (www.aukcar.ac.uk/public-involvement).

2.2. Study Design, Data Sources and Study Period

A longitudinal retrospective observational study, involving at least two nested case control studies, using a national primary care linked database will be carried out. These studies will inform our further studies. The period of analysis will range from 01/01/2013 to 31/12/2015. The respective health (including key outcomes), environment, education and census datasets that will act as the exposure variables and confounding factors/effect modifiers are described subsequently and the process of data linkage illustrated in Figure 1.

2.3. Health Data

Almost all individuals resident in Scotland are registered with a general practice, which provides and co-ordinates healthcare services, including prescriptions, free of charge. We will use data from about 500,000 patients from 75 primary care practices from across Scotland. The main health outcomes of interest are daily asthma hospitalisations and daily oral steroid prescriptions. These variables, together with other patient characteristics such as sex, body mass index (BMI) and previous attacks, which are linked to the encrypted unique patient identifier - the Community Health Index number (CHI), will be extracted. The anonymised primary care data will be linked (using CHI) to the national unscheduled care DataMart, which includes Accident and Emergency (Emergency Room known as A&E in the UK) and ambulance data, and hospital/mortality datasets (held by the Information Services Division (ISD) for the purpose of linkage and research).

2.4. Environmental Data

We will utilise the EMEP4UK atmospheric chemistry transport modelling system (http://www.emep4uk.ceh.ac.uk/), which comprises the state-of-the-art Weather Research and Forecasting (WRF) meteorological model, and the EMEP4UK Chemistry Transport Model. EMEP4UK generates geospatially and temporally resolved data on key air pollutants at a horizontal resolution of 5 km x 5 km at hourly time steps for nitrogen oxides (NO, NO_2), ground level ozone (O_3), sulphur dioxide (SO_2), and the main components of fine particulate matter ($PM_{2.5}$ and PM_{10}) as well as meteorological factors including temperature, wind speed/direction, and humidity.

2.5. Census and School Data

To enhance our exposure estimation for people with asthma we will link patient data from primary care practices to the 2011 Scottish national census and ScotXed, an education database with school attendance at the individual patient level. We will extract information relating to school and work location, hours spent at work, mode of transport taken when travelling to work as well as other individual level socio-economic and demographic data. This will generate greater contextual information about the patients [23-25]. We will use Data Zones as our area unit of analysis and calculate exposures at the locations provided by school and census data.

2.6. Data Linkage

We will perform data linkage involving cross-sectoral administrative data and environmental and NHS health data. Available information from pollution, census and education databases will be linked to the individual patient health data (using the individual CHI number and geocodes) to allow us to implement more detailed exposure models by linking to concentration estimates pertaining to residential, work and commuting locations (Figure 1).

2.7. Modelling Environmental Factors

We will use the EMEP4UK atmospheric chemistry transport modelling system (http://www.emep4uk.ceh.ac.uk/), and provide both historic model outputs for a multi-year period to input into a statistical analysis, and to develop and evaluate an operational Scotland-wide high resolution (~1 km x 1 km) atmospheric composition forecast for priority gaseous air pollutants (nitrogen oxides - NO, NO_2 and ozone - O_3) and the main components of $PM_{2.5}$ and PM_{10}. The following tasks will be undertaken:

I. Undertake a scoping study for setting up the model framework and input parameters for both the meteorology (Weather Research Forecast model, WRF) and the atmospheric chemistry (atmospheric chemistry transport model, EMEP4UK);

II. Compare model forecasts with hourly observations of air pollutant concentrations from existing regulatory monitoring networks to validate model output at the appropriate scale and resolution;

III. Develop a robust statistical method to test the EMEP4UK model forecast skills and use this information to quantify uncertainties in the model results, which feeds into the final statistical analysis.

2.8. Statistical Analysis

For the two outcomes specified – asthma hospitalisations and oral steroid prescriptions, we will construct from the cohort, two nested-case control studies whereby the cases are those with the relevant outcome. As not everyone with asthma may be identifiable (in terms of a risk group coding) from the data prior to an attack, we will include all individuals in the extraction process for cases and the subsequent matching of controls. All cases will be identified with corresponding date of exacerbation dependent on the outcome - either date of hospitalisation or date of prescription dispense. Controls will be assigned to cases matching on age (within two years) and gender. We will aim to assign up to six controls for each case. For inclusion in the nested-case control study, we will restrict to individuals who have been registered with their GP for at least one year prior to the case date (with controls assigned a pseudo-date matching the date of outcome for their case) to allow appropriate GP history to construct the health related variables. For all cases and controls, all health related and comorbidity variables will be calculated relative to the case date. Environmental exposure and metrological variables will be linked based on the appropriate datazones (place of residence, place of work, place of education) for the three days prior to the case date, to investigate the time dependent effects of short-term exposure to air pollution and temperature. We will also use a longer-term look back four to seven days previously, as well as eight to 35 days. Average exposure will be calculated for the longer durations. Average as well as maximum exposure will be calculated for the short-term durations. Associations between all covariates (health and environmental) will be measured using conditional logistic regression to account for the matched design.

