Queen Mary University of London

Barts and The London School of Medicine and Dentistry

Blizard Institute

Centre for Immunobiology & School of Mathematical Sciences

Developing a database and geostatistical methods to study the epidemiology of PIBD

Author:
Polychronis Kemos

Supervisors:
Professor Nick Croft
Dr Silvia Liverani

A thesis submitted for the partial fulfilment of the degree of
Doctor of Philosophy

1st March 2023
London, UK

Statement of Originality

I, Polychronis Kemos, confirm that the research included within this thesis is my own work or that where it has been carried out in collaboration with, or supported by others, this is duly acknowledged below, and my contribution is indicated. Previously published material is also acknowledged below.

I attest that I have exercised reasonable care to ensure that the work is original and does not, to the best of my knowledge, break any UK law, infringe any third party's copyright or other Intellectual Property Rights, or contain any confidential material.

I accept that the College has the right to use plagiarism detection software to check the electronic version of the thesis.

I confirm that this thesis has not been previously submitted for the award of a degree by this or any other university.

The copyright of this thesis rests with the author, and no quotation from it or information derived from it may be published without the author's prior written consent.

Signature: Polychronis Kemos

Date: **1st March 2023**

List of Abbreviations

(P)IBD: (Paediatric) Inflammatory Bowel Disease

AIC: Akaike Information Criterion

API: Application Programming Interface

AQSR: Air Quality Standards Regulations

CD: Crohn's disease

CMV: Cytomegalovirus

CRF: Case Report Form

DEFRA: Department for Environment, Food and Rural Affairs (UK)

EBI: Empirical Bayesian Interpolation

EBK: Empirical Bayesian Kriging

EDC: Electronic Data Capturing system

EEA: European Environmental Agency (EU)

E-PRTR: European Pollutant Release and Transfer Register

EPSG: European Petroleum Survey Group

GCS: Geographic Coordinate System

GIS: Geographic Information System

GISCO: Geographic Information System of the COmmission (EU)

GWR: Geographically Weighted Regression

IBD: Inflammatory Bowel Disease

IBD-U: Inflammatory Bowel Disease that is Unclassified (traditionally grouped with UC)

IDW: Inverse Distance Weighting

INSPIRE: INfrastructure for SPatial InfoRmation in the European community

IPR: Incidence-to-Prevalence Ratio

KNN: K-Nearest Neighbours

LAU: Local Administrative Units

LMM: Linear Mixed-effects Models

MAUP: Modifiable Areal Unit Problem

MERRA-2: Modern-Era Retrospective analysis for Research and Applications, version 2

NAEI: National Atmospheric Emissions Inventory (UK)

NASA: National Aeronautics and Space Administration (US)

NUTS: Nomenclature of territorial units for statistics (EU)

NSAIDs: Non-Steroidal Anti-Inflammatory Drugs

OK: Ordinal Kriging

PIBD: Paediatric Inflammatory Bowel Disease

PM: Particulate Matter

POD: Phytotoxic Ozone Dose

POWER: Prediction Of Worldwide Energy Resources

RBF: Radial Basis Function

REDCap: Research Electronic Data Capture

RLH: Royal London Hospital

SAR: Spatial Lag model

SEM: Spatial Error Model

SLX: Spatially Lagged X model

SK: Simple Kriging

SSE: Surface meteorology and Solar Energy

TLI: Triangulation with Linear Interpolation

TOMPS: Toxic Organic Micro-Pollutants

UC: Ulcerative Colitis

UK-AIR: United Kingdom Air Information Resource

VOC: Volatile Organic Compounds

WHO: World Health Organisation

Abstract

Inflammatory bowel diseases (IBD) are chronic idiopathic disorders that cause inflammation of the gastrointestinal tract. These chronic conditions affect over 2.5 million people in Europe alone, with a direct healthcare cost of €4.6-5.6 billion annually (Jairath, V., & Feagan, B. G., 2020). While the IBD burden has increased rapidly in recent decades in most Western countries, this rate of increase seems to be slowing down. Conversely, newly industrialised regions are currently seeing a spike in incidence rates. To date, paediatric inflammatory bowel disease (PIBD) incidence and prevalence information can be found in just over 100 publications that cite figures from 35 different countries reporting significant spatial and temporal variations with a few clear patterns. The current literature often combines mostly heterogeneous information from studies using various methods with limited spatiotemporal coverage. Despite over 200 genetic loci being linked to IBD, only 10-15% of the disease risk is attributed to genetic factors (Shouval, D. S., & Rufo, P. A., 2017). Therefore, we hypothesised that certain external risk factors might be linked to developing PIBD, similar to other chronic diseases, such as rheumatoid arthritis and asthma. In the paediatric population specifically, given the limited exposures and co-morbidities compared to the adult population, environmental and socio-economic factors may play a particularly significant role in the development of the disease.

In this prospective study, I developed several methods for collecting, processing and analysing PIBD incidence and prevalence data. The main six aims of the project were the following:

i. Develop and establish methods to collect international PIBD data uniformly
ii. Organise and prepare these data
iii. Develop the project-specific analytical methods
iv. Estimate the incidence and prevalence of the disease
v. Collect and prepare the risk factors data from multiple international datasets
vi. Study the effects of these factors on the disease epidemiology

The methods involved the development of two separate databases that were created to collect clinical and epidemiological data, refining the calculation of disease incidence by adjusting for different reporting centres, and compiling a novel dataset of suspected risk factors for the disease while focusing on interpolated pollutants.

After designing an automated Electronic Data Capturing system (EDC), I conducted five annual data collection rounds, gathering over 500 responses from 140 PIBD experts worldwide. The estimation and mapping of the incidence and prevalence revealed significant spatiotemporal trends with recent years and latitudes closer to the north presenting a significant positive correlation with the incidence. The risk factors were extracted from validated sources and used as predictors after being processed and transformed using various computationally intensive methods. The processing and harmonisation of the predictor data, which involved handling misaligned information formats and diverse types of territorial units, was one of the most challenging aspects of this project. The final geostatistical analysis of the PIBD incidence revealed that the most significant findings were the particulate matter (PM10) and carbon monoxide, followed by carbon dioxide and chlorine with inorganic compounds (HCl). The latter pollutant is a novel finding as it has not been reported previously in the literature. Further geostatistical analysis of the phenotype ratio of the disease, Crohn's and Ulcerative colitis, revealed the importance of sun exposure, population density and three pollutants from the broader family of volatile organic compounds in the development of the disease. The PIBD phenotype ratio analysis was repeated based on the environmental questionnaire of the inception cohort clinical study and showed that the patient demographics, the preferred source of water, and surprisingly, contact with certain animals were also potentially important predictors of the type of diagnosis.

Our incidence and prevalence findings are consistent with most of the partially established spatiotemporal patterns of PIBD in the literature. This provides validation of the methods used and strengthens our findings. The statistical methods developed in this project were based on several spatial interpolation models, disease mapping techniques, Linear Mixed-effects Models (LMM), spatial regression, model fit diagnostics and the development of study-specific functions used to adjust the reported data. The spatial interpolation methodology was optimised with a series of simulations, while the LMM and spatial regression were employed to address the autocorrelation in our data.

The following chapters provide detailed insights into the methods and results of this thesis followed by a discussion of our findings.

Project structure, ongoing studies and useful information

This work is part of the PIBD-SETQuality project, a Horizon 2020 multinational project funded by the European Commission as part of the Research and Innovation action scheme (Grant agreement ID: 668023). The PIBD-SETQuality supports two registries, the Inception Cohort and the Safety Registry.

The Inception Cohort is a prospective observational study recruiting children with newly diagnosed inflammatory bowel disease since early 2017. The maximum number of recruitment partners reached 21 sites in the United Kingdom, Italy, Germany, France, Israel, Malaysia, Japan, South Korea and the Netherlands. Besides clinical data and regular follow-ups, the Inception Cohort collects detailed exposure information from each newly diagnosed patient using several Electronic Case Report Forms (eCRFs) and questionnaires. Examples of the environmental questionnaires collecting the information used in the analysis of the phenotype ratio of the disease are provided in the appendix. For each subject, this includes geographical information, socio-economic data, previous medications and vaccinations, as well as dietary and lifestyle information. This study collects a large set of environmental exposures from the time of birth to the time of diagnosis. Thus far, the Inception Cohort has recruited 771 patients, with 376 to be based in the United Kingdom (UK).

The Safety Registry is a separate entity and an entirely epidemiological database initially designed to estimate the incidence and prevalence rates of rare and severe complications in the PIBD population. However, we expanded this project to ensure it also collects PIBD incidence and prevalence information from the general population, providing insights into the spatiotemporal distribution of PIBD. These data are unique as we have estimated incidence and prevalence rates across multiple countries in parallel using identical collection and validation methodologies and the most current data based on country (or centre)-specific databases and registries when available. Currently, the project has exceeded 140 active PIBD experts and is collecting information from areas covering over 30 million individuals under the age of 19. The Safety Registry provides the core datasets used to estimate the incidence and prevalence of PIBD in this PhD.

It is important to clarify that the inception cohort is a phase IV clinical study that offers individual patient data that we used to analyse the risk factors that affect the disease phenotype. In contrast, the safety registry is an epidemiological study that collects aggregated data, over specific patient catchment areas as discussed in the methods. In 2.1.2, the specific forms and data collection methodologies are discussed in further detail. Finally, the data from both the clinical and epidemiological studies were combined with several additional datasets of risk factors which were extracted from the European Environment Agency, NASA and EUROSTAT to prepare the datasets for the final analyses.

Work undertaken by the PhD candidate

The synopsis of my role in this project includes the database design and setup of the data management methodologies, the support of the ongoing studies, the collection of additional external data, the development of analytical methods and finally, the data analysis itself.

At the beginning of this project, I was responsible for designing the 2 study databases with the support and supervision of Professor Nick Croft and our collaborators from Erasmus MC University in the Netherlands, the associate Professor Lissy de Ridder and Dr Martine Aardoom. The finalised electronic data capturing (EDC) systems included 9,000 fields for the Safety Registry and 4,016 for the Inception Cohort. Much work was dedicated to developing system automations on the EDC server based on R Application Programming Interfaces (API). These automations aimed to streamline data quality checks and enable real-time interaction between the system, study participants, and data entry staff, thereby reducing errors and queries.

I was also involved in the study support, developing the Safety Registry study network, and adding new participants. Beyond the technical requirements of these tasks, I received the help and support of Professor Nick Croft and the team in the Netherlands, including Dr Renz Klomberg. Based on the clinical input of the group, I was also responsible for updating the system regularly in terms of the structure, content and the eCRFs.

The external data collection was also a significant volume of the work in this project, as the predictor data were available in various sources and formats. The complexity of this step increased after several of the datasets that the initial analytical methods were based on were updated, changed, or even discontinued by Eurostat. Exporting the data was also challenging in some cases due to the large volume of the required files. Frequently, this task required the development of APIs or several repetitive steps when the APIs were not an available option.

The data preparation and management were likely one of the project's most challenging and time-consuming aspects. I have applied multiple interpolation methods on more than a hundred external datasets that had to be interpolated as part of the change of spatial format to match our outcome data. Also, the spatial information from different sources was often based on incompatible geodetic formats, meaning that I had to reproject the data which was often a complex task for specific grid-based geodetic systems.

Very importantly, with Dr Silvia Liverani's guidance, I have developed a series of steps that combine several methods, allowing us to calculate and validate the disease incidence and prevalence in the studied areas and combine it with many investigated predictors. These steps were later programmed as functions in R, leading to the development of an R-based set of functions that receives patient data with geographical information in various formats as input and performs spatial analysis using the processed risk factors.

The model testing and optimisation was a multi-step process during which I received support from Silvia Liverani in multiple instances. Again, with the support and help of Silvia Liverani and Nick Croft, I have carried out all the analyses in this project, including the generation of tables and figures. In addition, my supervisors provided guidance for interpreting the results which set the foundation for the discussion.

Finally, during my PhD, I supported the team with the data management and statistical analysis required for several posters and publications, listed in the following section. The publication *"Incidence and Characteristics of Venous Thromboembolisms in Paediatric-Onset Inflammatory Bowel Disease"* (Aardoom et al., 2022), where I was responsible for the data management and statistical analysis, is a very relevant example to the Safety Registry.

Details of relevant publications

Peer-reviewed publications related to the Inception Cohort and Safety Registry studies:

Aardoom, M. A., Klomberg, R. C. W., **Kemos, P.**, Ruemmele, F. M., van Ommen, H., de Ridder, L., & Croft, N. M. (2021). The incidence and characteristics of venous thromboembolisms in paediatric-onset inflammatory bowel disease; a prospective international cohort study based on the PIBD-SETQuality Safety Registry.. *J Crohns Colitis*. doi:10.1093/ecco-jcc/jjab171

Aardoom, M. A., **Kemos, P.**, Tindemans, I., Aloi, M., Koletzko, S., Levine, A., . . . de Ridder, L. (2020). International prospective observational study investigating the disease course and heterogeneity of paediatric-onset inflammatory bowel disease: the protocol of the PIBD-SETQuality Inception Cohort study. *BMJ open*, *10*(7), e035538. doi:10.1136/bmjopen-2019-035538

Joosse, M. E., Aardoom, M. A., **Kemos, P.**, Turner, D., Wilson, D. C., Koletzko, S., . . . de Ridder, L. (2018). Malignancy and mortality in paediatric-onset inflammatory bowel disease: a 3-year prospective, multinational study from the paediatric IBD Porto group of ESPGHAN. *Alimentary Pharmacology and Therapeutics*, *48*(5), 523-537. doi:10.1111/apt.14893

Klomberg, R., Kaper, D., Barendregt, D., Vuijk, S., Schreurs, M., Charrout, M., . . . de Ridder, L. (2023). P243 The association between serological markers ASCA and ANCA and response to induction treatment in therapy-naïve children with inflammatory bowel disease: an international prospective study. *Journal of Crohn's and Colitis*, *17*(Supplement_1), i395-i396. doi:10.1093/ecco-jcc/jjac190.0373

Klomberg, R. C. W., Aardoom, M. A., **Kemos, P.**, Rizopoulos, D., Ruemmele, F. M., Croft, N. M., . . . Walker, M. (2022). High Impact of Pediatric Inflammatory Bowel Disease on Caregivers' Work Productivity and Daily Activities: An International Prospective Study. *Journal of Pediatrics*, *246*, 95-102.e4. doi:10.1016/j.jpeds.2022.04.014

Posters related to the Inception Cohort and Safety Registry studies:

Kemos, P., Aardoom, M., Ruemmele, F., de Ridder, L., & Croft, N. (2018). P785 Designing the first pan-European paediatric IBD database to allow the study of incidence and prevalence of UC, CD, and their rare and severe complications. In *Journal of Crohn's and Colitis* Vol. 12 (pp. s508). doi:10.1093/ecco-jcc/jjx180.912

Aardoom, M., **Kemos, P.**, Ruemmele, F., van Ommen, C. H., Croft, N. M., & de Ridder, L. (2020). Mo1900 THE INCIDENCE AND CHARACTERISTICS OF VENOUS THROMBOEMBOLISMS IN PEDIATRIC-ONSET INFLAMMATORY BOWEL DISEASE. In *Gastroenterology* Vol. 158 (pp. s-970). doi:10.1016/s0016-5085(20)33097-3

Aardoom, M., **Kemos, P.**, Ruemmele, F. M., van Ommen, C. H., Croft, N. M., & de Ridder, L. (2020). P116 The occurrence of venous thromboembolisms in paediatric-onset IBD. In *Journal of Crohn's and Colitis* Vol. 14 (pp. s194). doi:10.1093/ecco-jcc/jjz203.245

Aardoom, M. A., **Kemos, P.**, Ruemmele, F., Croft, N., & de Ridder, L. (2019). P596 An ongoing Safety Registry to identify rare and severe complications in children with paediatric-onset IBD. In *Journal of Crohn's and Colitis* Vol. 13 (pp. s413). doi:10.1093/ecco-jcc/jjy222.720

Aardoom, M. A., **Kemos, P.**, Ruemmele, F., Tindemans, I., Samsom, J. N., Croft, N., & de Ridder, L. (2019). P165 A global prospective observational study in children and adolescents with paediatric-onset IBD: the PIBD-SETQuality Inception Cohort. In *Journal of Crohn's and Colitis* Vol. 13 (pp. s171-s172). doi:10.1093/ecco-jcc/jjy222.289

Aardoom, M., **Kemos, P.**, Ruemmele, F., Croft, N., & de Ridder, L. (2018). P584 Rare and severe complications in children with paediatric-onset IBD; the international PIBD SETQuality Safety Registry by PIBDnet. In *Journal of Crohn's and Colitis* Vol. 12 (pp. s403-s404). doi:10.1093/ecco-jcc/jjx180.711

Costes, L., Tindemans, I., Joosse, M. E., Aardoom, M. A., Jongsma, M. M. E., Raatgeep, R. H. C., . . . Samsom, J. N. (2020). DOP09 IgG responses to multiple bacterial flagellins identify a subgroup of paediatric therapy-naive Crohn’s disease patients with increased microbiota-specific T-cell reactivity. In *Journal of Crohn's and Colitis* Vol. 14 (pp. s046). doi:10.1093/ecco-jcc/jjz203.048

Heredia, M., Charrout, M., Klomberg, R. C. W., Aardoom, M. A., Jongsma, M. M. E., **Kemos, P.**, . . . Samsom, J. N. (2022). P061 Circulating inflammatory protein and cellular profiles at time of diagnosis classify inflammatory bowel disease patients according to their underlying immune response and clinical disease course. In *Journal of Crohn's and Colitis* Vol. 16 (pp. i171). doi:10.1093/ecco-jcc/jjab232.190

Heredia, M., Costes, L. M. M., Tindemans, I., Aardoom, M. A., Klomberg, R. C. W., **Kemos, P.**, . . . Samsom, J. N. (2022). P045 TIGIT expression differentiates regulatory from inflammatory Th1* gut-homing effector CD4+ T cells in inflammatory bowel disease patients. In *Journal of Crohn's and Colitis* Vol. 16 (pp. i161). doi:10.1093/ecco-jcc/jjab232.174

Heredia, M., Costes, L. M. M., Tindemans, I., Aardoom, M. A., Klomberg, R. C. W., **Kemos, P.**, . . . Samsom, J. N. (2021). TIGIT expression differentiates regulatory from inflammatory non-classical Th1-like gut-homing effector CD4+T cells in inflammatory bowel disease patients. In *EUROPEAN JOURNAL OF IMMUNOLOGY* Vol. 51 (pp. 29). Retrieved from https://www.webofscience.com/api/gateway?GWVersion=2&SrcApp=PARTNER_APP&SrcAuth=LinksAMR&KeyUT=WOS:000753366400063&DestLinkType=FullRecord&DestApp=ALL_WOS&UsrCustomerID=612ae0d773dcbdba3046f6df545e9f6a

Heredia, M., Charrout, M., Klomberg, R. C. W., Aardoom, M. A., Jongsma, M. M. E., **Kemos, P.**, . . . Samsom, J. N. (2021). Circulating inflammatory protein and cellular profiles at time of diagnosis classify inflammatory bowel disease patients according to their underlying immune response and clinical disease course. In *EUROPEAN JOURNAL OF IMMUNOLOGY* Vol. 51 (pp. 275). Retrieved from https://www.webofscience.com/api/gateway?GWVersion=2&SrcApp=PARTNER_APP&SrcAuth=LinksAMR&KeyUT=WOS:000753366401348&DestLinkType=FullRecord&DestApp=ALL_WOS&UsrCustomerID=612ae0d773dcbdba3046f6df545e9f6a

Heredia, M., Charrout, M., Klomberg, R., Aardoom, M., Jongsma, M., **Kemos, P.**, . . . Samsom, J. (2021). P021 Circulating inflammatory protein and cellular profiles at time of diagnosis classify inflammatory bowel disease patients according to their underlying immune response and clinical disease course. In *Journal of Crohn's and Colitis* Vol. 15 (pp. s139). doi:10.1093/ecco-jcc/jjab076.150

Heredia, M., Costes, L., Tindemans, I., Aardoom, M., Klomberg, R., **Kemos, P.**, . . . Samsom, J. (2021). P012 Low frequencies of circulating inhibitory TIGIT+CD38+ effector T cells identify an immunologically distinct subgroup of pediatric patients with severe Crohn’s disease. In *Journal of Crohn's and Colitis* Vol. 15 (pp. s134). doi:10.1093/ecco-jcc/jjab076.141

Klomberg, R., **Kemos, P.**, Chondrou, P., Croft, N., & de Ridder, L. (2022). P678 Rare and severe complications in paediatric Inflammatory Bowel Disease: results of 5 years of international collaboration through the PIBD-SETQuality Safety Registry. In *Journal of Crohn's and Colitis* Vol. 16 (pp. i582). doi:10.1093/ecco-jcc/jjab232.799

Klomberg, R., van der Wal, H., Sarbagili, C., **Kemos, P.**, Ruemelle, F., Croft, N., . . . Levine, A. (2022). P146 Initial response to induction treatment more important than type of induction treatment for achieving sustained steroid free remission at 1 year in new-onset paediatric Crohn’s disease. In *Journal of Crohn's and Colitis* Vol. 16 (pp. i226). doi:10.1093/ecco-jcc/jjab232.274

Klomberg, R., **Kemos, P.**, Chondrou, P., Croft, N., & de Ridder, L. (2022). Rare and severe complications in paediatric Inflammatory Bowel Disease: results of 5 years of international collaboration through the PIBD-SETQuality Safety Registry. In *JOURNAL OF CROHNS & COLITIS* Vol. 16 (pp. I582). Retrieved from https://www.webofscience.com/api/gateway?GWVersion=2&SrcApp=PARTNER_APP&SrcAuth=LinksAMR&KeyUT=WOS:000778573400800&DestLinkType=FullRecord&DestApp=ALL_WOS&UsrCustomerID=612ae0d773dcbdba3046f6df545e9f6a

Klomberg, R., van der Wal, H., Sarbagili, C., **Kemos, P.**, Ruemelle, F., Croft, N., . . . Consortium, P. I. B. D. -S. E. T. Q. (2022). Initial response to induction treatment more important than type of induction treatment for achieving sustained steroid free remission at 1 year in new-onset paediatric Crohn's disease. In *JOURNAL OF CROHNS & COLITIS* Vol. 16 (pp. I226). Retrieved from https://www.webofscience.com/api/gateway?GWVersion=2&SrcApp=PARTNER_APP&SrcAuth=LinksAMR&KeyUT=WOS:000778573400274&DestLinkType=FullRecord&DestApp=ALL_WOS&UsrCustomerID=612ae0d773dcbdba3046f6df545e9f6a

Klomberg, R., **Kemos, P.**, Aardoom, M., Rizopoulos, D., Croft, N., de Ridder, L., & Neyt, M. (2021). P114 Frequent work productivity and activity impairment in caregivers of children with inflammatory bowel disease: a project of the prospective PIBD-SETQuality Inception Cohort study. In *Journal of Crohn's and Colitis* Vol. 15 (pp. s206-s207). doi:10.1093/ecco-jcc/jjab076.241

Klomberg, R., Vuijk, S., **Kemos, P.**, Aardoom, M., Ruemmele, F., Mulder, J., . . . De Ridder, L. (2021). Su525 THE INCIDENCE AND CAUSES OF RENAL FAILURE IN PAEDIATRIC INFLAMMATORY BOWEL DISEASE - A PROJECT OF THE PIBD-SETQ SAFETY REGISTRY. In *Gastroenterology* Vol. 160 (pp. s-726-s-727). doi:10.1016/s0016-5085(21)02447-1

Klomberg, R., Aardoom, M., **Kemos, P.**, Ruemmele, F., Van Ommen, C. H., Croft, N. M., & De Ridder, L. (2021). 730 INCREASED RISK OF VENOUS THROMBOEMBOLISM IN PAEDIATRIC-ONSET INFLAMMATORY BOWEL DISEASE - A STUDY OF THE INTERNATIONAL PROSPECTIVE SAFETY REGISTRY. In *Gastroenterology* Vol. 160 (pp. s-146). doi:10.1016/s0016-5085(21)01103-3

Rao, A., **Kemos, P.**, Kamperidis, N., Naik, S., Croft, N. M., & Sanderson, I. (2019). Su1797 – Children with Crohn's Disease Who Respond to an Enteral Diet are Twice As Likely to Remain in Clinical Remission Over 5 Years. In *Gastroenterology* Vol. 156 (pp. s-615-s-616). doi:10.1016/s0016-5085(19)38434-3

Thorn, N., Vallorani, M., Kirupananthan, R., **Kemos, P.**, & Croft, N. (2021). P20 Emergenci: a UK prospective survey of severe GI bleeding (requiring upper GI endoscopy) and emergency endoscopy in under 16s. In *Frontline Gastroenterology* Vol. 12 (pp. a22-a23). doi:10.1136/flgastro-2021-bspghan.30

Vavilikolanu, R., Sanderson, I., Naik, S., Croft, N. M., & **Kemos, P.** (2019). G118(P) The risk of nephrotoxicity from 5-ASA in children with ulcerative colitis (UC) is not dose related. In *Archives of Disease in Childhood* Vol. 104 (pp. a48). doi:10.1136/archdischild-2019-rcpch.114

Acknowledgements

I want to take this opportunity to express my gratitude to all those who have contributed to my PhD. Your guidance, support, and encouragement have been invaluable throughout this journey.

Firstly, I would like to thank my primary supervisor, Nick Croft, for his unwavering support and invaluable guidance throughout my research. Nick’s expertise and insight greatly impacted my work and progress. I would also like to thank my second supervisor, Silvia Liverani, for her valuable input in my methodology and continuous support.

I would also like to extend my sincere appreciation to my examiners. I look forward to receiving constructive criticism and suggestions to improve the quality of my research and writing.

I also want to thank my colleagues for their support and input during my PhD. Their enthusiasm, encouragement, and friendly competition have kept me motivated and focused.

Very importantly, I would like to express my heartfelt gratitude to my family for their unwavering love and support.

I would also like to thank my friends for their constant encouragement and support, which has been a major help for me. Kate Waller, a friend and former colleague, was extremely supportive with proofreading and formatting.

Finally, a big thank you to Evija for her unparalleled support, help and kindness.

Once again, I would like to thank you all for your invaluable support, encouragement, and guidance. I could not have achieved this without your help and expertise. Thank you!

Sincerely, Chronis

Thesis Structure

Table of Contents

List of Figures

List of Tables

1. INTRODUCTION

1.1. PIBD: General Notes

Paediatric-onset Inflammatory Bowel Disease (PIBD) refers to the manifestation of Inflammatory Bowel Disease (IBD) occurring in individuals before the age of 18. It includes the conditions of Crohn's disease (CD), ulcerative colitis (UC) and IBD-Unclassified (IBD-U). When it is not possible to be certain whether the phenotype is consistent with Crohn's or UC, it is more likely to be UC in most cases. The distinction of Paediatric IBD is crucial because the disease's onset in children and adolescents often presents with different clinical features, disease distribution, and genetic predispositions compared to adult-onset IBD. Variation in PIBD can be considerable due to several factors: genetic variations influencing susceptibility and disease presentation, environmental factors such as diet, exposure to infections, and use of antibiotics, which may impact disease onset and progression. Moreover, the clinical presentation of PIBD can vary significantly; some patients may have mild symptoms and localised disease, while others experience severe, extensive, and often more aggressive disease activity. This variability poses challenges for diagnosis, treatment, and management, underscoring the need for a tailored and multidisciplinary approach to care for paediatric patients with IBD.

This body of work is focused on the epidemiology of IBD in the paediatric population. The incidence of PIBD has risen dramatically in recent decades (Elis et al., 2012; Schwarz et al., 2017), with the prevalence in some countries expected to reach 1% of the population by 2030 (Kaplan and Windsor, 2021). In this thesis, PIBD incidence refers to the number of new disease cases that develop in a specific paediatric population during the study period. The latter is the population at risk and is presented as person-years. In contrast, the PIBD prevalence is the total number of individuals who have the disease in the paediatric population at a specific time or over a specified time, regardless of when they first developed the condition. Currently, there is solid evidence that countries with increasing trends of urbanisation and industrialisation present the highest increase in IBD incidence. Moreover, fully "westernised" countries have the highest but relatively plateaued IBD incidence (Kaplan and Windsor, 2021). Despite some known suspected risk factors, the exact cause of IBD/PIBD remains unknown.

Thus far, a large-scale epidemiological study that investigates the potential causes, examining a large number of possible predictors, is not available. Several projects have attempted to study the epidemiology of PIBD and identify causal factors. These studies, however, are arguably limited in their geographical and population coverage. At the same time, in most systematic reviews and meta-analyses, the datasets were collected at different times and under different methodologies, which poses a significant limitation in the statistical analysis due to the inhomogeneity of the final dataset. Therefore, developing our PIBD epidemiological knowledge is crucial for evaluating existing and new etiological hypotheses to better define how environmental and demographic factors might influence the onset of the disease.

1.2. PIBD: Current views on the incidence

To date, PIBD incidence information has been reported in just over 100 publications including figures from 35 different countries (Benchimol et al., 2011a; Hong et al., 2018; Huang and Aw, 2020; Lopez et al., 2018; Schwarz et al., 2017; Sýkora et al., 2018). However, comprehensive incidence data from multiple studies and with adequate regional coverage are available only for a few countries. The current literature suggests that Canada, Germany, Slovenia, Sweden, Finland, Germany and the United Kingdom are countries with better characterised PIBD rates, compared to most other countries (Ashton et al., 2014; Benchimol et al., 2011a, 2009; Grieci and Bütter, 2009; Henderson et al., 2012; Lehtinen et al., 2016; Ludvigsson et al., 2017; Malmborg et al., 2013; Urlep et al., 2015, 2014; Wittig et al., 2019). A review of the currently available publications and systematic reviews reveals great spatial and temporal variations, yet with a few clear patterns that I will discuss in the following paragraphs. These studies were retrospective and exploratory in their majority.

1.2.1. Temporal trends and data sources

In their majority (>80%), the discussed epidemiological studies provided substantial evidence for an increasing PIBD incidence within each country. Finland, Sweden, Scotland, the United Kingdom, France, the Republic of Ireland, Spain, Greece, Italy, Denmark, Slovenia, several states and provinces in the US and Canada, Czech Republic, Singapore, Saudi Arabia, South Korea and Singapore were some of the countries which reported a continuously increasing incidence over time for an approximate period spanning from 1985 to 2015 (Benchimol et al., 2014; Bequet et al., 2017; Castro et al., 2008; Coughlan et al., 2017; Dimakou et al., 2012; el Mouzan et al., 2014; Fernández et al., 2015; Hammer et al., 2016; Henderson et al., 2012; Hong et al., 2018; Hope et al., 2012; Jakobsen et al., 2011; Lehtinen et al., 2011; Malmborg et al., 2013; Martín-de-Carpi et al., 2014, 2013; Ong et al., 2018; Schwarz et al., 2017; Sýkora et al., 2018; Urlep et al., 2015, 2014; Virta et al., 2016a). Therefore, the temporal trends of PIBD incidence are well-characterised and consistent among several countries across different continents. In addition to the cited regions in West Asia, Europe and North America, the observed trends seem to be global, as increasing rates for both CD and UC have been reported from Australia and New Zealand in 2014 (Day et al., 2014), repeatedly from South America in 2015 to 2019 (Selvaratnam et al., 2019), across Asia and East Asia in particular in 2008 (Thia et al., 2008) and Africa in 2020 (Hodges and Kelly, 2020) with the latter to be the least characterised region in terms of PIBD and IBD incidence and prevalence.

1.2.2. Spatial trends and North-South gradient

Expanding our current understanding of the temporal PIBD trends to the geographical distribution of the disease is a more complicated task due to the lack of standardised practices employed by different PIBD studies and therefore their subsequent heterogeneity. These projects vary greatly in their data collection and analysis methods, completion date, diagnostic criteria, exclusion criteria and, in some cases, even clinical practices, depending on the country and clinical setting. In regards to the data collection methodologies used in previous studies of PIBD incidence, four main categories emerge i) the use of data from insurance companies, which we have seen recently in studies from Germany, Canada and South Korea, ii) the use of databases for medication expenditure claims, which are mostly used in the Baltic countries and in Finnish studies, iii) access of patient hospital records, which is common in most countries, iv) the use of dedicated IBD and PIBD registries and surveys completed by an active network of PIBD and IBD specialists. The latter data collection approach is used in this project and has also been used with success in several other countries previously. For example, one of the biggest survey-based PIBD registries was the SPIRIT registry (1996-2009), which revealed the incidence of the disease in Spain for the first time (Martín-de-Carpi et al., 2014). Additional concerns when reviewing the spatial distribution of the disease across different regions and studies are emerging as a result of inconsistencies with age cut-offs used in different studies. Most commonly, <18 years of age was the preferred cut-off in approximately half of the reviewed studies, while <19 and <16 were also used as cut-offs by several research groups. Lastly, to further underline the complexity of summarising the geographical trends of PIBD incidence, different incidence standardisation methods, including in some cases, none, have been used by different studies.

Despite all the obstacles mentioned above that inflate the uncertainty in any comparisons between regions, admittedly, two main geographical patterns emerge from the available literature. For the northern hemisphere, countries and regions closer to northern latitudes are reported to have up to 10-fold higher PIBD incidence rates compared to the countries located further to the south. It is imperative to underline that any incidence comparisons between different regions should occur only for similar periods due to the rise of PIBD incidence in the last 40 years. Focusing on Europe and for the 2000-2010 time period, in 2008, Scotland reported a PIBD incidence of 7.82 per 10^5 person-years (Henderson et al., 2012), which was the highest at the time for the United Kingdom, whilst Finland reported 15 new cases per 10^5 (Lehtinen et al., 2011). Germany also reported similar PIBD incidence rates of 13.65 in 2009 (Wittig et al., 2019), whilst Sweden reported 12.8 in 2007 for the county of Stockholm (Malmborg et al., 2013). In addition, Denmark also reported a high PIBD incidence in 2009 (Jakobsen et al., 2011). Although the reported incidence from Denmark was 6.4 per 10^5 person-years, the included population was <15 years of age and therefore, the expected incidence in the overall paediatric population is significantly higher. The Netherlands reported a PIBD incidence of 5.2 from 1999 to 2001 (van der Zaag-Loonen et al., 2004). However, considering the reporting year, this figure should be adjusted and increased to match the results from the other countries, given the well-recognised continuous increase in the annual disease incidence observed in all European countries. After grouping the higher latitude countries in Europe (above 53^{o}N) with available PIBD incidence data in the 2000 to 2010 period, Poland was the only outlier, while the paediatric incidence of the disease is unclear in the Baltic countries. For the 2002 to 2004 period, the reported incidence in Poland was 2.7 cases per 10^5 person-years, which, however, was still higher than any available report from the south of Europe for the same time period (Karolewska-Bochenek et al., 2009). This discrepancy may be explained by the size of this country as it spans from 49 to 53^{o}N. It is possible that countries of that size may have a within-latitude variation in the incidence of PIBD. In contrast to the figures from northern Europe, all countries with available PIBD incidence data in the south of Europe reported significantly lower rates. Italy reported a PIBD incidence of 1.39 in 2003 (Castro et al., 2008), and Spain reported 2.8 in 2009. In Greece, the average reported PIBD incidence for the 2000 to 2011 period is estimated at 1.9 patients per 10^5 person-year (Dimakou et al., 2012). This estimate was based on denominator data for the reporting unit that is known to us. Lastly, as expected, countries in between reported incidence figures that were significantly higher compared to the South but lower than the average reported incidence from the countries of the North. Specifically, Slovenia reported 5.14 for the 2000 to 2005 period (Orel et al., 2009), northern France reported 4.4 to 9.5 for the 1998 to 2011 period (Bequet et al., 2017), and the Czech Republic reported 3.8 in 2002 (Jabandziev et al., 2020). As discussed, these countries also reported a sharp increase in the incidence over time.

The main reason for our primary focus on the northern hemisphere is that approximately 90% of the global population is located above the equator. This fact, combined with the similar latitudes of the populated Australian regions and countries of New Zealand and South America (regions with available PIBD incidence data), poses difficulties in using data from these areas to investigate possible latitude effects on the PIBD incidence. Future studies in South America may confirm an equivalent pattern in the southern hemisphere as well.

1.2.3. Outliers and differences between East and West

Although the rapid incidence increase over time and the latitude effect are well-recognised patterns, some countries and regions, particularly in Eastern Europe and outside North America, appear to follow a different pattern. Eastern European countries are a good example of this deviation, while even at the northern latitudes, they present lower rates compared to the expected rates based on their location. Interestingly, these countries also present the steepest increase in their incidence rates. Hungary reported 0.7 PIBD incidence in 1981, subsequently increasing by a staggering 1770% increase in the following three decades (Lovasz et al., 2014). Although the reports for the Baltic countries are scarce, they appear to follow a similar pattern with the rest of Eastern Europe. From 1993 to 1998, for all age groups, only 29 patients were diagnosed in Estonia, suggesting that the annual incidence of PIBD was practically non-existent at the time and certainly much lower than the European average (Salupere, 2001). A more recent study in Lithuania, again not specific to paediatrics, reported an IBD incidence of 8.12, suggesting that the PIBD rates should not exceed 2.5 cases per 10^5 in the PAED population (Schwarz et al., 2). Similarly, the incidence rates in Romania and Russia are also particularly low and mainly driven by UC cases instead of CD (Goldiș et al., 2019; Khalif and Shapina, 2017). The current consensus is that this pattern is linked to a less ‘westernised’ lifestyle in these countries, and I will discuss the suspected risk factors in PIBD in the following sections. Lastly, outside Europe and North America, the PIBD incidence also follows the ‘westernisation hypothesis’. An example is the paediatric incidence in Australia and New Zealand which is significantly higher compared to Asian and Latin American countries (Ng et al., 2017). Interestingly, several studies supporting the latter also reported a sudden incidence increase, especially in Asian countries (Huang and Aw, 2020).

1.2.4. Spatio-Temporal trend plateau

An additional element that needs to be considered when describing the global trends of PIBD incidence is a 'stopping rule' for the continuously increasing disease rates. In several countries that have been traditionally reporting very high PIBD incidence rates, a plateau and in some cases, even a minor decrease has been observed in recent years (Huang and Aw, 2020). PIBD incidence reports from Slovenia suggest a possible decline in the country’s incidence rates (Urlep et al., 2015), while in most provinces in Canada, a plateau and, in some cases, even a decline has been observed over the last decade (Kaplan et al., 2019). It is noteworthy that Canada has been the country with the highest IBD incidence in the world over the last decade (Qin, 2011). Therefore, it should not be a surprise that it is also one of the first countries to report stable or declining incidence rates. Furthermore, similar trends have been reported in the US, with Wisconsin as a representative example (Adamiak et al., 2013), where the PIBD incidence has been stable at 9.5 new cases per 10^5 person-years for almost a whole decade.

1.2.5. Trends in IBD phenotypes

A thorough study of PIBD epidemiology should also factor in the different phenotypes of the disease. As discussed in 1.1, IBD includes CD, UC and, less frequently, IBD-U, which in clinical practice is often grouped with UC due to the similar location of the presenting symptoms. In a very simplistic analogy, one can picture this as an epidemiological study of the Influenza virus ('the flu'). Depending on several factors, in some seasons, the flu spreads by type B and in other seasons is spread by type A (Zhang, 2015). The CD and UC diagnoses are, in fact, two very distinct IBD phenotypes with major differences in their presentation, pathogenic mechanisms and therapeutic approaches. Very importantly, differences may also be found in some of the risk factors for each phenotype as several studies have previously suggested. For instance, smoking has been reported to be a protective factor for UC, while it is deemed a risk factor for CD (Carbonnel et al., 2009). According to the literature, it is widely accepted that CD is the most prevalent IBD and PIBD phenotype. In our prospective PIBD clinical study, the Inception Cohort (ClinicalTrials.gov identifier: NCT03571373), the CD, UC and IBD-U rates are 58%, 33% and 9%, respectively. Most reports from North America, Asia and Europe are in agreement with these ratios. However, in eastern European countries and Finland, the UC incidence and, subsequently, prevalence are higher than the CD. Specifically, in Finland, the paediatric UC incidence has been repeatedly reported to be much higher compared to the CD incidence (Lehtinen et al., 2011; Virta et al., 2016b), while in Russia, Romania and Estonia, a similar rate, favourable to UC has also been reported by the majority of available studies from these regions (Goldiş et al., 2019; Khalif and Shapina, 2017; Salupere, 2001). Older studies in Greece reported a very low CD incidence, however, this trend has been changing for this region in the last two decades (Archimandritis et al., 2002; Economou et al., 2007; Tsianos et al., 1994).

1.2.6. PIBD and Demographics

The demographics of PIBD are well-established and reported in a great number of studies (Johnston and Logan, 2008; Ashton et al., 2014; Fernández et al., 2015; Ludvigsson et al., 2017; Urlep et al., 2015; Forss et al., 2022; Kern et al., 2022; Kaplan et al., 2019). Although the patterns vary between countries, several studies have reported that both the incidence of IBD and the ratio of the diagnoses between males and females change with age (Ludvigsson et al., 2017; Urlep et al., 2015). Several studies have reported that in the young paediatric population, the CD phenotype is more common in males, while the UC phenotype is prevalent in females. However, as the paediatric population enters the later stages of life, this pattern tends to exhibit reduced prominence; in some instances, it may even reverse for the adult population (Johnston and Logan, 2008). The impact of age on the incidence of PIBD and IBD overall is substantial. Several studies have reported a marked rise in disease occurrence, at times resembling an exponential function, from the initial years of life up to the age of 18. This pattern has been observed in Germany, the United Kingdom, Sweden, Spain and Canada (Ashton et al., 2014; Fernández et al., 2015; Forss et al., 2022; Kern et al., 2022; Kaplan et al., 2019).

Understanding the disease distribution by age is an essential step for the precise incidence calculation, which is subsequently necessary for the study of the epidemiology of the disease. This also applies to the geographical and temporal PIBD incidence characteristics. Understanding the distribution of the disease over time and space is central to the knowledge of its development, spread, and dynamics.

1.3. PIBD: Risk factors

Inflammatory Bowel Disease is known to have a genetic component, with research suggesting that genetic factors contribute to a person's susceptibility to the condition. Although studies have identified genetic variants that are associated with IBD, the precise mechanisms by which these genetic factors contribute to IBD development are not yet fully understood. It is important to note that while genetics plays a role, IBD is a complex condition with various contributing factors, including environmental and lifestyle effects. Only a very small percentage of the disease cases, including primarily early onset cases, is reported as a monogenic disease (Loddo and Romano, 2015). Studies recruiting homozygotic twins reported a 20-50% concordance for CD and percentages as low as 10% for U confirming the significant role of the environmental effects on the disease development (Halme, 2006; Quigley, 2012). These studies also identified several genes with different expression levels between twins, which are associated with different previous environmental exposures. Such differences between individuals with identical DNA provide significant evidence underlying the importance of the environmental effects on the incidence of the disease.

1.3.1. Paediatric Population Epidemiology

This project focuses on the collection of paediatric incidence and prevalence of IBD data. Although obtaining paediatric data often presents limitations, it also has several advantages, especially regarding risk factor identification. The paediatric population tends to have limited exposure to substances and medications that might influence the results (for example, smoking, alcohol, blood pressure medication and others). Moreover, the paediatric population has significantly lower rates of comorbidities that often influence the incidence and progression of the studied disease. Lastly, from a geospatial perspective, individuals in a paediatric population are also expected to have fewer relocations prior to presenting the disease, and therefore, their environmental exposures can be followed up with much less effort compared to adults. However, the disadvantages of a paediatric population in an epidemiological study are the comparatively limited sample size and the risk of missing cases that might have been diagnosed in adult clinical settings instead of paediatric units.

1.3.2. Risk factors in IBD, a critical approach to previous studies

Several studies investigating the effects of risk factors in IBD and PIBD can be found in the current literature. However, there are limitations in the previously published research on this topic which may have an impact on the validity of their results. Therefore, we should use strict filters on the currently available list of suspected risk factors before choosing the predictors that will be used to analyse our data. In the following paragraphs, I will provide a few examples of such studies that I have reviewed to optimise the methods used in this project.

In 2005 a Belgian study reported a significant association between the month of birth and Crohn's disease incidence (Joossens et al., 2005). Specifically, a higher risk was reported for a four-month season peaking in April and August. In particular, June was reported as the month with the most potent protective effect. However, this study did not have pre-defined hypotheses and endpoints and also did not adjust for multiple comparisons. This study suggests that any combination of months (seasons) could be a different level of the predictor, while the individual months were also included as individual predictors. Assuming that the investigators defined a season as four months, four seasonal combinations are available, while 12 single months are also available predictor options. This inflates the type 1 error, and the chances for a false positive increase from 5% to 56%, invalidating the p-value cut-off of the study. Moreover, the four months period seems to be inconsistent as half of the individual months in each peaking period had an odds ratio in favour of the opposite effect. It should be noted that our Inception Cohort data do not confirm these findings while a year earlier, a similar Israeli study with the same limitations suggested a peak in the winter season (Chowers et al., 2004).

A French study in 2011 identified lower rates of sun exposure as a predictor of a higher Crohn's Disease prevalence rate (Nerich et al., 2011). Although this might be an accurate finding, since many other studies have also suggested it, the observed correlation was directly linked to the geographic latitude. As with many other countries, the latitude in France is frequently linked to several socio-economic factors and health indicators that might be important confounding factors. For instance, in the book *Health in France 2002*, an almost identical pattern was reported for the health inequalities and disparities and the standardised mortality ratios within the country. Arguably, this finding could be linked to health inequalities instead of sun exposure.

A last example to challenge a negative result in contrast to the previous examples is a Canadian study that found no associations between measles vaccinations and the onset of paediatric IBD (Shaw et al., 2015). The study included patients and controls to a 1:7.13 ratio, with 97% and 94% vaccination rates, respectively. For this recruitment ratio, the study is severely underpowered and could only detect a significant difference if the vaccination rate in the control group was lower than 92%. This percentage seems unrealistic, making the design of the study problematic.

Additional issues with the current literature are related to the generation of subgroups that reduce the validity of the results from a statistical standpoint. It is quite common in IBD research to separate the IBD sample into its two main phenotypes, UC and CD, while less frequently, further divisions based on age groups are applied. The great number of subgroups, in combination with the number of questions and examined risk factors, can often lead to false-positive findings if the necessary adjustments are not made. Following these rules is often challenging in practice and for "real-world-evidence" studies. However, due to the lack of a specific protocol and pre-defined hypotheses, it is crucial that the basic statistical rules are followed in order to maintain the global type I error within a reasonable range (Chen et al., 2017).

1.3.3. Defining environmental exposure and risk factors

In the last decade, a new concept that describes the interaction of environmental exposures with an individual's health emerged in the field of epidemiology. The exposome is a term used to describe the complementary role of the environment to the genome (DeBord et al., 2016). A literature search reveals that this notion, although not new in its entirety, is gaining popularity exponentially as it formally defines the totality of exposures that individuals experience from conception until death and its impact on chronic diseases. Hence, the definition of exposome can be used as a framework for summarising the risk factors for this study. This step will enable us to classify all possible exposures and determine which domains of the exposome we can investigate and where this project's limitations lie regarding accounting for all the potentially important stressors. Simply put, I will attempt to identify all possible routes in which non-genetic exposures may contribute to the IBD onset and I will specify the ones that we can account for in the data analysis.

The exposome is divided into three domains, the internal, the specific external and the general external (DeBord et al., 2016; Vrijheid, 2014). The internal factors are unique to the individual, including metabolic factors, gut microflora, inflammation, oxidative stress, and more (Wild, 2012). This is the type of exposure we cannot include in our analysis as such an approach would require us to collect biomarkers from a vast healthy population starting from birth until a large enough sample of patients emerges within the monitored population. Specific external factors include contaminants, diet, physical activity, tobacco use, infections, and lifestyle factors. This domain is covered by the Inception Cohort as we collect a very wide range of specific external factors from each newly diagnosed patient. Considering that we also collect detailed geographical information from each patient, we can also estimate the exposure of each individual to a large number of additional environmental factors including pollutants, sun irradiation and more. The latter are the general external exposures including outdoor pollutants, urban environment information, climate information and overall factors that are usually measured at the community and regional levels. Additional examples of the latter are the health and lifestyle indicators in a specific area, such as diabetes and average sugar consumption. Both the specific and general external factors can be powerful predictors in our analysis. The advantage of using specific factors is preciseness, while in the use of general external factors, the advantage lies in the much larger sample size. Nevertheless, it is possible to combine the findings using both approaches and validate our results.

1.3.4. Risk factors in IBD, current knowledge

1.3.4.1. Specific external factors

In the myriad of suspected and studied risk factors in IBD, as our starting reference point, we will use the comprehensive topical review on the Environmental Factors and Predisposition to IBD published by the European Crohn’s and Colitis Organisation (ECCO) in 2016 (Maaser et al., 2016). This report includes several suspected risk factors in IBD, frequently separated for the phenotypes of CD and UC. In the following paragraphs, the focus will be on the specific external factors as described by the exposome definition.

Several studies have considered prenatal and perinatal factors as risk factors in IBD in the last two decades. In particular, caesarean section has been reported as a risk factor, but several studies also contradict this finding. ECCO associated this possible effect with the intestinal microbiota composition encountered at birth. A recent study expanded this to the hygiene hypothesis, which will also be discussed as a specific factor (Beaugerie et al., 2018). Given the contradicting results and considering that the rate of caesarean section deliveries depends on the region and the local medical practices, we will only include this factor in the exploratory outcomes. Breastfeeding has also been studied previously, again, with contradicting results from several studies. According to more recent studies, the results might have been contradicting because the duration response effect means that at least 3 or 6 months (depending on the study) are required to reduce the risk (Gearry et al., 2010). Since the duration of breastfeeding is not collected by the Inception Cohort, we will not include this factor in the exploratory outcomes. Additional factors include the gestation period, birth weight, birth length, as well as infections, age and smoking status of the mother during pregnancy. Of these factors, smoking is a well-known factor of importance in IBD and therefore the maternal smoking status will be included in the primary analysis.

Childhood vaccinations have been studied repeatedly in the IBD population; to date, there is no solid evidence to demonstrate an association. Some reports, such as Burisch et al., 2014b, suggest a difference in the vaccination rates for Diphtheria and Polio between the two different IBD phenotypes. However, the dataset was obtained from Eastern and Western Europe, and this effect could be attributed to the fact that the phenotype and vaccination rates are quite different between the two regions, suggesting an expected correlation and not causation (Burisch et al., 2014b). In fact, comparing the vaccination rates between UC and CD patients in Eastern and Western Europe shows no significant differences. Therefore, the vaccination data from the Inception Cohort will also be included in the exploratory outcomes.

The **hygiene hypothesis** is, according to many scientists, one of the most critical factors in the development of IBD (Gearry and Dodgshun, 2012; Koloski et al., 2008). According to this hypothesis, childhood exposure to very hygienic conditions may be linked to the incidence of the disease. Extremely hygienic environments could impair immune development and predispose the population to immunological diseases. This notion fits almost perfectly with the timeline between a country's development status -which, as expected, is linked to sanitation improvements- and its IBD rates. Interestingly, countries that are classified as "developed" by the conventional metrics have very high IBD rates, while developing countries present low but rapidly increasing IBD rates. Nevertheless, we cannot exclude the possibility that other factors linked to a region's development status may be the fundamental driving forces of the IBD incidence. Adding this theory as a specific external factor in our study is somewhat complicated as it is a composite variable that includes many factors. For this reason, it is also challenging to quantify this measure as an arbitrary scoring system would introduce bias. However, since this is a well-recognised factor that may affect the immune system, we will be using proxy factors as suggested by the literature to investigate its influence on PIBD incidence as one of the primary outcomes (Saidel-Odes and Odes, 2014). The proxy factors that we will use include information about exposure to pets, other animals, and the number of siblings and household details.

Dietary habits are also an important specific external exposure. There is a plethora of studies on this topic, with many of them suggesting that very high fat and sugar intake could increase IBD incidence (Maaser et al., 2016). Studies using animal models also suggested similar results. Capturing such patterns is a particularly complicated task, and therefore, in the Inception Cohort, we are collecting limited information about any types of food that our patients exclude from their diet. This simplified factor will be included as a potential risk and examine if the patients exclude anything in particular from the diets (i.e., vegetarian). Expanding on dietary habits, caffeine and alcohol consumption will not be included in the analysis as our population's expected exposure rates are particularly low.

Supplements (Vitamin D), antibiotics and pain relief medications (NSAIDs and Aspirin) are also important in the study of IBD risk. Vitamin D deficiency has been reported multiple times as a risk factor for IBD. Low vitamin D levels in newly diagnosed IBD patients are frequently reported, while a recent study published late in 2018 identified 79% of all recently diagnosed IBD patients as vitamin D insufficient or deficient (Chetcuti Zammit et al., 2018). The same study also presented findings suggesting a correlation between the severity of the symptoms and Vitamin D. Additionally, this is consistent with the multiple reports that suggest a negative correlation between sun exposure and IBD. These reports also fit the geographical gradients of the disease from south to north. This will be further discussed in the general external factors paragraphs.

Antibiotics also seem to be positively associated with the onset of IBD. In 2011, a Canadian study with a substantial sample size of 24,.580 subjects and a 1:10 case-control design reported a significant association between new IBD cases and antibiotic administration (Shaw et al., 2011). Recent use of Antibiotics may trigger IBD, and it is also an indicator of a recent infection that could also be the trigger of IBD. More studies and meta-analyses suggest this connection, especially for the CD phenotype (Ungaro et al., 2014; Virta et al., 2012). However, antibiotics may also be given to the patients to treat the symptoms prior to diagnosis, making this a difficult predictor to study.

Aspirin and NSAIDs (Non-Steroidal Anti-Inflammatory Drugs), in particular, have also been linked to new IBD cases. In recent years, retrospective and prospective studies have reported this finding with exceptionally high odds ratios of 1.87, 2.96 and 6.2 indicating a large effect size (Chan et al.,2011; Felder et al., 2000; Gleeson and Davis, 2003). Aspirin and NSAIDs (Non-steroidal anti-inflammatory drugs), in particular, have also been linked to new IBD cases. Considering the substantial evidence supporting the association of vitamin D, antibiotics, NSAIDs and Aspirin with new cases of IBD, we will also include these factors in the primary outcome analysis. Arguably, the pain medication and antibiotics may not be directly linked to the aetiology of the disease and may be proxy measures of an underlying condition associated with the onset of IBD.

Additional factors are also considered as specific external factors in IBD. As discussed, smoking and alcohol do not apply to our study due to the low exposure rates in the paediatric population. Similarly, recreational drug use, oral contraceptives, hormone replacement therapy and occupation will also be excluded. Lastly, although stress, anxiety and depression are included in the Inception Cohort data collection, the information is gathered at diagnosis, meaning that the presenting symptoms will affect the results. These items will also be excluded since we can only obtain this information after diagnosis.

1.3.4.2. General external factors

Unlike the specific external factors, which include the exposures of each individual in our analysis, the general external factors include the exposures over an area and, subsequently, a group of individuals. Any underlying associations between the disease and the studied exposures are expected to be observed across different areas, establishing a pattern. Some of the already discussed factors can also be studied on this higher level.

Prenatal and perinatal factors are not consistently documented in detail in all European countries. However, the mother's age at birth is available on the regional level in Europe and will be included in the primary analysis complementing the prenatal analysis since this information is not available in the Inception Cohort.

Solar irradiance is expected to be a significant predictor in our analysis, and it can also be used as a proxy for vitamin D levels. Since sun exposure is the most recognised risk factor, we can use it as a validation metric in our project with the expectation that we will also be able to replicate these findings (Fletcher et al., 2019; Ghaly et al., 2019; Holmes et al., 2019, 2018; Limketkai et al., 2014; Lu et al., 2015; Nerich et al., 2011; Olmedo-Martín et al., 2019).

Pollutants are a less-studied group of general external risk factors in IBD, with the first available study published in 2010. However, almost all relevant studies have identified significant associations between air pollutants and the incidence of IBD (Ananthakrishnan et al., 2011; Beamish et al., 2011; de Silva et al., 2017; Opstelten et al., 2016; Salim et al., 2014). Our study offers an up-and-coming platform for studying pollutants and their connection with PIBD. The paediatric population is expected to have fewer relocations, allowing us to estimate their exposure level to each pollutant more accurately. This characteristic of the paediatric population could explain why some studies have observed these effects specifically in the young population (Kaplan et al., 2010). Also, our population is expected to have limited exposure to other chemicals, known for their toxicity, compared to adults. Lastly, the data collection in the Safety Registry covers a sizeable paediatric population from several European countries. As part of the EU legislation, each member country must provide detailed information about every pollution source from factories, businesses, and other organisations, while sampling stations also provide hourly measurements for 50,000 locations across Europe.

This results in a database with multiple million entries including all sources of pollutants in the last few years combined with the source details, duration, validation status and more. For these reasons, this project will thoroughly investigate the link between pollution and PIBD.

Due to the pollution datasets' size and level of detail, we can obtain a very accurate picture of their spatiotemporal distribution for the European region. However, since approximately 500 pollutants are available in the European Environmental Agency (EEA) databases, we must select the specific compounds that will be included in our analysis. According to the WHO, the pollutants with the most substantial evidence of health effects are particulate matter (PM), ozone (O_3), nitrogen dioxide (NO_2) and sulphur dioxide (SO_2) (Brook et al., 2010; Goshua et al., 2022). According to the United Kingdom Air Information Resource (UK-AIR and the British Department for the Environment, Food and Rural Affairs, toxic organic micro-pollutants (TOMPS), Benzene, 1,3-Butadiene, Carbon monoxide (CO), as well as Lead (Pb) and other heavy metals, should also be included in the list. The US Environmental Protection Agency also includes Carbon monoxide and Lead in this list. Compared to every previous study of IBD/PIBD, the novelty in this project is that it incorporates the most detailed pollution dataset, which we wish to utilise and validate previous findings regarding the effects of heavy metals, O_3, CO, NO2 and PM on the gut (Beamish et al., 2011) and PIBD specifically. We have also investigated the influence of Butadiene, Volatile Organic Compounds (VOC), toxic organic micro-pollutants and other pollutants on disease incidence for the first time. The additional pollutants that will be included in our analysis will be selected based on their coverage. Only pollutants that are present in at least 10% of the studied area will be added to our predictor dataset.

Socioeconomics and **demographics** have always been factors of interest in real-world data studies. Although associations between socio-economics and IBD have been suggested by many investigators (Bernstein et al., 2001; Blanchard, 2001; Farrokhyar et al., 2001; Piovani et al., 2019), these are expected to be spurious relationships rather than causal. For instance, unemployment cannot cause disease but could be associated with the prevalence of disease since it affects the lifestyle, quality of life, access to certain services and more. Although we will include the demographics and socioeconomic factors in our analysis, we will be critical of any relevant results as health disparities that might be present (Benchimol et al., 2011b).

Geographical location is a profound factor closely linked to IBD and PIBD incidence, as discussed in the previous paragraphs. However, this information is not useful in the aetiological study of the disease. In essence, a geographical region cannot be a risk factor but only an indicator of the true risk factors that may be present in the region or a common characteristic in the group of individuals who live at this location. One of the metrics used in spatial analysis aims at eliminating any spatial correlations in the residuals of the final model, meaning that the model can explain the difference in the sample using the predictors that have a causal relationship with the disease, which is detailed in the methods section. The same rationale applies to the urbanisation status, which has been repeatedly suggested as a risk factor by numerous studies and comprehensive meta-analyses (Soon et al., 2012).

Lastly, the **water quality and source** are a less straightforward general external risk factor previously documented in the literature since 1990 (Aamodt et al., 2008; Hermon-Taylor, 1993; van Kruiningen and Freda, 2001). This is included under the geographical location since it has been proposed that the evidence of clusters may suggest a shared exposure for certain groups that share the same water supply. Using the Inception Cohort data, we can identify phenotype patient clusters, and while using the Safety Registry data and pollution data from EEA, we can investigate potential associations on the regional level.

1.4. Maps and Exposures: Territorial units, maps and predictor data sources

This section is included in the introduction to discuss the source, details and characteristics of the datasets and variables included in the study.

1.4.1. Eurostat (NUTS, GISCO and INSPIRE datasets)

Eurostat is the statistical office of the EU. It is responsible for collecting, processing, and disseminating statistical data for the EU and its member states. Eurostat's main tasks include collecting and compiling statistical data from EU member states. Eurostat works with national statistical institutes to ensure that data is collected and reported consistently and comparably across the EU. Eurostat also processes data using statistical methods and produces various statistical publications and databases. Disseminating data and information is an essential function of the agency. It ensures that the data and publications become available to the public through its website and other channels, including data visualisations and interactive tools. Eurostat's data is widely used by researchers, policymakers, and the general public, and it is considered a reference source of information for the EU. The statistical authorities of Switzerland and the European Economic Area are also included in the available datasets. During the design of this project, the United Kingdom was still included in the Eurostat summaries.

The maps we used throughout the project for data collection, disease and pollution mapping and analysis were provided by GISCO (the Geographic Information System of the COmmission). The primary responsibility of GISCO is to offer geospatial reference data and related services to Eurostat, the Commission, and the general public of Europe. Its objective is to encourage and stimulate the use of geographic information within the European Statistical System and the Commission. It also organises Commission-wide geographic information operations and shared policies. The GISCO maps provided administrative boundaries with their statistical units in the NUTS format (Nomenclature of territorial units for statistics). This Nomenclature uses four levels in a hierarchical structure, meaning that the smaller territories are included in the larger territories (**Table 1**). The first level is NUTS0 which is the country level, followed by NUTS1 which includes major socioeconomic regions (e.g.. Scotland). The NUTS2 level is assigned to basic regions for the application of regional policies (i.e., Oxfordshire), and it includes the NUTS3 level which is the most detailed level of the NUTS classification and was the territorial unit that was used in our project for data collection and analysis. Specifically, for the UK, the Office for National Statistics has integrated the NUTS framework, providing UK-specific datasets and areas of higher resolution. The United Kingdom was one of the countries that also developed NUTS4 and NUTS5, also known as LAU1 and LAU2, respectively. These Local Administrative Units (LAU) are essential for future work that may require using maps with higher resolution and precision. The following Table 1 summarises several areas for the NUTS0 to NUTS5 levels for the United Kingdom.

Table 1 The EUROSTAT territorial units in the UK

The six levels of NUTS units that were available up to May 2026 for the United Kingdom

Level (UK)	Number of territories/Administrative units	Description
NUTS0	1	Country level
NUTS1	12	Population range: 3-7 million
NUTS2	40	Population range: 0.8-3 million
NUTS3	173	Population range: 0.15-0.8 million
LAU1 (NUTS4)	415	Population range: 1,000-2.2 thousand
LAU2 (NUTS5)	10126	Population range: 0-36.4 thousand

The NUTS maps are available to download by Eurostat in 5 different formats, 5 versions from 2001 to 2019 in 3 geometry types, several resolution levels and 3 different coordinate reference systems. The maps used in this project were released in 2013 and 2016 and are based on the shapefile format using the polygon geometry type and 3 different coordinate reference systems (EPSG: 3035, EPSG: 4326 and EPSG: 3857).

An additional valuable source of spatial information for European countries is Infrastructure for Spatial Information in the European Community (INSPIRE). An EU legislation known as INSPIRE sought to improve public sector organisations' exchange of environmental spatial information and public access to environmental data throughout Europe. The pan-government equivalent of INSPIRE in the UK is called UK Location, and it aims to enhance the sharing and reuse of location data in the public sector. Although the INSPIRE and UK Location datasets were initially used in this project, they were excluded from the final analysis due to the great heterogeneity regarding the available formats, spatial units and areas covered. Depending on the findings of our study, these datasets may be necessary for a more detailed review of the effects that specific exposures may have on the incidence of PIBD.

1.4.2. European Environmental Agency (EEA) and Department for Environment, Food and Rural Affairs (DEFRA) datasets

The European Environment Agency (EEA) is tasked with disseminating reliable, unbiased environmental information from 41 European or adjacent to Europe countries, providing information to decision-makers and the general public. The EEA provides several million data points with information on over 400 pollutants, detailed location information, type of measurements, verification information and additional information, including the time and method of measurement. The datasets include pollutants found and released in the air, water and soil while distinguishing pollutants with local effects from long-range transboundary pollutants. Much of the data included in the EEA databases are available due to the existence of articles requiring organisations and businesses to declare and submit their pollutant release and transfer to relevant European registries. This results in a comprehensive list of databases that include sampled pollution data, emission data from individual industrial plants and emission data from several diffuse sources.

The agency also provides a variety of complementary datasets that may be important for specific ecological and epidemiological studies. For instance, EEA provides maps with the spatial distribution of *Anopheles maculipennis*, a mosquito species responsible for malaria transmission in European countries until 1970. In addition, the environmental agency provides a series of additional maps with health indices summaries, climate, transport, agriculture, waste, and land cover data.

The Department for Environment, Food and Rural Affairs (DEFRA) offers access to various environmental spatial data in the UK. The United Kingdom's air quality regulations require that the UK undertakes air quality assessments on an annual basis under the Air Quality Standards Regulations 2010 (AQSR). Therefore, several datasets become available annually, including essential pollutants such as particulate matter, Benzene, Nitrous Oxides, Sulphur Oxides, Ozone and others. Several of these datasets are publicly available and have been pre-processed and derived using interpolation methods (in some cases also based on scaling of the measurements) by UK-AIR, which is hosted and maintained by Ricardo Energy and Environment on behalf of DEFRA.

1.4.3. National Aeronautics and Space Administration (NASA)

The Prediction of Worldwide Energy Resources (POWER) project provides solar and meteorological data sets from National Aeronautics and Space Administration (NASA) research to support renewable energy, building energy efficiency and agricultural needs. They are supported by NASA Earth Science's Applied Sciences Program.

One of the first initiatives the Applied Science Program supported to encourage the use of NASA's data assets was the Surface Meteorology and Solar Energy (SSE) project. When first launched in 1997, the SSE data-delivery website was designed to simplify retrieving the data sets. The surface insolation measurements are derived from satellite observations. The Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA-2) assimilation model serves as the foundation for the meteorological parameters.

Inception Cohort

In addition to the large-scale epidemiological approach in the Safety Registry study, we also analysed data collected prospectively from the Inception Cohort PIBD study, which provided patient-level information and more detailed exposome information. An Inception Cohort is a study design in which a group of patients diagnosed with a specific disease or condition is enrolled at the time of diagnosis and followed over time. The goal of an Inception Cohort study is to understand the natural history of the disease, identify risk factors for progression or complications, and evaluate the effectiveness of different treatment options. Therefore, during the PIBD Inception Cohort, we collected detailed information on the patient level, covering a comprehensive set of questions from the environmental exposures prior to diagnosis to the clinical follow-up data.

1.5. Managing spatial information: Common practices and challenges

Disease mapping and spatial data analysis are processes that include several steps and require multiple techniques and methodologies. The following and final sections of the introduction will provide information on the types of spatial format and geodetic reference systems that are relevant to our project, followed by an expansion on the problem of spatial misalignment.

1.5.1. Types of spatial information and geodetic reference

Spatial data can be available in different formats and geometry types. In this project, the processed data were available as point data or point data grid, lattice format or a polygon grid, and continuous surfaces known as rasters. The spatial point data are coordinates with longitude and latitude information combined with the quantity of the measured variable at each point.

Often, point data may be organised in a grid, meaning that all data points are placed at an equal distance from each other. Most environmental datasets from EEA are available as point data, while the NASA meteorological data can be obtained via APIs in a point data grid format. Datasets based on the lattice format include observations over several spatial polygon regions supplemented by a neighbourhood structure. Each unique polygon represents a distinct territory, and they may be either regular and equally spaced following a grid format or irregular. The latter format is often based on the administrative boundaries of each area. The NUTS3 Eurostat maps are based on an irregular lattice format. Lastly, the raster maps are images where each pixel represents an area with an assigned value for the variable of interest. Raster maps with a high resolution present the spatial information as a continuous surface.

Spatial information is linked to a geodetic reference system which is also known as the geodetic reference datum. This allows geodetic coordinates to accurately represent a location's position on the map using a specific reference frame. There are two types of datum, horizontal and vertical. Although it may not seem intuitive, the horizontal datum is used to determine the latitude or longitude (or different values depending on the coordinate system used). Conversely, the vertical datum is used to determine the elevation of a point, usually from the sea level (often, alternative methods are used as the sea elevation may vary depending on the time and location). In this project, we have encountered a great variety of geodetic reference systems, which we have converted to the Geographic Coordinate System (GCS), as discussed in the methods section. The GCSs are spherical and, most frequently, ellipsoidal coordinate systems using latitude and longitude. The GCS specifications are listed in the EPSG Geodetic Parameter Dataset, named after the European Petroleum Survey Group (EPSG) that was responsible for creating the registry. Some of the most common EPSG codes are EPSG:4326, EPSG:3035, EPSG:3857 and EPSG:7789. Any Eurostat shapefile NUTS map is currently available in the first three coordinate reference systems. The EPSG:3857, also known as Web Mercator or WGS 84/Pseudo-Mercator projection and is now available on google maps and other online platforms [**Figure 1**]. EPSG:3857 is based on the Mercator projection, which is widely used worldwide, although this projection overestimates the areas closer to very high and very low latitudes.

Due to the limitations of the rather popular EPSG:3857, the GCS of choice in this project is EPSG:4326, also known as WGS 84. EPSG:3857 uses a coordinate system projected from the surface of the sphere, and in previous versions, it treated the Earth as a perfect sphere. However, EPSG: 4326 (WGS 84) uses a coordinate system on the surface of a sphere or an ellipsoid of reference. In a simplistic summary, the Mercator projection uses a coordinate system based on a flat surface projection. In contrast, the WGS 84 uses a reference system based on a curved surface, a globe. As shown in **Figure 1**, the coordinate origin of WGS 84 is meant to be located at the Earth's centre of mass with an uncertainty of fewer than 5 centimetres. The longitude is the angle between the Prime Meridian (longitude of 0 degrees) and the point of interest. Any locations west of the Prime Meridian have a negative angle, and areas to the east have positive degrees. The maximum and minimum longitudes meet and overlap in the Pacific Ocean and are 180^{o} and -180^{o} respectively. The latitude angle starts at the equator (latitude of 0 degrees) and receives positive and negative values for the northern and southern hemispheres. Areas in the Northern Hemisphere can have latitudes ranging from 0^{o} to 90^{o}, while those in the Southern Hemisphere can range from 0^{o} to -90^{o}.

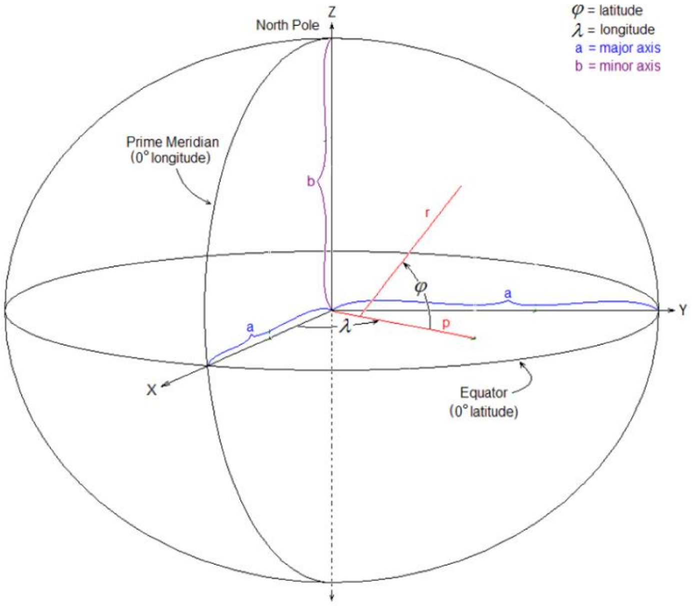


Figure 1. The WGS 84 coordinates system.

Because Earth is an imperfect ellipsoid, area-specific datums are often considered more accurate for specific areas of coverage that they have been designed for compared to WGS 84. For instance, the DEFRA datasets use the Ordnance Survey National Grid reference system, a Britain-specific reference grid. Instead of longitude and latitude, this system uses easting and northing values. These values are based on the distance between each point on the grid. This means that by estimating the distance between two points using the WGS 84 projection and having a commonly known point on the map, we can reproject the UK-specific grid code to WGS 84 coordinates. This was one of the necessary geodetic conversions in our project. Given the longitude and latitude of two locations, the haversine formula calculates their great-circle distance. The law of haversines, a more general formula in spherical trigonometry that connects the sides and angles of spherical triangles:

$$D(x, y) = 2 \arcsin \left[\sqrt{\sin^2 \left(\frac{x1 - y1}{2} \right) + \cos(x1) \cos(y1) \sin^2 \left(\frac{x2 - y2}{2} \right)} \right] \quad (1)$$

The use of the globally standardised WGS 84 system is prevalent in many epidemiological studies, and it was also the preferred choice for our project.

1.5.2. Spatial autocorrelation

Spatial autocorrelation is a measure of the association that observations may have depending on their proximity to each other. Most test statistics rely on the assumption of independent observations, which is one of the key reasons why understanding and measuring spatial autocorrelation is crucial. The assumption that observations are independent of one another is violated if autocorrelation is present in datasets with geographical information. Spatial autocorrelation, when present, can be negative or positive. A hypothetical instance of negative autocorrelation could be illustrated by observing a prosperous region that reduces the likelihood of its neighbouring regions becoming wealthy, as competition between the regions restricts overall economic growth. An example of positive autocorrelation would be observing wealthy regions, due to collaboration, to increase the probability of their neighbouring regions to also get wealthy. This information is essential for the mapping, and the analysis of the study outcome, as a major part of the observed variance might be explained by autocorrelation instead of the analysed predictors. A widespread measure of autocorrelation is Moran's index, a test that considers both the location and measured values of the spatial observations. For a measured outcome, the test investigates whether observations that are closer to each other tend to have similar, higher, or lower values compared to the mean. In order to perform the test, the criteria for the observations considered to be neighbours need to be defined. This is possible by assigning weights for each region or data point. The matrix of weights defines the neighbouring relationship that each region or data point has with any other region or data point in our sample. The formula for the Moran's I statistic is:

$$I = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\left(\sum_{i=1}^{n}\sum_{i=1}^{n} w_{ij}\right)\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad (2)$$

- where I is the Moran's I statistic,
- x_i is the value of a variable at location i,
- x_j is the value of a variable at location j
- while $w_{ij}$is the weight that determines the spatial relationship between i and j.

As we can see in (2), the numerator depends on the difference between point i from the mean and point j and the mean multiplied by the corresponding weight that we have assigned in the combination of points i and j. The denominator allows us to standardise the value I. The null hypothesis of this test is that the outcome is randomly disbursed geographically; said another way, under the null hypothesis of no spatial autocorrelation, the observations are independent identically distributed, while I is normally distributed with a mean equal to:

$$\mathrm{E}[I] = -\frac{1}{N-1}$$

The Moran's I statistic was used repeatedly in our analysis and in section 2.4.1 the spatial weights are discussed in further detail.

1.6. The challenge of spatial information misalignment

Spatial information misalignment refers to the problem of mismatch or inconsistency between spatial datasets that are used in a particular analysis or application. This can occur when data sets have different spatial resolutions, projections, or coordinate systems. Misalignment can also occur when data sets have different temporal resolutions or are collected at different times. Frequently, this occurs due to the use of different sampling methods, which ultimately leads to having multiple spatial scales for the same area (Fuentes et al., 2006; Gryparis et al., 2009; Hund et al., 2012; Jandarov et al., 2017; Ntirampeba et al., 2018; Sumetsky et al., 2020; Utazi et al., 2019; Zhang et al., 2016). Subsequently, the spatial information misalignment can lead to errors or inaccuracies in the analysis or application of the data. It can also make comparing or combining data from different sources particularly challenging. For example, if datasets that are used to map air pollution have different spatial resolutions, the resulting map may not accurately reflect the actual distribution of pollution. There are several scenarios of spatial misalignment depending on the location of observations, type of measurement and intended use of each variable in the statistical analysis.

In this study, we have employed a mix of methods to address the issue of spatial misalignment. We have reprojected datasets to a standard coordinate system and resolution, and utilised interpolation techniques and aggregation to combine data from multiple sources and correct the misalignment.

1.6.1. Cases of spatial misalignment

Similar to our study, the most common type of misalignment is the spatial type which occurs when the sampling locations are different between the outcome and predictor covariates. For instance, if we want to study the association between air pollution and IBD in a region, we could survey 10,000 households and record the observed cases of IBD while gathering the air pollution data from each available monitoring site in the same region. Subsequently, the covariate of air pollution will be misaligned with the health outcome, given that we do not have an air quality measurement at the location of each household. In this example, the mismatch is caused by the different locations of the disease cases and the measured risk factor. However, in studies with multiple risk factors, this could also occur for different covariates that present disparities in their sampling locations.

The second most frequent type of misalignment, also present in our study, is the spatial support type which occurs when using different scales and types of measurement for the outcome and covariates of interest. A typical scenario of spatial support mismatch occurs when areal data are provided by different sources using different administrative units. An example of discrepancies in spatial support that we have encountered can be observed for regions between the Eurostat datasets from 2013 and 2016 due to boundary shifts and the discontinuation of some administrative units. The spatial support misalignment also includes cases of surface-to-point, area-to-point, and other types of scale mismatches. Even the spatial type mismatch discussed in the previous paragraph can be considered a particular case of spatial support misalignment.

The third type of misalignment is the modifiable areal unit problem which is linked to spatial aggregation and the grouping effect (Pacifici et al., 2019). The spatial aggregation problem occurs when the outcome or a covariate, especially for count data, is aggregated to larger areas to match the other variable in the dataset. Spatial aggregation may change the results of inferences for estimated parameters (Wittig et al., 2019). Lastly, the grouping effect is related to differences in the size and shape of the areal units used in the data collection and analysis (Pacifici et al., 2019). For example, the average area size of Scotland's NUTS3 (2016) territorial units is 33 times larger than the average NUTS3 area size of greater London. Therefore, it is more likely that an environmental factor will influence a whole NUTS3 territory in London than a vast NUTS3 territory in Scotland.

Hence, prior to the geostatistical analysis, we need to assess the i) sampling locations, ii) level and scale of measurements, iii) use of each variable as an outcome or a covariate, iv) groupings of observations v) and the desired spatial support for the inference. In this study, the outcome of incidence is collected as the rate of aggregated cases for the paediatric population of each NUTS3 area. As discussed in the methods 2.2.2, the NUTS3 regions covered by a reporting unit will be assigned an incidence value as an aggregate rate of the new PIBD cases for this unit, which underlines that the modifiable areal unit problem is also present in our dataset. The socioeconomic, health, and environmental covariates employed in this study are based on the NUTS3 spatial support. However, the main volume of environmental covariates was provided by NASA and EEA, and they are available in a point-location spatial support for different sampling locations. Therefore, the data misalignment for this project is challenging as both location and spatial support misalignment issues are observed among the covariates and between the covariates and the outcome.

1.6.2. Spatially misaligned data and change of support

The current literature provides a limited number of approaches to address the spatial misalignment problem (Hund et al., 2012; Jandarov et al., 2017; Liang and Kumar, 2013; Sumetsky et al., 2020). Some of these approaches are based on interpolation methods, such as Kriging, for areal and point data that we can employ to predict variables for the under-sampled locations. Gryparis 2007, also has suggested extensions of Kriging interpolation (Benchimol et al., 2009), while others have suggested Gaussian process modelling and Bayesian smoothing (Grieci and Bütter, 2009; Ludvigsson et al., 2017) as well as kernel smoothing (Lehtinen et al., 2011). Other methods for areal data also include up or downscaling. Going back to the PIBD surveying example, these methods would be employed to estimate the pollutant variable at the location of each household with a known outcome and then estimate the regression parameter of interest using the predicted covariates and the known outcome.

While a wide range of methods is available for the misaligned data, the modifiable areal unit problem (MAUP) still remains a more complex issue in spatial statistics. Davis (2004) suggested using as much disaggregate data as possible to avoid aggregation-related issues, which is a preferred approach when population density data are available (Hope et al., 2012). Wong 1991 and Fotheringham et al. 2000, have also suggested the sensitivity analysis approach, which can identify the MAUP but with a minimal contribution to correcting the issue (Jakobsen et al., 2011; Martín-de-Carpi et al., 2013).

Overall, the choice of methods for changing spatial support to align the study data should not introduce bias and should maintain as much information as possible to preserve statistical power. As underlined by Gryparis et al., 2009, in exposure assessment epidemiological studies, such as this one, the exposure should not be assumed to be constant over the region of interest in order to reduce exposure measurement error and maintain high power. The second point made by Gryparis that was relevant to our study was that for studies of chronic diseases, the analyses rely mainly on exposure heterogeneity induced by spatial variability that should be maintained as much as possible during the change of special support.

1.7. Overarching hypothesis, aims and objectives

After reviewing the literature and finding evidence of notable environmental aspects in the development of PIBD, this project's overarching hypothesis is that specific environmental factors are linked to the incidence of Paediatric Inflammatory Bowel Disease in the areas under investigation.

To investigate this hypothesis, we must address the following questions:

1. *Can the developed set of methodologies be employed successfully to identify the disease incidence in the studied regions?*

To address this question, it is necessary to collect a sufficient amount of data that enables the estimation of disease incidence in the regions of interest. Additionally, it is crucial to ensure that the incidence reported from the same adjacent regions has been consistent over the different collection years since this is a critical indicator of data quality. Finally, comparing our findings with established patterns in the literature, such as the PIBD incidence latitude trend, can serve as an additional validation step in answering this question.

2. *Are there any spatial and temporal effects present?*

To verify the presence of spatial patterns, it is necessary to demonstrate significant differences in the incidence of the disease between different countries and regions. These differences must be reported consistently over time to ensure their validity. Similarly, to validate any potential temporal effects, the time-related trends should be observable within each reporting region over multiple collection years.

3. *Are any environmental exposures associated with the disease incidence rates?*

This requires establishing and validating the analytical methods to combine and analyse our data while considering the underlying spatiotemporal structures. Employing these methods should help us detect the presence of significant associations between the observed disease incidence and certain risk factors.

4. *Are any of the variables included in the recruited patients' exposome affecting the disease phenotype?*

This would require applying the appropriate methods that will provide statistical evidence of strong associations between the patients' specific characteristics and the probability of presenting a particular disease phenotype.

To answer these questions, five aims were set:

1. Design PIBD databases:
    - The first database was designed to store, manage and perform quality controls on data collected from the Inception Cohort clinical study.

    - The second database was designed to store and manage data collected from the PIBD safety epidemiology registry.
2. Develop analytical methods:
    - The estimation of disease incidence required the development of algorithms to process the received information and manage duplicates, overlaps, changes of reporting experts, the difference in recruitment age limits and other centre-specific variables in order to estimate the disease incidence rates for each covered area.
    - The alignment of spatial data required the development of methods that allowed us to combine information that was collected in various spatial formats.
    - The disease incidence analysis required the development of analytical processes to investigate the effects of the studied risk factors, geography and time on the frequency of the disease cases.
3. Estimate and analyse disease incidence:
    - Using the developed methodology, we have calculated the incidence and prevalence of PIBD in total and for each phenotype in various areas in Europe.
    - The spatiotemporal patterns of the disease were analysed to help us understand the disease distribution and validate our methodology.
4. Collect and prepare the external risk factors data:
    - Identified, extracted and, in some cases, calculated risk factors such as pollutants, population density, demographics and other predictors.
    - Aligned all spatial information to prepare for the analysis using the developed methodologies that included extraction of information based on the location and spatial interpolation.
5. Study the effects of these factors on the disease incidence and phenotype:
    - This required the validation of the developed methods based on simulated examples.
    - Based on the geostatistical analysis and modelling, we investigated the effects of space, time and risk factors on the incidence of the disease.

2. METHODS

2.1. Data: collection and management

Data collection and management are two of the most critical processes of clinical research, as they help ensure that the data collected is accurate, complete, and reliable. In this section, I will detail all the data sources utilised, the demographics of the population, methods used to gather data, techniques applied in processing, the purpose of collecting the data, and how they were utilised. Data collection in clinical research typically involves using standardised forms, questionnaires, and other tools to collect information from study participants. This information can include demographic information, medical history, and outcome measures. Data can be collected in various formats, such as observations, interviews, surveys, and clinical measurements. In environmental clinical studies, the data collection process expands to the collection of data from environmental factors such as air pollution and exposures that may be linked with health and disease outcomes.

In this project, as discussed in the introduction, we have collected data from the Inception Cohort, a prospective clinical study, and the PIBD Safety Registry, a large-scale epidemiological study which we combined with data from various agencies, including Eurostat, EEA and NASA. Furthermore, we have also used data from PIBD health records from patients diagnosed in the Royal London Hospital (RLH). The RLH PIBD health records were not part of the data analysis but were used as supplementary material for the refinement of our methods. This dataset provided additional evidence and information for the adjustment of the incidence that was reported by centres with an upper age limit of patients different to 18 years (2.2.1).

While the Inception Cohort dataset includes a plethora of variables, for this project, we have extracted the environmental questionnaire data only, which provides information about the previous exposures each patient has had up to the time point of diagnosis. Furthermore, the PIBD Safety Registry collected information from multiple centres worldwide regarding the covered areas, new and existing patients seen, and the rare and severe complications seen in their practices. Lastly, additional data from Eurostat provided the maps and administrative units used for data collection and analysis while the EEA, NASA, DEFRA and other sources were used to extract the risk factors information, as shown in **Table 2**.

Table 2 Summary of the data sources and main variables extracted.

This table summarises the source of information and the data extracted, highlighting the important variables and their use.

Source	Data	Collection method	Storage
PIBD Inception Cohort	Environmental Questionnaire	REDCap EDC – entered by clinical research staff and patients	REDCap EDC – stored at the external servers of a data centre in Liverpool, Exported directly in csv format
PIBD Safety Registry	Annual Denominator data form	REDCap EDC – entered by participating centres (PIBD experts)	REDCap EDC – stored at the Queen Mary University servers in London, Exported directly in csv format
PIBD Safety Registry	Monthly report form	REDCap EDC – entered by participating centres (PIBD experts)	REDCap EDC – stored at the Queen Mary University servers in London, Exported directly in csv format
Hospital Records	Hospital PIBD records	Infoflex EDC – entered by PIBD clinicians at Royal London Hospital	Infoflex EDC – Kept at NHS servers, Exported directly in csv format
EEA	Pollution data	Uploaded by from European Environment Agency	Extracted from the European Environment Agency from AirBase and the European Pollutant Release and Transfer Register (E-PRTR)]
Eurostat	NUTS Maps	Uploaded by GISCO - the Geographic Information System of the COmmission	Extracted from the Eurostat website under the GISCO: Geographical information and maps
Eurostat	NUTS Data	Uploaded by Eurostat	Extracted from the Eurostat website under the data/database navigation tree
Eurostat	Population Density	Uploaded by GISCO - the Geographic Information System of the COmmission	Extracted from the Eurostat website under the GISCO: Geographical information and maps
NASA	Sun radiation and climate data	Uploaded by NASA Earth Science's Applied Sciences Program	Downloaded using R and APIs
DEFRA and NAEI (National Emissions Inventory)	Pollution data	Uploaded by DEFRA and the NAEI team	Downloaded using the Defra Data Services Platform, the Defra pollution inventory and NAEI website under the Spatial emissions and maps
INSPIRE	Pollution data	Uploaded by the European commission as part of the INSPIRE Directive	Downloaded from the INSPIRE knowledge database
Recent use of Antibiotics	Specific	REDCap EDC – entered by clinical research staff and patients	Inception Cohort
Recent use of NSAIDs and Aspirin	Specific	REDCap EDC – entered by clinical research staff and patients	Inception Cohort
Exposure to air pollutants	Specific/General	REDCap EDC – entered by clinical research staff and patients (& EEA)	Safety Registry/Inception Cohort
Exposure to solar irradiance	Specific/General	REDCap EDC – entered by clinical research staff and patients(& EEA)	Safety Registry/Inception Cohort

Socio-economics - Parents' employment	Specific/General	REDCap EDC – entered by clinical research staff and patients(& EUROSTAT)	Safety Registry/Inception Cohort
Socio-economics - Parents' income	Specific/General	REDCap EDC – entered by clinical research staff and patients	Inception Cohort
Demographics – Parents’ ethnic background	Specific/General	REDCap EDC – entered by clinical research staff and patients	Inception Cohort
Recent Migration (1st generation)	Specific	REDCap EDC – entered by clinical res earch staff and patients(&EUROSTAT)	Safety Registry/Inception Cohort
Previous infections	Specific/General	REDCap EDC – entered by clinical research staff and patients(& EUROSTAT)	Safety Registry/Inception Cohort
Geographic Location	Specific/General	REDCap EDC – entered by clinical research staff and patients(& EUROSTAT)	Safety Registry/Inception Cohort
Urbanisation	Specific/General	REDCap EDC – entered by clinical research staff and patients(& EUROSTAT)	Safety Registry/Inception Cohort
Water quality and source	Specific/General	REDCap EDC – entered by clinical research staff and patients(& EEA)	Inception Cohort

The collected data must be managed and stored to ensure its integrity, security, and accessibility. This typically involves the use of databases to store and organise the data. Data management procedures should be in place to ensure that the data is accurate and complete and that any errors or inconsistencies are identified and corrected. In the table below, the collection and storage of each dataset is presented in detail **Table 3**.

Table 3 Summary of the data source, collection, processing, and storage.

Source	Data	Collection method	Storage
PIBD Inception Cohort	Environmental Questionnaire	REDCap EDC – entered by clinical research staff and patients	REDCap EDC – stored at the external servers of a data centre in Liverpool, Exported directly in csv format
PIBD Safety Registry	Annual Denominator data form	REDCap EDC – entered by participating centres (PIBD experts)	REDCap EDC – stored at the Queen Mary University servers in London, Exported directly in csv format
PIBD Safety Registry	Monthly report form	REDCap EDC – entered by participating centres (PIBD experts)	REDCap EDC – stored at the Queen Mary University servers in London, Exported directly in csv format
Hospital Records	Hospital PIBD records	Infoflex EDC – entered by PIBD clinicians at Royal London Hospital	Infoflex EDC – Kept at NHS servers, Exported directly in csv format
EEA	Pollution data	Uploaded by from European Environment Agency	Extracted from the European Environment Agency from AirBase and the European Pollutant Release and Transfer Register (E-PRTR)]

Table 4 Summary of risk factors included in the primary analyses with exposome domain and analysis plan.

Exposure	Type	Dataset
Caesarean section	Specific	Inception Cohort
Breastfeeding	Specific	Inception Cohort
Smoking status during pregnancy	Specific	Inception Cohort
Age of mother at birth	General	Inception Cohort
Vaccination history	Specific	Inception Cohort
Hygiene hypothesis - Pets	Specific	Inception Cohort
Hygiene hypothesis – Birth order	Specific	Inception Cohort
Hygiene hypothesis – Siblings	Specific	Inception Cohort
Hygiene hypothesis – Family size	Specific	Inception Cohort
Diet - Exclusions	Specific	Inception Cohort
Vitamin D levels at diagnosis	Specific	Inception Cohort
Recent use of Antibiotics	Specific	Inception Cohort
Recent use of NSAIDs and Aspirin	Specific	Inception Cohort
Exposure to air pollutants	Specific/General	Safety Registry/Inception Cohort
Exposure to solar irradiance	Specific/General	Safety Registry/Inception Cohort
Socio-economics - Parents' employment	Specific/General	Safety Registry/Inception Cohort
Socio-economics - Parents' income	Specific/General	Inception Cohort
Demographics – Parents' ethnic background	Specific/General	Inception Cohort
Recent Migration (1st generation)	Specific	Safety Registry/Inception Cohort
Previous infections	Specific/General	Safety Registry/Inception Cohort
Geographic Location	Specific/General	Safety Registry/Inception Cohort
Urbanisation	Specific/General	Safety Registry/Inception Cohort
Water quality and source	Specific/General	Inception Cohort

The stored data were subjected to quality controls. The quality of the data is essential for the validity of the research. The methods used to ensure the data integrity of the project are discussed in the following paragraphs. Data security was also critical in our clinical and epidemiological research, as the data collected contained sensitive participant information. All patient data in our studies were protected from unauthorised access, alteration, or deletion. Lastly, the study has received approval from the Health Research Authority in the UK and local regulators for each country where patients were recruited in the Inception Cohort. The Safety Registry did not require ethics approval since it did not collect any patient-identifiable information used in this PhD, and only aggregate data were collected and analysed for the study of PIBD epidemiology.

2.1.1. Inception Cohort

As discussed in the previous paragraphs, the environmental eCRF data from the Inception Cohort were used in our analysis. This is an invaluable dataset as it provides broad exposome information from PIBD patients at the time of disease diagnosis. The high level of exposome detail is related to the great number of questions that the participants or their guardians have answered in conjunction with the availability of the patient's residence location prior to the diagnosis. Of the 770 prospectively recruited patients, 598 have answered the environmental questionnaire (**Table 5**). 86% of the recruited patients submitted their postcode partially or fully. Thus, information on the patient location was available for 513 patients with 300 postcodes located in the UK and 123 in the Netherlands.

Table 5 The demographics of the patients who filled the environmental questionnaire.

The patients who had their environmental questionnaire filled and were considered for inclusion in the PIBD phenotype analysis.

Sex	Count
Male	354
Female	244

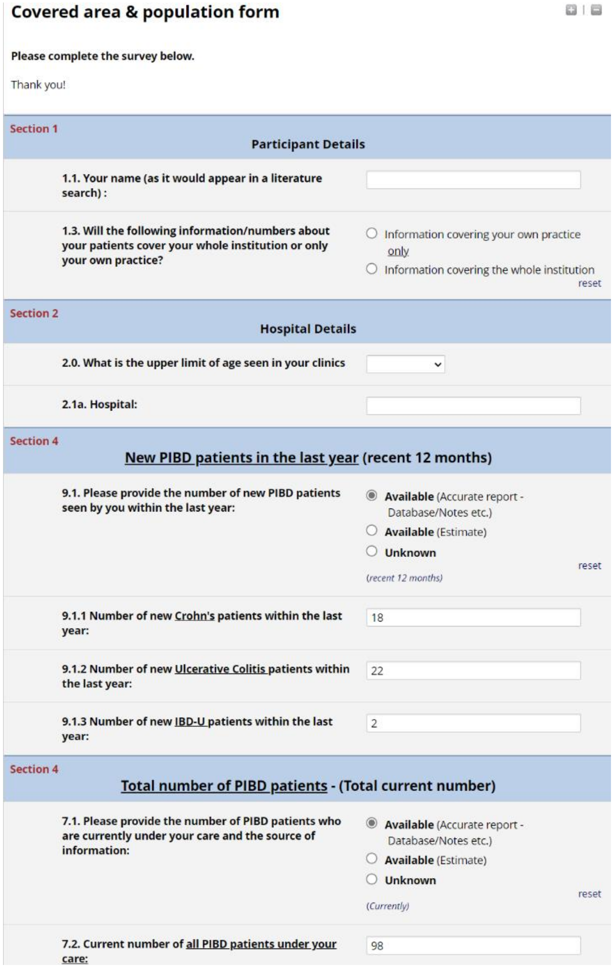
Diagnostic Impression Following Investigation	Count
CD	350
UC	198
IBD-U	47

Country	Count
UK	325
NL	126
IT	43
IL	33
RS	29
FR	14
MY	11
JP	8
JP	7
UAE	2

Age	Count
0	1
1	1
2	3
3	5
4	4
5	5
6	12
7	11
8	22
9	20
10	43
11	39
12	55
13	65
14	91
15	100
16	75
17	38
18	9

2.1.2. Safety Registry and Reporting network

The PIBD-SETQuality Safety Registry is an electronic registry that tracks rare and severe complications in children and adolescents with IBD. The registry was created by PIBD-NET and began in the UK in October 2016 and was soon after expanded in the Netherlands, followed by several other countries in Europe, the Middle East, Asia, Oceania and North America. Participating physicians are asked to report any complications seen monthly and to fill out an annual survey to with information about their practice and the number of patients under their care. We named this annual survey "Denominator data form" and used it to collect information about the number of new and current patients seen in each centre, the age of the patients who are being transferred to adult care and the catchment area for the referrals (**Figure 2**). The catchment area of a clinic refers to the geographic area from which it draws its patients. In this project, the catchment areas were selected by the responders using the Eurostat map of territories called NUTS combined with census data. The selected NUTS regions that were reported under the catchment areas combined with the population information, provided the estimate of the denominator population for each reporting centre. Therefore, this registry provided the data source for estimating the general population covered by each centre and the number of patients needed to estimate the corresponding incidence and prevalence per centre. The general population covered was estimated based on the catchment areas in conjunction with their population density and was used as the denominator for estimating the disease incidence and prevalence.



Covered area & population form

Please complete the survey below.

Thank you!