3. ETHICS AND DISSEMINATION

The study received National Health Service Research Ethics Committee (NHS REC) approval (16/SS/0130). It has also obtained permissions via the Public Benefit and Privacy Panel for Health and Social Care (PBPP) in Scotland to access, collate and use the following datasets: population and housing census for Scotland; Scottish education data via the Scottish Exchange of Data (ScotXed) and primary care data from general practice Data Custodians. We will conduct all data linkage, management and analysis using the secured environment provided by the Scottish Government's electronic Data Research and Innovation Service (eDRIS) hosted at the Farr Institute in Scotland. The standard protocols provided by eDRIS will be adopted for our data management.

The study protocol will be registered with the European Union electronic Register of Post-Authorisation Studies (EU PAS Register) as a non-interventional post-authorisation study (PAS) before the commencement of data analysis. The results of this study will be published in international peer reviewed journals. We shall also aim to publish our analytic codes in the open source GitHub website (https://github.com/asthmalhs).

A dissemination workshop will be held in which key stakeholders will be invited to participate. These stakeholders include asthma patients and representative patient group (PPI), clinicians, leading researchers, policy makers and industry. The meeting session will aim to leverage support from affiliated organisations and from individuals who are experts in this area of research to help chart new research lines that will extend our pilot study.

In order to further disseminate our research outputs to a wider audience, we will make presentations at international conferences and local workshops. Media channels of the University of Edinburgh (the Usher Institute Newsletter), Health Data Research UK (e.g. The Farr Scotland Newsletter) and the AUKCAR will be used to share publications resulting from this study. We will also engage with policy-makers by participating in organised events and meetings where we can showcase the outputs (findings and publications) of this study, and interact with policy-makers and the general public.

Data Statement

The technical appendix and the dataset used for this study will be made accessible via the eDRIS secured platform under the project number: 1516-0489.

4. CONCLUSIONS

Findings from this project will aid the future development of predictive algorithms for asthma outcomes and be used to form the basis for our learning health system.

ACKNOWLEDGEMENTS

This project is funded by the Natural Environment Research Council (NERC), the Medical Research Council (MRC) and the Scottish Government Chief Scientist Office (CSO); Grant Ref: R8/H12/83/NE/P011012/1 under the Innovation grant scheme on Improving Health Care with Environment Data with support from the Farr Institute and the Asthma UK Centre for Applied Research.

CONFLICTS OF INTERESTS

The authors have no competing interests.

Author Contributions

The research was conceptualized by CRS, AS, INS, SR, MV and KK. Methods were written by CRS, AS, INS, SR, MV, KK and CR. Supervision was provided by CRS, SR and CR. The original draft was produced by INS with inputs from CRS. All authors (CRS, AS, INS, SR, KK, MV, TC, EJC, JP, AK, RCB, HW, CD, CR) contributed to reviewing & editing.

REFERENCES

1. Patel SP, Jarvelin M-R, Little MP. Systematic review of worldwide variations of the prevalence of wheezing symptoms in children. *Environ Health* 2008;7(1):57.
2. Anandan C, Gupta R, Simpson C, et al. Epidemiology and disease burden from allergic disease in Scotland: analyses of national databases. *Journal of the Royal Society of Medicine* 2009;102(10):431-42.
3. Mukherjee M, Stoddart A, Gupta RP, et al. The epidemiology, healthcare and societal burden and costs of asthma in the UK and its member nations: analyses of standalone and linked national databases. *BMC Medicine* 2016;14(1):1-15. doi: 10.1186/s12916-016-0657-8
4. Levy ML. The national review of asthma deaths: what did we learn and what needs to change? *Breathe* 2015;11(1):14.
5. Gauderman WJ, Urman R, Avol E, et al. Association of Improved Air Quality with Lung Development in Children. *New England Journal of Medicine* 2015;372(10):905-13. doi: doi:10.1056/NEJMoa1414123
6. Guarnieri M, Balmes JR. Outdoor air pollution and asthma. *The Lancet* 2014;383(9928):1581-92. doi: http://dx.doi.org/10.1016/S0140-6736(14)60617-6
7. Lim SS, Vos T, Flaxman AD, et al. A comparative risk assessment of burden of disease and injury attributable to 67 risk factors and risk factor clusters in 21 regions, 1990–2010: a systematic analysis for the Global Burden of Disease Study 2010. *The Lancet* 2012;380(9859):2224-60. doi: http://dx.doi.org/10.1016/S0140-6736(12)61766-8
8. Wu Y-C, Batterman SA. Proximity of schools in Detroit, Michigan to automobile and truck traffic. *Journal of exposure science and environmental epidemiology* 2006;16(5):457-70.
9. Kopnina H. asthma and air pollution: connecting the Dots. *A Companion to the Anthropology of Environmental Health* 2016:142.
10. Son J-Y, Kim H, Bell ML. Does urban land-use increase risk of asthma symptoms? *Environmental Research* 2015;142:309-18. doi: http://dx.doi.org/10.1016/j.envres.2015.06.042
11. Dick S, Friend A, Dynes K, et al. A systematic review of associations between environmental exposures and development of asthma in children aged up to 9 years. *Bmj Open* 2014;4(11) doi: 10.1136/bmjopen-2014-006554
12. Saulyte J, Regueira C, Montes-Martinez A, et al. Active or passive exposure to tobacco smoking and allergic rhinitis, allergic dermatitis, and food allergy in adults and children: a systematic review and meta-analysis. *PLoS medicine* 2014;11(3):e1001611-e11. doi: 10.1371/journal.pmed.1001611
13. Rufo CJ, Madureira J, Fernandes OE, et al. Volatile organic compounds in asthma diagnosis: a systematic review and meta-analysis Corrigendum (vol 71, 175, 2016). *Allergy* 2016;71(5):733-33.
14. Nurmatov UB, Tagiyeva N, Semple S, et al. Volatile organic compounds and risk of asthma and allergy: a systematic review. *European respiratory review : an official journal of the European Respiratory Society* 2015;24(135):92-101. doi: 10.1183/09059180.00000714
15. Bousquet J, O'Hehir RE, Anto JM, et al. Assessment of thunderstorm-induced asthma using Google Trends. *Journal of Allergy and Clinical Immunology* 2017;140(3):891-93.e7. doi: https://doi.org/10.1016/j.jaci.2017.04.042

16. Bush A, Kleinert S, Pavord ID. The asthmas in 2015 and beyond: a Lancet Commission. *The Lancet* 2015;385(9975):1273-75.
17. Pinnock H. Supported self-management for asthma. *Breathe* 2015;11(2):98-109. doi: 10.1183/20734735.015614
18. Jerrett M, Arain A, Kanaroglou P, et al. A review and evaluation of intraurban air pollution exposure models. *J Expo Anal Environ Epidemiol* 2004;15(2):185-204.
19. Reis EJ, Guzmán RM. An econometric model of Amazon deforestation. 2015
20. Willocks LJ, Bhaskar A, Ramsay CN, et al. Cardiovascular disease and air pollution in Scotland: no association or insufficient data and study design? *BMC Public Health* 2012;12(1):1-6. doi: 10.1186/1471-2458-12-227
21. Dibben C, Clemens T. Place of work and residential exposure to ambient air pollution and birth outcomes in Scotland, using geographically fine pollution climate mapping estimates. *Environmental Research* 2015;140:535-41. doi: http://dx.doi.org/10.1016/j.envres.2015.05.010
22. Butland BK, Armstrong B, Atkinson RW, et al. Measurement error in time-series analysis: a simulation study comparing modelled and monitored data. *BMC Medical Research Methodology* 2013;13(1):136. doi: 10.1186/1471-2288-13-136
23. Strickland MJ, Klein M, Darrow LA, et al. The issue of confounding in epidemiological studies of ambient air pollution and pregnancy outcomes. *Journal of Epidemiology and Community Health* 2009;63(6):500-04. doi: 10.1136/jech.2008.080499
24. Woodruff TJ, Parker JD, Darrow LA, et al. Methodological issues in studies of air pollution and reproductive health. *Environmental Research* 2009;109(3):311-20. doi: http://dx.doi.org/10.1016/j.envres.2008.12.012
25. Clemens T, Dibben C. A method for estimating wage, using standardised occupational classifications, for use in medical research in the place of self-reported income. *BMC Medical Research Methodology* 2014;14(1):1-8. doi: 10.1186/1471-2288-14-59

Figures

Figure 1: Data linkage framework in secured environment

(See attached)

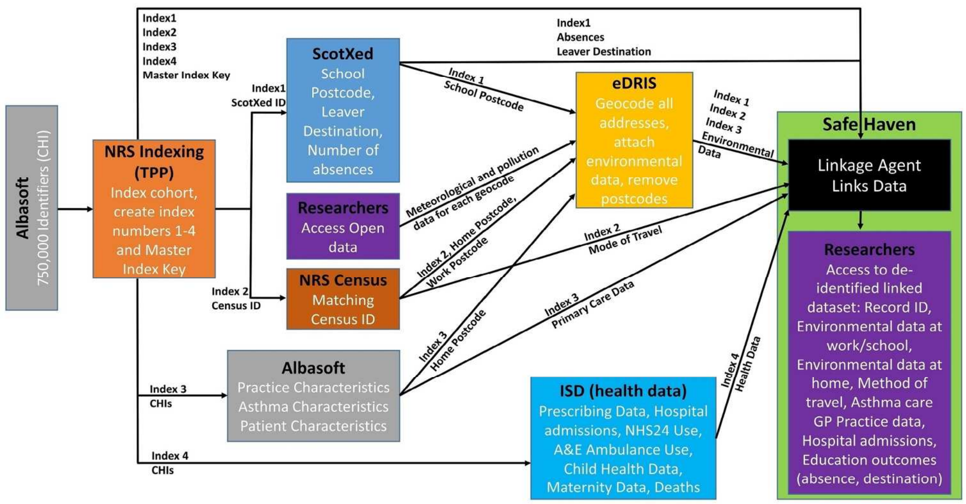

Figure 1

90x90mm (600 x 600 DPI)