Section 1

Participant Details

1.1. Your name (as it would appear in a literature search) :	
1.3. Will the following information/numbers about your patients cover your whole institution or only your own practice?	○ Information covering your own practice only ○ Information covering the whole institution reset

Section 2

Hospital Details

2.0. What is the upper limit of age seen in your clinics	
2.1a. Hospital:	

Section 4

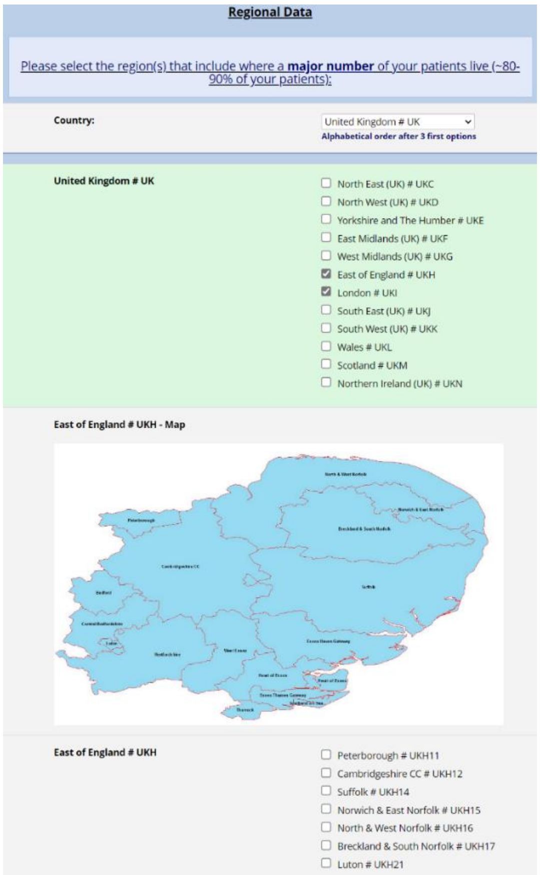
New PIBD patients in the last year (recent 12 months)

9.1. Please provide the number of new PIBD patients seen by you within the last year:	● **Available** (Accurate report - Database/Notes etc.) ○ **Available** (Estimate) ○ **Unknown** reset *(recent 12 months)*
9.1.1 Number of new Crohn's patients within the last year:	18
9.1.2 Number of new Ulcerative Colitis patients within the last year:	22
9.1.3 Number of new IBD-U patients within the last year:	2

Section 4

Total number of PIBD patients - (Total current number)

7.1. Please provide the number of PIBD patients who are currently under your care and the source of information:	● **Available** (Accurate report - Database/Notes etc.) ○ **Available** (Estimate) ○ **Unknown** reset *(Currently)*
7.2. Current number of all PIBD patients under your care:	98



Regional Data

Please select the region(s) that include where a **major number** of your patients live (~80-90% of your patients):

Country:	United Kingdom # UK **Alphabetical order after 3 first options**
United Kingdom # UK	☐ North East (UK) # UKC ☐ North West (UK) # UKD ☐ Yorkshire and The Humber # UKE ☐ East Midlands (UK) # UKF ☐ West Midlands (UK) # UKG ☑ East of England # UKH ☑ London # UKI ☐ South East (UK) # UKJ ☐ South West (UK) # UKK ☐ Wales # UKL ☐ Scotland # UKM ☐ Northern Ireland (UK) # UKN

East of England # UKH - Map

East of England # UKH	☐ Peterborough # UKH11 ☐ Cambridgeshire CC # UKH12 ☐ Suffolk # UKH14 ☐ Norwich & East Norfolk # UKH15 ☐ North & West Norfolk # UKH16 ☐ Breckland & South Norfolk # UKH17 ☐ Luton # UKH21

Figure 2. Example of the survey used to collect denominator data.

Details about the number of new patients, areas covered and the practice of the PIBD experts were requested.

As the Safety Registry increased in size, new centres joined at various times while a few centres stopped reporting. These changes posed a challenge to the estimation of the disease incidence. To address these changes, we have calculated all estimates per reporting unit and area and not per participant. Each area that was covered by a centre with at least one report is included in the incidence calculations. Finally, any reporting participant who was not dormant for more than three months was considered active. **Figure 3** below shows the total number of active participants during the study period.

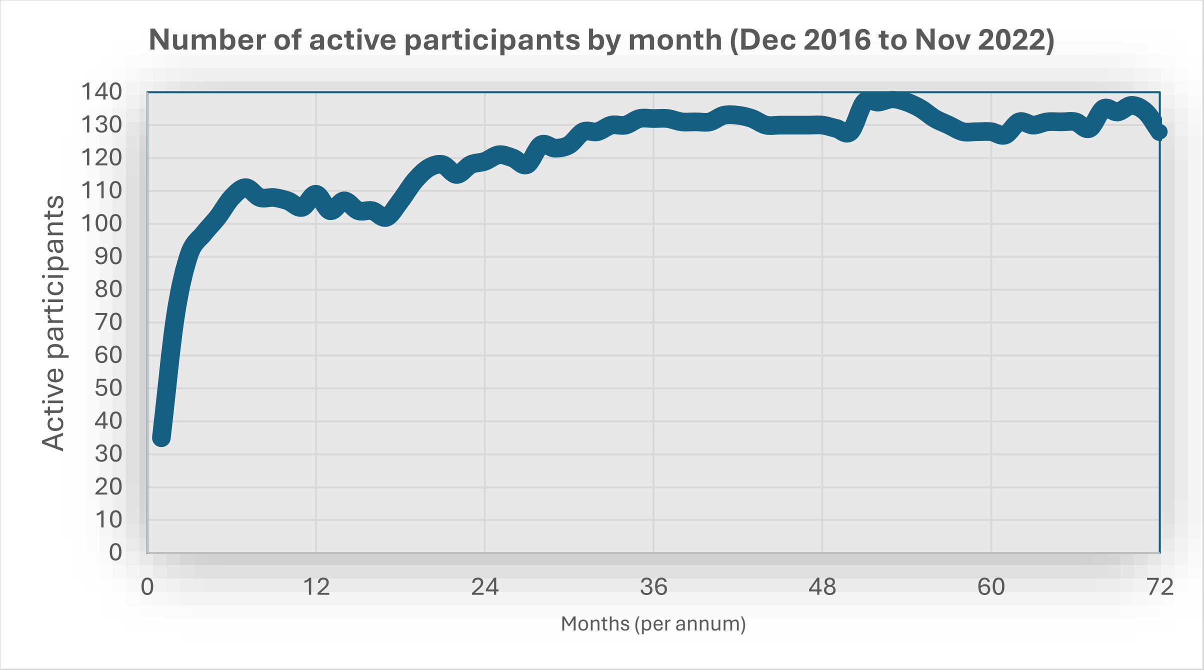
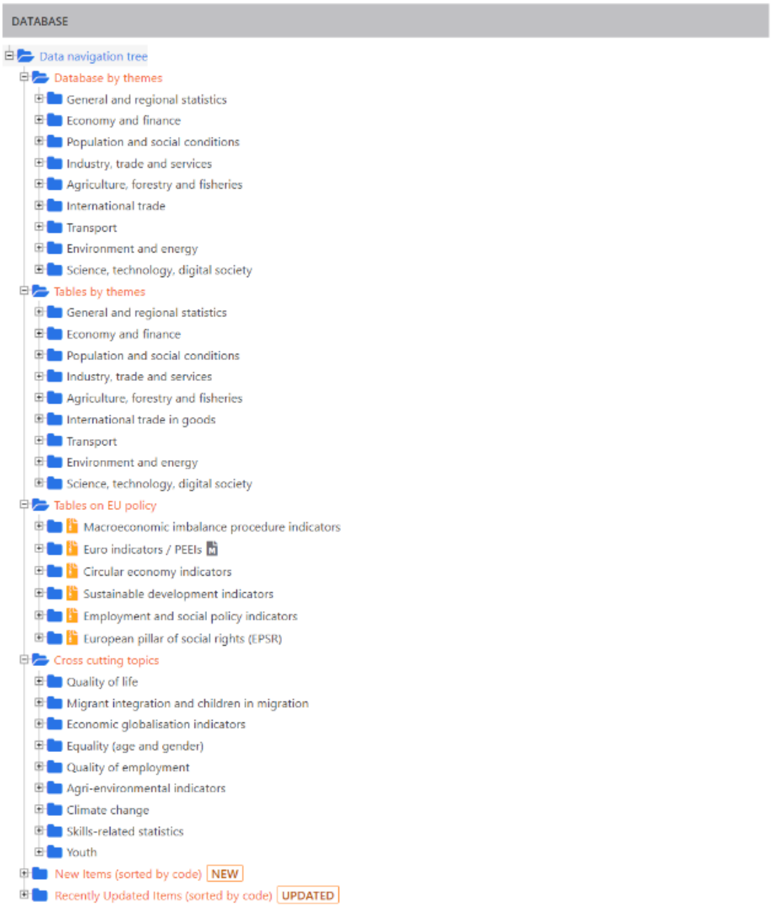


Figure 3. Network development and number of active participants.

The number of active participants in the Safety Registry has been increasing since 2016, expanding its coverage.

2.1.3. NUTS data integration of and misalignment handling

The Eurostat data, which are based on the NUTS system, were exported and combined with the NUTS maps that GISCO has made available also via Eurostat. These are polygon maps based on the shapefile format. Using the same administrative units has allowed combining the information from the Eurostat database with the shapefiles. The following paragraph describes the process of merging the extracted Eurostat data with the GISCO maps. The first step is to select the dataset of interest from the Eurostat database, as shown in **Figure 4.**



DATABASE

- Data navigation tree
  - Database by themes
    - General and regional statistics
    - Economy and finance
    - Population and social conditions
    - Industry, trade and services
    - Agriculture, forestry and fisheries
    - International trade
    - Transport
    - Environment and energy
    - Science, technology, digital society
  - Tables by themes
    - General and regional statistics
    - Economy and finance
    - Population and social conditions
    - Industry, trade and services
    - Agriculture, forestry and fisheries
    - International trade in goods
    - Transport
    - Environment and energy
    - Science, technology, digital society
  - Tables on EU policy
    - Macroeconomic imbalance procedure indicators
    - Euro indicators / PEEIs
    - Circular economy indicators
    - Sustainable development indicators
    - Employment and social policy indicators
    - European pillar of social rights (EPSR)
  - Cross cutting topics
    - Quality of life
    - Migrant integration and children in migration
    - Economic globalisation indicators
    - Equality (age and gender)
    - Quality of employment
    - Agri-environmental indicators
    - Climate change
    - Skills-related statistics
    - Youth
  - New Items (sorted by code) NEW
  - Recently Updated Items (sorted by code) UPDATED

Figure 4. Example of themes for the NUTS datasets that are available to download from Eurostat.

Despite the great number of themes, only datasets with NUTS3 information were included in our study.

The second step is ensuring that the dataset complies with the NUTS hierarchy desired for the mapping and analysis. Most frequently, the information is available in NUTS0, 1 and 2 formats but not in NUTS3 as the latter requires significantly greater resources to gather. The third step is to ensure that the correct filters have been applied and export the data. Depending on the request, these filters may include age groups, sex groups and years of data collection. At this step, ensuring that the extracted dataset uses the same version of the NUTS classification is vital. The NUTS maps were updated in 2006, 2010, 2013, 2016 and 2021. These updates were partial or minor for most regions. However, they have also introduced some major changes affecting the maps of Scotland, a few regions of England, France, and a small number of other regions in Europe. The third step is merging the data with the base shapefile using the NUTS3 id as the unique identifier to perform the data join. In this step, it must be clarified that several supportive files are needed to map and analyse a shapefile. A shapefile is the file that includes the feature geometry with the extension ".shp”, which requires an additional file with the ".shx” extension containing the indexing of the feature geometry. However, another file with a “.dbf” extension is also required to store any information about the different areas that are included in the feature geometry files. This is the attribute information file and essentially is the database containing all values of the variables that correspond to each area of the map. Lastly, a “.cpg” extension file identifying the character set to be used and a “.prj" extension file that provides the coordinate system and projection information are optional but essential files used for mapping a shapefile. The attribute information file needs to be amended using specialised software to merge the data with the shapefile. **Figure 5** below shows the results of merging the 2016 population density data with the 2013 version of the NUTS shapefile. This example was selected to demonstrate the potential compatibility issues between the NUTS maps and datasets depending on the release year and version. The NUTS0 mapping is complete, while the NUTS1 level shows incompatibility issues in east Poland and France. The NUTS2 level requires further attention as the incompatibility issues affect France, South of Scotland and the Republic of Ireland. Lastly, the NUTS3 level reveals additional nomenclature changes for Lithuania. This is an example of spatial misalignment that we have encountered in our data.

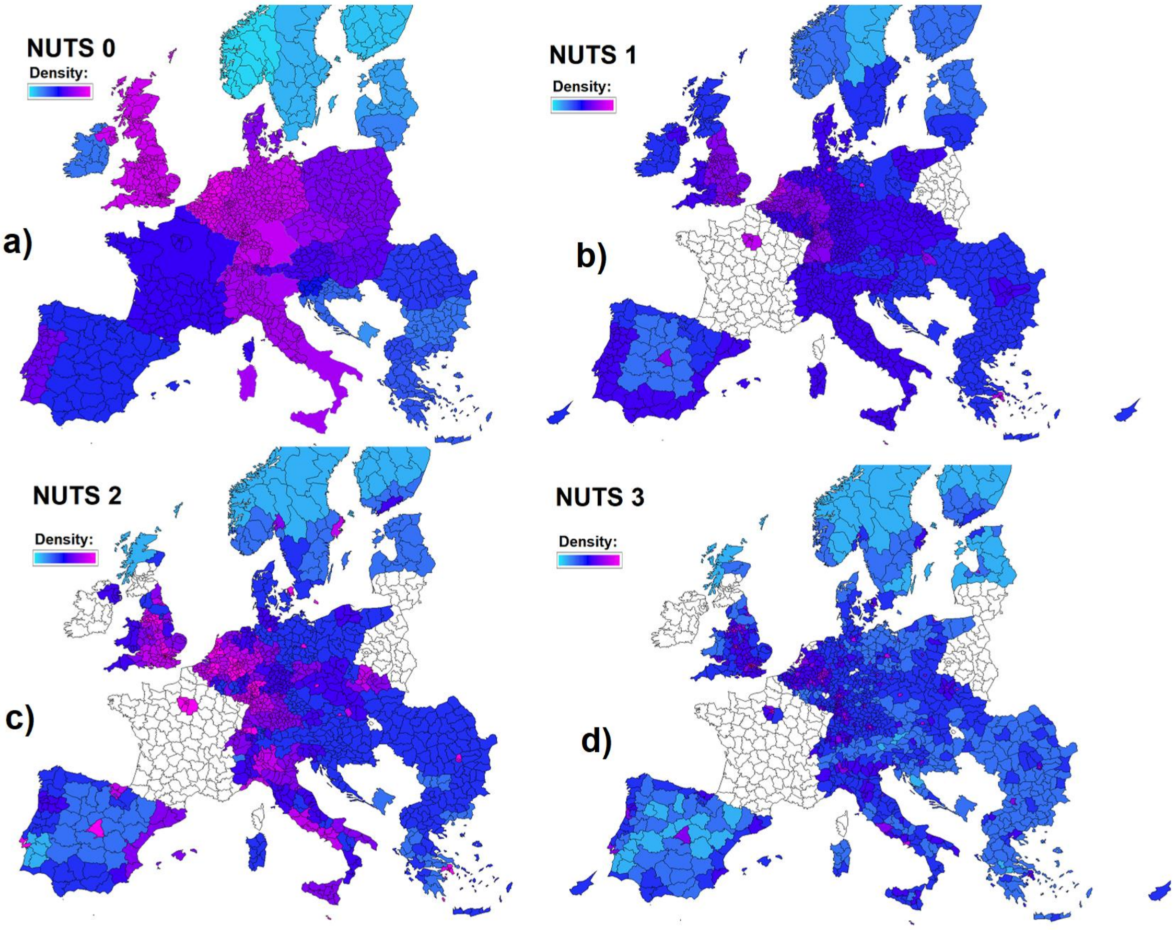


Figure 5. Choropleth map of Europe shows the population density for the NUTS0 to NUTS3 regions.

Population density was one of the predictors in our analysis. In this scenario, the information is fully compatible with the NUTS2016 map, but the map we have to use is based on the NUTS2013 format and is partially compatible with the population density data. Areas with no data due to incompatible administrative units appear white. a) When plotting the population density on the NUTS0 level (country level), the information is available in all areas. b) At the NUTS1 (i.e., Scotland, England, Wales) level, some incompatibilities appear. France is a great example of this, with only the greater Paris area to show compatibility between our data and the map. c) and d) as we proceed to lower tiers for smaller areas the number of incompatible regions increases.

For most Eurostat variables, this incompatibility issue can be addressed by appropriately filtering and matching the Eurostat data with the compatible GISCO maps. In continuation of the previous example of the population density mapping, **Figure 6** shows the complete population density NUTS3 map of Europe. This map was created by extracting the population density and assigning the values to the 2013 and 2016 maps which are then combined in **Figure 6.**

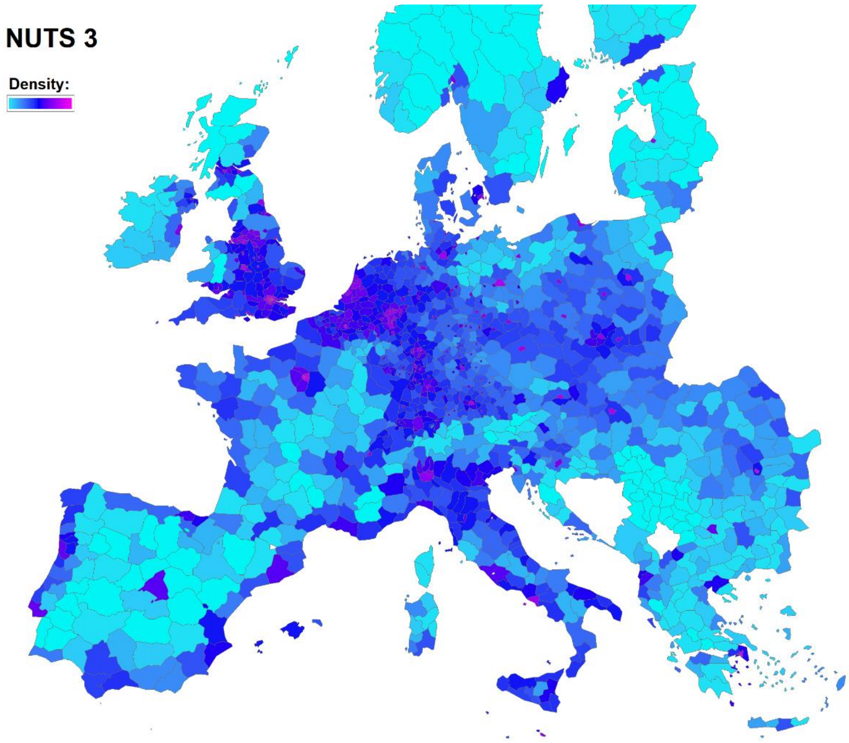


Figure 6. Choropleth map of Europe shows the population density for the NUTS3 regions in 2013 and 2016.

Whilst filtering and matching the Eurostat data with the compatible GISCO maps is a helpful approach for presentation and visual inspection purposes, in several cases, the incompatibility observed between the administrative units used for the data collection of our study and some of the Eurostat predictors remains an obstacle. This was the first major spatial data misalignment problem encountered in this project, and the following paragraph describes the methods used to address it.

In the population density example, the Safety Registry data collection was performed using the NUTS2013 polygons. Therefore, the predictor of population density must become available in the NUTS2013 format as well. The first step was to identify and use the map (shapefile) that is fully compatible with the variable of population density. For this example, the compatible shapefile was the 2016 version. The second step was the interpolation of the variable that converts the NUTS2016 population density map into a continuous raster map. In the third step, the raster file was converted to point data of high density. The creation of point data allowed step 4 to take place, as in this step, all the generated spatial points were summarised for each misaligned NUTS2013 polygon. This allowed us to transfer the population density information from the NUTS2016 map to the regions of the NUTS2013 map that had incompatibility issues. Although this is a computationally intensive method, it is a robust approach for estimating the values in the misaligned regions. However, this approach requires the selection of an appropriate interpolation method. For variables where the data misalignment is present only in some regions -similarly to this example- the common NUTS2013 and NUTS2016 polygons where the information is aligned can be used to validate the selected interpolation technique and the broader methodology. In our example for the common NUTS2013 and NUTS2016, we compared the original and interpolated values to conclude that the inverse distance-weighted interpolation with high-resolution settings had the best performance. The performance of the interpolation methods is shown in 2.4.2. In our example, where some NUTS2013 areas have missing values, we have produced a complete NUTS2016 map, as shown in **Figure 6** which was subsequently interpolated, as shown in **Figure 7**. The values of the interpolated maps were then assigned to the NUTS2013 maps resulting in the choropleth shown **Figure 8**. The final NUTS2013 map remains the same for the areas that did not have misaligned data, and it has incorporated new values based on the interpolation of the NUTS2016 map only for the areas with missing data due to misalignment. The original NUTS2013 map and the corrected version are compared in **Figure 8**.

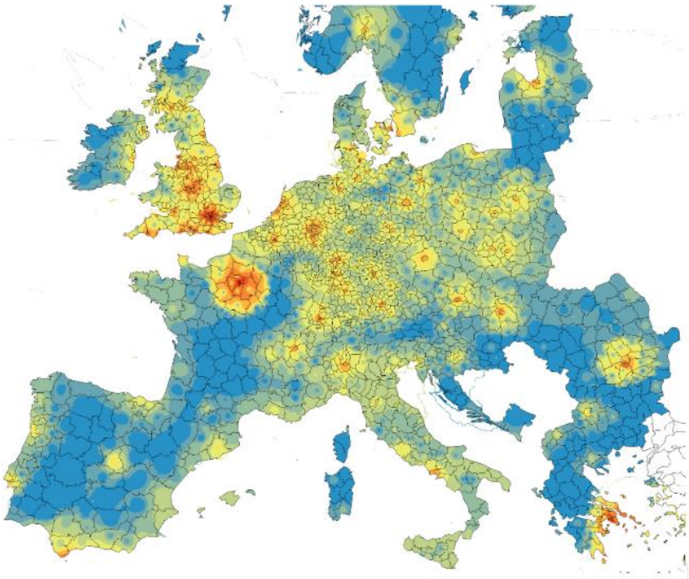


Figure 7. The interpolated map of the population density in Europe in 2016 is based on the NUTS2016 map.

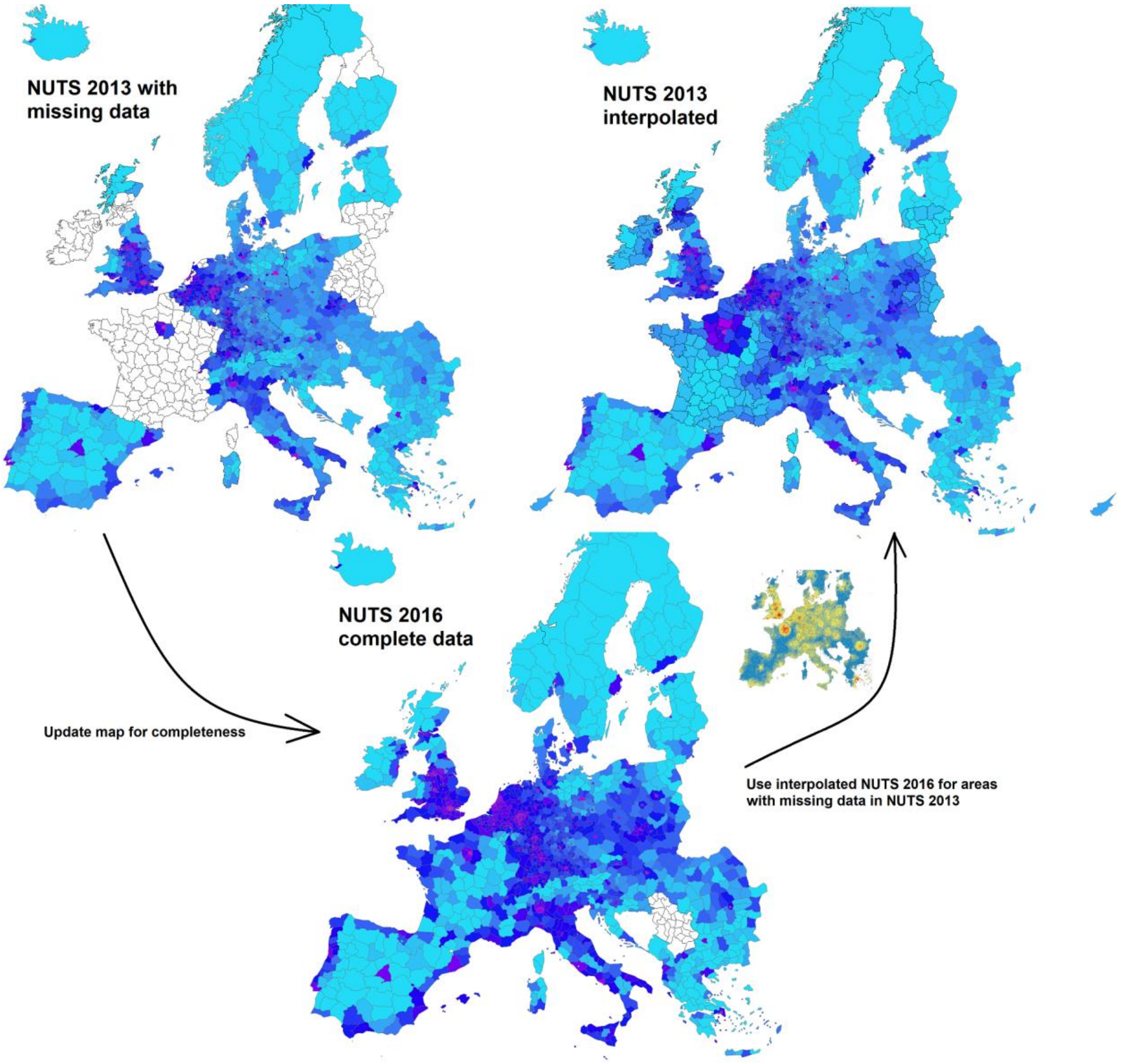

NUTS 2013 with
missing data
NUTS 2013
interpolated
NUTS 2016
complete data
Update map for completeness
Use interpolated NUTS 2016 for areas
with missing data in NUTS 2013

Figure 8. The NUTS 2013, 2016 interpolated and combined map of the population density in Europe.

The NUTS2016 map is used to produce an interpolated map that we use to fill in the information in the misaligned areas of the 2013 map.

A potential limitation of the proposed method is related to the interpolation step. Extracting information from one support vector (NUTS 2016 in our example) and merging it with another (NUTS 2013 in our example) relies on the assumption that the interpolated dataset will provide the true values for the queried regions. To investigate the accuracy of the interpolation using the 2016 map, we generated the interpolated dataset and then re-assigned the interpolated values to the same support vector (NUTS2016 map). Performing a linear regression for the regional data we set the original values as the outcome and the interpolated, re-assigned values as the predictor. This way we estimated the extent of disagreement between the observed and predicted values for each region. The analysis returned an R squared of 96.76%. Further analysis showed that the R squared value was improved significantly by the smaller cell size and a reduction of the neighbouring regions' settings in the Inverse Distance Weighting (IDW) interpolation.

The Eurostat predictors that were included in the analysis were selected based on their relevance with the researched topic and the availability of information on the NUTS3 level. Although the relevance of several variables may not be apparent, these were included as proxy predictors that may indicate an effect from another variable not included in our data. For instance, fertility is unlikely to be linked with the incidence of IBD. However, certain environmental factors that influence fertility (i.e., exposure to chemicals) may also affect the incidence of IBD. Hence, the variable of fertility may work as a proxy predictor for the chemical with the hypothetical effect of this example. All Eurostat variables that were included in the study are summarised in the table below **Table 6**.

Table 6 Codes and description of Eurostat datasets with NUTS3 information.

Variable (EUROSTAT Code)	**Description**
(demo_r_pjanaggr3), (demo_r_pjangrp3), (demo_r_d3dens), (demo_r_gind3), (demo_r_pjanind3) , (cens_11ag_r3)	Population on 1 January by age, sex and NUTS 3 region, and additional data on demographics and population
(cens_01rsctz)	Population by sex, citizenship and NUTS 3 regions
(cens_11ms_r3), (cens_11fs_r3)	Population by marital status, by family status and NUTS 3 region
(cens_11fts_r3)	Families by type, size and NUTS 3 region
(cens_11dwob_r3)	Conventional dwellings by occupancy status, type of building and NUTS 3 region

(cens_01rdhh)	Dwellings by type of housing, building and NUTS 3 regions
(cens_01rheco), (cens_01rhagchi)	Private households by composition, size, age group of children and NUTS 3 regions
(demo_r_gind3)	Population change - Demographic balance and crude rates at the regional level (NUTS 3)
(demo_r_births), (demo_r_fagec3)	Live births (total) by NUTS 3 region and live births by age group of the mothers and NUTS 3 region
(demo_r_find3)	Fertility indicators by NUTS 3 region
(demo_r_deaths), (demo_r_mweek3), (demo_r_magec3)	Deaths (total) by NUTS 3 region and by week, sex, 5-year age group and NUTS 3 region
(cens_01reisco)	Employed persons by sex, age group, educational attainment level, occupation (ISCO-88) and NUTS 3 regions
(reg_area3)	Area by NUTS 3 region
(aei_fm_ms)	Manure storage facilities by NUTS 3 regions
(aei_pr_soiler)	Estimated soil erosion by water, by erosion level, land cover and NUTS 3 regions
(ef_r_nuts)	Structure of agricultural holdings by NUTS 3 regions - main indicators
(bd_esize_r3), (bd_hgnace2_r3), (bd_size_r3)	Employer business demography by size class and NUTS 3 regions plus Business demography and high growth enterprise by NACE and NUTS 3 regions plus Business demography by size class and NUTS 3 regions
(crim_gen_reg)	Crimes recorded by the police in NUTS 3 regions

2.1.4. EEA, NASA, DEFRA and NEI data integration

The disease incidence and prevalence results from the PIBD Safety Registry were combined with the environmental and NUTS data for the epidemiological analysis. As mentioned in the previous paragraphs, the environmental data were extracted from several validated sources. In the following paragraphs, the methods used for each source type are presented.

The European Pollutant Release and Transfer Register Regulation (E-PRTR) has been publishing data on many pollutant releases across Europe over more than 15 years. In this project, we have selected the 2016-2019 air quality datasets that contain information on 458 air pollutants. Although some pollutants are also included in the INSPIRE datasets, as discussed in the following paragraphs, other pollutants such as Benzene, Cadmium, Arsenic and Nickel are unique for this project and should be investigated thoroughly. Benzene, in particular, is a volatile organic compound that has been linked to potential changes to the human microbiome and could also be linked to the development of IBD.

The EEA E-PRTR datasets required interpolation and special handling compared to the other predictor datasets of this project. The information in these datasets is organised in point data at the locations where pollutant releases have been reported. However, considering that the spatial interpolation methods have been developed to predict the values of a variable at unknown locations, a random sampling approach to gathering the data is an essential assumption that must be met. In contrast, in the E-PRTR interpolation, every unknown location is surrounded by point locations of reported emissions. Thus, the interpolation will return an over-inflated prediction for each interpolated point on the map. **Figure 9** shows an example of this limitation, where a few sources of a pollutant seem to affect an unrealistic vast area on the map simply because every observation of the dataset has a high value of the pollutant quantity since this is an emissions-only report. The proposed solution to this problem was to introduce additional point data locations on the map where the interpolated pollutant is either zero or at the global average for the examined regions. This interpolation design aligns with the assumption that the E-PRTR report is complete and that no additional pollutant releases exist in the examined region.

The INSPIRE Directive established an infrastructure for the dissemination of spatial information in Europe to support environmental policies and policies or activities which may have an impact on the environment. The available INSPIRE data are high-detail and resolution datasets and include NO2 and NOx, O3, PM10, PM2.5 and information for the phytotoxic ozone dose (POD) related to different kinds of vegetation. These high-resolution maps can be converted to our standard NUTS format by simply producing summary statistics for each NUTS3 polygon. Below are two examples showing the NOx and PM2.5 conversion for the INSPIRE raster format to the NUTS3 of the project (**Figures 9 & 10**).

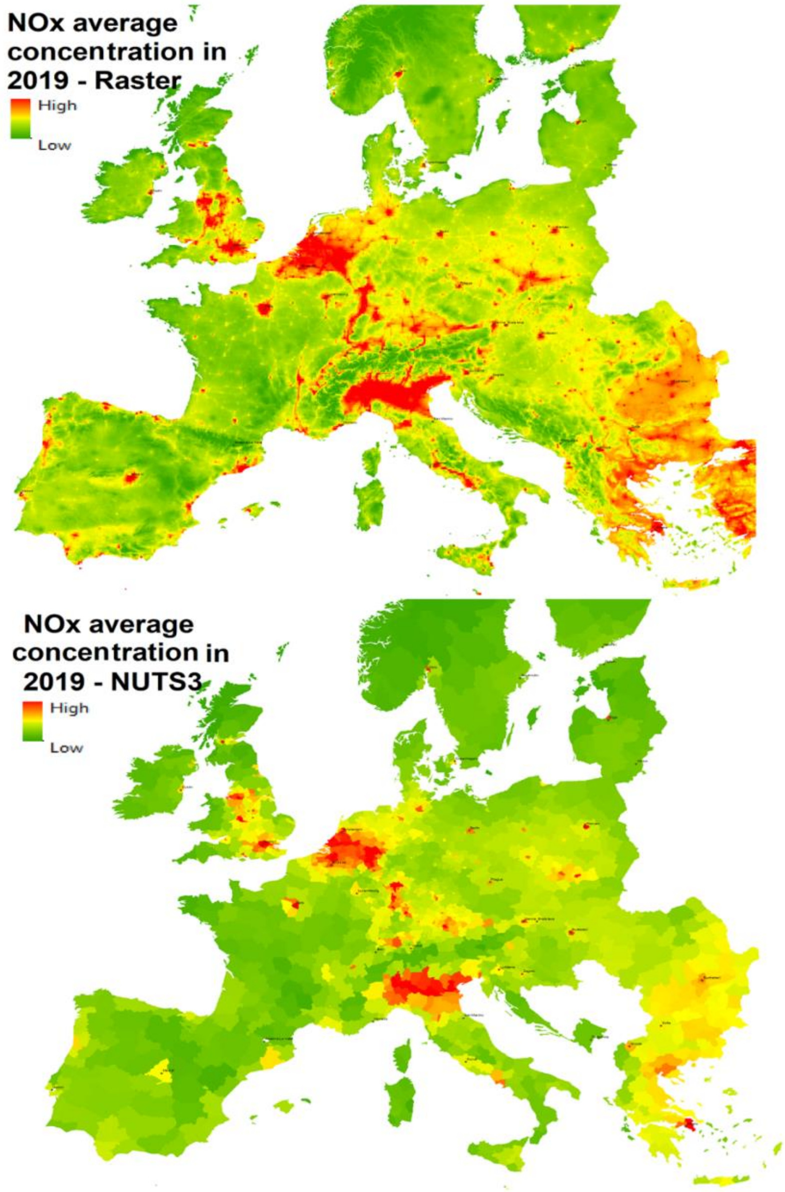

NOx average concentration in 2019 - Raster
High
Low
NOx average concentration in 2019 - NUTS3
High
Low

Figure 9. The interpolated NOx (continuous surface-raster) in Europe for 2019, followed by aggregation into the NUTS3 format.

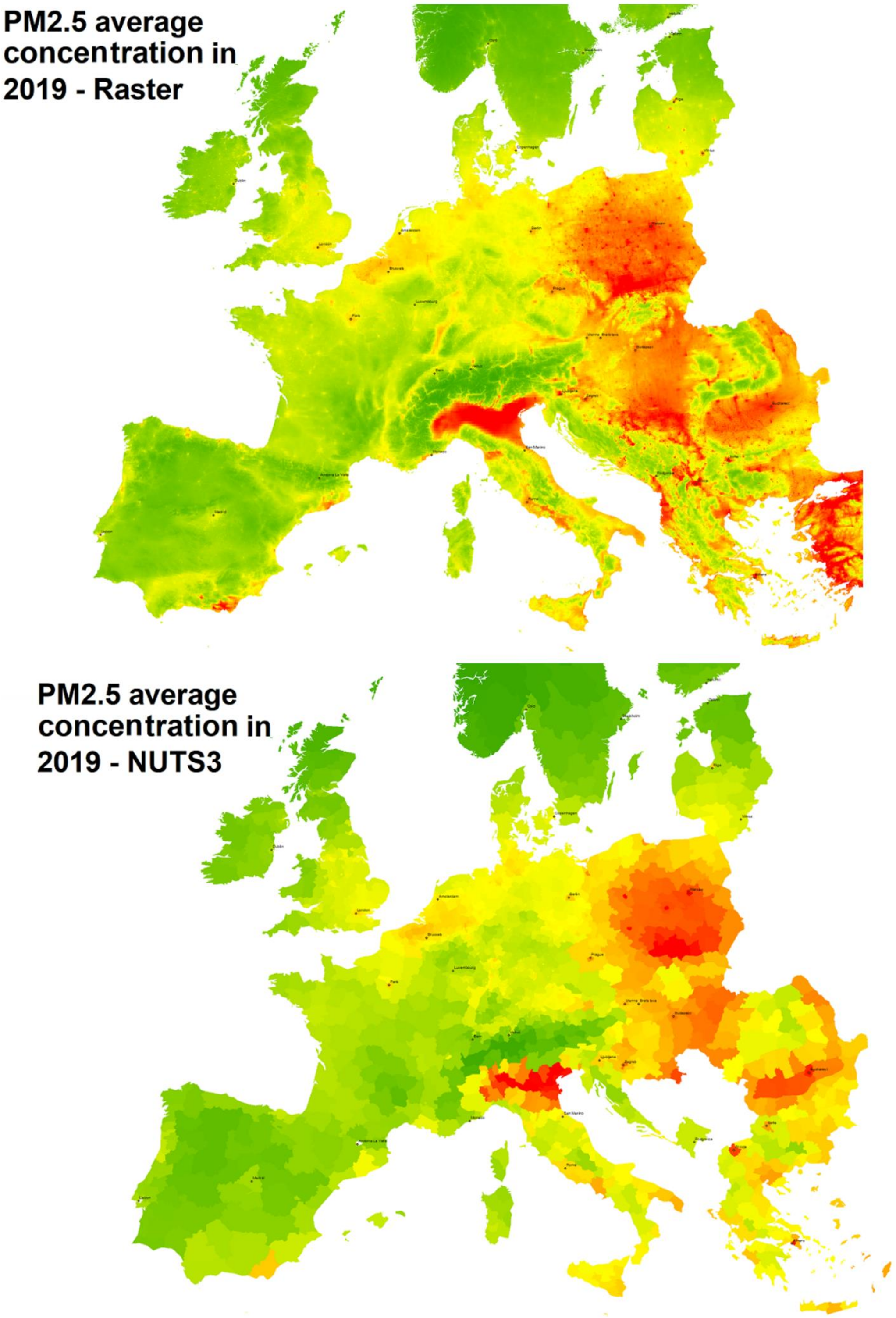

PM2.5 average concentration in 2019 - Raster
PM2.5 average concentration in 2019 - NUTS3

Figure 10. The interpolated PM2.5 (continuous surface-raster) in Europe for 2019, followed by aggregation into the NUTS3 format.

DEFRA collects and maintains datasets providing crucial information about the pollution levels in the UK's environment. These datasets contain information about air pollution, water pollution, soil pollution, and others, along with their geographical locations. The datasets include several pollutants such as Benzene, CO2, NOx, PM, SO and other sources of pollution. Although these datasets are UK-specific, they were included in this project as they can provide detailed predictor datasets for UK-specific subgroup analyses. The UK, followed by the Netherlands, had the highest participation levels and may reveal results with smaller effect sizes in country-specific subgroup analyses. The processing of these datasets required three steps. The first step required the change of the geodetic reference used by DEFRA from the British National Grid used by the Ordnance Survey to the WGS84 coordinate system. The second step was the interpolation of the point data to a continuous raster map, followed by the last step, which was integrating the interpolated information into the NUTS3 map (**Figure 11).**

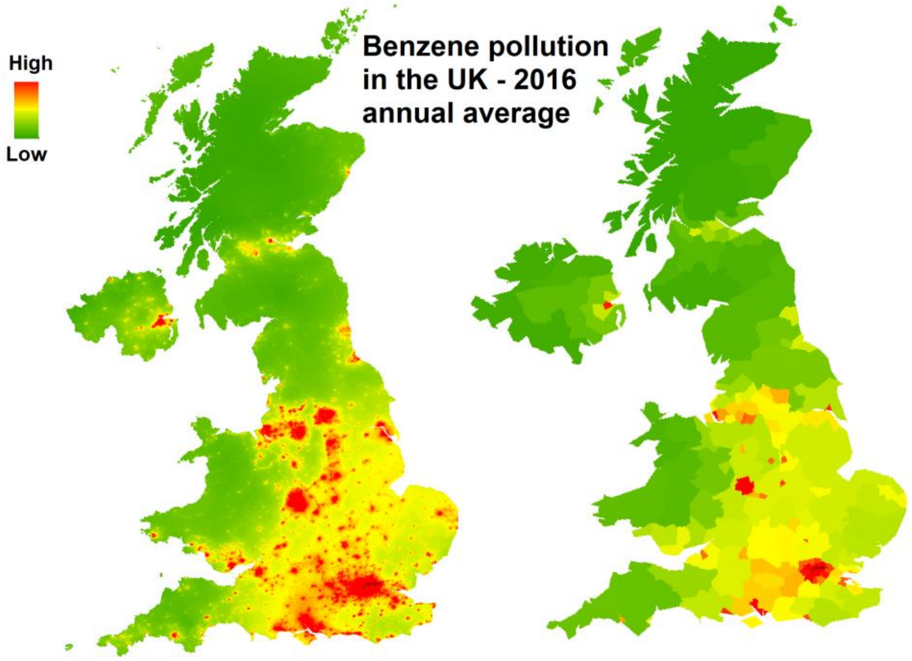


Figure 11. Conversion of interpolated DEFRA Benzene pollution dataset using the NUTS3 territory format.

The NAEI dataset, similarly to the DEFRA, is a UK-specific set of pollutants with a very high resolution. Although the very high resolution of this dataset may be redundant given the size of the NUTS3 polygons on which our analysis is based, the NAEI data include additional variables that allow an even more thorough investigation of the pollutant effects on the PIBD incidence in the UK.

2.1.5. Database design and maintenance

A significant portion of the workload in this project was devoted to database design and management throughout the study. Both the PIBD Inception Cohort and Safety Registry databases were designed manually, *de novo*. Both systems were built as relational databases using REDCap, our web-based electronic case report form data capture platform REDCap that is particularly versatile while providing features such as data validation and audit trails, making it a compliant data management system. The relational design of the databases allows data storage in the form of related tables. Subsequently, the data management and manipulation can be completed in a structured way. The relationships between the tables were established using the unique patient (Inception Cohort) or participant ID (Safety Registry) in conjunction with the time and arm of the study.

The Inception Cohort database is an extensive clinical database with 133 users, 780 records, and 4014 fields spread in 41 repeated forms organised in 5 arms, including multiple study visits. For this PhD project, three forms were extracted from two different arms for a single visit, the baseline. These were the screening and recruitment form, the Race and Origin Information Form and the Environmental questionnaire. The screening and recruitment form captures information on the diagnosis, age and inclusion criteria, followed by the race and origin information form that captures detailed information about the ethnic background of the patient. The third form is the environmental questionnaire, a 183 field eCRF collecting information about the patient's location and exposome, including a broad range of questions from a type of heating and washing practices to pets, vaccination history and water supply.

The Safety Registry database is separated into two different databases, and it is based on the same system and principles as the Inception Cohort. The registry's central database is based on two forms, the monthly surveys and the rare and severe complication follow-up forms. The supplementary database contains the denominator data form. The monthly surveys are being sent to the participants to collect "Yes/No" responses about any rare and severe complications that the reporting centre might have encountered in the previous month. This form is fully automated to ensure it is sent to the participants at the beginning of each month. Depending on the complications reported as "seen", it also triggers specific follow-up forms. The participants also receive up to three reminders for the monthly surveys and the follow-up forms as required. The automations of the database are based on over 100 calculated fields that were added on REDCap. Using "if" statements that were designed for this project, the system assigns scores to each participant. Depending on the syntax of the automated survey invitation rules in the system, these scores can trigger several forms.

The denominator data form contains detailed spatial information for over a thousand regions from 41 different countries with 38 maps and instructions to assist the participants in selecting the areas that their practices are covering. When the participants choose their country, the system will reveal the NUTS2 regions (or equivalent for regions outside Europe). Similarly, depending on the NUTS2 selections, the system will reveal the relevant NUTS3 options for the participants to choose from. The form has been designed with interactive features allowing the participants to provide accurate reports from their databases, estimates when the database information is not available and the ability to save the form and return it later. This combined with more features that were intergraded into the system, was in place to maximise the participants' engagement and survey completion rate. Lastly, REDCap was linked to an API system that we developed and ran in R programming language, feeding us live data and reports to recognise data discrepancies in real-time. This allowed for a shorter follow-up time, increasing the chances of a successful query resolution.

According to the action log of the REDCap system, the database design required over 50,000 actions (adding/removing/amending fields and rules). Despite the effort and time needed for its design, the major advantage of this system is that it requires very little maintenance. The Inception Cohort database currently does not require any action. In contrast, the Safety Registry requires a small number of actions to maintain the contact details of the participants and update the automation rules annually.

2.2. Data: preparation, validation and analysis

Data validation in clinical research refers to the process of ensuring that the data collected during a study is accurate, consistent, and reliable. This involves verifying the data for completeness, consistency, and accuracy before entering the information into a database or using it for analysis. The goal of data validation is to minimise errors and increase the validity and reliability of the results. Data validation procedures may include checking for missing or inconsistent data, reviewing data for accuracy, comparing data to external sources, and conducting statistical checks to identify outliers. The following paragraphs present the methods used to validate our data, followed by several sections discussing the data preparation methods used in this project.

The datasets of this project have been extracted from over 10,000 submitted eCRFs. Due to the high volume of data that I had to process, individual eCRF checks were not feasible, and the validation approach in the Safety Registry was focused on discovering inconsistencies and following up on missing data. In the Inception Cohort, I communicated regularly with the sites, and the entered data was monitored live using APIs scanning the data. In contrast, the correspondence with the sites in the Safety Registry was limited as more than 100 sites have been participating, and these have not got allocated data managers on site. Thus, a major part of the data quality checks was based on validation tests and automations. The primary purpose of the validation tests on the Safety Registry data is the detection of inconsistencies and outliers. The methods employed to achieve this were based on using funnel plots, the Tietjen-Moore test, and z scores to detect outliers.

A funnel plot is a graphical representation usually applied in meta-analysis to detect publication bias and small-study effects. This study used funnel plots to assess bias and significant variations in the reported incidence and prevalence from the participating sites. A funnel plot is a scatter plot of the effect size against its standard error. If the results from different sites are symmetrically distributed around the mean effect, this would suggest the absence of reporting bias. This ensures that centres that tend to see fewer or more new and current patients than expected are not more or less likely to report the results to the study. For instance, if the plot is asymmetrical, with smaller sites having larger effect sizes and wider confidence intervals, it suggests the presence of reporting bias or "small-study effects”. As shown in **Figure 12**, funnel plots also have two-sided confidence intervals (usually set at 95%) that become stricter as the standard error increases. In practice, this means that sites with larger sample sizes are expected to have higher precision. Therefore, only minor deviations from the expected mean will be considered acceptable and not exceed 95% C.I margins for larger sites. In contrast, the results from smaller sites are more likely to vary as they will have smaller precision. This means that even an extreme deviation from the mean may not be statistically significant if the site's sample size is particularly small. Therefore, the funnel plots can aid us in summarising if the reported results vary significantly per site and if there is a specific trend of reporting bias favouring lower or higher incidence values. The effect size is calculated as the proportion (p) of PIBD patients (k) in the sample of the general paediatric population (n).

- The proportion $p = k/n$ is the effect size.

While the standard error is estimated as follows:

- Standard error: $\frac{\sqrt{\sigma^2}}{\sqrt{n}} = \frac{\sqrt{p(1-p)}}{\sqrt{n}}$
- with variance $\sigma^2 = p(1-p)$ and standard Deviation $\sqrt{\sigma^2}$

A funnel plot example of the new-to-current patient ratio is performed on several proportions and summarised in the table below. It should be noted that all the analysed outcomes in the results section are proportions.

Table 7 Definitions of incidence and prevalence outcomes for PIBD, CD, UC and IBDU and CD/UC and IBDU.

N	**Numerator**	**Denominator**	**Compared per site**
1	New PIBD cases	General Paediatric population	PIBD Incidence
2	Current PIBD cases	General Paediatric population	PIBD Prevalence
3	New Crohn's cases	Total new cases	New CD to total Ratio
4	Current Crohn's cases	Current new cases	Current CD to total Ratio
5	New Ulcerative Colitis and IBDU cases	Total new cases	New UC to total Ratio
6	Current Ulcerative Colitis and cases	Current new cases	Current UC to total Ratio
7	New PIBD cases	Current PIBD cases	PIBD Incidence/Prevalence

Considering the number of funnel plots for each proportion, number of participating sites and multiple data collection rounds, a significant deviation in one assessment should not justify excluding a site from the analysis as random deviations are very likely. Only sites with significant deviations in multiple funnel plots were investigated further **Figure 12**.

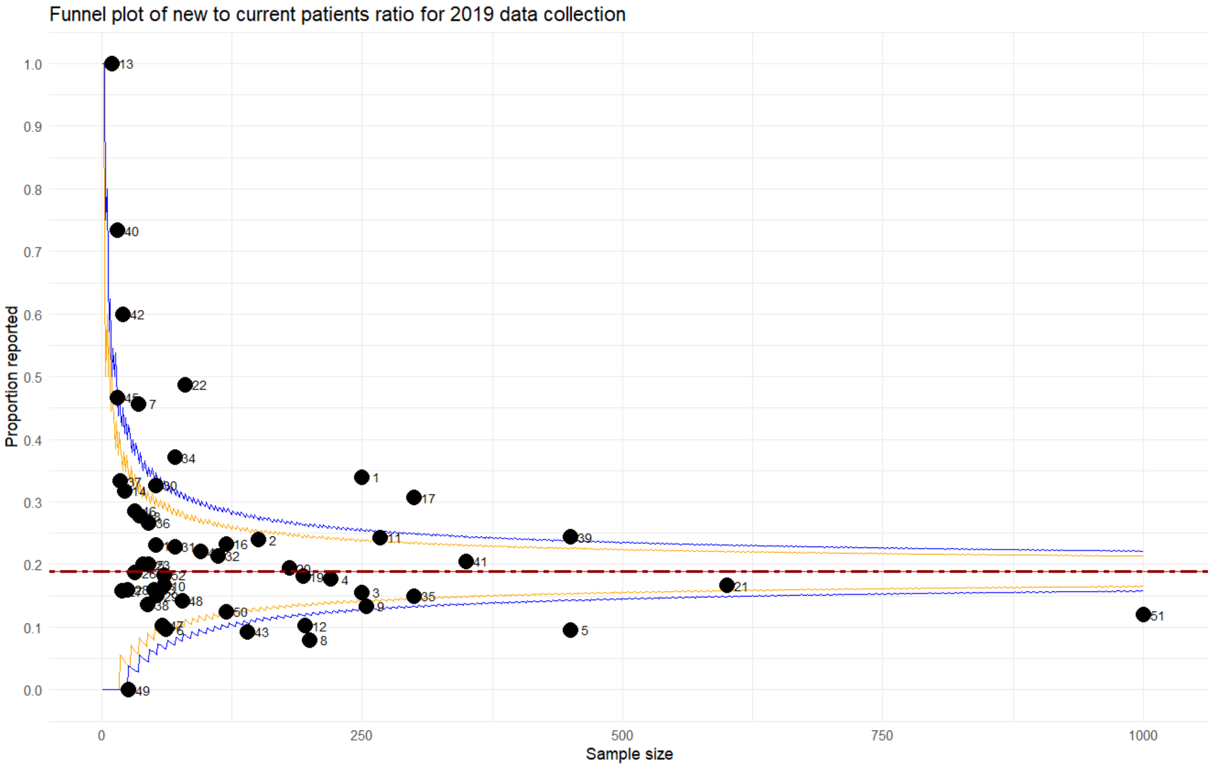


Figure 12. Funnel plot of the incidence-to-prevalence ratio reported by 52 sites in 2019. The yellow and blue curves are the 95% and 99% confidence intervals. Twelve sites were outside the margins, with sites 51, 5, 17, 1 and 22 deviating significantly from the expected ratio.

In conjunction with the funnel plots, t-scores, the Tietjen-Moore and z scores tests were performed to detect outliers. The latter tests were performed for variables that we found to be approximately normally distributed in the Safety Registry and Inception Cohort. Results with a Z score exceeding the value of 3 were considered outliers. Due to sample size limitation, the t-scores were used only for the within–centre comparisons to ensure that the multiple reports over the several data collection years are consistent. The Tietjen-Moore test was used to test specific hypotheses of a certain number of outliers to be present. The first step of the test requires the calculation of the absolute residuals:

- $r_i = |y_i - \bar{y}_i|$

In the second step of the test, the y_i values are sorted based on their absolute residuals r_i in ascending order. In the third step, we calculate the test statistic where the y_i are sorted. The $\bar{y}$ is the overall mean for all observations and $\bar{y}_k$ is the mean after the k most extreme observations have been removed. The number of the most extreme observations must be selected when formulating the hypothesis for the test.

The test statistic: $E_k = \frac{\sum_{i=1}^{n-k}(y_i - \bar{y}_k)}{\sum_{i=1}^{n}(y_i - \bar{y})}$, with the critical region for this test to be determined by a simulation based on the sample size of the dataset.

2.2.1. Safety Registry analysis: PIBD Incidence adjustment

In the Safety Registry, the denominator data form filled out by the participating centres also included information about the age of the patients transitioning to adult care services. This information is essential for adjusting the disease incidence as our final incidence figures were produced for the 0-18 years population. Therefore, centres with a patient age limit of 17 or lower may underreport the disease incidence as they are expected to encounter fewer newly diagnosed patients compared to centres that see patients up to 18 . In contrast, centres with an upper age limit above 18 are expected to report an inflated number of cases and incidence. To address this issue, we have adjusted the denominator population (the general paediatric population) to reflect the upper age limit for the patients seen in each clinic. For instance, considering the stable birth rate in Europe (Sobotka et al., 2011), over an extended period, a clinic that treats patients only up to the age of 10 will have roughly half the number of the patients compared to another clinic in the same location that treats patients up to the age of 20. While this issue can be solved with a proportional adjustment of the general paediatric population that was included as the denominator, the variations in the upper age limit of each centre are introducing additional complications. Using data from the Inception Cohort study, the Royal London Hospital PIBD database, and several sources from the literature (1.2) we have identified that the relationship between PIBD incidence and age does not follow a uniform distribution as shown in **Figure 13**.

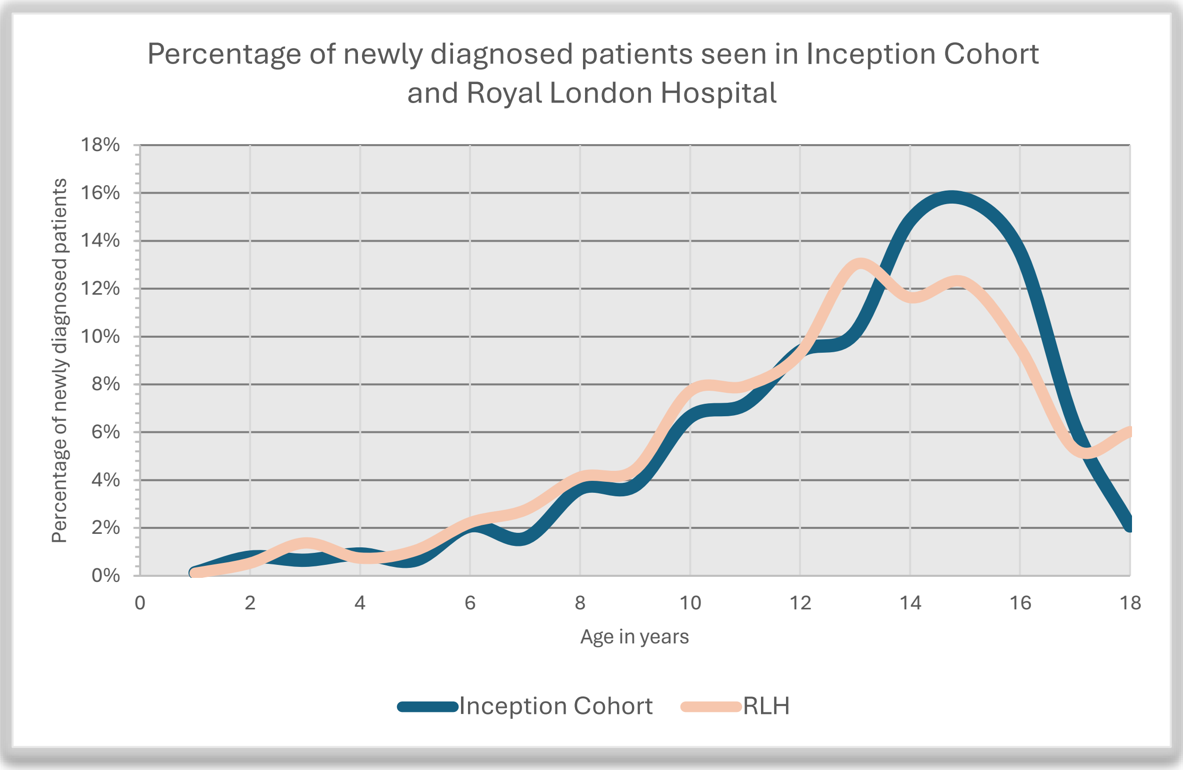


Figure 13. The probability distribution function of new patients diagnosed with PIBD from birth to the age of 18.

The two distributions suggest a rapid increase in the incidence from the age of 8 onwards. Please note that the two populations share approximately 10% of the patients, a percentage that is relatively small to explain the high correlation observed. The high correlation reinforces our assumptions about the exponential increase in the incidence.

Based on these findings, even after adjusting the denominator data to match the age limit of each clinic, a further correction is required to ensure that the final incidence estimate reflects the expected results for the 0-18 population. Based on **Figure 14**, a site with a 0–10-year-old population will report an incidence reduced by 5-fold compared to another site with a 0–18 population. Thus, appropriate statistical modelling should be used to determine the necessary correction.

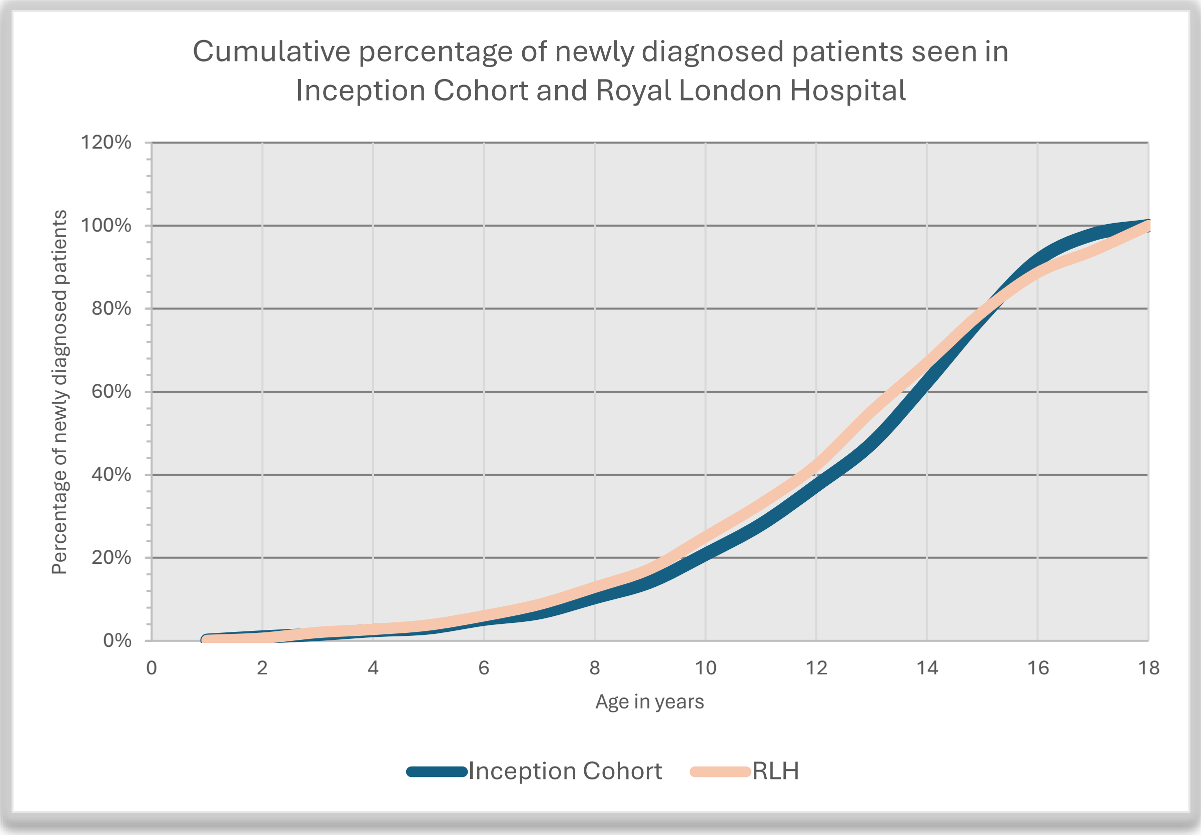


Figure 14. The cumulative distribution function of new patients diagnosed with PIBD from birth to the age of 18.

Both distributions follow a very similar trend.

Seo-Hee Kim et.al 2022, Wittig et.al 2019, Anders Forss et.al. 2021 and others have conducted PIBD incidence studies using administrative health records or insurance data and reported an incidence increase of PIBD by age. Similarly, to our data, these authors also reported a decline in the incidence following the rapid increase observed after age 8. However, the age that this incidence decline starts remains to be determined, as Forss suggests, the incidence may decline close to the age of 20, while Kim suggests figures closer to the age of 14. Based on a dataset of 37,555 cases of IBD, Kim also proposed that in recent years, the peak incidence of PIBD has shifted significantly towards younger age groups, which is in line with our data from the Inception Cohort and Royal London Hospital (**Figures 15 & 16**).

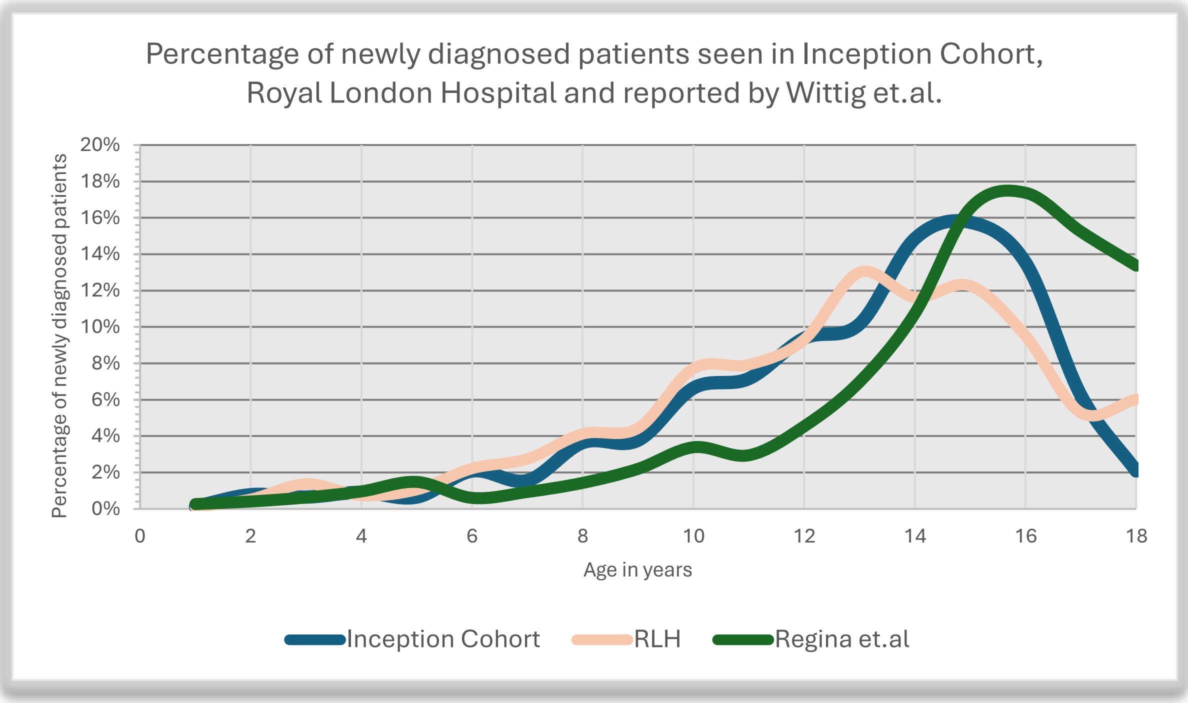


Figure 15. The probability distribution function of new patients diagnosed with PIBD from birth to the age of 18.

The distribution of patients reported by Wittig et al. was reported in age groups (1-5, 5-10, 10-15 and 15-18). The trends within the reported groups have been approximated.

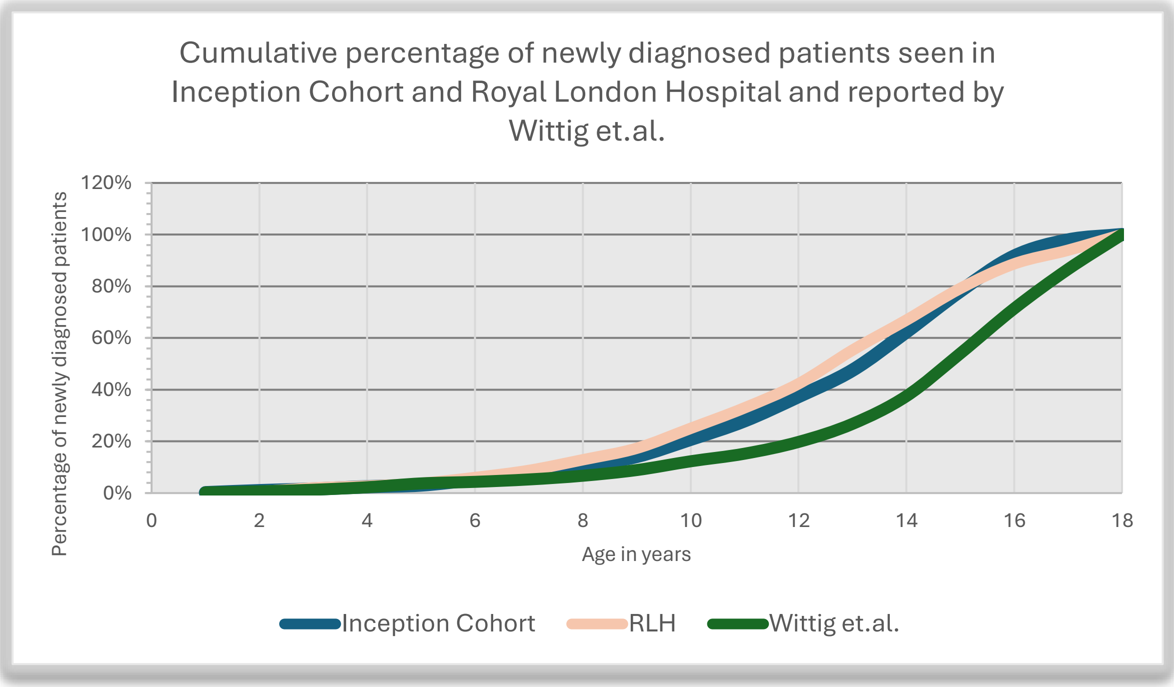


Figure 16. The cumulative distribution function of new patients diagnosed with PIBD from birth to the age of 18.

To adjust the incidence for the clinic age limit, we have generated a function that explains the observed relationship between incidence and age that we have discussed in the previous paragraphs. The fitted model is a cubic regression model, a type of polynomial regression model in which the relationship between the incidence and the age is modelled as a third-degree polynomial. The equation for the line of best fit is:

$$\text{Incidence} = 0.4482 - 0.1443 \cdot \text{age} + 0.01494 \cdot \text{age}^2 - 0.000449 \cdot \text{age}^3$$

This model reached significance for all coefficients and was used to fit a curved line through a set of data points, which in this case were the average of the three sources presented in **Figures 15 & 16** allowing for a more complex relationship to be modelled (**Figures 17-19**).

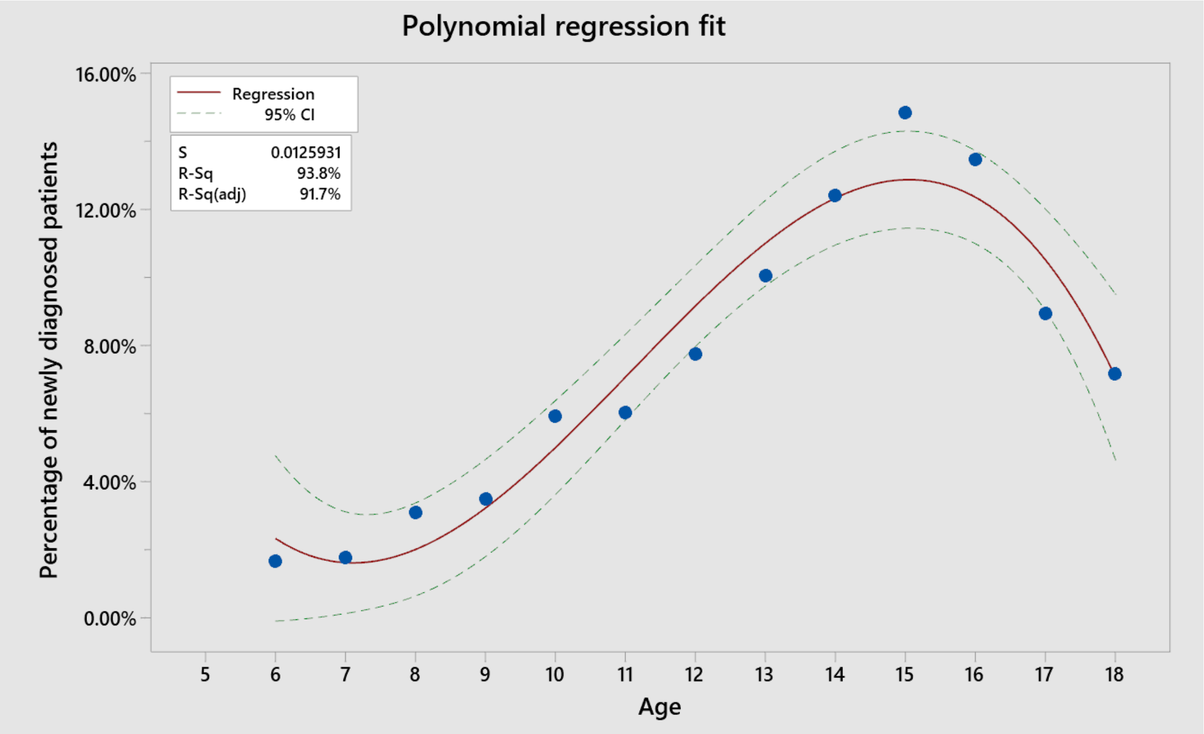


Figure 17. Fitting the incidence with the age adjustment.

The model was developed using the average of the observations from the literature, Royal London Hospital, and Inception Cohort to predict the expected PIBD incidence per age group.

A comparison of the predicted against the observed incidence values (**Figure 18**) showed that the performance of the model was acceptable, with an R squared exceeding 90% allowing us to use it for the adjustment of the incidence.

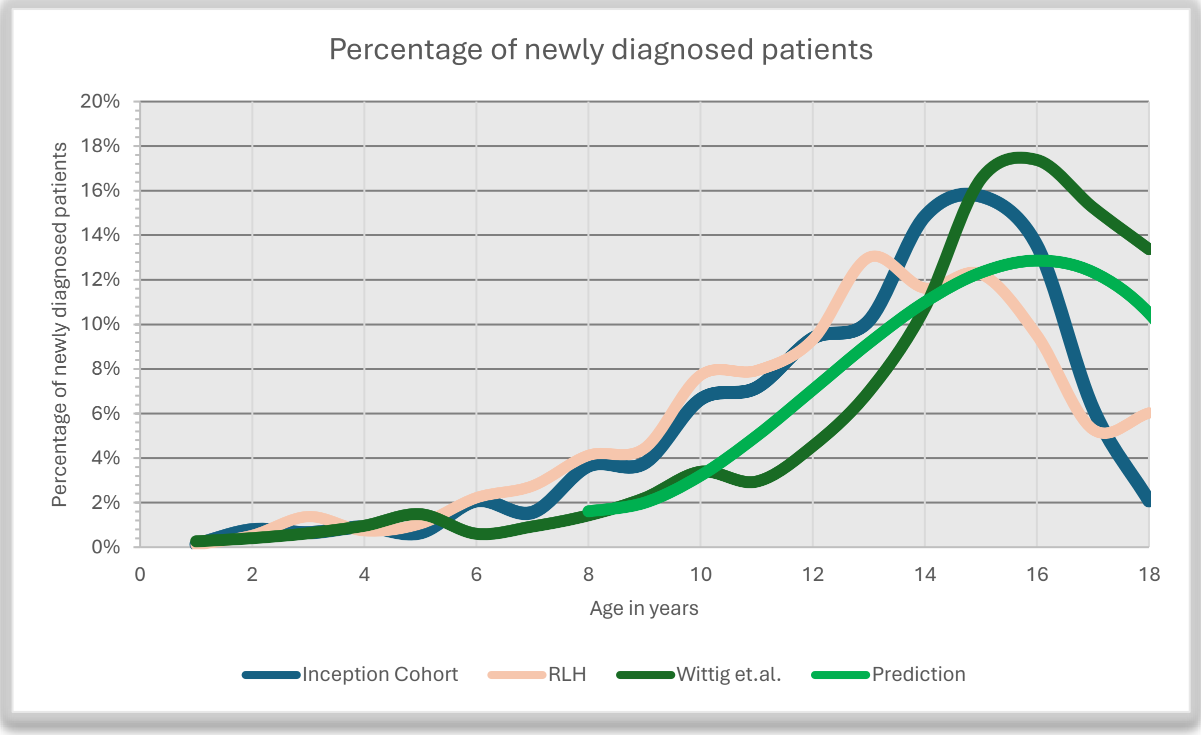


Figure 18. The probability distribution function of previous studies and available data compared to the developed model.

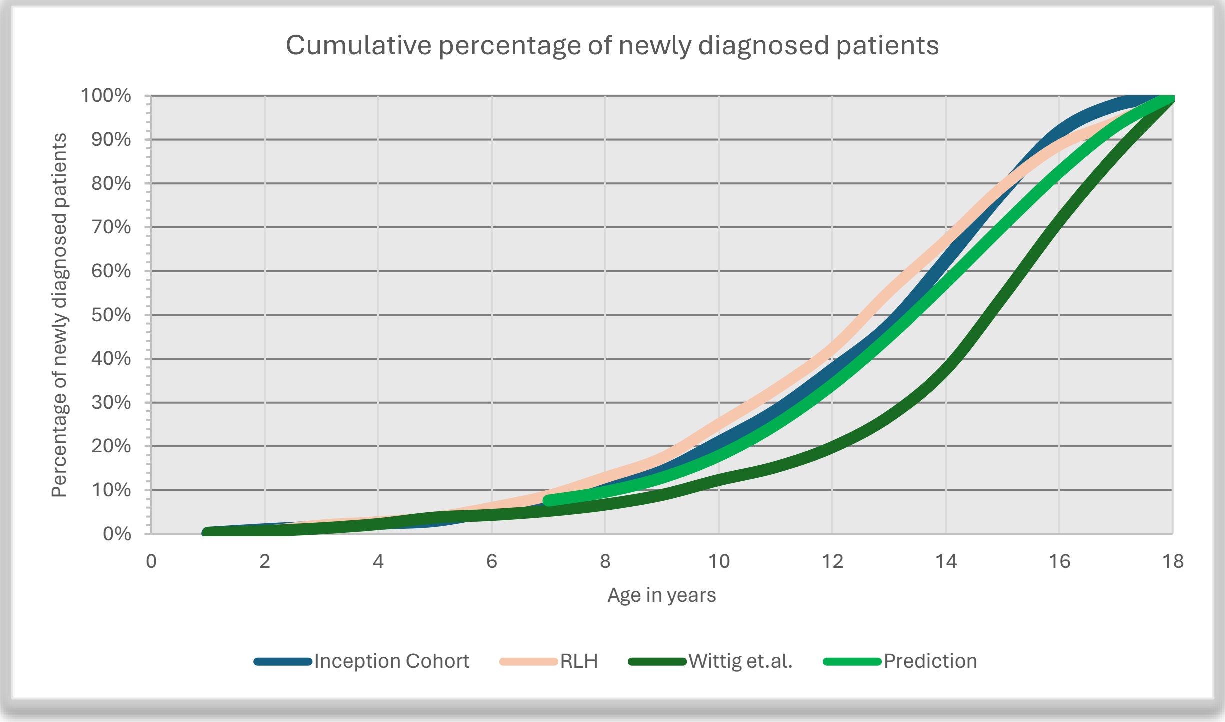


Figure 19. The cumulative distribution functions of previous studies and available data compared to the developed model.

Based on the developed model, we adjusted all results from clinics that were not using an 18 upper-age limit according to the **Table 8 Instructions with the steps required to adjust the reported incidence in centres that do not have 18 as the upper age limit. Table 8** below. This ensured that the results from each participating centre were now comparable.

Table 8 Instructions with the steps required to adjust the reported incidence in centres that do not have 18 as the upper age limit.

Please note that the corrections for the sites seeing patients older than the age of 18 assume that the incidence remains stable from 18 onwards.

Upper age limit	Adjustment required based on modelling	Step 2: Final Formula
7	13.13752791	Incidence x 13.14
8	10.40681928	Incidence x 10.41
9	7.798638149	Incidence x 7.8
10	5.613555363	Incidence x 5.61
11	4.020177474	Incidence x 4.02
12	2.938905296	Incidence x 2.94
13	2.221151822	Incidence x 2.22
14	1.743875065	Incidence x 1.74
15	1.424398328	Incidence x 1.42
16	1.211389397	Incidence x 1.21
17	1.074670171	Incidence x 1.07
18	1	Incidence x 1
19	0.985990005	Incidence x 0.986
20	0.973712469	Incidence x 0.974
21	0.962864753	Incidence x 0.963
22	0.953210830	Incidence x 0.953

2.2.2. Safety Registry analysis: Incidence and Prevalence calculation

The incidence and prevalence are both measures used in epidemiology to describe the frequency and distribution of a disease in a population. In our study, incidence referred to the number of new cases of IBD that occurred in the paediatric population up to the age of 18 per annum. In our study, this is expressed as a rate, and in almost all cases, we reported the PIBD incidence as the number of new cases per 100,000 individuals up to the age of 18 per annum. We also referred to this as 10^5 person-years.

Prevalence, on the other hand, refers to the total number of cases of a disease that exist in a population at a particular point in time. In our study, prevalence referred to the number of all IBD cases, new and previous, in the paediatric population up to the age of 18 at the data collection time. This was also expressed as a rate, and in almost all cases we reported the PIBD prevalence as the number of all cases per 100,000 individuals up to the age of 18. To calculate the incidence and prevalence, first, we collected information on the number of new and current patients seen by the PIBD experts participating in the Safety Registry in the last year. In the next step, we also collected information on the areas that each PIBD expert and their clinic were covering. The available options for the selected areas in Europe are based on the NUTS3, a nomenclature of territories discussed in the introduction and previous paragraphs. Using the available demographics from Eurostat, we have assigned the total paediatric population in each NUTS3 area, which allowed us to estimate the population that each clinic is covering. The sum of the paediatric population in the areas that a centre is covering is the denominator data for that centre. The denominator data allows the calculation of the incidence and prevalence per 100,000 individuals in the paediatric population. As shown in the results (3.1), the overlap of clinics that claimed the same regions as covered was low in our study and varied with the collection year. When a NUTS3 area is shared between two or more clinics, the denominator population was split accordingly. The incidence assigned to the shared regions was the average of the incidence calculated for the clinics that claimed the same region. Lastly, in several cases, when the same clinic submitted multiple responses due to multiple PIBD experts participating from the same site, one response was kept. The selected response was preferred based on the data quality, where responses that were based on local databases were favoured against estimates. Also, the reporting consistency of the PIBD experts and other metrics (including CD/UC and incidence to prevalence ratio) were also used. In some cases, the selection was based on correspondence with the experts.

2.2.3. Safety Registry analysis: Conversion of lattice data to centroids

To handle the multiple spatial misalignment issues in our data, there were instances whereby the aggregated patient data for each NUTS area needed to be converted to point data. The Inception Cohort disease phenotype analysis was the only analysis in this work that was based on the individual point data for each patient. All other analyses were based on the population centroids. The centroids were selected as the most appropriate representation of where the population lies within a polygon. The motivation for estimating the centroids is to identify the location with the minimum distance from all residents within a territory. By calculating the population-weighted mean centroids for each polygon region, we can obtain a point-location estimate that represents most of the population at risk within a region. Consequently, risk factors that are close to the centroids will also be close to most residents within that region and are expected to have a more substantial influence on the measured outcome and vice versa. This addresses the modifiable areal unit problem and grouping effect, in particular. By using this approach, for a small territory such as Westminster in London, most of the population is expected to be exposed to a risk factor within that region. However, for a large territory such as the Highlands in Scotland, the number of individuals that will be exposed to a risk factor within that region depends on the exact location of the covariate within this region (a pollutant in the north should have a minimal effect as it is located hundreds of kilometres away from the 95% of the population in that area). However, this method was used for particularly small areas based on postcodes in the Netherlands.

The general estimate for the median centre is calculated by the following function:

$$d_i^t = \sqrt{(X_i - X^t)^2 + (Y_i - Y^t)^2 + (Z_i - Z^t)^2}$$

Where:

- x, y and z are the coordinates for each feature
- i and t is each candidate location within the examined area

The centroid estimation is an iterative algorithmic process with each step (t) being a candidate centroid location (feature). The algorithm converges when the location (Xi, Yi, Zi) that minimises the Euclidean distance d to all features (i) is found. In our case, we are interested in the 2-dimensional centroids (longitude, latitude), and we also want to weigh the centroids according to the population density.

Therefore, the previous estimate becomes:

$$d_i^t = \sqrt{\left(((X_i - X^t)^2 + (Y_i - Y^t)^2)\right) * Wi}$$

Where:

- x and y are the coordinates for each location i
- W is the population density at each location
- t is each candidate centroid location within the examined area

2.2.4. Inception Cohort analysis: Obtaining the coordinates of patients

In the environmental questionnaire of the Inception Cohort study, the residence information was submitted by 549 patients. 80% of these patients were residing in the UK and the Netherlands, while the remaining 20% were based in eight different countries. Interestingly, 5 of the patients that were recruited in the study were residing in Spain (1), Austria (3) and Denmark (1), which are countries not participating in the study. Considering the small number of patients in some of the participating countries, the patient location was obtained only for the United Kingdom and the Netherlands.

In the Netherlands (NL), the postcodes are based on an alphanumeric format with four digits followed by two letters. Most NL-based patients followed the data protection study instructions and provided only the first part of their postcode. In the United Kingdom, the postcodes are also based on an alphanumeric format and contain two parts separated by a single space, the outward and the inward code. Similarly, to the patients in the Netherlands, most study participants in the UK provided only the outward code of their postcode. The environmental questionnaire provides patient-level data with location information. Each patient was assigned a map location, and the risk factor factors were subsequently added to this location. Thus, the dataset of the Inception Cohort is point locations of patients with individual patient and environmental exposure data. The UK postcodes were only available with coordinates, while the NL postcodes were available as areas. These areas were converted into centroids, and the following processing steps were identical for both countries. Each centroid was assigned the value of the underlying continuous surface of each rasterised pollutant. The information available in the NUTS format was also assigned to the overlapping centroids. The steps required to convert the Dutch lattice postcode data into point data and align them with the UK Inception Cohort patients are shown in **Figure 20** below.

It should be noted that this dataset was not used for the study of the PIBD incidence but for the study of the differences in the rates between Crohn's Disease and Ulcerative colitis grouped with IBD-U. As discussed in the introduction, several reports in the literature suggest that certain environmental factors affect these conditions differently, and this dataset offers an excellent opportunity to investigate these effects. The study of the PIBD incidence is not possible using the Inception Cohort dataset as the patient point data are insufficient without control data.

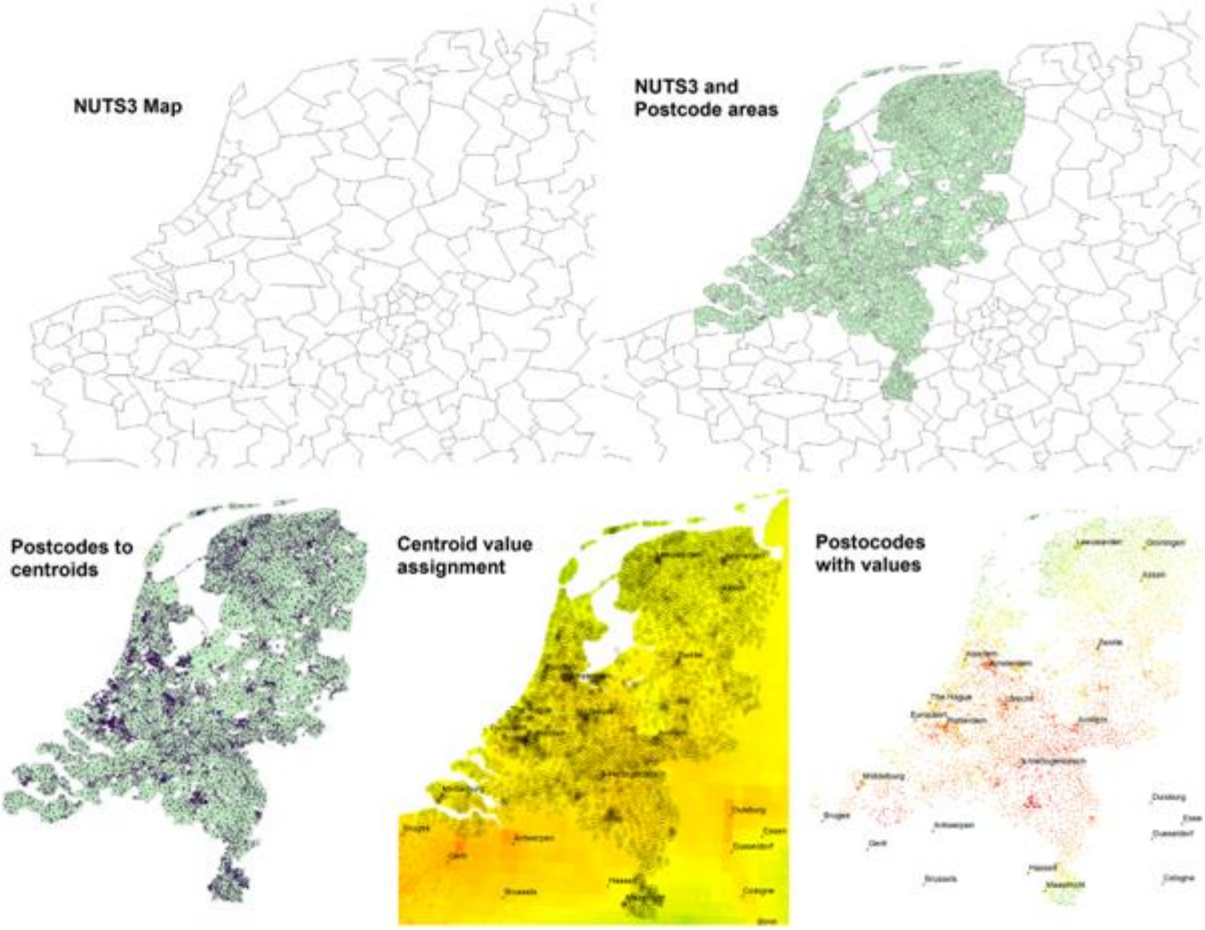


Figure 20. The steps of preparation of the Dutch patients with location information

This dataset was used in combination with the UK Inception Cohort data to analyse the disease phenotype. The steps from the first image on the top left are described as follows:
the base NUTS3 map is used; 2. the postcode areas are merged with the base map; 3. each postcode area is converted to a centroid; 4. the centroids receive the values of the interpolated pollutants; 5. the postcodes with the exposure information are extracted ready for the analysis. The last panel shows how each postcode centroid had been assigned a value for the exposure of the example.

2.2.5. Interpolation of pollutants

As the incidence information is collected using the NUTS3 lattice dataset format, it can be combined with several predictors that have also been collected on the same territory level. However, most environmental factors have been sampled or reported at specific point locations that are not aligned with the locations of the NUTS3 regions. Therefore, we needed to estimate the exposure of each population centroid or NUTS3 region to these factors. To estimate the exposure, we interpolated the predictors geographically. The literature suggests several interpolation methods, which may vary depending on the type of exposure that we wish to interpolate. The preciseness of the interpolation at a location without a measurement depends on several factors, including the distance from the other locations with available measurements, their number, variance, and geographical distribution.

Since collecting data from all locations within a study area to observe a phenomenon or measure a variable of interest is usually difficult or impractical, it is common to measure the quantity of interest at selected sample sites and use predicted values to estimate values at all other locations. These sampling sites may be distributed randomly or follow a specific sampling strategy. By combining the information from the measured locations and predicting the quantity of the variable of interest in the remaining locations, we can create a continuous surface of the variable. Since a truly continuous map would require an infinite number of spatial units, in this context, the term means that the spatial units with the variable values follow a regular pattern (grid) and are sufficiently dense to create an effectively continuous surface. This section will discuss some of the interpolation tools that allow us to make predictions from sample measurements for all the locations required to create a continuous measurement for the spatial variables of interest. The importance of interpolation for this project has been paramount as these methods allowed us to combine information from different sampling sites and with various spatial formats and geometries. Hence, spatial interpolation was essential in solving most of the spatial misalignment problems of our study.

Inverse distance weighted methods and spline

Inverse Distance Weighting (IDW) is a method of interpolation that estimates the value of a point based on the values of nearby points. The IDW algorithm assigns a weight to each nearby point based on the inverse of the distance between the target point and each data point. Points that are closer to the target point have a higher weight and thus have a greater influence on the final estimate. IDW is commonly used in environmental modelling to estimate values for un-sampled locations.

A spline is a piecewise polynomial function to approximate a set of data points. The polynomial functions used in spline interpolation are chosen to minimise the overall curvature of the spline. This results in a smooth and continuous function that passes through or closely approximates the set of data points. There are several types of splines, including natural splines, cubic splines, and B-splines, each with different properties and use cases.

Kriging

Kriging is a method of spatial interpolation that estimates a point’s value based on nearby points' values. The method is based on spatial autocorrelation, which is the idea that nearby points are more similar than points that are farther apart. Kriging uses statistical models to estimate the spatial structure of the data and uses this information to make predictions about the values at unsampled locations. In our study, Kriging was used to create continuous surfaces from point data, such as pollution data from several sampling locations. Kriging can be classified into two main types. The most common type is Ordinary Kriging (OK) which assumes the mean of the underlying studied variable is constant over the area of interest.

The formula for ordinary Kriging is as follows:

$$Z(u) = \hat{Z}(u) + \lambda(u)\ (Z(u) - \hat{Z}(u))$$

Where:

- Z(u) is the estimated value of the variable at location u

- $\hat{Z}(u)$ is the expected value of the variable at location u
- $\lambda(u)$ is the kriging weight for location u

In contrast, Universal Kriging allows for a trend in the data, and it is used when there is a linear or non-linear trend (hence it is called universal) in the studied variable. The formula for universal Kriging is similar to the ordinary Kriging formula but includes a trend component:

$$Z(u) = \hat{Z}(u) + \lambda(u)\,(Z(u) - \hat{Z}(u)) + T(u)$$

Where:

- $T(u)$ is the trend component of the variable at location u

In both cases, the estimates of $\hat{Z}(u)$ and $\lambda(u)$ are obtained from the spatial structure of the data using a statistical model such as a variogram. The trend component, $T(u)$, is usually obtained using a regression model or a polynomial function.

Kriging is considered one of the most accurate interpolation methods available, but it can be computationally intensive and requires a sufficient number of data points to produce reliable results. In our work, although Kriging was favoured in some cases of spatial interpolation. However, it was rejected in case that the interpolated variables were returning problematic variograms.

A variogram is used in Kriging as a statistical measure that describes the spatial variability of a variable. It is used to model the spatial correlation structure of the variable of interest, such as an air quality measure. A variogram is calculated as the variance of the differences between values at two locations, as a function of the distance between those locations. In other words, it tells us how much the variable's values change as the distance between two locations increases. The result is a graph showing how the variable's variance changes with distance. The variogram is typically calculated by selecting pairs of data points that are a certain distance apart and calculating the variance of the differences between their values. These variances are then plotted against the distance between the points. The resulting graph is called the experimental variogram. The experimental variogram is then fitted to a mathematical model that describes the spatial structure of the variable. The most commonly used models are the parametric spherical, exponential, and Gaussian models. The parametric nature of these models allows for a structured approach to modelling spatial correlation. The choice of model depends on the data and the specific application. The exponential covariance model decreases the correlation between points at an exponential rate as the distance between them increases. It is defined as:

$$C(h) = \sigma^2 \exp\left(-\frac{h}{\varphi}\right)$$

Where:

- $C(h)$ is the covariance for lag distance h, σ^2 is the variance, and φ is the range parameter.

The exponential model is suitable for processes with a gradual decline in spatial correlation. It is often used when the spatial process is believed to be more random or when the data exhibits a rougher spatial structure. The squared exponential covariance model, also known as the Gaussian covariance model, decreases the correlation between points at a rate that is proportional to the square of the distance between them, leading to a smoother decline in correlation. The definition of this model is:

$$C(h) = \sigma^2 \exp\left(-\frac{h^2}{2\varphi^2}\right).$$

The squared exponential model is appropriate for processes with a smooth spatial variation, where changes occur more gradually over space. Subsequently, this model assumes that spatial correlations diminish more smoothly compared to the exponential model. Lastly, the spherical covariance model unlike the exponential and squared exponential models, introduces a hard limit beyond which there is no spatial correlation. This makes it particularly suitable for datasets where spatial dependence is very likely present up to a certain distance and then drops to zero. The definition of this model is 0 for $h > \varphi$, with φ to be the range limit for the lag distance h. Beyond α, we assume no spatial dependence is present. For $h < \varphi$, the definition becomes $C(h) = \sigma^2(1 - \frac{3h}{2\varphi} + \frac{h^3}{2\varphi^3})$. In practice the model selection for each pollutant was based on the visual inspection of the variogram, the prediction and accuracy metrics of each model and expectation of the spatial extent of the correlation of the processed variable. Once the variogram model is chosen, it can be used to estimate the values of a variable at unsampled locations using interpolation methods such as Kriging. In the following subchapter, a thorough example of Kriging application is shown.

Empirical Bayesian

Empirical Bayesian interpolation (EBI) is a method that combines the principles of Bayesian statistics and spatial interpolation to estimate the values of a variable at unsampled locations. It is instrumental when the data is sparse or has a high degree of measurement error. This method works by first estimating the parameters of the spatial interpolation model using the observed data and then using these parameters to make predictions about the variable at unsampled locations. The key difference between EBI and other interpolation methods is that EBI estimates the spatial model parameters using a Bayesian framework, allowing for the incorporation of prior information and uncertainty. The process of EBI can be broken down into three main steps. The first step is to estimate the parameters of the spatial model using the observed data. This is done by specifying a prior distribution ($p(\theta)$) for the parameters and then using Bayesian inference to update the prior distribution based on the data. The posterior distribution of the parameters is given by:

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)}$$

Where:

- θ is the vector of spatial model parameters
- y is the vector of observed data
- $p(y|\theta)$ is the likelihood function, which describes the probability of the observed data given the parameters
- $p(\theta)$ is the prior distribution of the parameters
- p(y) is the marginal likelihood, which is a normalising constant

Once the spatial model parameters have been estimated, the variable can be predicted at unsampled locations using interpolation techniques such as Kriging. In the final step, we estimate the uncertainty or error in the prediction. This can be done by simulating the model with the estimated parameters and comparing the simulated values to the observed data. EBI is a powerful method that can provide more accurate and reliable predictions than traditional interpolation methods, mainly when the data is sparse or has a high degree of measurement error. In our simulations, Empirical Bayesian Kriging (EBK) was the best performing algorithm for certain datasets.

Additional techniques

Triangulation with Linear Interpolation (TLI): This method uses the Delaunay triangulation of the known data points to create a set of non-overlapping triangles. The variable's value is then estimated at the unsampled location by interpolating between the values at the vertices of the triangle containing the location.

Natural Neighbour Interpolation: The natural neighbour interpolation method is based on using the neighbours closest to the unsampled location to estimate its value. It uses the Voronoi diagram of the known data points to estimate the variable value at the unsampled location.

Radial Basis Function (RBF) interpolation: This method uses a set of basis functions, such as Gaussian or inverse multiquadric, that are centred on the known data points. The values at the unsampled locations are estimated by a weighted sum of the basis functions.

Interpolation example, sun irradiance with Kriging

In this section, I describe the required steps for the interpolation of the average solar radiation in the United Kingdom as the average recorded values for June. The gridded dataset was extracted using the NASA APIs for R and included the average satellite measurements captured from 2010 to 2018 (**Figure 21**). These measurements were used for our interpolations. Radial basis functions, Empirical Bayesian Kriging, Inverse Distance Weighting, simple and ordinal Kriging were some of the methods we used to interpolate the sun exposure. Special attention is given to the kriging methodologies that we review in detail. The final model was selected under four criteria; overall accuracy, avoiding the use of a deterministic model, considering the autocorrelation and our ability to incorporate additional information to maximise the validity of each interpolation (**Figure 22**).

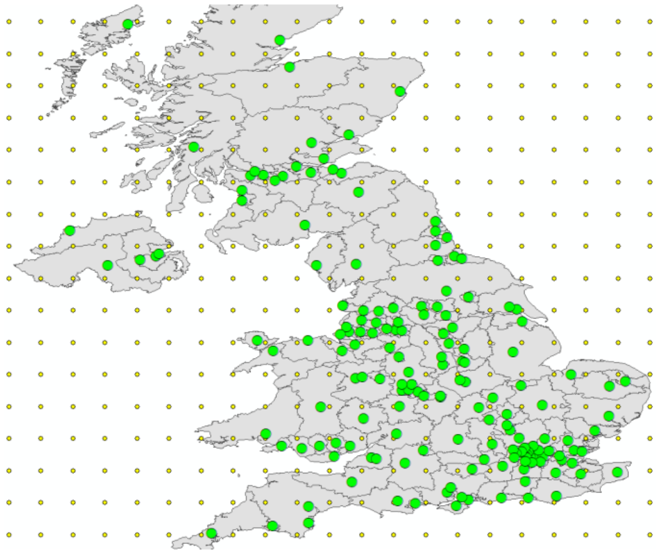


Figure 21. Preparation of the sun exposure data from NASA.

The green dots are the centroids of each NUTS3 area, and the dots on the grid are the locations with extracted sun radiation exposures used in the interpolation.

i) In completely deterministic models, one can interpolate by estimating a continuous surface which is a simple calculation of an unknown function. In contrast, a stochastic process also allows for errors and uncertainty in our estimates. The latter is preferred, especially for the dataset of pollutants where regional variations can be substantial. It is more beneficial to consider such variations as random fluctuations, which result in the addition of the error in the deterministic function, which is equivalent to the mean of the stochastic function, assuming that the error is randomly distributed. The three best-performing methods satisfy these criteria.

ii) The spatial autocorrelation can also hold vital information that we must use for the final estimation of the interpolated surface. To underline the importance of spatial correlation, we can compare noise pollution to carbon dioxide. Both sources of pollution often originate from the same sources, such as traffic; however, the way they are distributed in space is quite different. After analysing the spatial correlation of the CO2, we can see that a sampling location can easily be interpolated in the surrounding area for a few kilometres, while our approach to noise pollution will be very different.

iii) Different predictors, such as carbon monoxide and carbon dioxide, frequently have similar characteristics, sources and geographical distributions. Therefore, combining the information from different sources may improve the accuracy and preciseness of the interpolation.

In the following paragraph we show the results from four different interpolation methods (**Figure 22**) and emphasize on the Kriging methodology.

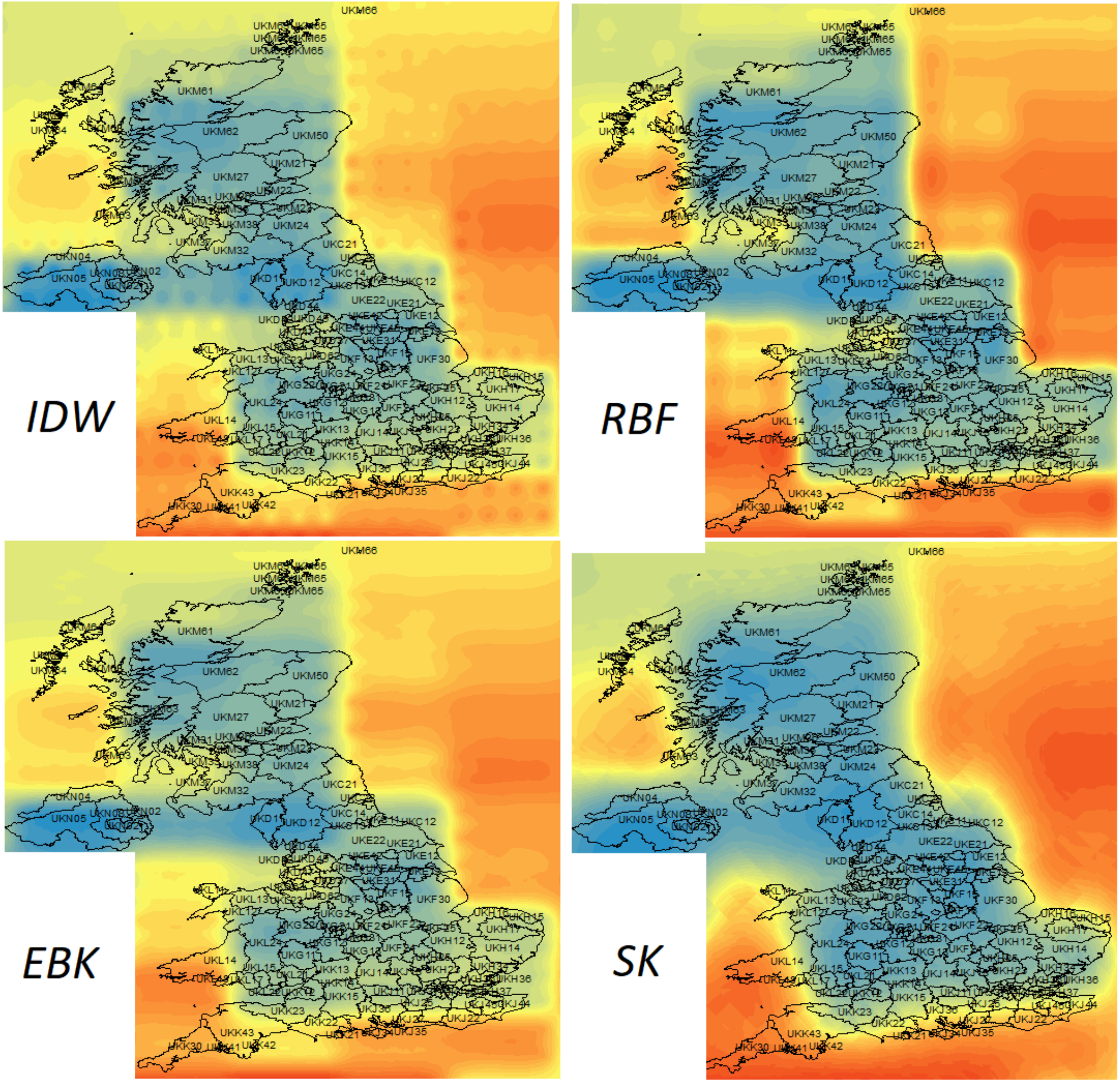


Figure 22. Comparison of four different interpolation approaches for solar radiation in the UK.

The interpolation results of the Inverse Distance Weighting (IDW), Radial Basis Functions (RBF), Empirical Bayesian Kriging (EBK) and Simple Kriging (SK). Please note that the final interpolation of sun exposure was performed with a higher resolution dataset (a more dense grid extracted from NASA).

The basis of the Kriging approach is that we treat spatial variability as a model with two main components: the trend on a large scale and spatial autocorrelation on a smaller scale. Therefore, the most simplistic calculation for this is:

$Z_s = \mu_s + \varepsilon_s$ (1)

Where:

- Z_s is the value at the location s
- μ is the conditional mean
- ε is the error.

In a general least square setting, we can obtain the prediction and error, and in combination with the prediction errors, we can estimate the following:

$$y^* = \mathrm{x}^*\, b + c'C^{-1}(y - Xb)(2)$$

Where:

- the b term is the coefficient(s) from the model
- y* and x* are the outcome and predictor(s) values at the new unknown locations
- $(y - Xb)$ is a vector with all the residuals from the regression (mean of 0) which is pre-multiplied by C^{-1}
- C^{-1} which is the inverse of the variance-covariance matrix (precision metric).
- This term of the equation is pre-multiplied by c', which is a row vector of covariances between the error at the new locations (e*) and the observed residuals (e).

It should be noted that the residuals at the new locations (e*) are unknown; however, we do know how they covary, given that they covary as a function of distance. The latter is estimated using information from the Covariogram. Hence, we can calculate the term c' in (2) for every unknown location if we know its distance from the known locations and have an available variogram for our dataset. In other words, assuming that some autocorrelation is present in the residuals, we exploit this additional structure in the error term to obtain the covariance between the observed and predicted points.

Ultimately, the simple Kriging prediction at a point Si is given by the expression:

$$Prediction\colon Y(s_i) = \mu + c(s_i)'C_i^{-1}\varepsilon_i(3)$$

Where:

- Y(S$_i$) is the value of interest at the location
- μ is the constant mean
- c', C and ε, as described in (2)

Furthermore, the prediction variance is defined as follows:

$Prediction\ variance\colon \sigma_p^2 = \sigma^2 - c^T(s)C^{-1}c(s)$ (4)

As we can see from (4), the larger the covariance, the higher the precision of the estimate since it will reduce the variation and the standard error for each prediction. Expanding on that, having a robust and well-characterised autocorrelation present and being close to the sampled locations will increase the preciseness of the predictions at each new unsampled location. In contrast, higher variance in the dataset and being farther apart from any sampled locations will inflate the standard error of the predictions.

$SE\colon \sigma_i = \sqrt{s - c'(s_i)C_i^{-1}c(s_i)}$ (5)

In terms of the Semi-variogram calculations:

$(semi)Covariogram\colon C(h) = s - \gamma(h; r, s, a)$ (6)

Where:

- h is each distance and
- r, s, and α are the parameter estimates (discussed in the following paragraphs)

The semi-variogram for the selected method, simple Kriging (and ordinary Kriging), when used to estimate the solar radiation surface in the UK from the gridded dataset is shown in **Figure 23**. The first step is the estimation of covariance depending on the distance, which reveals strong spatial correlation patterns, as we can see in the following semi-variogram. As already mentioned, the semi-variogram is an important tool that helps us measure the spatial dependence and provides the parameter estimates that explain the exact autocorrelation mechanisms, which may differ for each studied predictor. The rationale of a variogram is that the relative location of two points may influence their spatial relation in addition to the effects from their absolute geographical location.

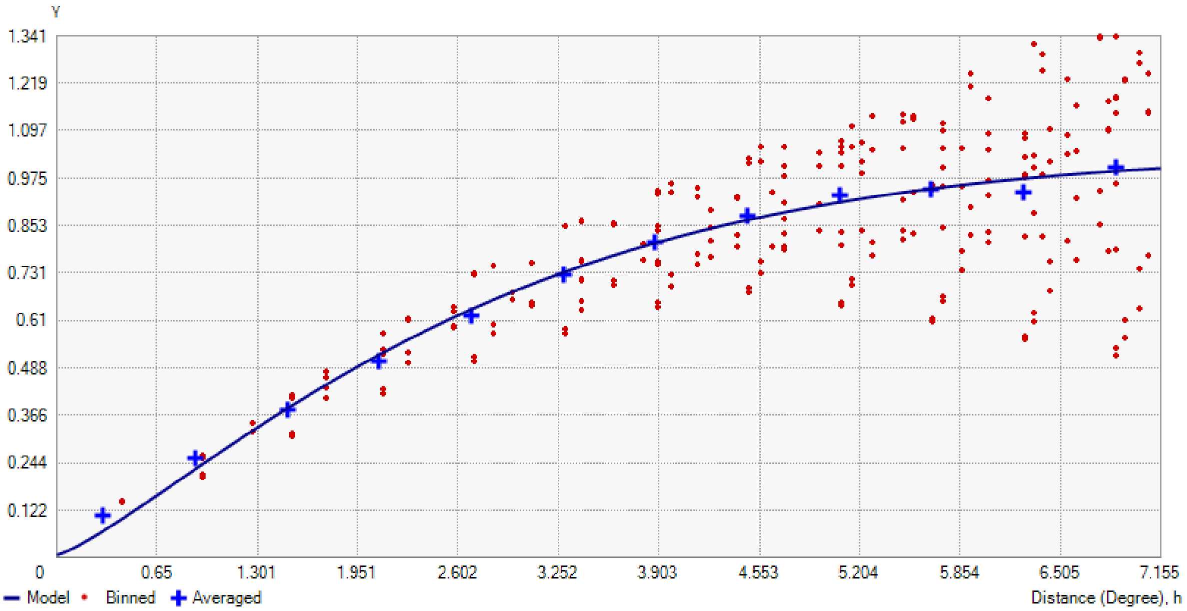


Figure 23. Semi-variogram of solar radiation.

The covariance increases with distance, suggesting, as expected, that the adjacent locations tend to share similar values.

Before producing the components of the semi-variogram, we need to follow the steps of creating a correlogram. For a set of locations on the map and a given direction, we can calculate the correlation coefficient between the data points at each step of distance increment (lag h=0,1,2,3...). Understandably, this will be 1 for h=0 and should decrease as we move further away from the data points.

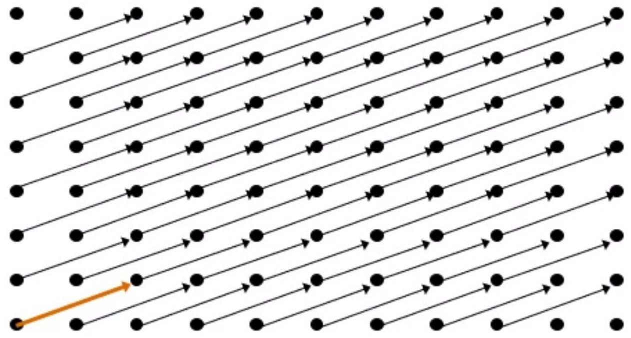


Figure 24. Data points on a regular grid in connection with their pairs at a distance (h) for a given direction.

This is used to develop a correlogram. Please note that one arrow is equivalent to one lag, two arrows in a row are equivalent to two lags and so forth.

Expanding this on an irregular grid, we proceed on the same principle by adding additional parameters. These are the angle tolerance, as shown in **Figure 25**, the direction of the vector, the lag distance, and the lag tolerance, which is usually ½ of the lag distance. Lastly, we also define the bandwidth as the distance that limits the surrounding points that will be included in the calculations. **Figure 25** below shows the lag vectors with a specific start and end (tail and head, respectively). Each vector has a length (h), and for the location of the tail (u), the location of the head is u+h. Hence, the difference in the data values of interest between the head and the tail will be z(u)-z(u+h).

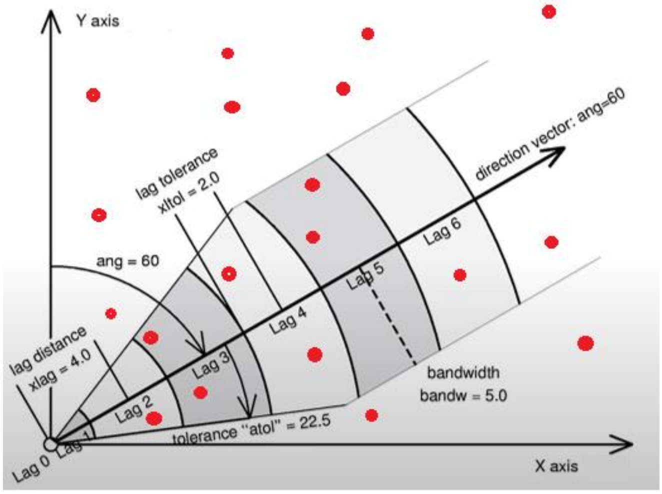


Figure 25. The lag vectors have a specific start and end.

An irregular grid of points (red), the selected rules and tolerance settings we use to estimate the correlations between the data points.

Therefore, this is the motivation for developing the semi-variogram, which is calculated as follows:

$$Dissimilarity\ \gamma(h) = \frac{1}{2N(h)} \sum_{a=1}^{N(h)} \big(z(u_i)\big) - z(u_i + h|x)^2$$

This is essentially, the average squared difference over lag distance (h) for all possible pairs of data, halved. This calculation returns the semi-variogram, as shown in **Figure 23** for the example of solar radiation. The parameters that we must obtain from the semi-variogram are; the sill, which is the maximum value of the semi-variogram, and represents the level of spatial autocorrelation that is reached at a distance beyond which there is no further spatial dependence, the range, which is the distance h at which the sill is observed (when we move far enough to safely assume that correlations are present) and the nugget effect which is observed when the level of variability is very high at a very low h values (h close to 0). Often, the latter can be a measurement error as well.

With all the requirements in place, the kriging process can now be run for each point on the map. For an unsampled location, simple Kriging will return the global mean with the necessary adjustment (based on what was described in the previous paragraphs) according to the surrounding observations and the autocorrelation information. **Figure 2** shows an example of this estimation when five observations are available at a close distance from the unsampled location. If no observations are available, the prediction will be the global mean.

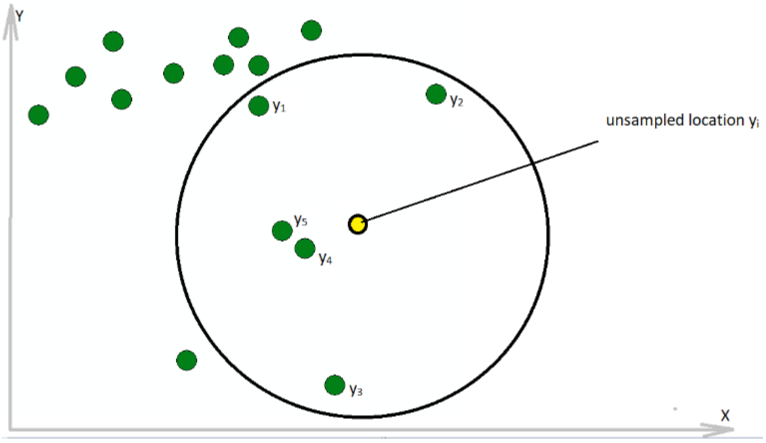


Figure 26. Kriging interpolation based on the adjacent locations.

The Kriging interpolation will estimate the value of the unsampled location using every data point within the predefined bandwidth.

Simple Kriging is rarely used in practice, especially for irregular point data. The Kriging method that is most frequently used is ordinary Kriging, which calculates the local mean instead of the global mean for each prediction. **Figure 27** shows the difference in the performance of the two interpolation methods when used on the Nitrous oxide point data for the Greater London area.

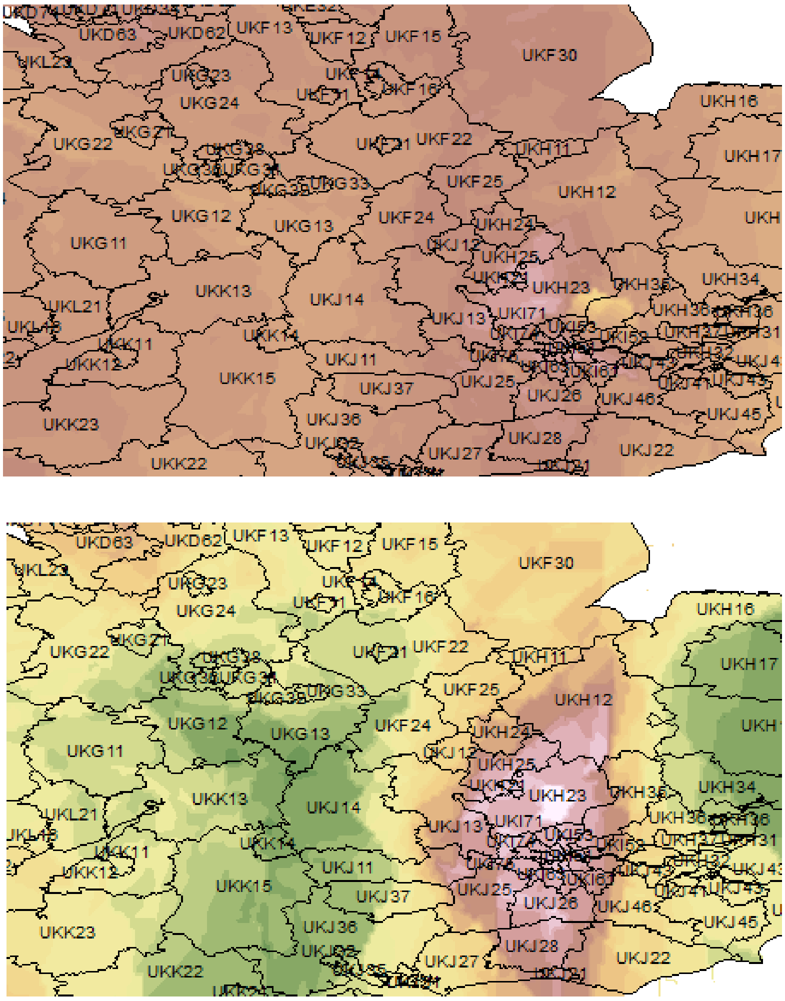


Figure 27. Comparison of simple (universal) Kriging (top) with ordinary Kriging.

Interpolation of the Nitrous monoxide air samples taken in London and its surrounding areas between 2018 and 2019.

Returning to the sun radiation example, the final estimates using ordinary Kriging are very similar to the surface provided by the governmental bodies regarding the solar radiation in the UK. This demonstrates our ability to convert point data to a continuous surface with prediction error information. The accuracy of our methods was subsequently assessed with simulations as discussed in the simulations section.

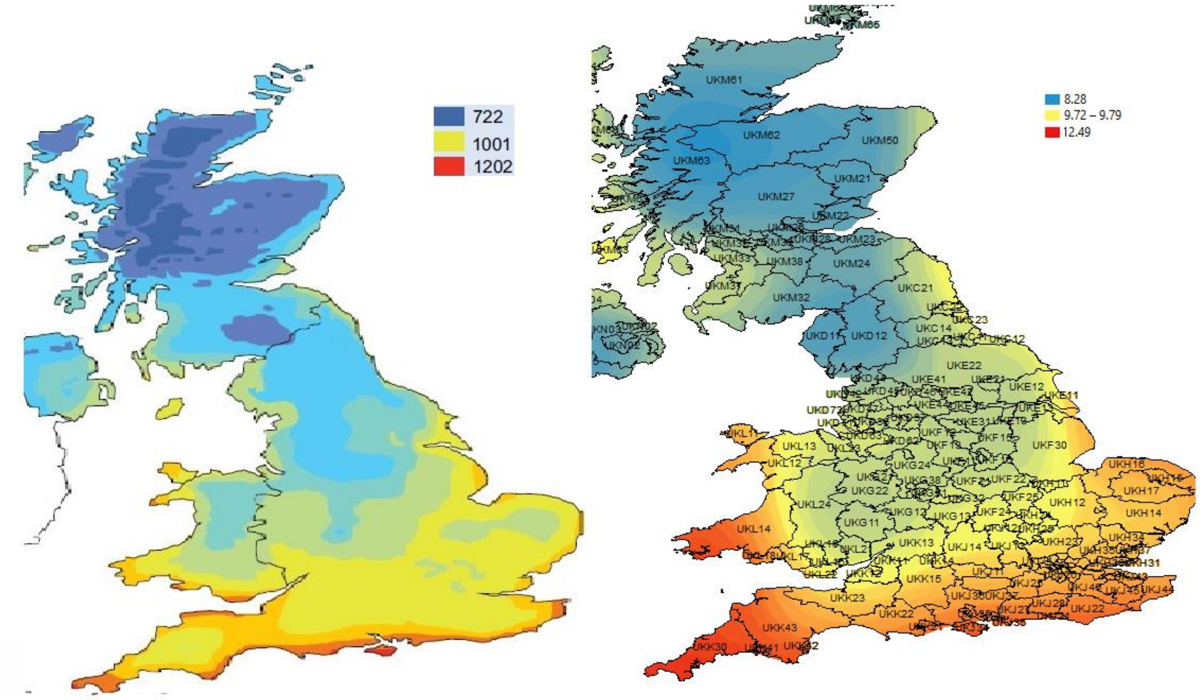


Figure 28. The UK sun irradiance map provided by Microgeneration Certification Scheme (left) compared to our estimates using interpolation methods (right).

This figure shows the high similarity between the interpolated and the map that we consider the ground truth for solar sun irradiance.

2.3. Spatial disease mapping and analysis methods

This chapter contains the background and some considerations regarding common issues and statistical models that were encountered repeatedly during the data analysis. The following paragraphs are complementary to concepts covered already, such as autocorrelation, outlier handling, funnel plots and mapping methods.

2.3.1. Aggregation methods

To analyse the epidemiology of a disease and, in several cases, to inform and influence health policies, we need to understand its geographic distribution. Frequently, this requires aggregating our findings within each area included in the overall region of interest. Although this step seems straightforward, it may introduce bias as the spatial aggregation level used may alter the results. Roquette et.al. tested the effect of the aggregation level in the analysis of the geographical distribution of cancer mortality (Roquette et al., 2018). The researchers included three levels, NUTS3, municipalities (10-fold increase in the area count) and parishes (additional 15-fold increase in the area count). Not surprisingly, the findings supported that a lower number of areas and higher spatial aggregation led to more reliable results with the caveat of a decreased capacity to identify small local clusters as larger geographic territories are more likely to mask the underlying heterogeneity. However, in terms of clustering and based on the Moran's I test, the mortality was found to be clustered for all three spatial levels that were used in the study. Burghardt et al., 2022; Jeffery et al., 2014 and others also conclude that the higher level of aggregation may mask certain phenomena and interactions while it improves the reliability and robustness of the disease distribution. According to the current literature, a specific proper scale for mapping disease is absent. The decision must be study specific, balancing the trade-off between increased noise and bias. In our case and based on the simulations in 2.4.3, given the very high level of detail in our predictors, a low level of aggregation is desired, and a greater number of regions increases the statistical power of the study.

2.3.2. Spatial Weights

Determining the spatial weight is a necessary step for the Moran's I value estimation. As mentioned repeatedly in the literature, Tobler's first law of geography states that *"everything is related to everything else, but near things are more related than distant things"* (Tobler, 2004, 1959). This is the motivation for developing spatial weights matrices that can be used not only for Moran's I estimation but for other analytical processes, including a great number of spatial regression models that we will discuss in the next subchapter. Spatial weights are essentially model-specific adjacency definitions that describe the cases to consider neighbouring in our analysis. The spatial weights matrix W is an $N \times N$ dimensional table with Wij elements specifying the connection between each pair of the units i and j.

$$W = \begin{bmatrix} W_{11} & W_{12} & \dots & W_{1n} \\ W_{21} & W_{22} & \dots & W_{2n} \\ W_{n1} & W_{n2} & \dots & W_{nn} \end{bmatrix}$$

This matrix describes all possible connections and all the diagonal elements to be set as 0 (diagonal elements of W_{11}, W_{22} ... W_{nn} refer to single units and not neighbouring ones).

Any predictor x should be treated as a vector that must be multiplied by the spatial weight matrix and produce a spatially lagged variable vector.

$$W = \begin{bmatrix} W_{11} & W_{12} & \dots & W_{1n} \\ W_{21} & W_{22} & \dots & W_{2n} \\ W_{n1} & W_{n2} & \dots & W_{nn} \end{bmatrix} \times \begin{bmatrix} x_1 \\ x_2 \\ x_n \end{bmatrix}$$

The Rook weights are the simplest variant of the spatial weight’s matrix. The matrix reflects the status of a neighbouring relation in a binary format using weights of 1 and 0. The presence of a common edge (boarder) connecting two spatial units serves as the rook criterion's definition of neighbours, as shown in **Figure 29**.

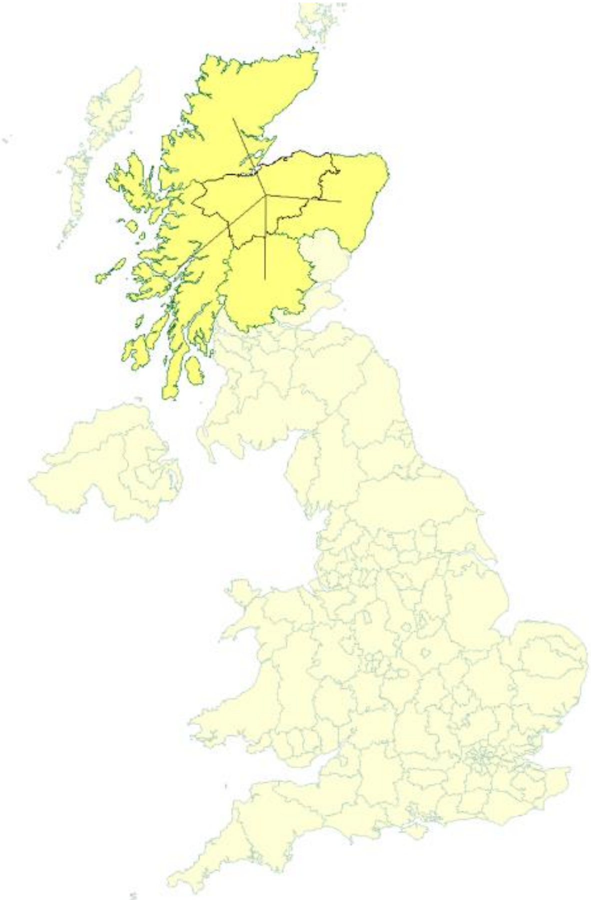


Figure 29. An example of Rook weights for the NUTS3 unit of Inverness, Nairn, Moray, Badenoch, and Strathspey

In this example, the Rook and Queen methods returned the same matrices containing 770 connections for the United Kingdom in total.

An expansion of the Rook weights is the Queen weights. Depending on the type of polygons on a map, the Queen's criterion may include other relationships. The neighbouring areas are defined as spatial units with a common edge or vertex where two or more polygon edges meet.

This may double the number of non-zero weights for a grid map, but it has little or no impact in a more real-world scenario.

Another popular weight-assigning method is the k-nearest neighbours (KNN), where every spatial unit must be linked to a specific number of its closest neighbours (**Figure 30**). Unlike the Rook and Queen methods, KNN considers the proximity of the areas prioritising the closest "k" spatial units (k is the number of units that will be assigned weights for their relationship with each unit in the sample). Most commonly, the distances are estimated using centroids. In this context, a centroid is the geometrical centre of each polygon area (we previously looked at population centroids).

Depending on the analysed outcome, not restricting the distance between areas that will be assigned weights may lead to inaccurate results. As shown in **Figure 30a**, several islands are receiving weights that link them with other islands and remote areas, which may not reflect the true connections between these areas. Also, as shown in **Figure 30b** and **Figure 30c**, when the spatial unit size changes depending on the location, the proximity of the k number of spatial units for certain locations might also vary significantly. When adjacency is a critical factor (for instance, a policy of an area affecting a policy of its neighbouring areas), this approach is appropriate. In contrast, this approach may be invalid when the distance is the key factor (for instance, the contamination of the air from pollutant emissions). Using large weights extensively can also reduce the power of the Moran's I test as the covariance term of the test statistic will be inflated in areas with no true spatial relationship thus diluting the effect of autocorrelation from the areas with true spatial relationship.

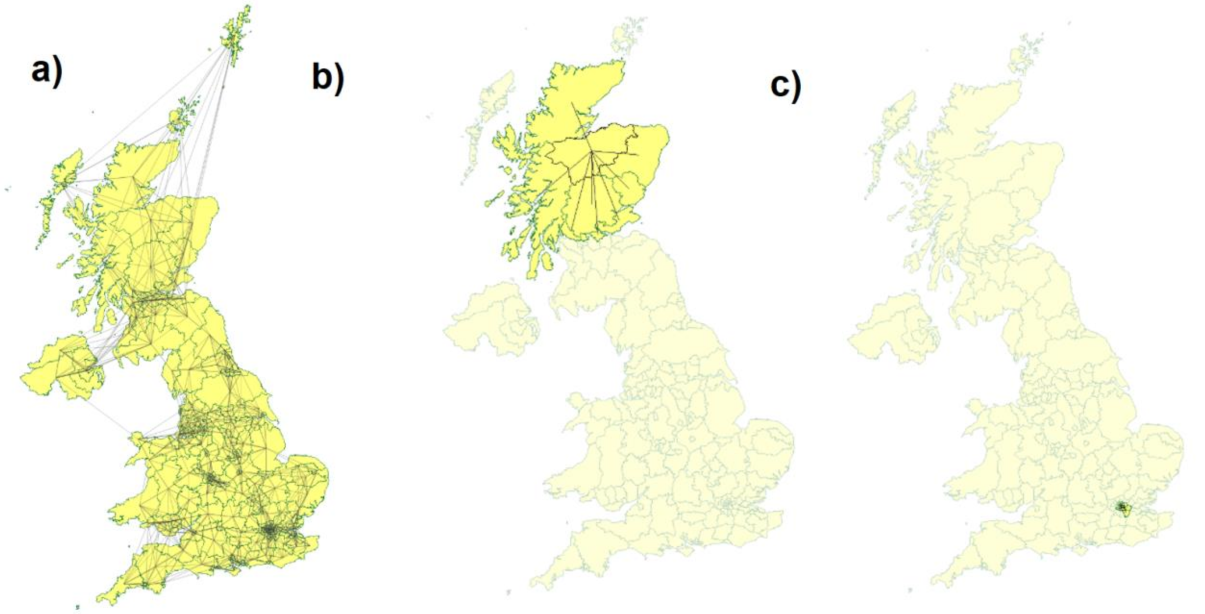


Figure 30. A visual presentation of the KNN spatial weight matrix appears as a network of spatial weights linking centroids of different areas.

a) The k number was set at 8, resulting in 173 x 8 = 1384 connections for the 173 NUTS3 areas b) and c) Examples of the KNN spatial weights for the NUTS3 of Tower Hamlets in London and the NUTS3 including Inverness in Scotland. The size difference between these territorial units and their neighbours results in connections with significantly variable distances.

A popular method that considers the distance between the spatial units proportionally is the inverse distance weights. In this approach, the weights are essentially transformations of the distances between the centroids, while we can also set a distance threshold for distant areas to receive a weight of 0 (**Figure 31**).

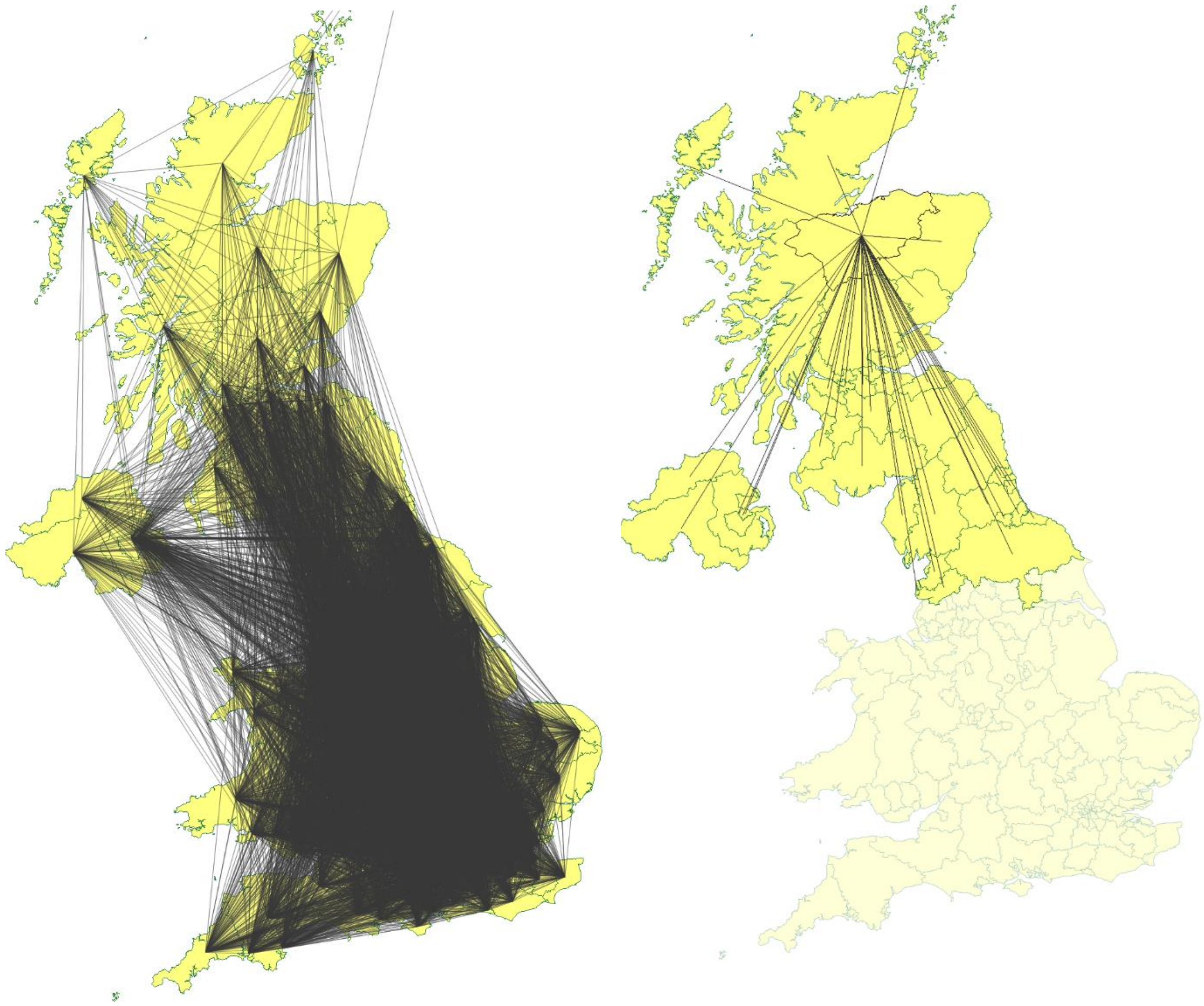


Figure 31. A visual presentation of the inverse distance weights matrix appearing as a network of spatial weights linking centroids of different areas.

In this example, the image on the left (**Figure 31**) shows the number of connections, which is 23,986, for the 173 NUTS3 areas. Each spatial unit has 138.65 no zero weight, suggesting that for the selected distance cut-off, each spatial unit was linked with 4 out of 5 available spatial units on the map. The figure on the right shows the NUTS3 territory including Inverness in Scotland and the non-zero weights for the areas that fall under the proximity threshold. Each assigned weight is calculated as the inverse of the distance.

The weight calculations for any distance weight method (inverse, KNN, Kernel) depend on the geodetic datum. For maps using projected coordinates (i.e., Mercator projection, see 1.5.1), the

Euclidian distance should be used, while for spherical, ellipsoidal coordinates, arc distance or similar approaches should be used.

2.3.3. Spatial Regression models

For any dataset $\{y_i, x_i\}_i^n$ with y as the outcome and x as the predictor variable, the simple linear regression model would be written as $y = x\beta + \varepsilon$ or $y = \beta_o + x\beta + \varepsilon$, with ε in $\{\varepsilon_i\}_i^n$ as the error term (a residual for each observation). It is important to note the error term is the random difference between the observed and predicted outcome that the predictor values cannot explain and the coefficient β (slope) is the effectively the effect size of the predictor on the outcome (for a unit change of x). When significant spatial relationships are present in the data, we expect to observe emerging patterns in the residuals when fitting the model. These patterns will indicate dependence in the residuals, which may be attributed to spatial dependence. Non-independence in the residuals invalidates the fit and suggests that the analysis can improve significantly by switching to a spatial regression model. A great variety of spatial models can address most issues of spatial dependence and autocorrelation while improving the fit with the inclusion of additional information. In this section, we will discuss the Spatially Lagged X model (SLX), the Spatial Lag model (SAR) and the Spatial Error Model (SEM).

The SLX model can be written as: $y = x\beta + Wx\theta + \varepsilon$, where W$x$ is the mean of x (predictor) of the neighbouring locations based on the prespecified spatial weights W. θ is the coefficient (slope) that accounts for the spatial dependency where the values of the predictor in the neighbouring locations can affect the value of the outcome in the location that is surrounded by these neighbouring locations. This relationship is described by the weight matrix and the value of θ. An example of the relevance and importance of this approach to this study is the case of air pollution effects. When investigating the link between disease incidence and pollution, we must factor in that even areas with minor sources of pollution, may still be affected by high pollution levels from their surrounding areas. This also explains why this model is categorised as a local spatial model in contrast to the SAR and SEM models.

The SAR model can be written as: $y = \rho Wy + x\beta + \varepsilon$, where W$y$ is the mean of y (outcome) of the neighbouring locations based on the prespecified weights W. ρ is the coefficient (slope) that accounts for the spatial dependency of the outcome since its values in the neighbouring locations can affect the outcome value in the location that is surrounded by these neighbouring locations. Based on the weight matrix and the adjacency definitions, ρ describes the relationship of the outcome in location i with the outcome in the surrounding and neighbouring locations. An example of the SAR model application is the study of hospital admissions per region. An area with low rates of the disease related to hospital admissions that are adjacent to areas with high disease incidence and admission rates is likely to be affected. This effect is described by many as "spatial spillover” (Li and Lv, 2021), where the observed outcome in some areas is not affected by the predictors, and it can be explained by autocorrelation effects. In contrast to SAR, this is a global spatial model as the value of y_1 may affect the value of y_2 $\dots y_i$.

The SEM model can be written as: $y = x\beta + u, u = \lambda Wu + \varepsilon$, where u is the residual that can be described as a function of the neighbouring location residuals with some random error ε. In this model, we have a "spillover" of the residuals and not the outcome as we described in the SAR model. This model increases the flexibility in the regression's error term, which may change over different locations (i.e., observed clusters). This variation in the u term can be observed when a spatially related predictor which is significant is not included in the model. This may lead to areas where the model consistently under or overestimates the outcome.

2.3.4. Mixed effects models

Mixed effects models, also known as hierarchical linear models, are also a type of statistical modelling that we can use to analyse our data given the non-independence of the observations. These models allow for the analysis of both within-group and between-group variability simultaneously. As a generalisation of the linear regression models, they allow for the inclusion of random effects, which will account for the variability that is not explained by the fixed effects. The fixed effects are the coefficients, which will be the investigated predictors in our case, while the random effects are the intercepts and/or the slopes for each level of the random grouping variable. The intercept random effects account for the variation in the mean response across groups. In contrast, slope random effects account for the variation in the relationship between the independent and dependent variables across groups. The estimation of mixed effects models involves finding the values of the fixed and random effects that maximise the likelihood of the data. This is typically done using iterative algorithms such as maximum likelihood estimation or restricted maximum likelihood estimation.

To deal with autocorrelation in our data, we considered adding random intercepts for the grouping variable of interest and account for the correlation among observations within each group. We can also include random slopes for the time variable. Suppose the autocorrelation is present due to a time-dependent trend in the data. In that case, the model can also include a random slope to account for the correlation among observations at different time points. Lastly, adding an autoregressive term in case an observation at time t is related to the value of the previous observation at time t-1 may also improve the model fit. The mixed effects models used in this project are shown in 6.5 and explained in the context of our dataset.

2.3.5. Geographically Weighted and nearest neighbour Regression

The Geographically Weighted regression (GWR) differs from the spatial regression methods discussed in the previous paragraphs as it allows the coefficients to vary across space. By fitting a regression equation to each feature in the dataset, GWR evaluates a local model to analyse the studied outcome. For a data set of $1 \dots i$ observations GWR will fit $1 \dots i$ weighted regression models. GWR generates these separate equations by including the outcome and predictor variables falling within the neighbourhood of each target feature. Each neighbourhood's size and scope are determined based on the neighbourhood type and selection method parameters. GWR should be applied to large datasets with at least a few hundred features, which is inappropriate for small datasets. This type of regression can be applied using multiple predictors and will return intercept and coefficient values that vary across different locations. One of the most popular applications of GWR is the assessment of the heterogeneity in the factors affecting the outcome (for instance, income affects chances of hospital admission in the area α but not in area β). The GWR models can be used on continuous, binary, and count outcome data. The neighbourhood selection, also called bandwidth, is the most important parameter when applying the GWR. The process involves fitting spatial kernel to the data. The weight W of a data point is greatest at the location of the selected regression point i (processed repeated for each point). As the distance between two points grows, the weight gradually decreases. Thus, a regression model is locally calibrated for each i by shifting the regression point throughout the area of interest. The data are weighted differently for each location so that the estimates that are produced are specific to that location. Two important parameters in the calibration process are the choice between adaptive or fixed kernel and the optimal bandwidth selection. The bandwidth is the limit that, beyond the weighted observations, is given a value of zero. With larger bandwidths, more observations are used to fit a local regression, and more observations receive non-zero weights. A fixed kernel uses a specific bandwidth, while an adaptive kernel uses a varying bandwidth to define the region around each regression. A fixed kernel may be problematic in datasets with areas (or points) with significantly varying density (similar to the example with the weights assignment in 2.3.3). In such a setting, the GWR will include many cases in areas of high density and a minimal number of cases in areas of low density. This imbalance may be corrected with an adaptive kernel, although these adjustments improve the precision while introducing bias, especially when the parametrisation occurs after we have seen the data.

2.3.6. Denominator and Population density

Population density measures the number of people in a given area, typically expressed as individuals per square kilometre or another grid-based scale. It describes the concentration of people in a region or country and can help identify areas of high or low population concentration. The population density maps are an essential element in the disease mapping work and contribute to a good understanding of the underlying population density. In this PhD, they are also necessary to estimate the disease incidence and prevalence. All estimates of the PIBD incidence and prevalence are based on using the population density in conjunction with the age NUTS3 demographics in each area.

2.4. Validation of geostatistical methodology

This chapter aims to assess and measure the effectiveness and potential of our methods in addressing the spatial misalignments while retaining the information in the gathered data and accurately depicting the true statistical correlation between environmental factors and disease incidence. This is achieved by testing the methods and settings used to determine the spatial weights, selection and refinement of the different interpolation methods and, finally, with the use of data simulations. These results are significant as they supported educated decision-making during the development and optimisation of the methods and helped with the validation of the final analysis.

2.4.1. Spatial weights selection

The selection of spatial weights is an essential step for the spatial regression models, as discussed in the previous sections. These model-specific adjacency definitions describe which cases we will consider as neighbouring in our analysis. According to our simulations, where a randomly distributed pollutant is affecting the disease incidence in the NUTS3 areas, the preferred weights should be based on the distance between the analysed NUTS3 areas. Regions in proximity are more likely to share a population with similar characteristics and be exposed to the same risk factors. However, in our analysis, the Rook's and Queen's contiguity weights returned a significantly better model fit. This finding may seem counterintuitive considering that the Rook and Queen weights do not consider the distance but are based on shared boarders between the areas on the map. The areas in our study present auto-correlation on the basis of adjacency and not proximity which explains why the contiguity weights perform better. The NUTS3 regions have been created on the basis of the population size, which results in highly variable polygon sizes. This means that low density areas will have much larger size and vice versa. Clusters of large areas that are covered by the same clinic are more likely to have the same incidence compared to clusters of small areas that are covered by different units despite the proximity of the small areas. As shown in **Figure 55a**, any distance-based model will overlook the relationship between large regions despite being adjacent. Even if the weights are adjusted to include large areas as shown in **Figure 55c**, this will inflate the number of regions included in parts of the map with smaller regions (**Figure 55d)**. In summary, although the distance-based weights are more appropriate for the analysis of incidence and pollution, in our study, the highly variable size of NUTS3 regions, in combination with the coverage patterns of participating units, require the use of contiguity weights.

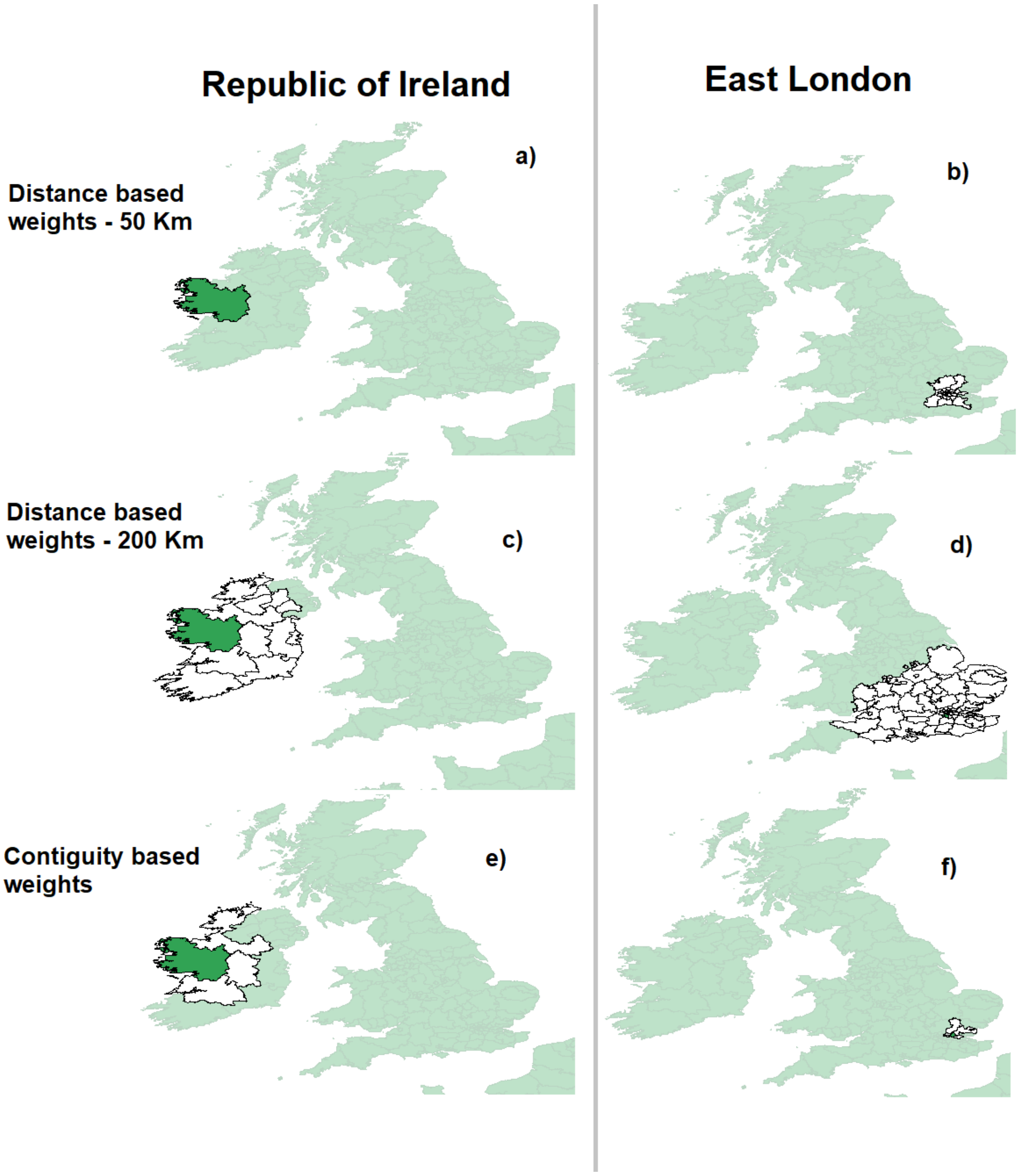


Figure 55. The contiguity weights (e/f) against two distance-based weight options (a/b, c/d).

In this example, in the Republic of Ireland, the c) and e) are capturing accurately the regions covered by the same unit while the a) shows no assigned weights with other regions which is expected to be inaccurate. In the same example, for the area of east London, both distance-based approaches, as shown in b) and d), have included a great number of areas in the spatial weight assignment which is also expected to be inaccurate in contrast to f) which is representative of the expected coverage by our unit in east London. Therefore, the contiguity weights as shown in e) and f) create the most accurate weight matrix for our study.

2.4.2. Interpolation methods

In the previous sections, several interpolation methods were discussed due to their relevance to our project. These techniques were used to align our spatial information into one harmonised dataset. However, any interpolation method comes with a degree of uncertainty regarding the produced estimates in the unsampled locations. Therefore, investigating the performance of our interpolations is an essential validation requirement. The pollution interpolation aimed at generating a continuous surface based on several available pollution measures at various locations in Europe. Therefore, the validation process must measure the accuracy and precision of the interpolated surface. In the following paragraphs, we describe the interpolation validation procedure.

For this study, a validated continuous surface was utilised as the reference point for three different pollutants. These maps were selected as the ground truth since they incorporate precise information from a range of sources, such as sampling, reported releases, and statistical modelling, and, therefore, are expected to be in close agreement with the true values of the pollutants examined. Nonetheless, regardless of whether these maps are accurate for all regions, the primary goal of our validation method remains unaffected. The focus of this validation is to assess the extent of information loss that occurs when we interpolate a pollutant and create a continuous surface based on a limited number of sampling points.

In the first step, the quantities of the continuous ground truth pollution maps were averaged and assigned to each NUTS3 region (**Figure 57**). This created the NUTS3 ground truth maps used in the estimating the validation metric, an R squared. Following that, we randomly sampled several point locations from the continuous ground truth map (original raster). Using different interpolation methods, we generated a continuous surface using only the information from the randomly sampled locations. The generated quantities of the interpolated pollution map were averaged and assigned to each NUTS3 region creating the predicted NUTS3 pollution estimates that were compared to the ground truth NUTS3 map. This allowed us to evaluate the proportion of the variance in the ground truth pollution map that was explained by our interpolated map which was generated based on the randomly sampled location. As shown in **Figure 56**, three pollutants with significant differences in their spatial distribution were included in the validation. These were the Nitrous oxides (NO_2), Sulphur oxides (SO_2) and Benzene in the UK.

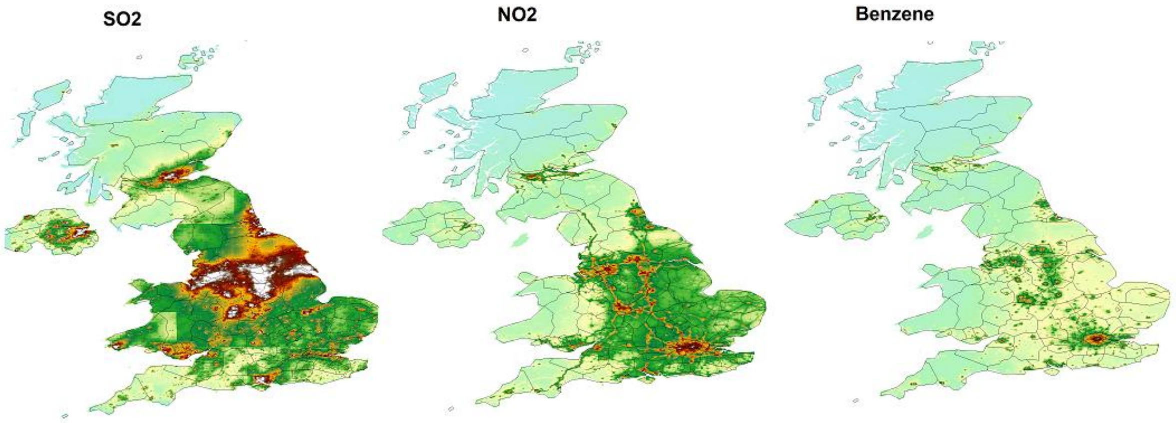


Figure 56. DEFRA maps with UK pollution data

The high-resolution ground truth DEFRA maps with UK pollution data used for the interpolation validation.

The simulations were performed for two sets of sampled locations reflecting different levels of sampling detail. The following results in **table 9** were produced from the lower detail sample where the sampled locations were required to exceed 20 kilometres of distance from each other while the minimum number of sampled locations for each NUTS3 area was set at one. This resulted to a dataset of 304 random sample points used for the interpolation. The Radial Basis Function, Inverse Distance Weight, ordinary Kriging and empirical Bayesian Kriging were the tested interpolation functions. The results for the highest-performing functions are summarised in the following table.

Table 9 The performance of IDW, empirical Bayes and ordinary Kriging interpolation.

R squared for Interpolation performance			
Pollutants	Benzene	NO2	SO2
IDW	87.00%	90.90%	86.75%
Kriging	76.61%	85.32%	77.12%
EBK	89.88%	92.09%	86.53%

For the tested dataset, the IDW and empirical Bayes Kriging (EBK) functions demonstrated the highest performance. However, in datasets with a higher degree of autocorrelation, the empirical and ordinary Kriging (OK) demonstrated the best performance (similarly to the sun exposure example). Therefore, the EBK and OK were the preferred methods when Moran’s I exceeded 0.5, a value suggesting that high spatial autocorrelation is present. One major limitation of the EBK method is the 100-fold increase in the required computations compared to the IDW method, which reduces the feasibility of its application on large datasets. However, this disadvantage of the EBK is linked to the reason for its superior performance. In contrast to the ordinal Kriging, the Empirical Bayesian Kriging does not assume a fixed variogram model. Instead, it uses a Bayesian approach to estimate the variogram parameters for each data point individually based on the local data structure. The estimated parameters are then used to interpolate values at unobserved locations. Finally, it is worth mentioning that the performance tends to improve as the number of sampled locations increases and the minimum distance between the sampled points is reduced. Specifically, a second set of simulations showed that decreasing the minimum distance between sampled locations from 20 km to 15 km can increase the R squared by at least 5%.

In the following **Figure 57**, the steps of the interpolation validation are shown for the SO_2, NO_2 and Benzene in the first, second and third column, respectively. The first row shows the ground truth continuous map as downloaded from DEFRA. These values are averaged per NUTS3 region into the lattice ground truth map as shown in the second row. Using randomly sampled locations on the maps of the first rows we create an interpolated surface for each pollutant that is averaged per NUTS 3 region. In rows 3, 4 and 5 the interpolation-based lattice map is shown for each method-pollutant combination. Although the maps appear to be similar under visual inspection, the EBK method in the last row presented the best performance overall.

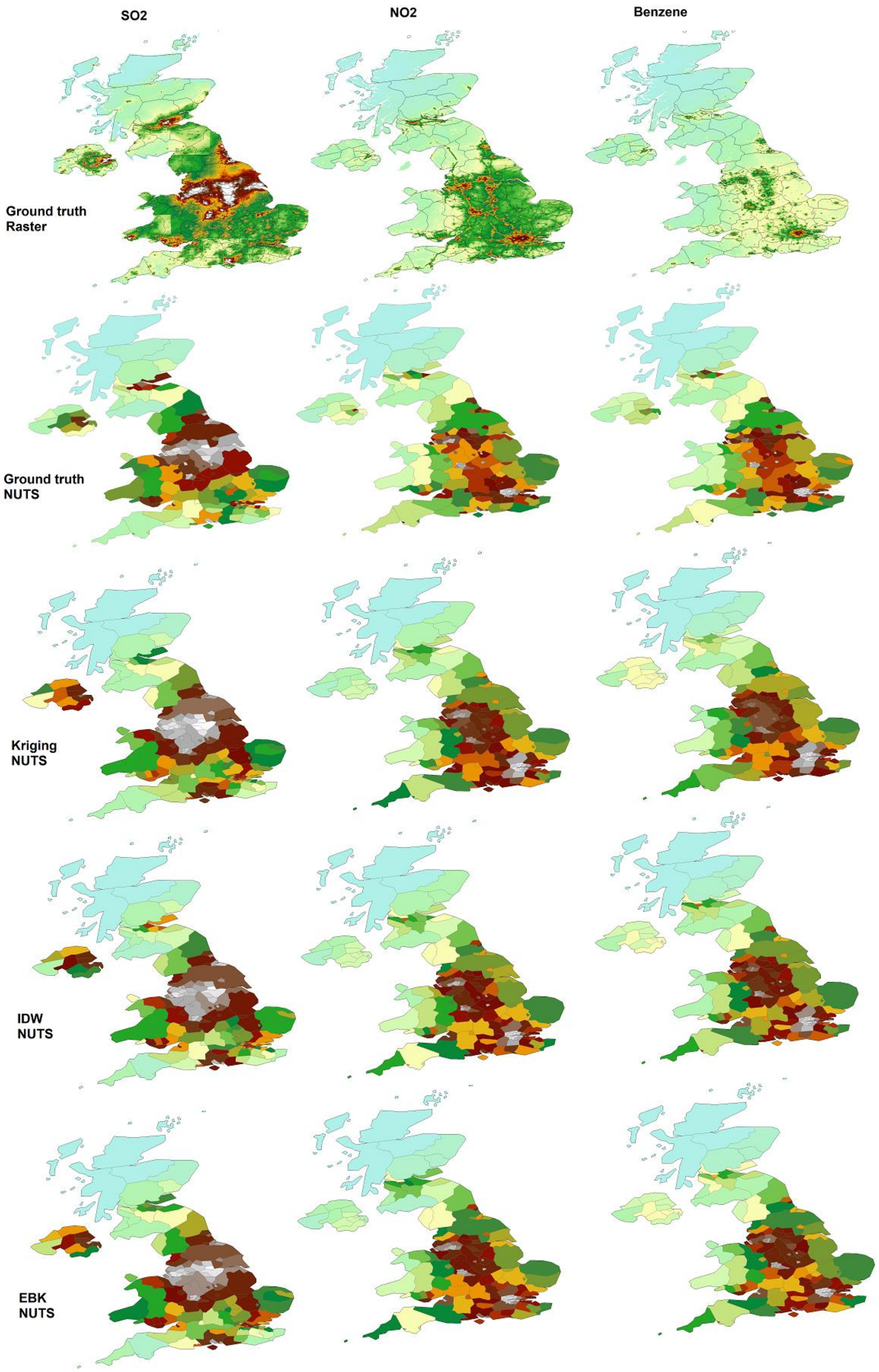


Figure 57. The ground truth maps (first and second rows) followed by the interpolated maps.

2.4.3. Simulation-based calculation of population exposure

In the previous paragraphs and methods section, I outlined the steps required to prepare and analyse our spatial data. In a synopsis, the PIBD incidence, which is the outcome of interest, is collected in a lattice format and thereafter is combined with a limited number of Eurostat predictors that are also available for the same spatial support. In contrast, most of the predictor variables, predominantly the environmental exposures, are collected in a point data format. As shown in the previous paragraphs, these data points are interpolated and converted to continuous surfaces. The values of each interpolated surface that fall in each lattice area (NUTS3) are averaged as a metric of the exposure to each pollutant for the population in this area. Due to the use of the interpolation methods, these average values reflect the rate of exposure to the environmental factors that not only fall within that area but also from the adjacent locations.

A key question is whether our approach is sensitive enough to detect the effects that the risk factors may have on the disease incidence. Under the alternative hypothesis, which is that the environmental factors (collected as data points) influence the incidence (collected from regions on the lattice map), the aim of this simulation is to determine whether we can capture this relationship accurately using interpolation methods combined with linear mixed effects models and spatial regression. Furthermore, the effects of the exposure mechanisms, sample and effect size requirements will be also investigated in the following paragraphs.

To evaluate our proposed methodology, we simulated the incidence of disease within each NUTS3 polygon area based on point data information from nearby risk factors and hypothesised a coefficient for their effects on incidence. These simulations allowed us to model scenarios where certain risk factors, such as pollutants, impact disease incidence in adjacent areas to a specified degree, represented by the coefficient used in the simulations. Thereafter, we attempted to predict this coefficient using a linear regression model based on the interpolated risk factors instead of using the original format of point data. This mimics the process used in our real-world data analysis. In a simplistic example, if carbon dioxide is correlated with the incidence of the studied disease, the population in each NUTS3 area that is close to a source of CO_2 pollution is expected to present higher incidence. In the following paragraphs we simulate this incidence-pollution relationship, creating a simulated outcome that we then analyse using linear regression and the NUTS areas with the goal to identify the hypothesised effect accurately.

Our simulations were based on certain assumptions that are summarised as follows:

- In this work, an important assumption is that the effect of a risk factor on the population cluster at a specific location can be described as the product of a mathematical function of the distance and quantity of that factor at the known, adjacent sampled locations. These functions may vary (i.e., exponential, linear, having a cut-off), depending on the type of factor analysed and its specific patterns of spatial distribution. Our function is described as the exposure z_i for the population cluste*r i in* the following paragraphs.
- The area surrounding an environmental factor is affected homogeneously. We assume that any two points on the map that are equally distant from an environmental factor source will have the exact same exposure to that factor. Hence, when converting the data points to a continuous surface, any physical barriers, wind direction and other factors that may influence the concentration of the pollutant will not be considered. However, this is not a major limitation for our analysis since several risk factors have detailed datasets available with thousands of sampling locations.
- The duration of the exposure was assumed to be the same for all risk factors and individuals in the population. Therefore, the location and quantity of the risk factors are the main variables that can influence the incidence in our simulations. As an example, in this study, for each pollutant, a six-month exposure to a quantity of 1 is equal to the annual exposure to a quantity of 0.5.
- The population is spatially distributed based on the population density map. Specifically, we used the UK population density map including just under 65-million-point locations with the proximity between the points to be based on the population density as reported in the 2011 census. The simulation was made feasible after sampling 1 in 10,000 point-locations. Considering that the population is a highly clustered variable, we can consider the sampled locations as population clusters of 10,000 individuals.

The simulations were run for the United Kingdom using the NUTS3 classification map (173 or fewer territories depending on the simulation) and the DEFRA CO_2 point data. The disease incidence was simulated for each population cluster in all NUTS3 areas as a linear function of the exposure to the pollutant with a coefficient representing the strength of the CO_2 effect. The probability of retrieving this coefficient in our analysis is the metric used to assess our model's performance in the simulations. The simulated incidence of the disease at the location of cluster *i* depends on its exposures to the pollutant and is simulated as follows:

$$\text{Incidence}_i = \mathrm{b}_0 + b_1 \cdot z_i + \varepsilon_i$$

Where:

- b_0 is the disease incidence without any environmental exposures
- b_1 the effect size of the pollutant 1 effect
- z_i the exposure of cluster *i*
- and ε_i the random error for the cluster *i*

This can also be expanded for multiple pollutants (z, y and x):

$$\text{Incidence}_\mathrm{i} = \mathrm{b}_0 + b_1 \cdot z_i + b_2 \cdot y_i + b_3 \cdot x_i + \varepsilon_i$$

The exposure for the population cluster *i* is calculated as follows:

$$\mathrm{z_i} = \left(\frac{q_j}{(d_{ij})^2}\right) + \left(\frac{q_{j+1}}{(d_{ij+1})^2}\right) + \cdots + \left(\frac{q_m}{(d_{im}\ )^2}\right)$$

Where:

- Z*i* is the pollution exposure estimate for the population cluster in location *i* and
- $\mathrm{d_{ij}}$ is the distance between population cluster *i* and pollution sources *j* ∈ *{1,2 ..m}*. (**Figure 58**).

In **Figure 58** below, each yellow line shows the distance between the population clusters and sources of pollution. Please note that for the estimation of the exposure, the distance is squared, as the exposure is expected to decrease exponentially. Inversely, for clusters closer to the pollution sources the exposure increases exponentially. This is also reported widely in the literature (Bian et al., 2020; Crumeyrolle et al., 2019; Iwata et al., 2019; Lv et al., 2021; Zhang et al., 2019) where most studies describe that the exposure increases exponentially in areas closer to the source of the pollution.

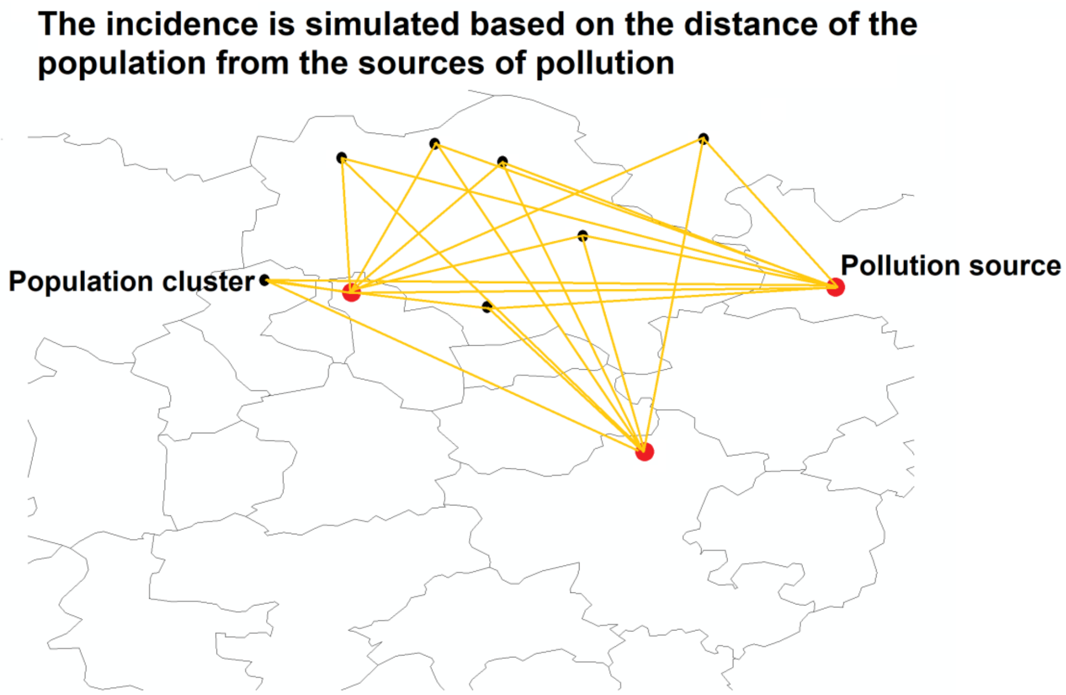

The incidence is simulated based on the distance of the population from the sources of pollution
Population cluster
Pollution source

Figure 58. The simulated interactions between the sources of pollution and population clusters.

The CO_2 sources of pollution are marked in red, and the population clusters are marked in black. The yellow lines show the distances d_{ij} that are used to calculate the simulated exposure Zi as described by the previous formula.

Based on the total quantity of exposures in each population cluster, the simulated incidence is now calculated as described in the previous paragraph. In the next steps, as shown in **Figure 59**, we replicate the analysis used for the real-world data. The sources of pollution were interpolated creating a continuous surface that we subsequently averaged over each NUTS3 area. The incidence was also averaged for all population clusters within each NUTS3 area (in our real-world data we collected the average incidence directly). Repeating this process for all regions provide us with the average estimates of the interpolated pollution and incidence for the NUTS3 regions. This is the dataset that we went on to use in the data analysis.

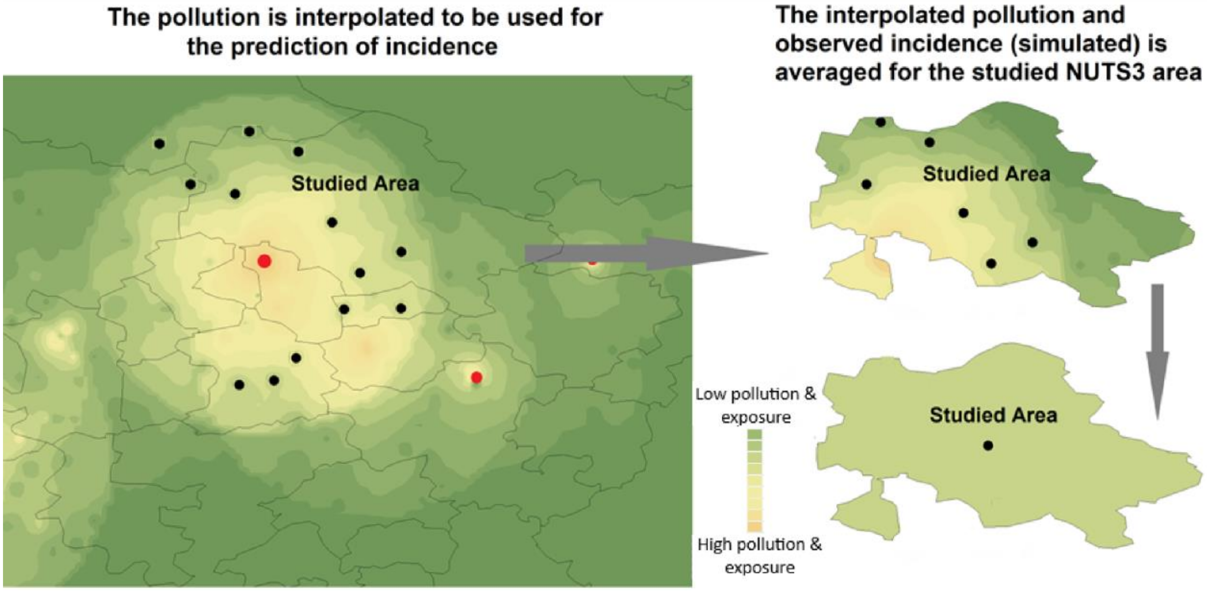

The pollution is interpolated to be used for the prediction of incidence
Studied Area
The interpolated pollution and observed incidence (simulated) is averaged for the studied NUTS3 area
Studied Area
Low pollution & exposure
High pollution & exposure
Studied Area

Figure 59. Pollutant interpolation in the wider region and a selected NUTS3 area.

The simulated incidence and interpolated exposure are averaged for each NUTS3 area which provides the dataset that we will use in our simulation analysis. Please note that in the steps described in the previous paragraphs, we have attempted to generate the simulated incidence in a manner that represents the observed incidence in our real-world study.

After preparing the simulated dataset, the following parameters were investigated in terms of their effects on the analysis validity and statistical power:

- Map and polygon size and shape
- Effect size
- Standard deviation of incidence
- Density of pollution data points
- The relationship between exposure and distance (such as linear or exponential)
- Sample size

The effect size is a very important parameter in the design of any clinical study as it directly affects the statistical power since large effects are easier to detect. Similarly, the standard deviation (SD) is equally important considering that lower SD values allow the detection of smaller effects. In this study, I will assess the effect size in conjunction with the SD using the standardised effect size which is a measure of the strength of the relationship between the pollution and incidence. This metric produces a quantity that can be used to compare the strength of different effects on a common scale and to interpret the practical significance of this effect in a meaningful way. A large effect size may not be detected if the SD is also very large while a smaller effect size might be detected if the SD is also very low. Therefore, this metric is important as we can only understand the true power of the study when considering the effects in conjunction with the variance and standard deviation. In the following **Figures 61 & 62**, we see the simulation results using three different UK maps (random missing areas introduced), and variable effect size-SD ratio (standardised effect size). The results show that for a standardised effect size of 5 or higher, the power of the study is nearly 100%. Very importantly, the power exceeds 80% for standardised effect sizes above 1.8. For a map with approximately 150 regions, this result means that the study is likely to detect a significant pollutant when 1 unit increase in exposure increases the incidence by approximately 12.6 patients per 100,000 individuals.

In the **Figure 60**, I provide an example of two population clusters that are exposed to different levels of pollution. Using the exposure formula, the estimates for cluster 1 and 2 respectively, are 0.035889 and 0.002088 meaning that the exposure difference is 0.0328. Considering that we can detect an incidence increase of 12.6 patients for 1 unit of exposure increase, in this example we can detect the increase of 0.031936*12.6= 0.41 patients per 100,000 individuals. Therefore, in this example, we can detect if the population of the cluster 1 has an incidence increase of 0.4 compared to the cluster 2 due to its the proximity to the sources of pollution. Finally, to put the CO_2 ppm difference used in this example, in perspective, 10 ppm is a small fluctuation occurring throughout the day depending on the studied area (García et al., 2008). It is important note that although the example provided shows 2 population clusters, our following simulations are based on the entire UK.

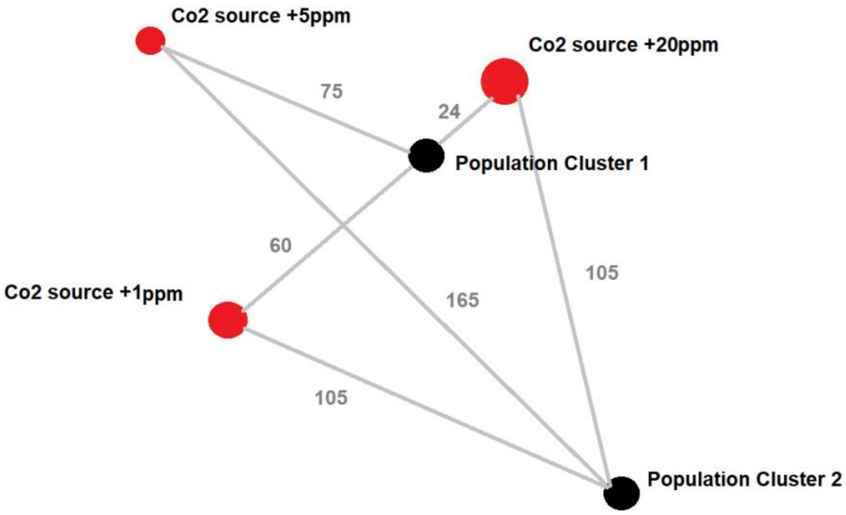


Figure 60. Example of two population clusters being exposed to 3 sources of CO_2 pollution that increase the CO_2 locally by 1, 5 and 20 ppm.

Intuitively, we can understand that the CO_2 levels are expected to be higher in the cluster 1 area. However, quantifying the exposure and analysing the effect is a complex task. Our simulations show that for this example, if the CO_2 has an effect on the incidence, we could detect a difference greater than 0.4 new patients per 100,000 person-years

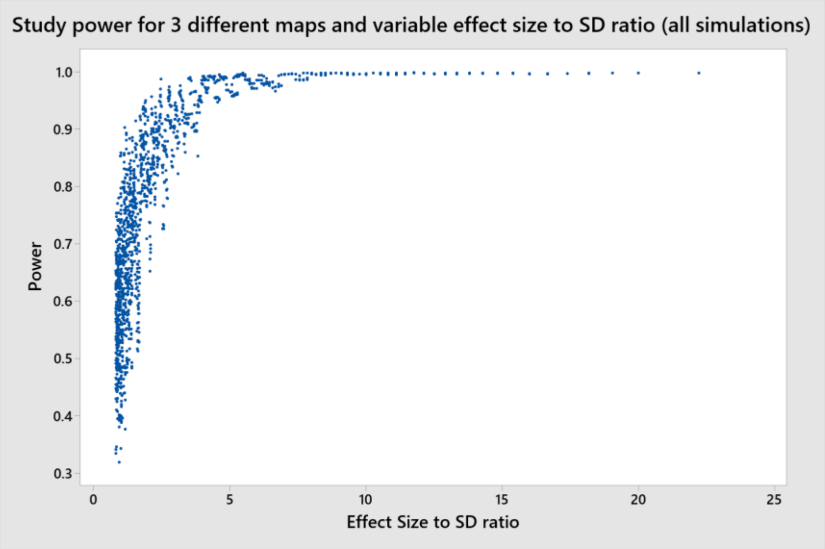


Figure 61. The probability of study success per simulation.

The estimated statistical power of the study (y-axis) for different combinations of map regions in the UK for the range standardised effect size (x-axis). It should be noted that the observed variation in the simulation results has two sources, the random error introduced in the incidence calculations and the random selection of NUTS3 regions in the UK map (min:150 and max: 173).

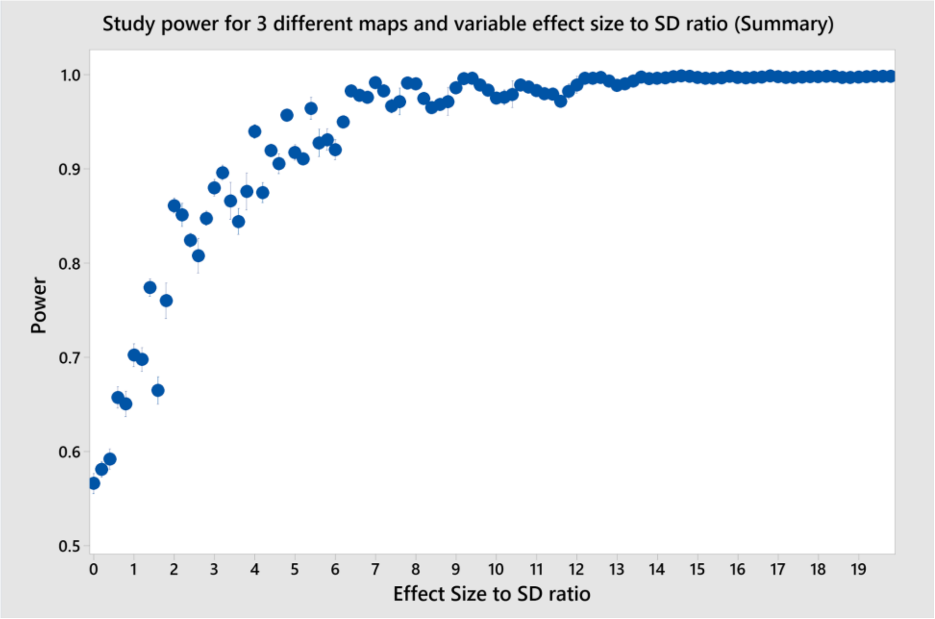


Figure 62. The grouped simulated probability of study success

The estimated statistical power of the study (y axis) for different combinations of map regions in the UK for groups of standardised effect size (x axis). This graph is the same as figure 61 but with the results grouped for easier interpretation.

Another important factor of the simulation was the investigation of the sample size. Although it is clear that a larger sample size would increase the power of the study, in the following analysis, we investigated the difference in the study power after increasing the sample size by 2 and 3-fold with the latter to reflect the expected sample size of our study. As shown in **Figure 63**, doubling the sample size from 150-173 regions to 300-346 regions improves the power significantly while a 3-fold increase improves the power even further but only marginally. To achieve a power of 80% or greater, the study requires a standardised effect size lower than 1.8, 1.55 and 1.42 for the three tested sample sizes of 150, 300 and 450 regions. Returning to the previous example shown in **Figure 60**, the decrease of the standardised effect size from 1.8 to 1.42 would reduce the minimum incidence difference that can be detected from 0.41 to 0.32.

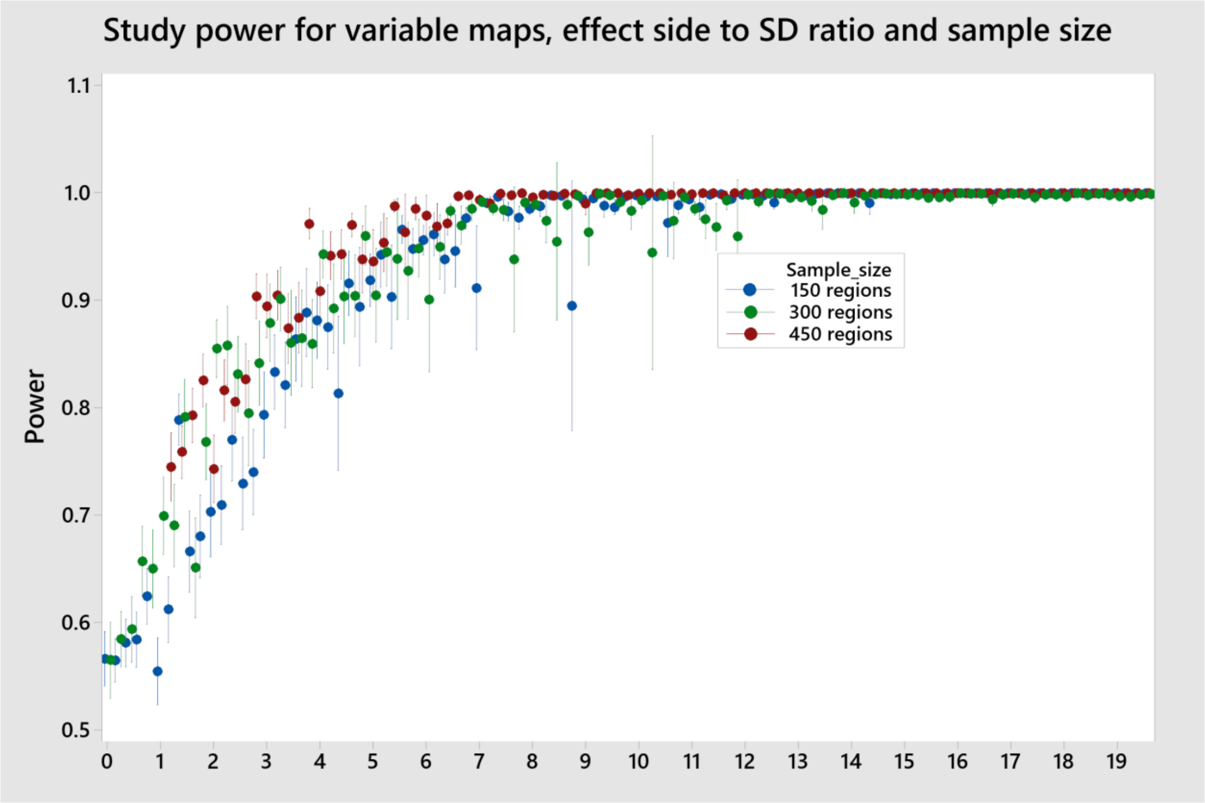


Figure 63. The simulated probability of study success for three sample sizes.

The estimated statistical power of the study (y-axis) for different combinations of map regions in the UK for groups of standardised effect size (x-axis) and three different sample sizes. The sample size reflects the size of the map in the analysis.

The final factors investigated in the simulation were the shape and size of NUTS3 areas, the density of pollution data points and the relationship between exposure and distance (such as linear and exponential). The density and availability of point locations with pollution information were found to be less significant compared to the other factors investigated. Quite possibly this is explained by the use of the interpolation methods. The shape and size of the NUTS3 areas also had a marginal effect introducing a variation of less than 3% in the power of the study. However, the relationship between the size of exposure and distance from the pollutant was found to be a particularly important factor as any pollutants which are distributed in space in a linear manner reduced the power significantly. In contrast, when the quantity of the pollutant was modelled for the population location using a double inverse square root adjustment of the distance, the power of the study exceeded 80% for a standardised effect size of 0.7 for 150-173 regions and 0.4 for the larger tested sample sizes. When the adjustment was based on a logarithm transformation of the distance the increase in the power was even greater.

This important finding shows that risk factors that are measured at exponentially higher levels in their immediate surroundings are more likely to be detected in our study.

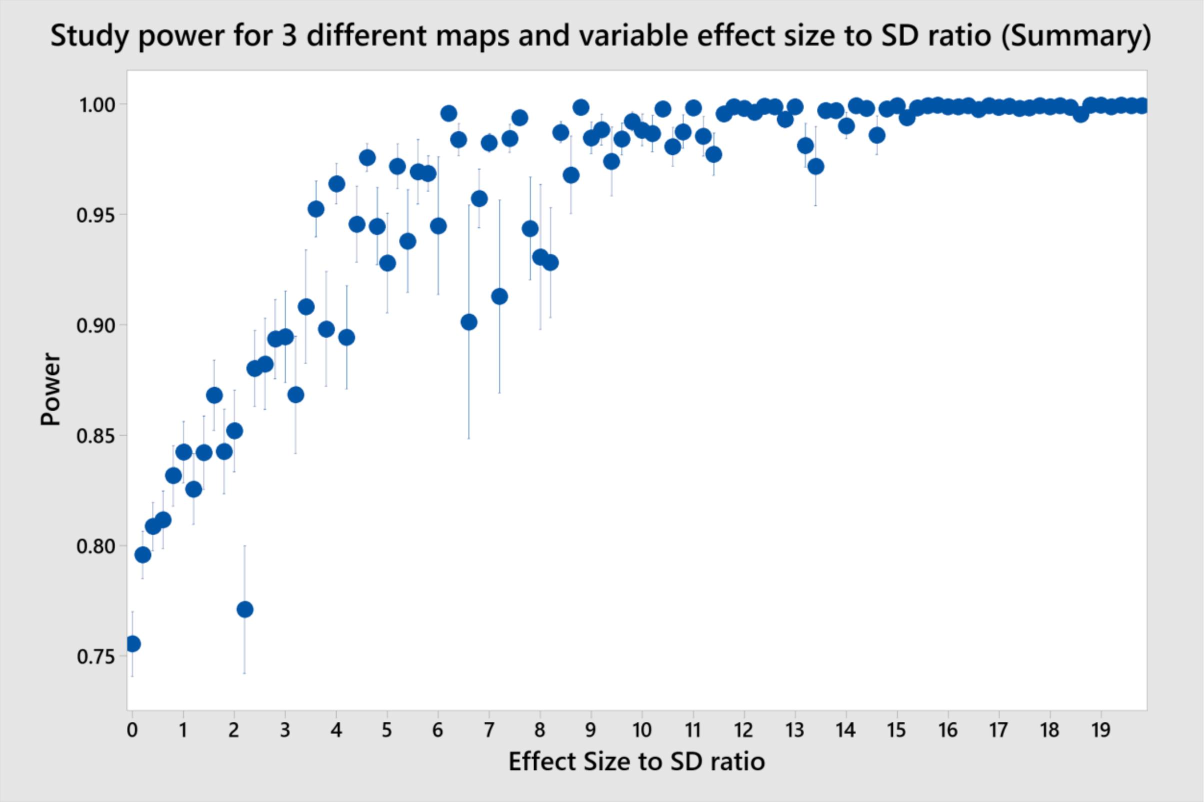


Figure 64. The simulated probability of study success for a linear different pollution–exposure mechanism.

The estimated statistical power of the study (y-axis) for different combinations of map regions in the UK for groups of standardised effect size (x-axis). Overall, the results showed a significant increase of power when the quantity of the pollutant was modelled for the population cluster location using the inverse of a double square root adjustment of the distance used in the exposure calculation of the simulated incidence.

2.4.4. Discussion of methods validation

This chapter investigates the expected performance of our methods while testing different strategies and parameters to optimise our study-specific analytical approach. The subsequent spatial analysis used contiguity spatial weights and the interpolation of the point data was based on the IDW and Kriging methods depending in the variogram and autocorrelation in each predictor dataset. Furthermore, the simulation study evaluated the performance of the methods overall and specifically for detecting the effects of risk factors on the disease incidence. Several factors were considered, including map and polygon size and shape, effect size, standard deviation, density of pollution data points, the relationship between exposure and distance, and sample size.

The simulations showed that the relationship between exposure and distance is a critical parameter in detecting the effects of risk factors. Risk factors that affect areas closer to them exponentially increase statistical power dramatically. This mechanism is consistent with the existing literature on the spatial distribution of pollutants which reinforces the methodology of the study. The simulations also suggest that the density of the pollution point locations and the shape/size of the map areas had a rather small effect on the performance. This indicates that these factors can be standardised since they do not have a significant effect on the study's results. In contrast, the sample size was found to be an important factor, as more observations can improve the confidence in the predictions. Lastly, the simulations suggest that the effect size adjusted by the standard deviation is a major factor in identifying risk factors. The less noisy the data are and/or the stronger the effects, the easier it is to identify risk factors. This indicates that the study's power and validity will increase after removing erroneous reports and outliers that have been confirmed as inaccurate.

Overall, the simulations suggest that the methods used in this study can detect risk factors with a standardised effect size of 0.4 or greater which as shown is considered to be a small difference. This is a promising finding, particularly because the study's final models explain part of the variance based on spatial models and therefore, the expected power of the study may be even higher.

2.5. Software and computational methods

The software used in this project includes QGIS, ARCMap, GEoda, Minitab and R.

QGIS (Quantum GIS) an open-source Geographic Information System (GIS) software that we use to visualise spatial data, process maps and update the geodetic information when possible. QGIS is built on top of the Python programming language, and it provides a Python API (Application Programming Interface) that allows you to extend the functionality of the software using Python scripts and plugins ("QGIS.org, %Y. QGIS Geographic Information System. QGIS Association. http://www.qgis.org," n.d.).

ArcMap is a GIS software developed by Esri, a leading provider of GIS software and solutions. ArcMap offers a Python scripting interface that can be accessed through the built-in Python window in GUI environment. The ArcPy module in Python is used to interact with ArcGIS tools, which allows users to automate various tasks and workflows in ArcMap. ArcMap was used for most of the plotted map in this thesis and for the initial testing of different interpolation methods ("ESRI 2011. ArcGIS Desktop: Release 10. Redlands, CA: Environmental Systems Research Institute," n.d.).

GeoDa is a free and open-source software designed for exploratory spatial data analysis and mapping developed by the Centre for Spatial Data Science at the University of Chicago. In this project we used GeoDa to investigate the presence of clusters, and investigate auto-correlation. Several maps in this thesis were also generated in GeoDa (Anselin et al., 2006).

Minitab was first developed by researchers at the Pennsylvania State University and is a statistical software package that is commonly used for data analysis and statistical quality control. In this project Minitab was used exclusively for visualisations (Abegunde et al., 2016; Gilat et al., 1987; Hansen et al., 2011; "Minitab, LLC, 2021. Minitab, Available at: https://www.minitab.com.," n.d.).

Lastly, the main volume of work and computations in this project was performed in R. Several libraries were used for data management, mapping, interpolation of multiple spatial datasets, update of geodetic information, extraction of spatial information from raster files to point data, averaging of raster information over areal maps and more. The R packages used were: readr, sf, ggplot2, sp, ggspatial, raster, sgo, rgdal, rgeos, dismo, gstat, terra and spatstat ("R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/).

3. RESULTS – Safety registry reporting metrics

3.1. Reporting summaries and metrics from the Safety Registry

In the following paragraphs, I present the Safety Registry summaries of collated reports and collected data for the 2018 – 2022 period. These summaries include participation, coverage, and data retention descriptive statistics.

3.1.1. Participating individuals and units

Since the launch of the Safety Registry in October 2016, 230 different email addresses have been added to our database. This number may seem high, considering that the maximum number of concurrently active participants in our project was 140 PIBD experts. However, this can be explained by the fact that in recent years, several of these reporting physicians have changed practices and, therefore, email addresses multiple times. In other cases, they have been retired or moved elsewhere and, frequently, were replaced by a new PIBD expert. In addition, there are cases where multiple experts have been reporting for the same units. This, combined with the changes in the clinician location and participation, created the need for a smaller, internal database where every email and participant name submitting data to our system is linked to a clinical unit (such as a hospital, clinic, or practice). At the time of the analysis in late 2022, 148 unique units were identified from 33 countries since the beginning of the Safety Registry. The number of active units per country is summarised in the **Table 10** below**.**

Table 10 The number of units with confirmed details reporting to the Safety Registry by country.

The count of participants and population numbers reflect all years of operation of the Safety Registry. Please note that the denominator information is not available for all countries.

Country	Count of participants	Population (all years)	Country	Count of participants	Population (all years)
Netherlands	31	3,330,571	Ireland	2	1,246,426
UK	18	10,219,257	Malaysia	2	N/A
Germany	13	3,456,810	Portugal	2	579480
Israel	13	N/A	Albania	1	N/A

Italy	11	8,107,268	Austria	1	241,872
Switzerland	7	1,017,216	Czech Republic	1	661,885
Canada	6	NA	Japan	1	N/A
France	5	916,581	Korea	1	N/A
Belgium	4	1,623,316	Lithuania	1	157,647
Spain	4	1,784,844	Luxembourg	1	1
Australia	3	NA	Poland	1	484,634
Slovenia	3	353,173	Romania	1	1,618,982
Sweden	3	655,994	Serbia	1	242,483
USA	3	N/A	UAE	1	N/A
Croatia	2	398,260			
Finland	2	581,259			
Greece	2	1,414,321			

Only data from participants who provided sufficient information were included in the analysis. As discussed in the methods, several forms were excluded according to the exclusion criteria Three levels of checks and exclusions were employed, the first level involved missing essential information (i.e., number of patients), the second level involved the unavailability of data required for calculating the incidence of PIBD in European countries, and the third level dealt with reports that covered the same unit. The total number of reports included in this project's analysis of the PIBD incidence and prevalence was 266, which was 62% of all the reports with denominator data submitted to the database. These reports cover a paediatric population that exceeds 30 million individuals. The number of reports at each review level, and the annual population coverage are summarised in **Tables 11 & 12** respectively.

Table 11 The number of forms received and kept after each review round.

Each eCRF has been submitted by a single participant who completed the denominator data form online. At each review stage a percentage of the submitted forms was rejected. Most forms were rejected at the first review level.

Year	Rate of kept eCRFs	Total Submitted (n)	Review 1	Review 2	Review 3
2018	63%	65	51	51	41
2019	63%	71	48	48	45
2020	59%	118	78	78	70
2021	62%	79	55	51	49
2022	62%	98	63	63	61
Total	62%	431	295	291	266
Lesions excluded at each level of review:			32%	1%	9%

Table 12 The reported annual paediatric population covered at each annual data collection round.

The covered population fluctuates depending on the participation metrics as captured in the denominator data form.

Year	**2018**	**2019**	**2020**	**2021**	**2022**
Paediatric population covered	19,552,288	16,850,641	30,478,376	19,426,346	26,955,434

The total population covered by the safety registry is summarised in **Table 12**. The minimum and maximum number of person-years per centre was 193,000 and 10,000,000 respectively while the median, lower (25%), and upper (75%) quartiles were 970,000, 500,000 and 1,780,000 respectively. The geographical reporting coverage of the Safety Registry in Europe for the 2018-2022 period included 26 European countries is shown in **Figure 32** below.

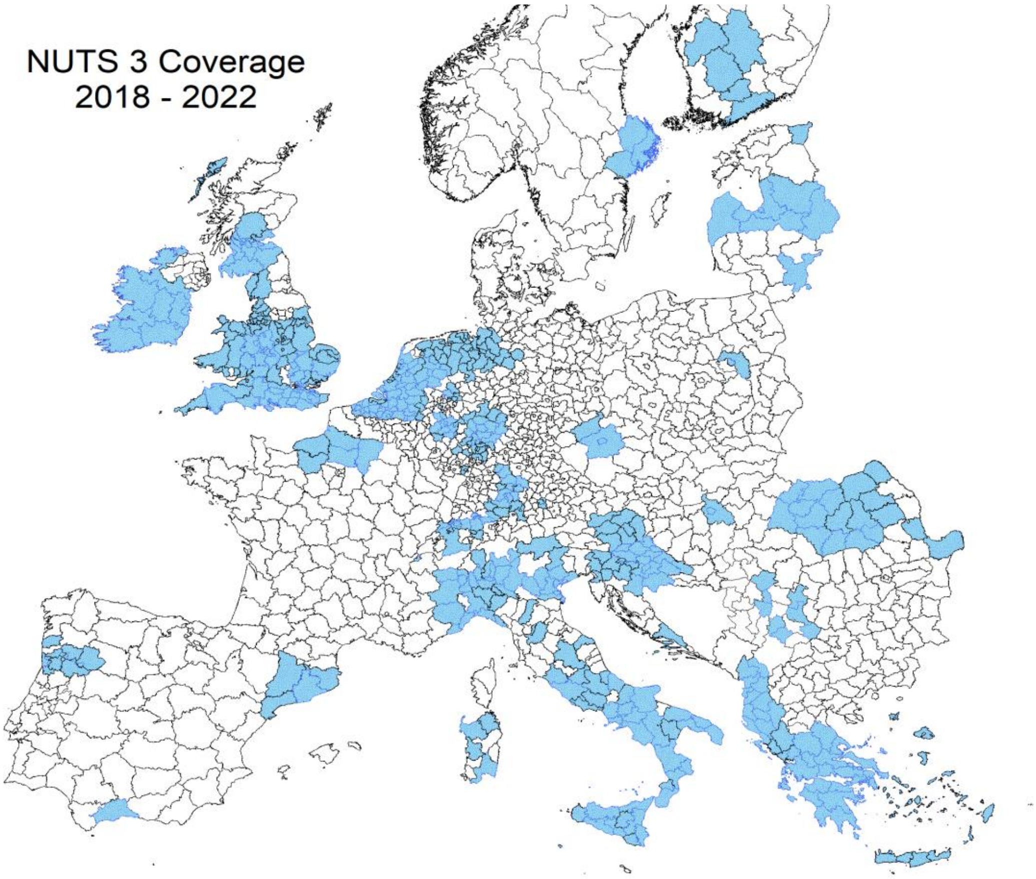


Figure 32. The map of Europe shows the coverage of the Safety Registry as reported from 2018 to 2022.

Note that the map shows the areas covered for at least one year during the reporting period.

As discussed in the methods, a fraction of the submitted forms encountered the problem of overlapping coverage. This was related to cases where multiple centres claimed coverage of the same NUTS3 region (See section 2.2.2). In such instances, the population of that region was divided according to our predefined methods. In 2018 and 2021, the rate of overlapping regions did not exceed 5%, whereas, in 2019 and 2020, the overlap rate was under 10%. However, the latest data collection round in 2022 showed a higher overlap rate that reached almost 14.5% between different centres. The average rate of regions with unique coverage with no overlap in our study was calculated to be 91.76%, as shown in the below in **Table 13**.

Table 13 The number and rate of the NUTS3 regions claimed by one or more centres per data collection year.

In 2018, 95.5% of the covered regions were claimed by a single centre, as two different units claimed only ten regions. However, in 2022, 38 regions were claimed by at least two or more different reporting units.

Number of centres claiming coverage of each region	**2018**	**2019**	**2020**	**2021**	**2022**	**Total/ Average**
1	213 (95.5%)	159 (94.1%)	295 (90.8%)	209 (95.0%)	227 (85.7%)	1103 (91.76%)
2	10 (4.5%)	10 (5.9%)	26 (8.0%)	11 (5.0%)	23 (8.7%)	80 (6.66%)
3	0 (0.0%)	0 (0.0%)	4 (1.2%)	0 (0.0%)	14 (5.3%)	18 (1.50%)
4	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	1 (0.4%)	1 (0.00%)
5	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.00%)

4. RESULTS – Safety registry Reported Incidence

4.1. International rates of PIBD incidence

The final dataset was used to calculate the disease incidence in the participating European countries. The following tables and plots present the incidence results by country (**Figures 33** & **34** and **Table 14**). These summaries also include the pooled reported incidence using fixed and random effects approaches as discussed in the methods. Given the high degree of heterogeneity and our knowledge that the true variation in the incidence between countries is beyond what would be expected due to chance alone, the random effects model is favoured as a metric of the average incidence observed in the study sample. The pooled incidence based on the random and fixed effects models was reported as 8.53 95% (95% CI, 6.35 - 11) and 6.46 95% (95% CI, 6.32 – 6.61) new cases per 100,000 paediatric person-years, respectively. However, the simple average of the reported incidence from all European sites was 7.25 new cases per 100,000 paediatric person-years.

Table 14 The PIBD incidence estimates with their respective confidence intervals and sample size as reported by each country for 2018 – 2022.

The number of reports that were included in the calculations in total and per year is also reported.

Country	Lower C.I. 95%	Estimate	Upper C.I. 95%	Sample size	2018	2019	2020	2021	2022	Million PAED-Years
Romania	**1.51**	**1.89**	**2.34**	**4**	**0**	**1**	**1**	**1**	**1**	**4.50**
Italy	**1.97**	**2.15**	**2.34**	**24**	**3**	**4**	**7**	**3**	**7**	**24.87**
Greece	**2.19**	**2.66**	**3.21**	**4**	**1**	**0**	**1**	**1**	**1**	**4.17**
France	**2.58**	**3.16**	**3.84**	**7**	**2**	**2**	**2**	**0**	**1**	**3.22**
Czechia	**2.83**	**3.52**	**4.34**	**5**	**1**	**1**	**1**	**1**	**1**	**2.50**
Lithuania	**2.58**	**4.23**	**6.53**	**3**	**0**	**0**	**1**	**1**	**1**	**0.47**
Switzerland	**3.89**	**4.55**	**5.3**	**19**	**2**	**4**	**5**	**4**	**4**	**3.69**
Portugal	**3.92**	**5.69**	**8**	**1**	**1**	**0**	**0**	**0**	**0**	**0.58**
Spain	**5.21**	**5.81**	**6.44**	**11**	**0**	**3**	**2**	**3**	**3**	**6.06**
Belgium	**6.19**	**7.01**	**7.92**	**13**	**3**	**1**	**3**	**2**	**4**	**3.72**
Croatia	**5.58**	**7.07**	**8.84**	**4**	**0**	**1**	**1**	**1**	**1**	**1.09**
Germany	**6.85**	**7.43**	**8.05**	**20**	**1**	**4**	**4**	**4**	**7**	**8.06**
Hungary	**7.48**	**9.13**	**11**	**2**	**0**	**0**	**0**	**1**	**1**	**1.16**

Netherlands	**8.8**	**9.33**	**9.87**	**64**	**12**	**12**	**17**	**11**	**12**	**12.77**
Slovenia	**7.71**	**9.50**	**11.6**	**6**	**2**	**1**	**1**	**1**	**1**	**1.02**
Sweden	**8.34**	**9.57**	**10.9**	**6**	**0**	**2**	**1**	**1**	**2**	**2.27**
United Kingdom	**10.2**	**10.60**	**11**	**53**	**11**	**5**	**17**	**10**	**10**	**26.19**
Serbia	**8.9**	**11.50**	**14.6**	**4**	**0**	**1**	**1**	**1**	**1**	**0.58**
Ireland	**13.9**	**15.10**	**16.2**	**4**	**0**	**1**	**1**	**1**	**1**	**4.43**
Finland	**15.9**	**18.50**	**21.4**	**5**	**1**	**1**	**2**	**1**	**0**	**0.99**
Austria	**23.7**	**28.30**	**33.6**	**4**	**0**	**1**	**1**	**1**	**1**	**0.47**
Poland	**41.4**	**46.00**	**50.9**	**3**	**1**	**0**	**1**	**0**	**1**	**0.80**

The confidence in the reported incidence was analogous to the sample size which was the product of the participants, sample size and the number of reporting years per country. **Figure 33** below shows how the confidence is reflected on the size of the confidence intervals per country.

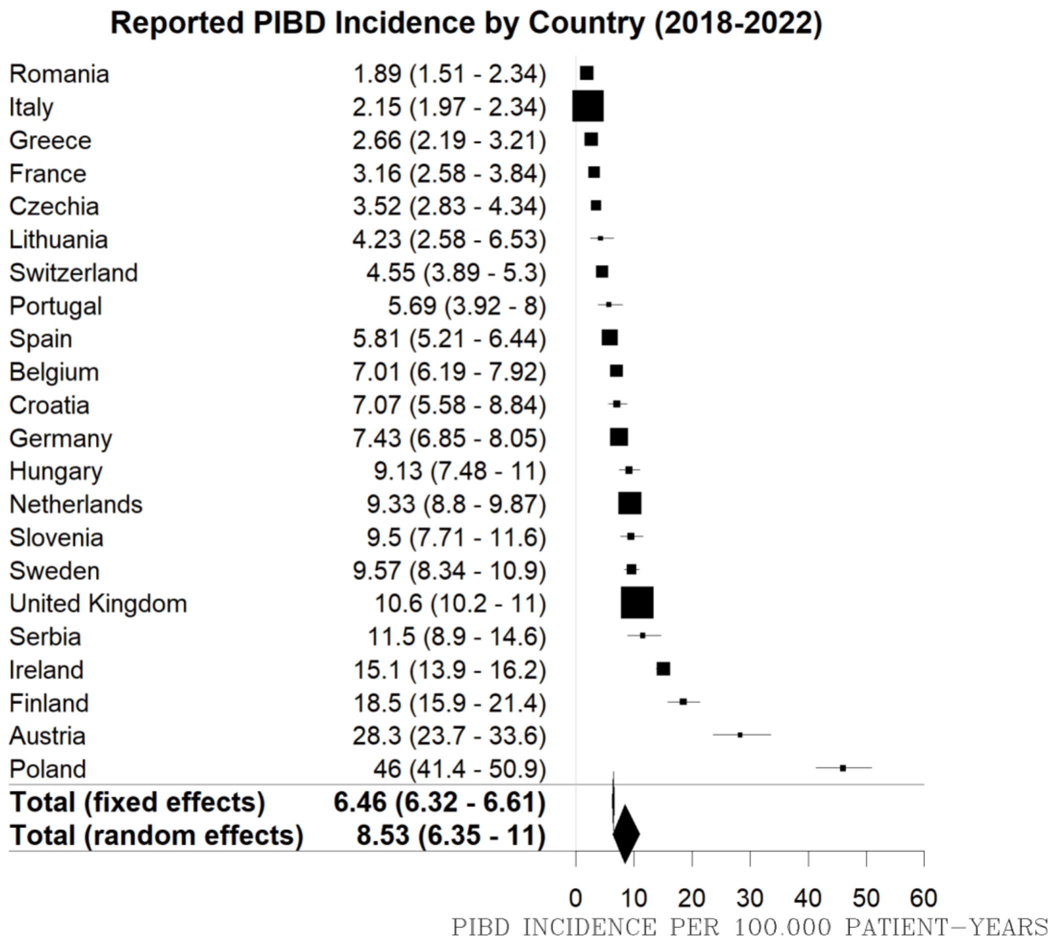


Figure 33. Forest plot of the PIBD incidence estimates with their respective confidence intervals as reported by each country for 2018 – 2022.

Each box's size is analogous to each country's sample size.

As shown in the funnel plot in **Figure 34** there is no significant reporting bias in the reported incidence rates by country considering that the deviations are not asymmetrical. However, the number and extent of the deviations suggest there is substantial heterogeneity in the reported incidence rates between countries. The fact that more than half of the countries reported incidence rates that deviate considerably from the mean value, as shown in the funnel plot, is expected considering the incidence rates among different regions were also explained to vary. However, the extent of the variation is an important indication for the quality of the data reported per country.

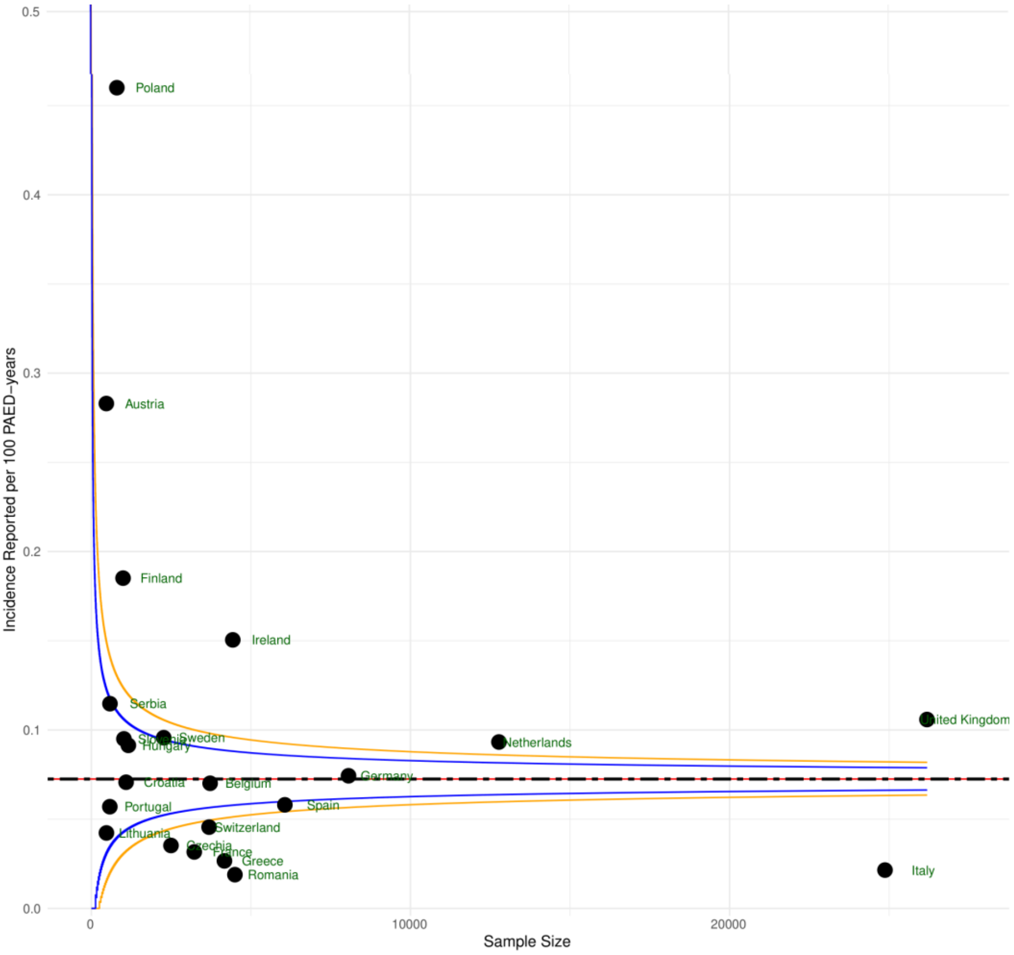


Figure 34. Funnel plot of the PIBD incidence results reported by country with 95% and 99% margins.

It should be noted that the sample is per 1000 paediatric person-years. Austria and Poland appear to be significant outliers, a finding confirmed by additional tests for outliers.

4.1.1. Spatial analysis of the phenotype ratio of PIBD

The effects of latitude and longitude were investigated with a Spearman correlation test between the reported incidence and the decimal degrees per country's centroid. Although the longitude analysis did not return any significant results, the incidence differences that depend on the latitude were significant and can be summarised in 3 latitude groups, as shown in **Figure 35**. The countries in central Europe (between 41° and 51° North) have reported an almost 2-fold increase in the incidence compared to the countries in the South (<41° North), while countries in the North (>51° North) reported an almost 4-fold increase compared to the countries in the South.

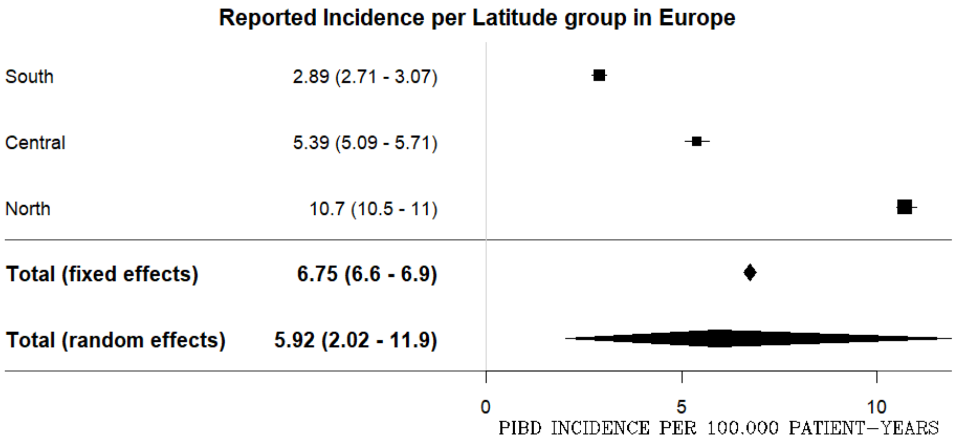


Figure 35. The PIBD incidence is grouped by the latitude of the reporting countries.

To better understand the reported data and investigate the presence of potential outliers, we replotted the funnel plot after adjusting for the expected differences based on the latitude of each country (**Figure 36**). Assuming that the latitude can explain a part of the observed variance, we can eliminate these differences and study the remaining, unexplained variance by an appropriate adjustment for latitude. The adjustment adds the average Central-South difference to the countries of the South and subtracts the average North-Central difference from the countries in the North. The following **Figure 36** shows the improvement of the results' homogeneity after the adjustment. Austria and Poland remain the main outliers of the sample, although the adjustment improved the deviation of the latter. The homogeneity of the results overall was improved substantially, however, Germany is now outside the 99% bounds, possibly because it spans geographically from the Central to the Northern latitude groups and it was included in the latter.

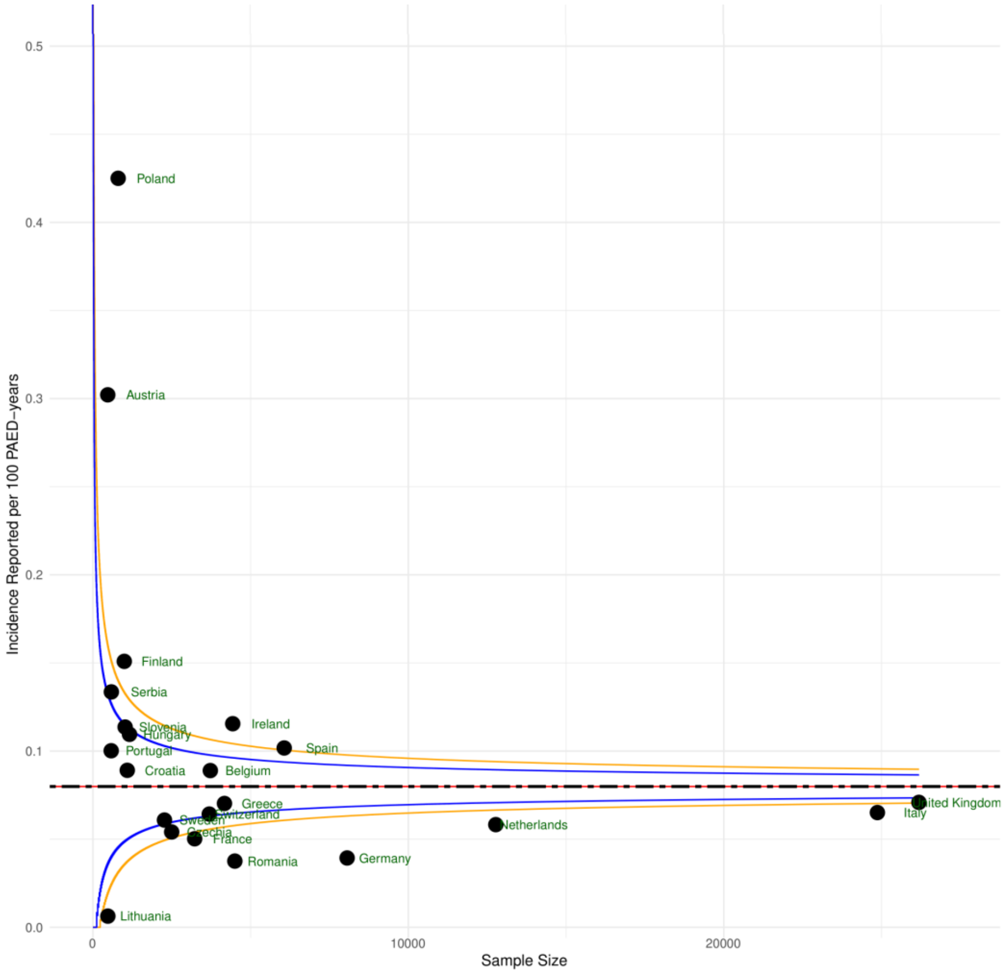


Figure 36. The funnel plot of the IBD incidence by country adjusted for the effect of latitude.

The funnel plot shows that if we account for the latitude bases difference the results tend to be homogeneous, yet with notable variation.

4.1.2. Temporal analysis of Incidence

Further analysis of the results by country and year suggests that the year of collection may also explain some of the observed variances in our data. After removing Portugal from the analysis dataset, as it was the only country with no multiple reporting over different years, we performed a mixed effects linear regression model with the year set as the fixed effect of the model and the reporting country as the random effect. The latter was set as a random effect to account for the correlation between the observations gathered in multiple years from the same countries. The used model is written as follows:

$$Incidence_{ij} = \beta_0 + \beta_{0i} + \beta_1 \cdot time_{ij} + e_{ij}$$

Where:

- $Incidence_{ij}$ is the observed incidence for the i*th* country at the j*th* collection time
- β_0 is the fixed intercept
- β_{0i} is the random intercept for the i*th* country
- β_1 is the fixed effect coefficient for the time, representing the fixed effect of time on the incidence
- e_{ij} is the residual error, which represents the variability in incidence that is not accounted for by the time or the random effects

The results suggested that the year of reporting had a significant effect on the reported PIBD incidence with a coefficient of 1.12 (95% C.I., 0.14, 2.11) and p-value of 0.025. It is worth noticing that the results were not affected when the random effect of the country was replaced with a random effect for the reporting unit. Finally, when introducing the covariate of the year as a factor, individual and specific to each year estimates were produced as shown in **Figure 37** and the incidence seems to increase gradually from 2018 to 2020, followed by a steep increase in 2021 and 2022.

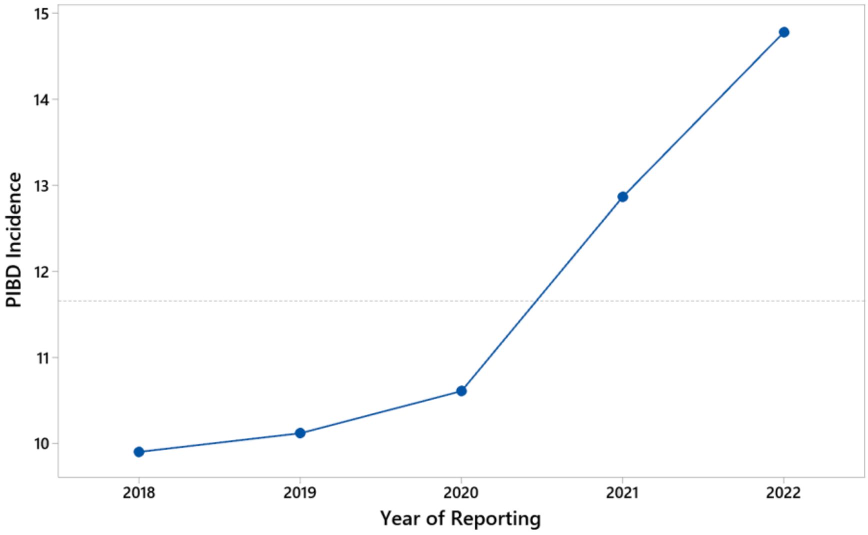


Figure 37. The effects of the year on the reported incidence.

The effects of the year on the reported incidence per 100,000 paediatric person-years, based on the mixed effects linear model with the year is introduced as a factor.

The incidence results are presented in **Figure 39** for 90 country-year combinations. Poland and Austria have been reporting exceptionally high PIBD incidence over multiple data collection years compared to the other countries. Furthermore, we can observe that the consistency of the reports varies between different countries. Comparing the average incidence reported per country in different years is an additional indication of data quality as described in the methods (2.1). In **Figure 38** below, the variance in the responses is shown for each country. Very importantly, both suspected outliers, Poland and Austria, present a significantly higher variance than the other countries.

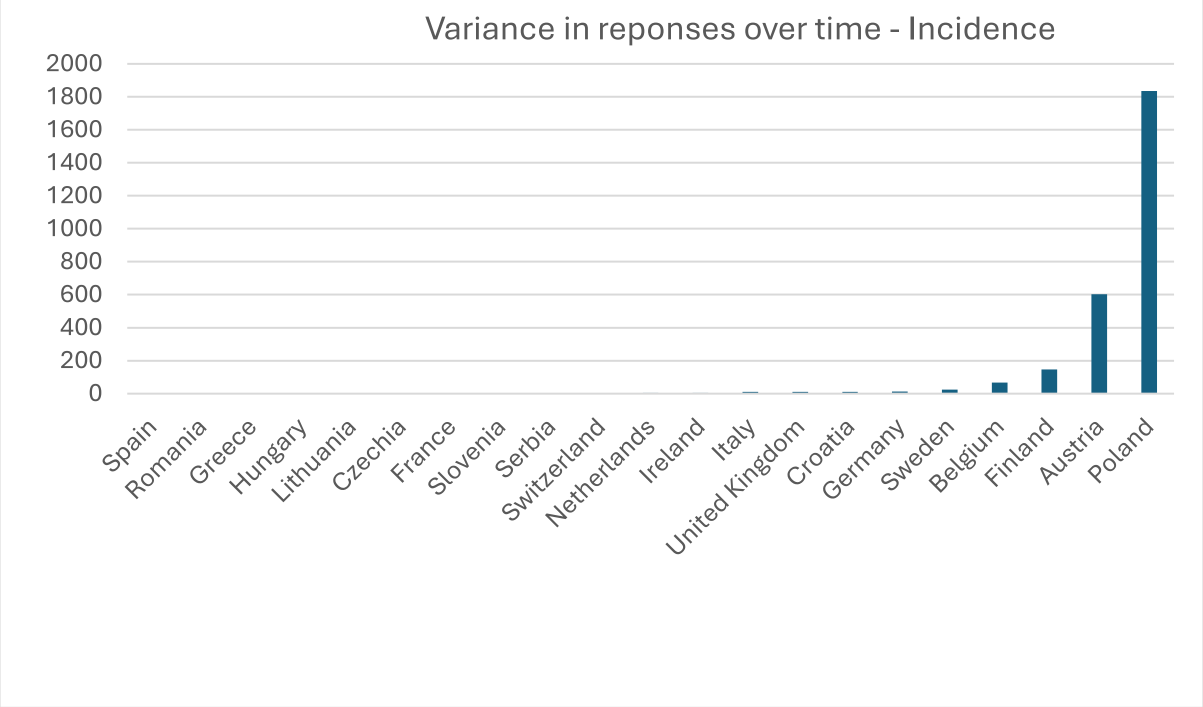


Figure 38. The variance of the annual mean of the reported incidence per county.

The high variance observed for Austria and Poland can be explained by the observed inconsistency in their reported numbers, as shown in the following **Figures 39 & 40**. **Figure 39** shows how the reports of these two countries vary compared to other countries while **Figure 40** shows the within variation per country. Very importantly, **Figure 40** also shows the incidence increase for most countries over different reporting years.

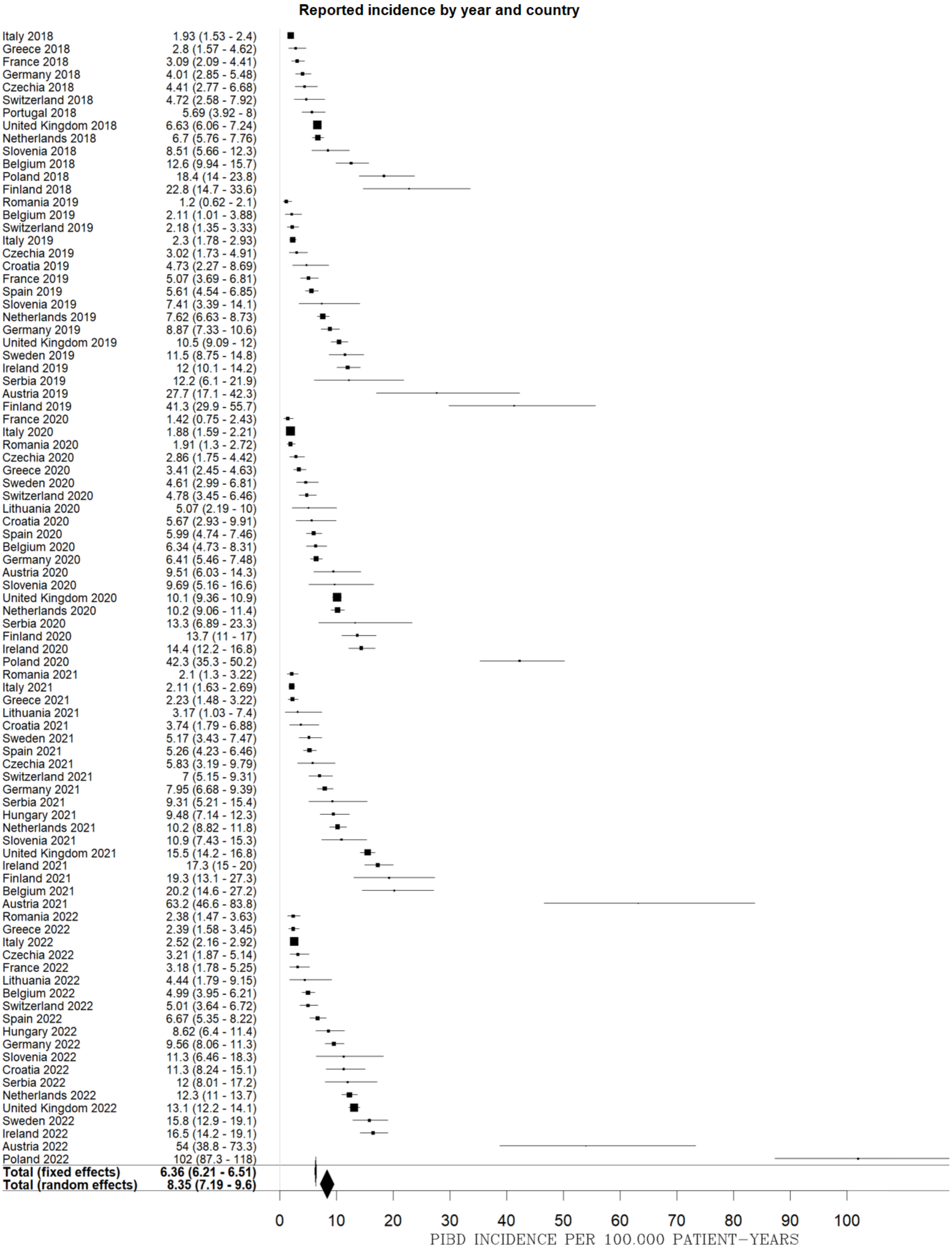


Figure 39. The longitudinal summary of the reporting by country.

The incidence results were ordered by the collection round and the incidence values.

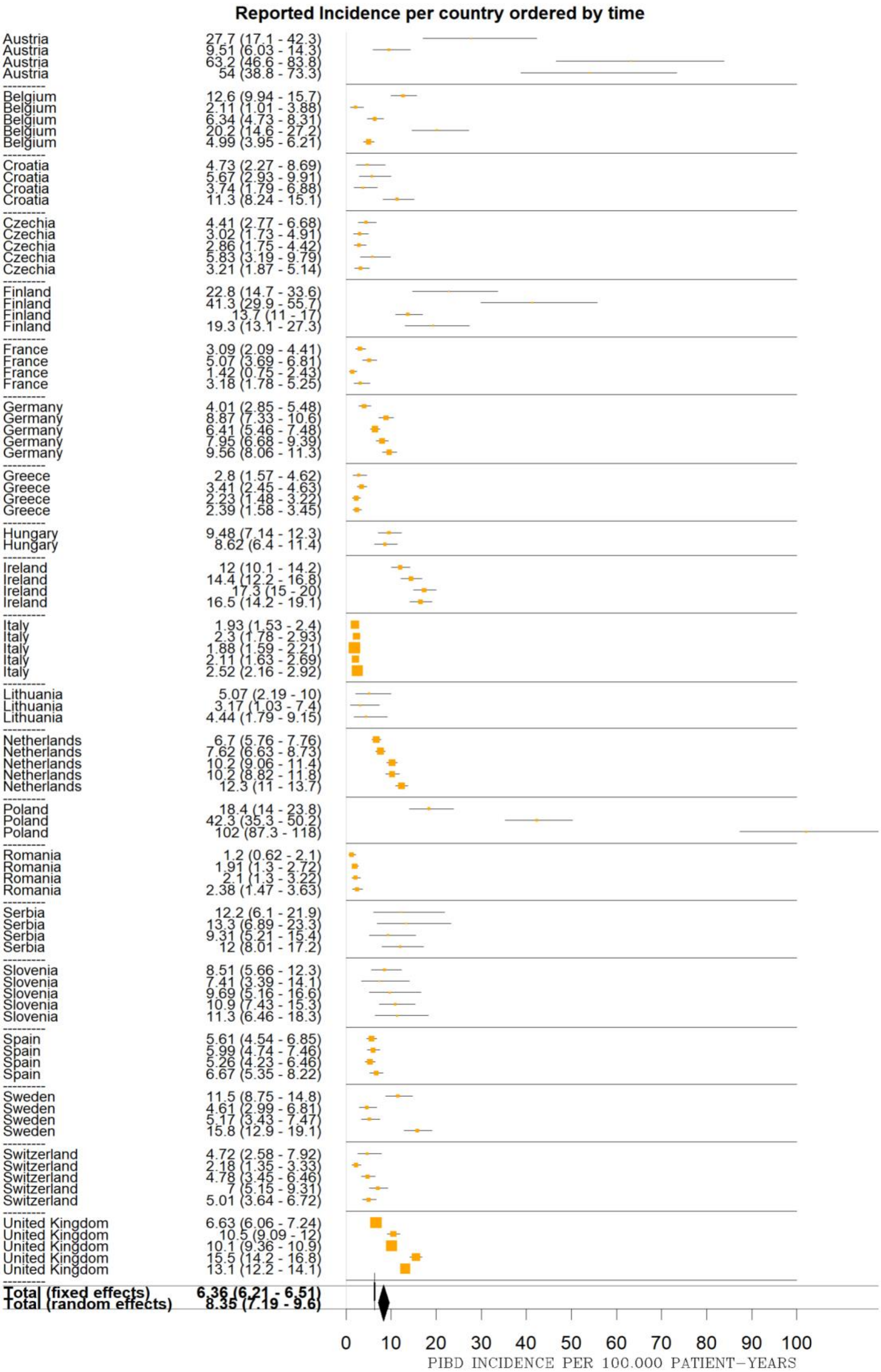


Figure 40. The reported incidence by country was ordered per data collection year.

In this forest plot, we can observe the increase in the incidence in several countries and the degree of consistency in their reporting.

4.1.3. Incidence of the clinical phenotype of IBD (Crohn's or UC/IBDU)

Expanding on the incidence analysis, it is important to consider the ratio of Crohn's disease to Ulcerative Colitis and IBD-U. This figure is vital because some risk factors affecting each subtype may be different and subtype-specific, which is a crucial element for the analysis in the following chapters. Additionally, the consistency of the reported ratio between different sites and within each site and country at different years is an important data quality indicator (erroneous report should be highly inconsistent). As shown in **Figure 41**, the reported rate of CD to all PIBD cases spans from 41% to 71% across different countries. However, in contrast to the disease incidence data, no outliers were present in this summary.

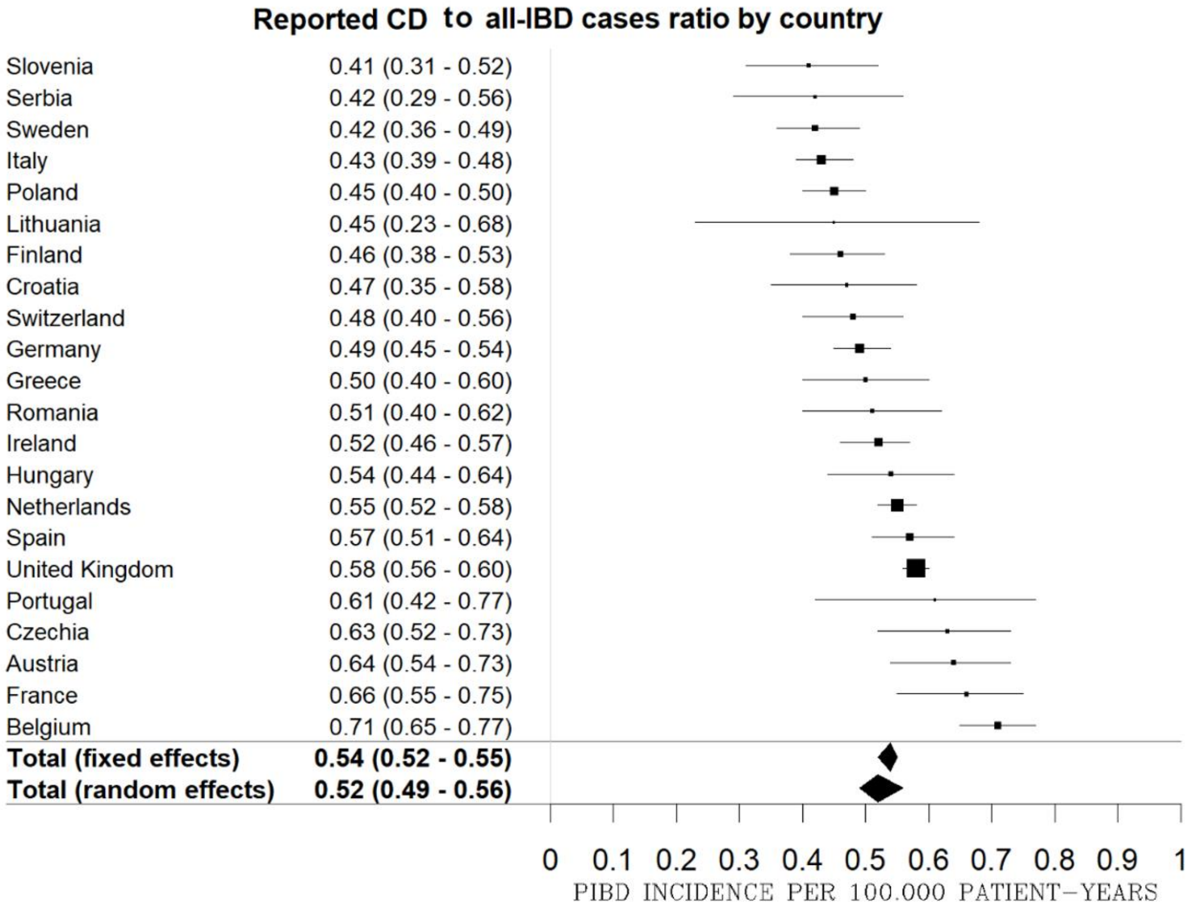


Figure 41. The ratio of the paediatric Crohn's disease incidence to the total PIBD incidence.

The phenotype ratio differences also show a trend based on the latitude shown in **Figure 42**. Specifically, the countries in central Europe have reported an 8% increase in the CD/PIBD ratio compared to those in the South. Countries in North Europe reported a 10% increase compared to those in the South. A chi-square test of the CD against the UC/IBDU cases for the three different latitude groups returned p = 0.19, while the Cochran - Armitage trend test returned p = 0.082. Therefore, the latitude trend for the phenotype ratio favouring CD in the North countries was present but was not found to be significant. In contrast, both the chi-square and the Cochran - Armitage trend tests of the CD against the UC/IBDU cases for the three different longitude groups returned p < 0.001. Therefore, the longitude trend for the phenotype ratio was found to be significant (**Figure 43**). The longitude group definitions are <3.5°East, 3.5°East to 15°East and <15°East. Even when the "traditional" definition of Eastern Europe (15°East) was used, the results remained significant.

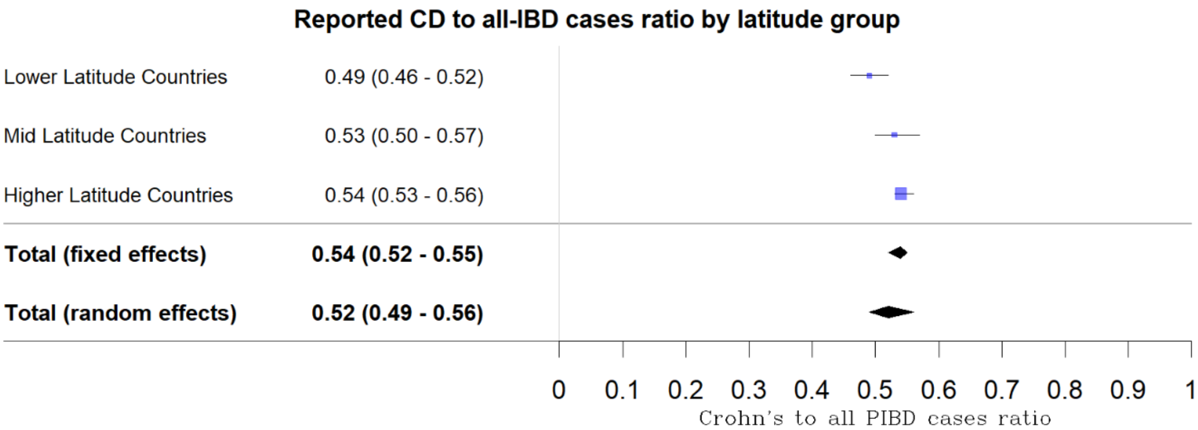


Figure 42. The ratio of Crohn's to PIBD incidence was grouped by the latitude of the reporting countries.

A latitude trend for the phenotype ratio of PIBD was present but the effect size was not found to be significant.

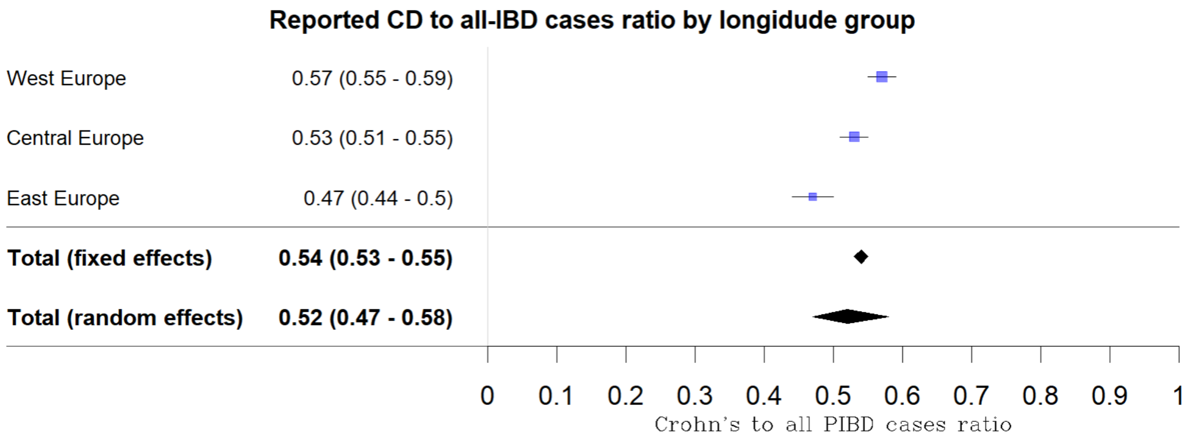


Figure 43. The ratio of Crohn's to PIBD incidence was grouped by the longitude of the reporting countries.

A significant longitude trend for the phenotype ratio of PIBD was present. This was a linear trend (3.5°East, 9.25°East and 15°East groups).

The CD to PIBD ratio results is presented in **Figure 44** for 90 country-year combinations. Most countries did not present a temporal trend for the disease phenotype ratio, although a few countries showed a decreasing pattern, with the UK presenting the strongest trend.

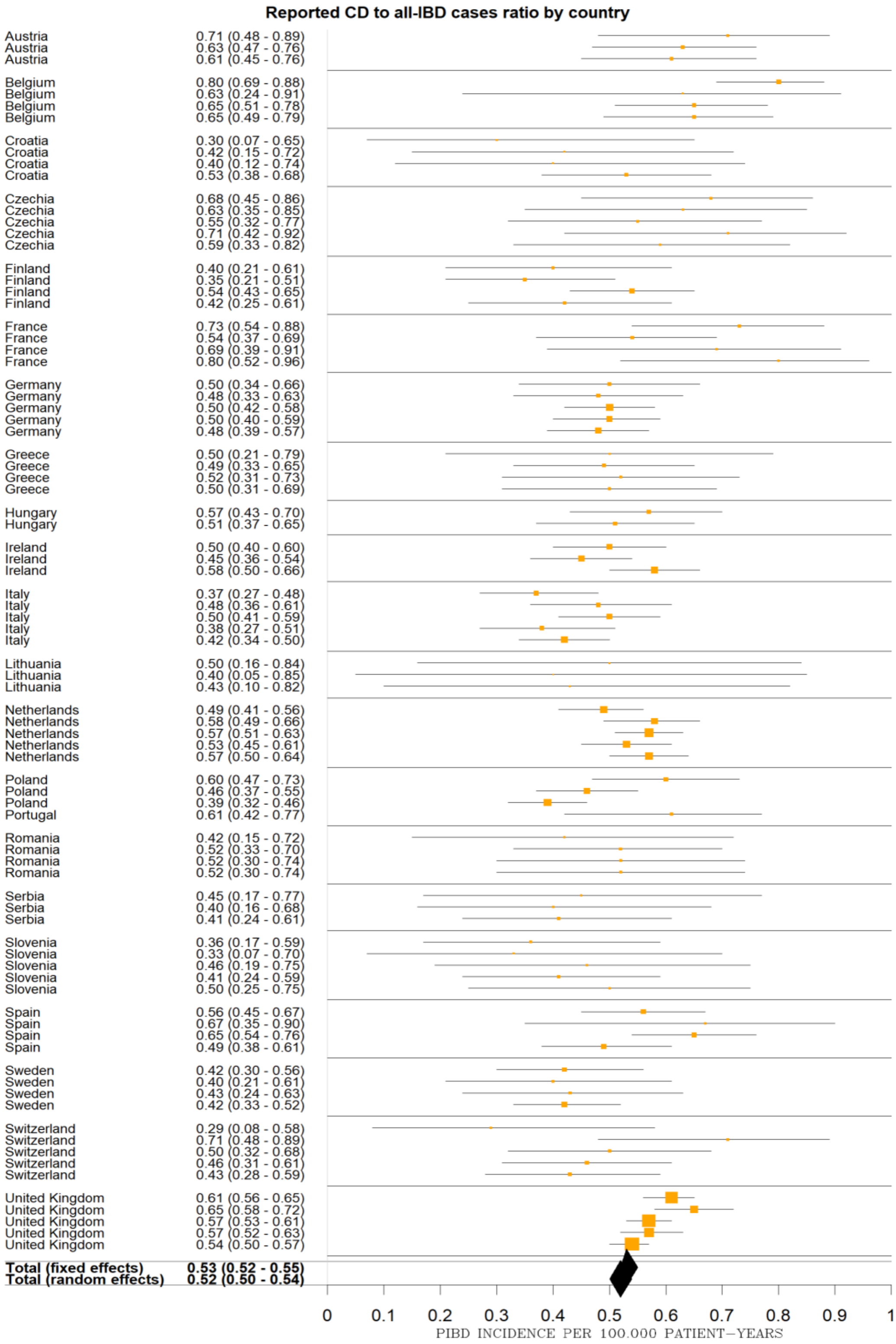


Figure 44. The ratio of Crohn's to PIBD incidence is grouped by country and ordered by the time of reporting.

Although in certain countries such as the UK, the CD to UC/IBDU ratio seems to decrease, overall, there are no evident temporal trends on most of the variation that appears to be random.

4.1.4. Discussion of PIBD incidence results

In this chapter, we have presented the results of our study on the incidence of paediatric inflammatory bowel disease (PIBD) based on five years of prospective data collection. Our findings indicate a strong latitude trend, which aligns with the previous literature, and therefore supports the validity of our results. Furthermore, our analysis reveals a consistent and significant increase in the pooled incidence of PIBD across Europe during the study period from 2018 to 2022. As shown in **Figure 40,** our data show a significant and steady rise in incidence rates in several countries such as the U.K., the Netherlands, Ireland, Romania, and Slovenia. In addition, an incidence increase was also present in Switzerland, Sweden, Spain, Poland, Italy, Austria, and Croatia although the results from these countries were less consistent. Only 4 out of the 21 countries reported a decreasing incidence, which was not statistically significant due to the small effect and sample size. Interestingly, the observed incidence increases linearly from 2018 to 2020, followed by a steeper increase in the 2020 to 2022 reporting period. Considering that the few outliers are not adequate to explain this increase, this raises the question of whether the COVID-19 pandemic may have influenced the referral pathways in a way that could inflate the reported incidence. Although there are several studies reporting the effects of IBD on the COVID-19 incidence we were not able to find any studies reviewing the incidence of Crohn's disease and ulcerative colitis before and after the beginning of the COVID-19 pandemic (Allocca et al., 2020; Ungaro et al., 2021). The exact reasons for this increase are not yet clear and further research is needed to understand the underlying causes. However, there are some possible explanations for this increase. It is possible that COVID-19 introduced a delay in the diagnosis and treatment of the disease. During the pandemic, many people may have delayed seeking medical care due to concerns about exposure to COVID-19 or due to changes in healthcare delivery. This delay could have led to a worsening of symptoms and a delay in diagnosis, resulting in a higher incidence of IBD. However, in our data, we have observed a stable increase for both 2021 and 2022 which questions this theory. Another explanation may be related to changes in lifestyle. Lockdowns and restrictions implemented to slow down the spread of COVID-19 have led to changes in lifestyle, including decreased physical activity, changes in diet, reduced sun exposure and increased stress as well as other factors that have been linked to an increased risk of IBD. It is important to repeat that, as reported in the literature, sun exposure is a protective factor in PIBD. In this study, we are also reporting the latter as one of our secondary findings.

Furthermore, changes in gut microbiota related to the COVID-19 infection and medications used to prevent or treat it may also be related to this temporal trend. Disruption in the gut microbiota which has been linked to the development of IBD and, therefore could potentially increase the risk of developing IBD or worsen existing symptoms. Lastly, the possibility of immune dysregulation should also be considered. The COVID-19 infection and the immune response it triggers can lead to dysregulation of the immune system, which is also thought to play a role in the development of IBD.

We also analysed the disease phenotype ratio, which can be used as an additional validation metric of the reporting quality and may reveal areas that are exposed to certain factors affecting one of the two phenotypes more than the other. The results varied between countries, and a clear longitude effect was identified, suggesting that Eastern Europe has higher levels of UC which we have also established from the literature. In addition, a latitude effect was also detected, but it was not significant. This can mean that the observed trend was either random or that the latitude effect on phenotype has a smaller effect size and requires a large sample size to be detected. The following data collection rounds will provide more clarity on this.

Another finding from our analysis is that the consistency and variability in the reported incidence per country strengthen our suspicion that Austria and Poland are outliers due to inaccurate reporting. These two countries failed most outlier tests and Poland specifically reported a 10-fold higher incidence compared to the average and three-fold compared to its own average in the most recently submitted eCRF. However, considering that the phenotype ratio for these two countries was within the expected range, we can assume that the submitted patient numbers are accurate and that the reporting issues are related to erroneous reporting of their catchment areas.

In summary, our study confirms the latitude trend in the PIBD incidence, and a significant longitude effect on the phenotype of the disease. Regarding the temporal effects, we have identified an apparent increase in the incidence of PIBD in Europe, with some countries reporting more consistent and significant increase than others. These results highlight the need for further research to understand this trend's underlying factors.

5. RESULTS – Safety registry Reported Incidence

5.1. International rates of PIBD prevalence

The prevalence results are presented by country in the following **Table 15** and forest plots (**Figures 45 & 46**) These summaries also include the reported prevalence using fixed and random effects approaches, as presented in the incidence paragraphs. Similarly to the incidence analysis, given the high degree of heterogeneity, the random effects model is favoured as the most accurate estimate of the pooled prevalence that was reported in this study. The pooled prevalence based on the random effects and fixed effects models was 31.4 (95% C.I.: 30 to 31.7) and 38.9 (95% C.I.: 30 to 48.8) cases per 100,000 paediatric person-years, respectively. However, the simple weighted average of the reported prevalence from all European sites was 34.13 new cases per 100,000 paediatric person-years.

Table 15 The PIBD prevalence estimates with their respective confidence intervals and sample size as reported by each country for 2018 – 2022.

The number of reports that were included in the calculations in total and per year was also reported.

Country	Lower C.I. 95%	Prevalence (/100,000)	Upper C.I. 95%	Sample size	2018	2019	2020	2021	2022	Million PAED-Years
Romania	4.93	5.60	6.34	4	0	1	1	1	1	4.50
Czechia	12	13.40	14.9	5	1	1	1	1	1	2.50
Italy	16.2	16.70	17.2	23	3	4	7	2	7	24.07
Greece	15.5	17.00	18.5	3	1	0	1	0	1	2.91
France	16	17.50	19	7	2	2	2	0	1	3.22
Lithuania	14.2	17.80	22	3	0	0	1	1	1	0.47
Spain	25.1	26.40	27.8	10	0	2	2	3	3	5.81
Switzerland	25.6	27.30	29	19	2	4	5	4	4	3.69
Croatia	29.3	32.60	36.2	4	0	1	1	1	1	1.09
Hungary	30.5	33.80	37.3	2	0	0	0	1	1	1.16
Slovenia	31	34.50	38.3	6	2	1	1	1	1	1.02
Portugal	29.9	34.50	39.6	1	1	0	0	0	0	0.58
Belgium	32.8	34.70	36.7	13	3	1	3	2	4	3.72
Germany	33.8	35.10	36.4	20	1	4	4	4	7	8.06
Sweden	36.8	39.30	42	6	0	2	1	1	2	2.27

Netherlands	38.9	40.00	41.1	62	12	11	16	11	12	12.48
UK	45.5	46.30	47.1	52	11	5	17	9	10	26.06
Serbia	44.1	49.70	55.8	4	0	1	1	1	1	0.58
Ireland	47.7	49.70	51.8	4	0	1	1	1	1	4.43
Austria	57.7	64.70	72.4	4	0	1	1	1	1	0.47
Finland	158	166.00	174	5	1	1	2	1	0	0.99
Poland	209	219.00	229	3	1	0	1	0	1	0.80

The confidence in the reported prevalence is almost analogous to the sample size, which was the product of the reporting participants, sample size and the number of reporting years per country. **Figure 45** below shows how the confidence is reflected in the size of the confidence intervals per country. Also, the prevalence estimates for Austria and Poland appear to deviate equally or more compared to the incidence results.

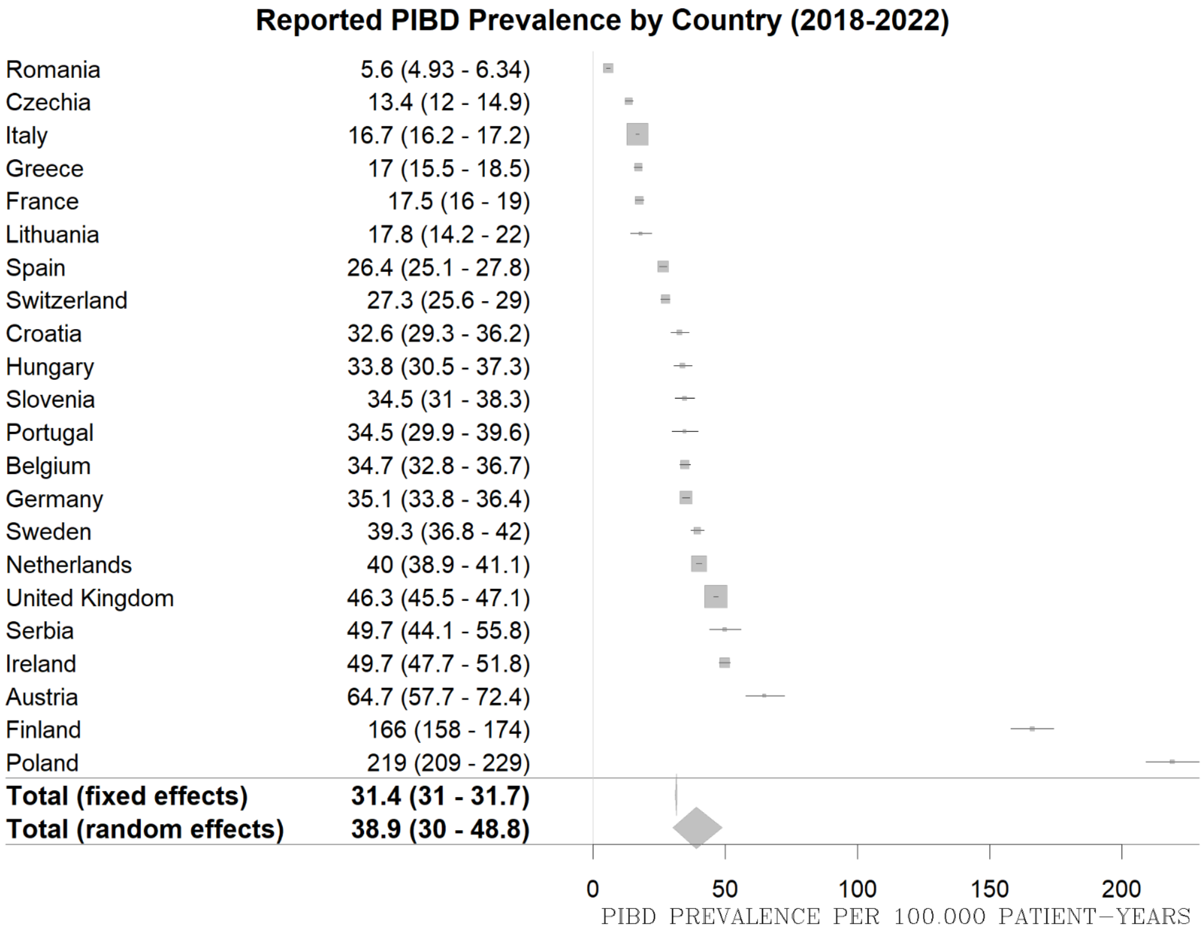


Figure 45. Forest plot of the PIBD prevalence estimates with their respective confidence intervals as reported by each country for 2018 – 2022.

Each box's size is analogous to each country's sample size.

As shown in the funnel plot in **Figure 46**, there is no significant reporting bias in the reported prevalence rates by country, considering that the deviations are not asymmetrical. However, the number and extent of the deviations suggest that heterogeneity is present in the reported prevalence rates between countries. Although, apart from Poland, the number and extent of the significant deviations from the median are noticeably lower for the prevalence than the incidence of the disease.

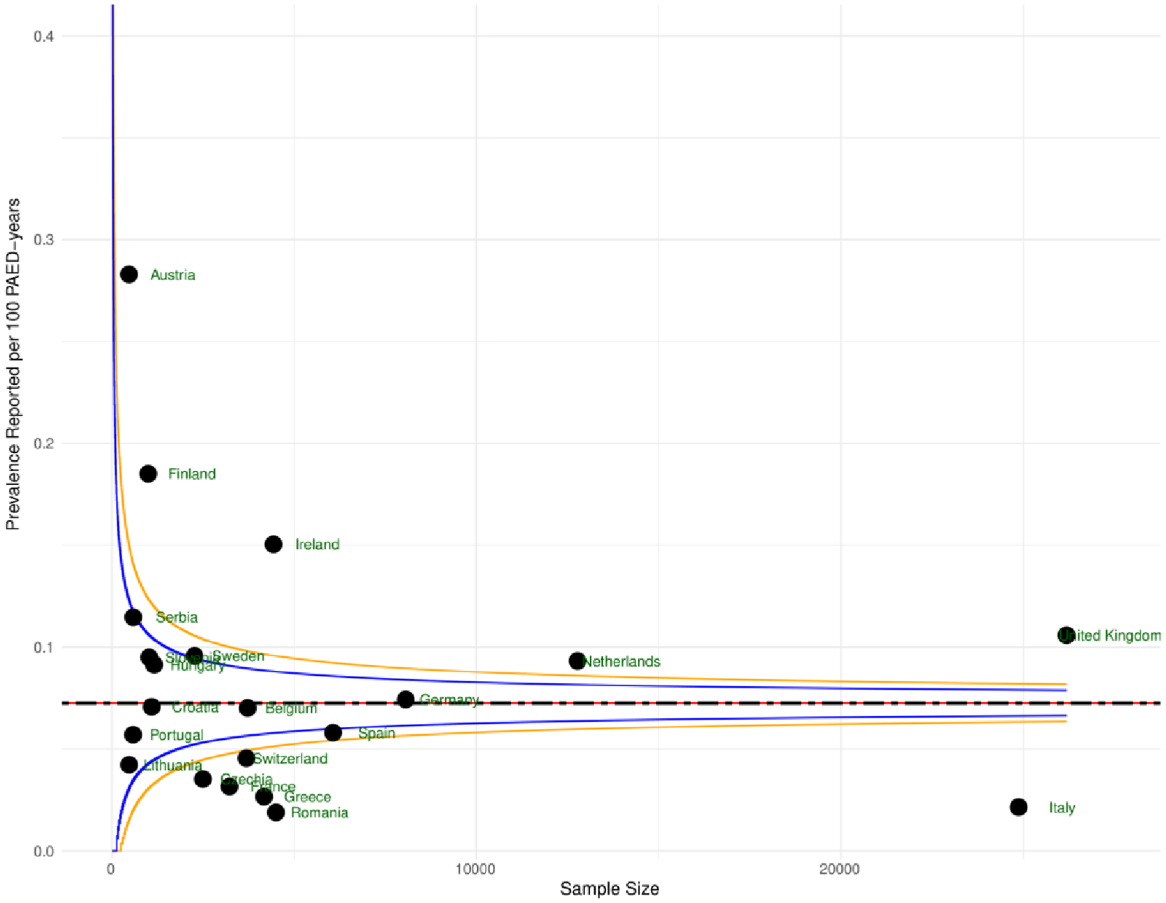


Figure 46. Funnel plot of the PIBD prevalence results as reported by country per 1000 PAED-years.

Please note that Poland was an extreme outlier and was not included in the plot as it would require a scale adjustment. Austria was also an outlier, similar to the incidence results. The overall consistency of the prevalence is very high, considering that these values are not adjusted for the latitude effect.

5.1.1. Spatial analysis of the phenotype ratio of PIBD

The prevalence differences that depend on the region can be summarised in latitude groups, as shown in **Figure 47**. These results are consistent with the incidence trends reported in the previous chapter with countries in central Europe (between 41° and 51° North) having reported an almost a 30% increase in the prevalence compared to the countries in the South (<41° North), while countries in North Europe (>51° North) also reporting an almost 3-fold increase compared to the countries in the South.

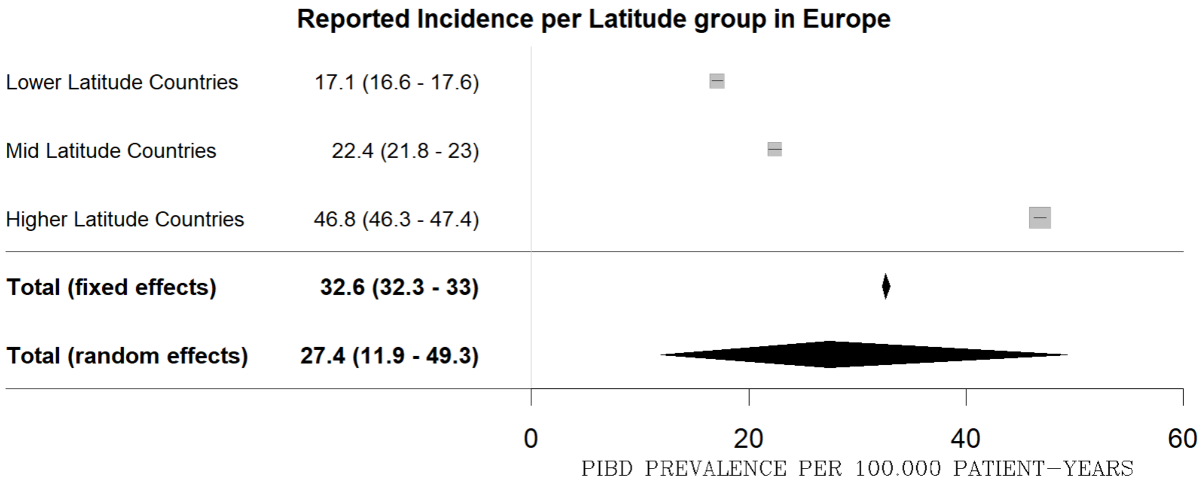


Figure 47. The PIBD prevalence was grouped by the latitude of the reporting countries.

5.1.2. Temporal analysis of Prevalence

Further analysis of the results by year and country suggests that the year of collection can also explain some of the observed variances in our data. As expected, the temporal trend of the prevalence also shows an increase over time. Similarly, to the incidence analysis, the mixed effects linear regression model suggested that the year has a significant effect (p value = 0.018) on the reported PIBD prevalence. After introducing the covariate of the year as a factor, separate specific to each year estimates were produced as shown in **Figure 48**. The prevalence increases by a small percentage from 2018 to 2020, followed by a steep increase in 2021 and 2022.

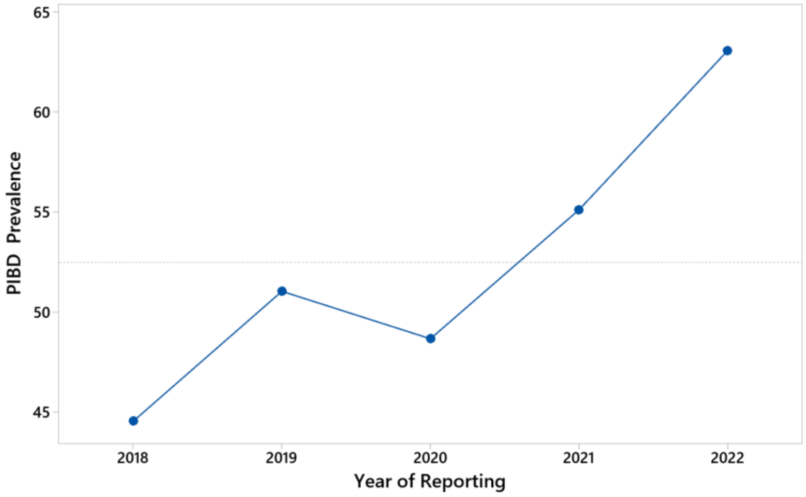


Figure 48. The reported increase in PIBD prevalence by year

The effects of the year on the reported prevalence per 100,000 paediatric person-years were based on the mixed effects linear model where time was introduced as a factor.

In **Figure 50**, the prevalence results are presented for 88 country-year combinations. Finland, Poland and Austria have reported higher PIBD prevalence over multiple data collection years and showed the highest variance compared to the other countries. Furthermore, we can observe that the consistency of the reports varies between countries. In **Figure 49** below, the variance in the responses is shown for each country. Poland and Austria present high variance consistent with the incidence results, however, Finland presents an extreme value which is caused by inconsistent reporting between 2018/19 and 2020/2021 with extreme values reported in 2018/19 as we can see in **Figure 51.**

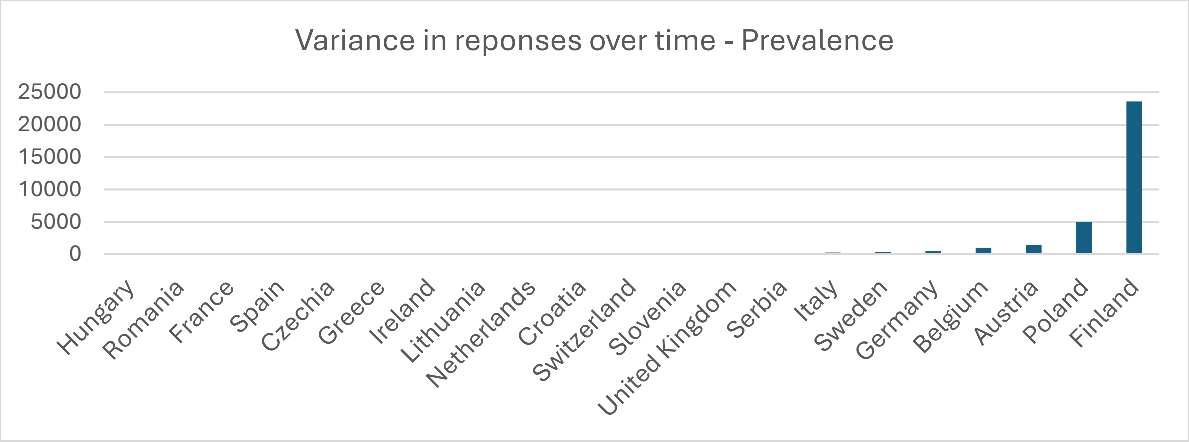


Figure 49. The variance of the annual mean of the reported prevalence per county.

The high variance observed for Austria, Poland and Finland can be explained by the observed inconsistency in their reported figures, as shown in the following **Figures 50 & 51**.

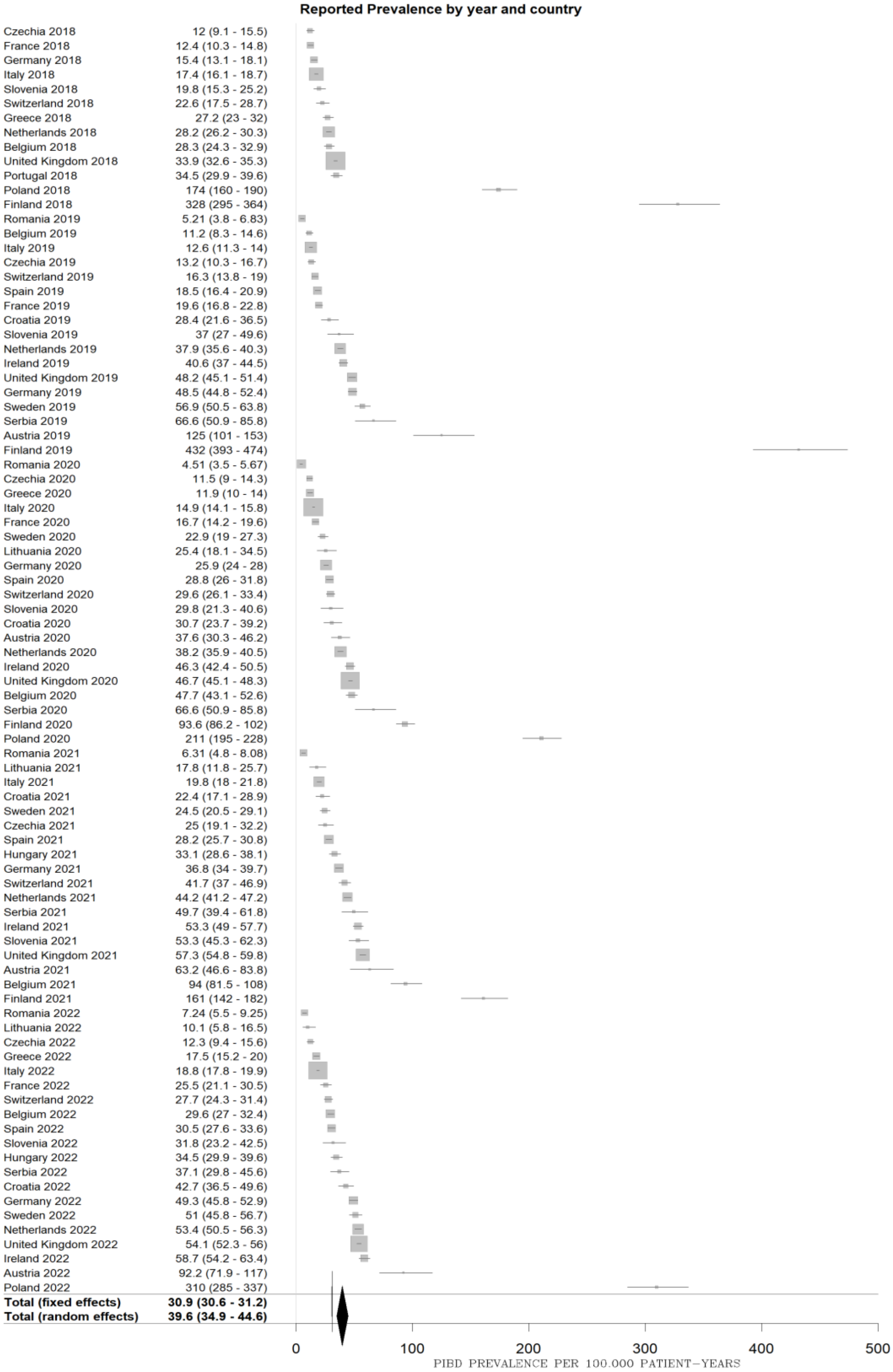


Figure 50. The longitudinal summary of the reporting by country.

The prevalence results were ordered by the collection round, followed by the incidence level.

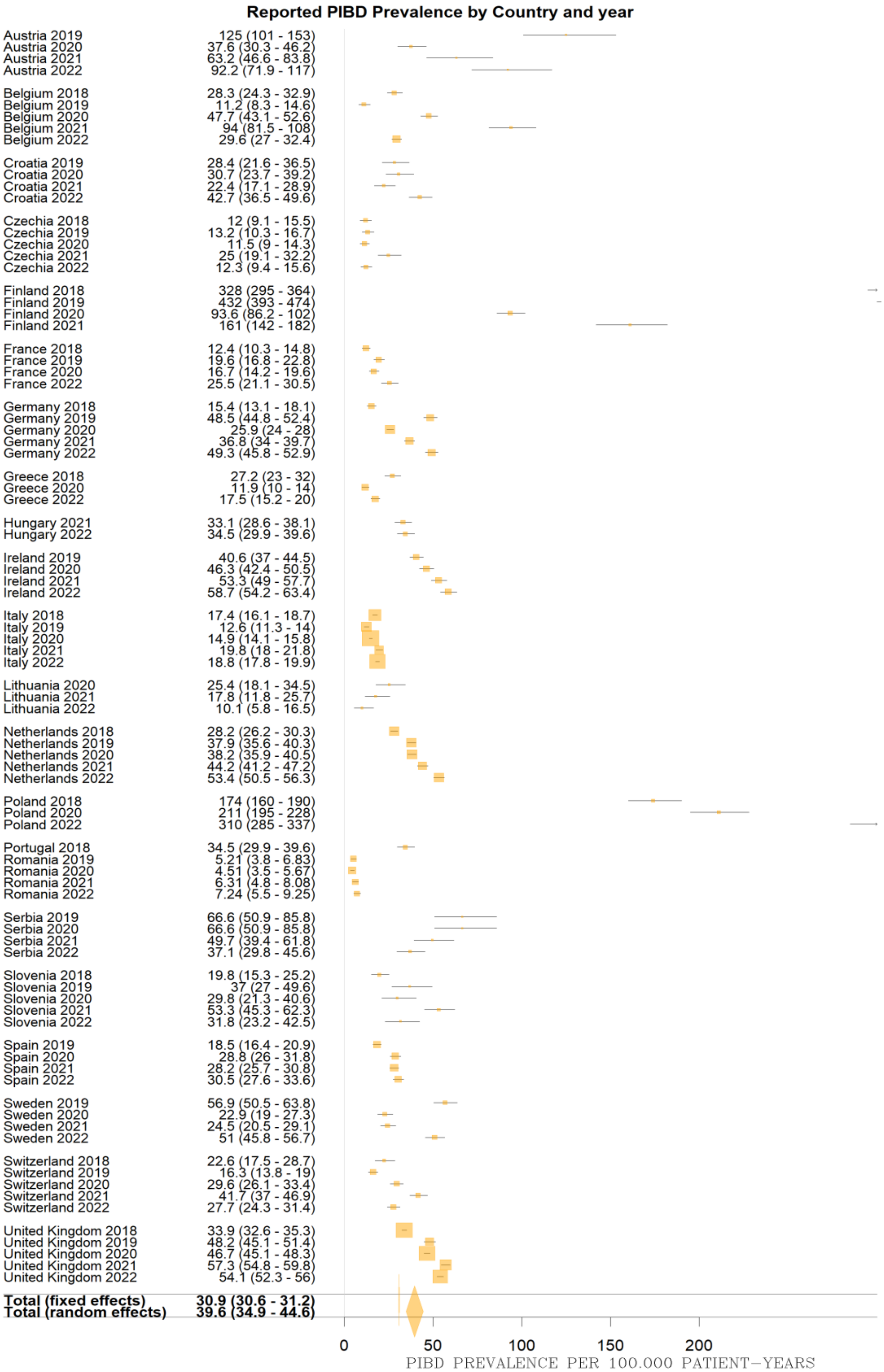
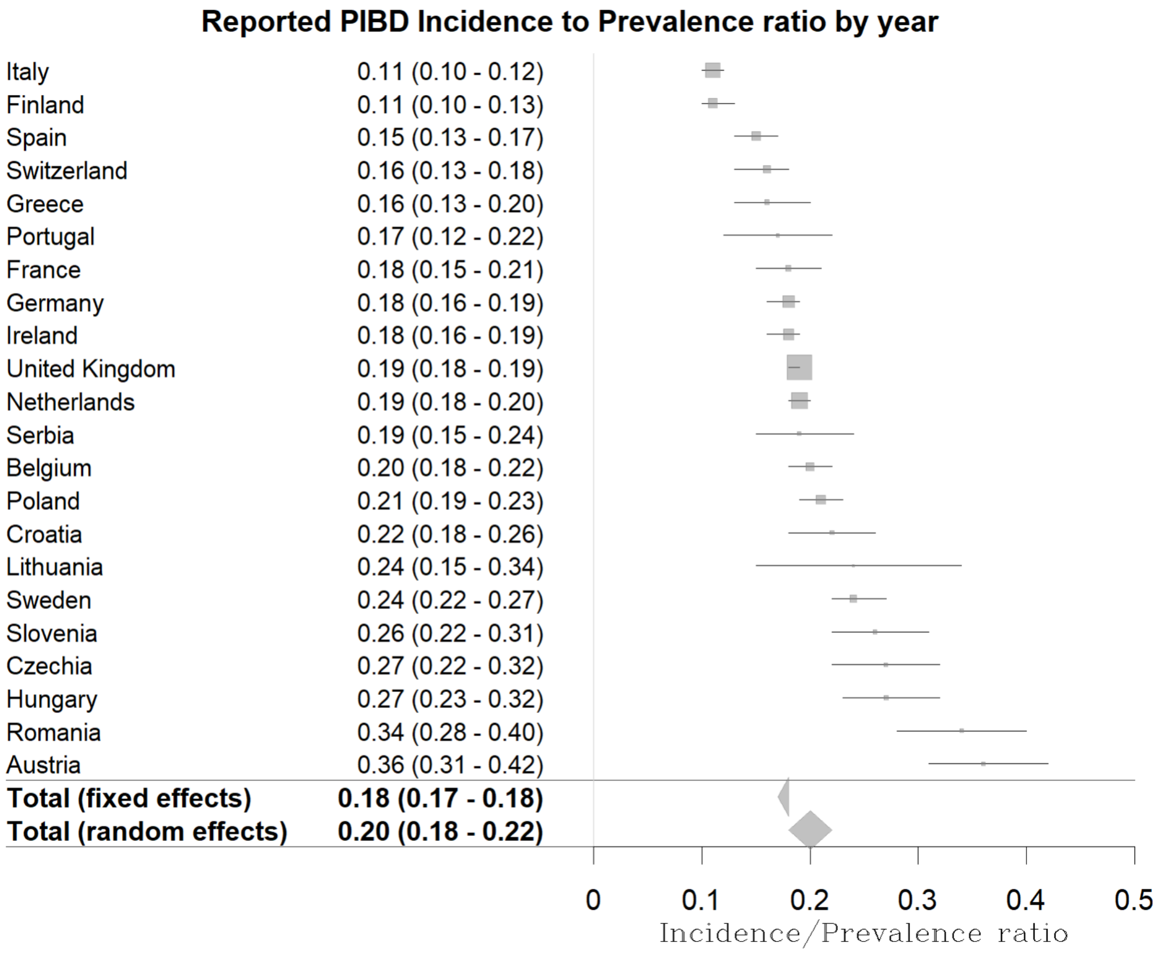


Figure 51. The reported prevalence by country was ordered per data collection year.

In this forest plot, we can observe the increase in the prevalence in several countries and the degree of consistency in their reporting.

5.1.3. Incidence to prevalence ratio of PIBD

The ratio of Crohn's disease to Ulcerative Colitis and IBD-U was not investigated for the prevalence results as this information was not collected in the Safety Registry. However, an important figure that was investigated was the incidence-to-prevalence ratio (IPR). This ratio is an important metric that can be used as an indicator of the incidence progression over time and as an additional validation metric to assess the quality of the reported data by country and year (**Figure 52**). In the following paragraphs we will report the observed IPR and estimate the expected IPR based in the methods described in 2.2.1 for comparison purposes. According to our data, the overall correlation between the reported incidence and prevalence was 0.776 (Spearman Rho p<<0.0001), while the pooled random effects estimate of the IPR was calculated at 20% (95% C.I.: 18% to 22%) as shown in **Figure 52**.



Reported PIBD Incidence to Prevalence ratio by year

Italy	0.11 (0.10 - 0.12)
Finland	0.11 (0.10 - 0.13)
Spain	0.15 (0.13 - 0.17)
Switzerland	0.16 (0.13 - 0.18)
Greece	0.16 (0.13 - 0.20)
Portugal	0.17 (0.12 - 0.22)
France	0.18 (0.15 - 0.21)
Germany	0.18 (0.16 - 0.19)
Ireland	0.18 (0.16 - 0.19)
United Kingdom	0.19 (0.18 - 0.19)
Netherlands	0.19 (0.18 - 0.20)
Serbia	0.19 (0.15 - 0.24)
Belgium	0.20 (0.18 - 0.22)
Poland	0.21 (0.19 - 0.23)
Croatia	0.22 (0.18 - 0.26)
Lithuania	0.24 (0.15 - 0.34)
Sweden	0.24 (0.22 - 0.27)
Slovenia	0.26 (0.22 - 0.31)
Czechia	0.27 (0.22 - 0.32)
Hungary	0.27 (0.23 - 0.32)
Romania	0.34 (0.28 - 0.40)
Austria	0.36 (0.31 - 0.42)
Total (fixed effects)	**0.18 (0.17 - 0.18)**
Total (random effects)	**0.20 (0.18 - 0.22)**

Figure 52. The reported incidence-to-prevalence ratio by country.

To study the reported IPR, we first needed to determine the expected IPR ratio. We can consider an example where we are following up a population of 100,000 individuals aged 0 to 18, split equally into 18 groups for each year of age. Assume that in this population, the probability of developing the disease each year is *p* and is the same for every year of age. Therefore, the expected number of new diagnoses in the youngest group of the population at year 1 of the follow-up is $\left(\frac{100,000}{18}\right) \cdot p$. This is simply the product of the population fraction and the probability to develop the disease. Similarly, the probability of a new diagnosis at the age of 2 is the same after excluding the patients already diagnosed with the disease in this age group (the previous year) which makes the calculation $(\frac{100,000}{18} - \left(\frac{100,000}{18}\right) \cdot p) \cdot p$. In rare diseases, given the small incidence, the latter can also be written as $\left(\frac{100,000}{18}\right) \cdot p$ since the number of patients who already have the disease is very small. Therefore for a rare disease, the prevalence in the population up to the age of 2 is $1 \cdot \left(\frac{100,000}{18}\right) \cdot p$ + $2\cdot \left(\frac{100,000}{18}\right) \cdot p$, while for a population up to the age of 3, this will be $1 \cdot \left(\frac{100,000}{18}\right) \cdot p$ + $2\cdot \left(\frac{100,000}{18}\right) \cdot p$ + $3\cdot \left(\frac{100,000}{18}\right) \cdot p$ and similarly for all age group upwards. This means that the prevalence in a population with *n* years of age is the *nth* triangular number multiplied by the expected number of cases per age group: $\frac{i(i+1)}{2} \cdot \left(\frac{100,000}{18}\right) \cdot p.$

For an examined age *i*, since the incidence at the age *i* for the population is: $i \cdot \left(\frac{100,000}{\mathrm{n}}\right) \cdot p$ and the prevalence is $\frac{i(i+1)}{2} \cdot \left(\frac{100,000}{n}\right) \cdot p$ the IPR is:

$$\frac{i\cdot\left(\frac{100,000}{\mathrm{n}}\right)\cdot p}{\frac{i(i+1)}{2}\cdot\left(\frac{100,000}{n}\right)\cdot p} = \frac{i}{\frac{i(i+1)}{2}} = 10.5\%$$

However, in our case, as discussed in the methods section (2.2.1), the probability of a new PIBD diagnosis varies by age. Therefore, the different years of age contribute significantly different numbers of patients to the total prevalence. In the **Tables 16 & 17** below, the calculations for both a constant (our previous example), and study-specific incidence by age group are presented. In **Tables 16**, the study-specific calculations are based on the function described in the incidence adjustment section (2.2.1) and assume that the PIBD incidence in our population is 8.73 per 100,000 paediatric person-years. This is based on our findings (Chapter 4) using the random effects pooled estimate per country.

Our calculations consider that each individual in the paediatric population is exposed to a variable risk annually, resulting in an expected incidence-to-prevalence ratio of 0.19 (**Table 16**). The estimated expected IPR is in agreement with the observed IPR in our data since the latter is between 0.18 and 0.20, which are the pooled fixed and random effects IPR estimates (**Figure 52**). This indicates that the incidence adjustment was accurate and strengthens the validity of the overall incidence and prevalence estimates of the study. The following **Figure 53** shows the study specific calculation of the expected incidence for different age cut offs.

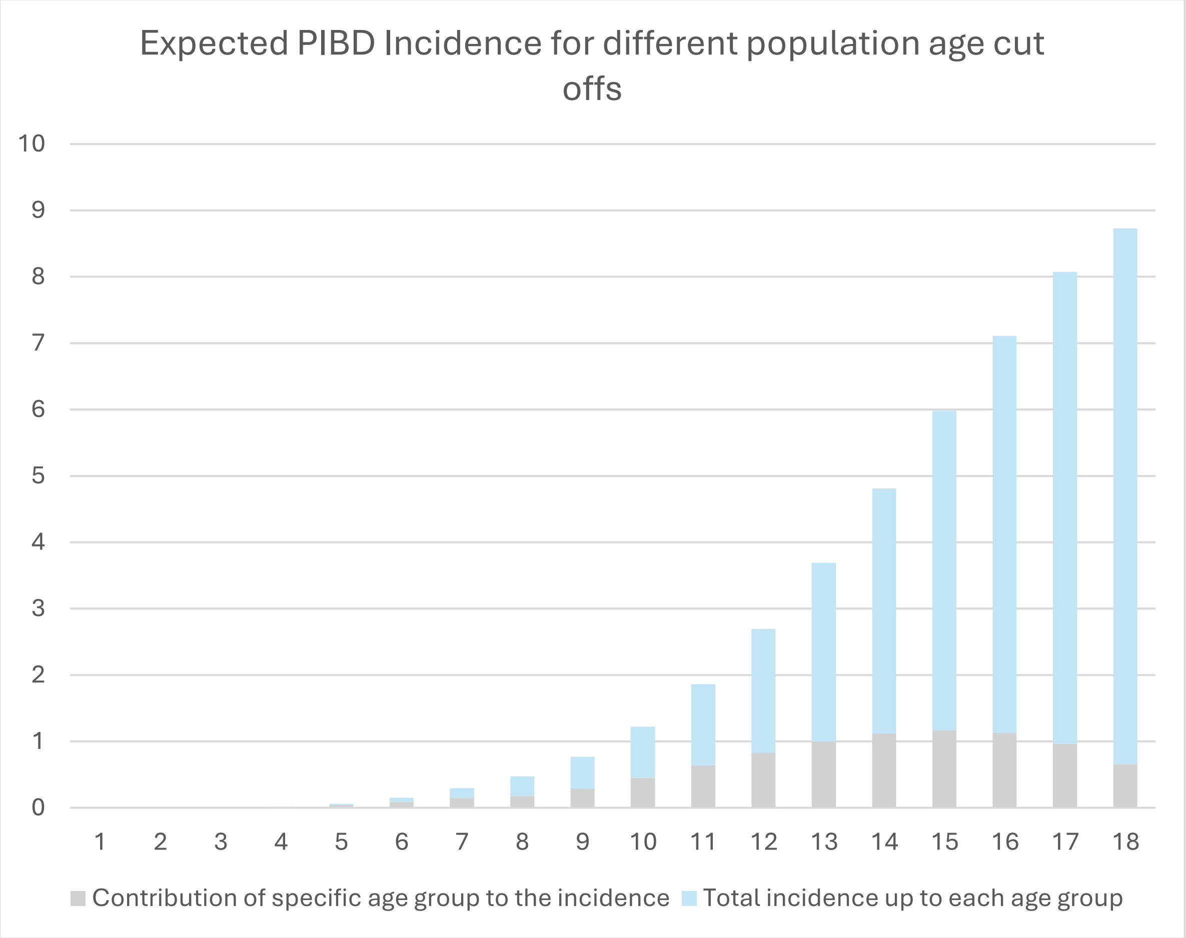


Figure 53. Overall incidence up to each year of age and individual contribution by year.

A stacked column plot showing how the expected incidence changes depending on the upper age limit of the population. Each column includes two parts. The lower part shows how many new patients are expected to present the disease at that age, while the upper part shows how many patients are expected to present the disease in all age groups before that year. For instance, the stacked columns for the age of 15 show that annually 6 patients per 100k *individuals are expected to present the disease in a population up to the age of 15. The lower part of the bar shows that the contribution of the age of 15 is 1 patient while the remaining 5 are from the 0-14 age groups.*

The study-specific calculations of the expected incidence for different age cut-offs shown in **Figure 53** are also summarised in **Table 16** below.

Table 16 The estimation of the expected incidence-to-prevalence ratio.

The study-specific expected contribution of new PIBD cases per age group results in a prevalence that increases in a sigmoid-like function manner. The expected incidence in this example was 8.73 per 100,000 PAED-years and 45.95 per 100,000 paediatric with an incidence prevalence ratio of 0.19.

Study-specific variable risk calculation Incidence and Prevalence expectations in a 0-18 population with 100,000 individuals			
Age group	Incidence per age group	New cases per age group	All new cases up to this age group
0-1	0.00000%	0.000	0.000
1-2	0.00002%	0.001	0.001
2-3	0.00008%	0.005	0.006
3-4	0.00016%	0.009	0.015
4-5	0.00081%	0.045	0.060
5-6	0.00162%	0.090	0.150
6-7	0.00262%	0.146	0.296
7-8	0.00326%	0.181	0.477
8-9	0.00525%	0.292	0.768
9-10	0.00815%	0.453	1.221
10-11	0.01153%	0.641	1.862
11-12	0.01495%	0.831	2.693
12-13	0.01798%	0.999	3.691
13-14	0.02017%	1.120	4.812
14-15	0.02108%	1.171	5.983
15-16	0.02029%	1.127	7.110
16-17	0.01735%	0.964	8.074
17-18	0.01182%	0.657	8.730
Total		8.73 (Incidence)	45.95 (Prevalence)
IPR	**0.190**		

The calculations of the expected incidence under the assumption of a constant risk for all ages are summarised in **Table 17** below.

Table 17 The expected contribution of new PIBD cases per age group

For an equal chance of developing the disease across all groups results in a prevalence that increases linearly. The expected incidence in this example was 8.73 per 100,000 paediatric-years and 82.94 per 100,000 paediatric-years with an incidence prevalence ratio of 0.11.

Constant risk calculation Incidence and Prevalence expectations in a 0-18 population with 100,000 individuals			
Age group	Incidence per age group	New cases per age group	All new cases up to this age group
0-1	0.00873%	0.485	0.485
1-2	0.00873%	0.485	0.970
2-3	0.00873%	0.485	1.455
3-4	0.00873%	0.485	1.940
4-5	0.00873%	0.485	2.425
5-6	0.00873%	0.485	2.910
6-7	0.00873%	0.485	3.395
7-8	0.00873%	0.485	3.880
8-9	0.00873%	0.485	4.365
9-10	0.00873%	0.485	4.850
10-11	0.00873%	0.485	5.335
11-12	0.00873%	0.485	5.821
12-13	0.00873%	0.485	6.306
13-14	0.00873%	0.485	6.791
14-15	0.00873%	0.485	7.276
15-16	0.00873%	0.485	7.761
16-17	0.00873%	0.485	8.246
17-18	0.00873%	0.485	8.731
Total		8.73 (Incidence)	82.94 (Prevalence)
IPR	**0.11**		

The calculated incidence from the time of birth up to each year of age is shown as calculated in **Tables 16 & 17**. Although in both scenarios, the total 0-18 incidence is the same (8.73 per 100,000 paediatric person-years), the difference in the rate of incidence across the age groups has substantial effects on the expected prevalence (**Figure 54**).

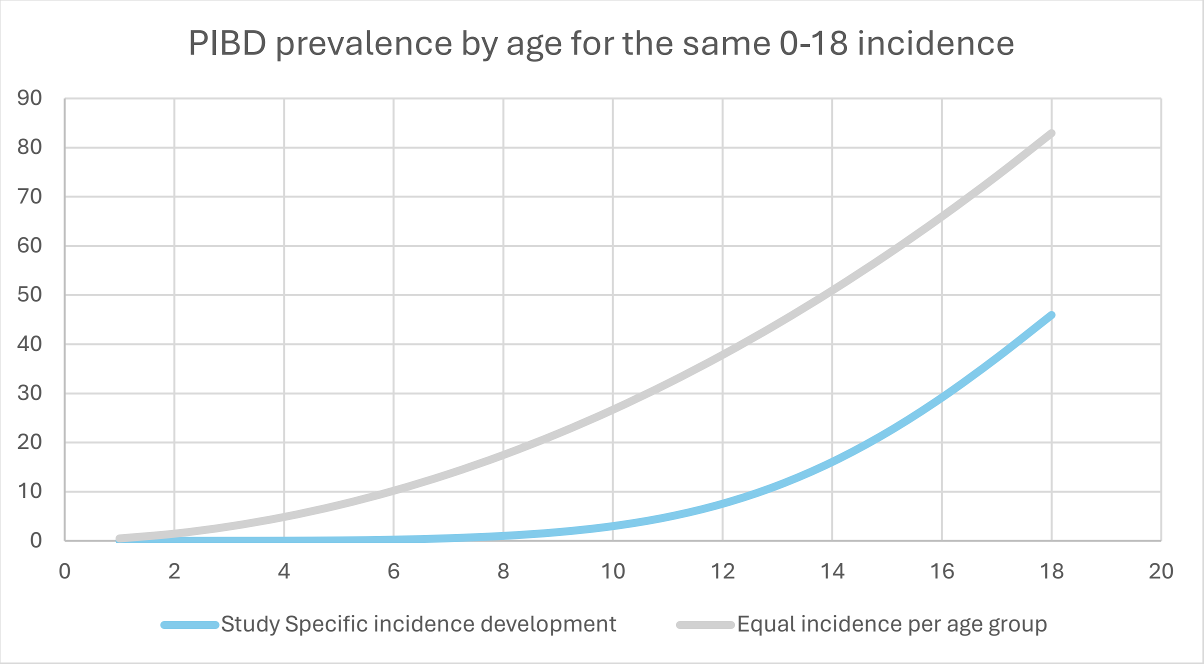


Figure 54. The expected disease prevalence by age for the two different incidence patterns.

A comparison of the incidence by year of age for a scenario with equal risk increase (grey) and increasing risk increase with older age (light blue).

5.1.4. Discussion of PIBD prevalence results

In this section, we have presented the results of our study on PIBD prevalence based on five years of prospective data collection. Our results indicate a strong latitude trend consistent with our incidence findings providing another validation mark. Our analysis also reveals a similar temporal trend to the incidence analysis, showing an increase in the pooled prevalence of PIBD across Europe during the study period from 2018 to 2020, followed by a steep increase in the 2020-2022 period. As shown in **Figure 51** our data demonstrate a significant rise in the prevalence rates for several countries, such as the U.K., the Netherlands, Ireland, France, and Poland. While the results were less consistent, a prevalence increase was also seen in Croatia, Finland, Germany, Hungary, Italy, and Spain. Only 3 of the 21 countries in our study reported a decreasing prevalence. These were Greece, Serbia, and Lithuania, and they were the 3 out of the 4 countries that had also reported a decrease in the incidence as well.

Furthermore, we noted that three countries, Finland, Austria, and Poland, were outliers for prevalence reporting. Considering that the phenotype and incidence-to-prevalence ratios for these countries were not outliers, our findings provide more evidence that the reported catchment areas of these countries may be inaccurate, especially for Poland and possibly Austria.

As a final point, we have also investigated the incidence-to-prevalence ratio, which was found to be within the expected range, supporting the validity of the incidence calculation methodology. Any evidence confirming the accuracy of the age-based incidence adjustment strengthens the robustness of the subsequent geostatistical analysis. The IPR and the phenotype ratio that was analysed in the previous chapter, are crucial metrics for the quality assessment of the submitted data since it is implausible that countries or centres reporting inaccurate results would report these two ratios within the expected margins.

6. RESULTS – Safety registry analysis of incidence rates using Safety Registry population

The aim of this chapter was to map the estimated incidence and predictors to get a better understanding of their spatial distribution, autocorrelation, clustering effects and the presence of outliers. The disease mapping and preparation of predictor data were followed by the geostatistical analysis, where a large number of suspected risk factors were investigated for their effects on PIBD incidence.

6.1. Disease mapping, characteristics and spatial distribution of incidence

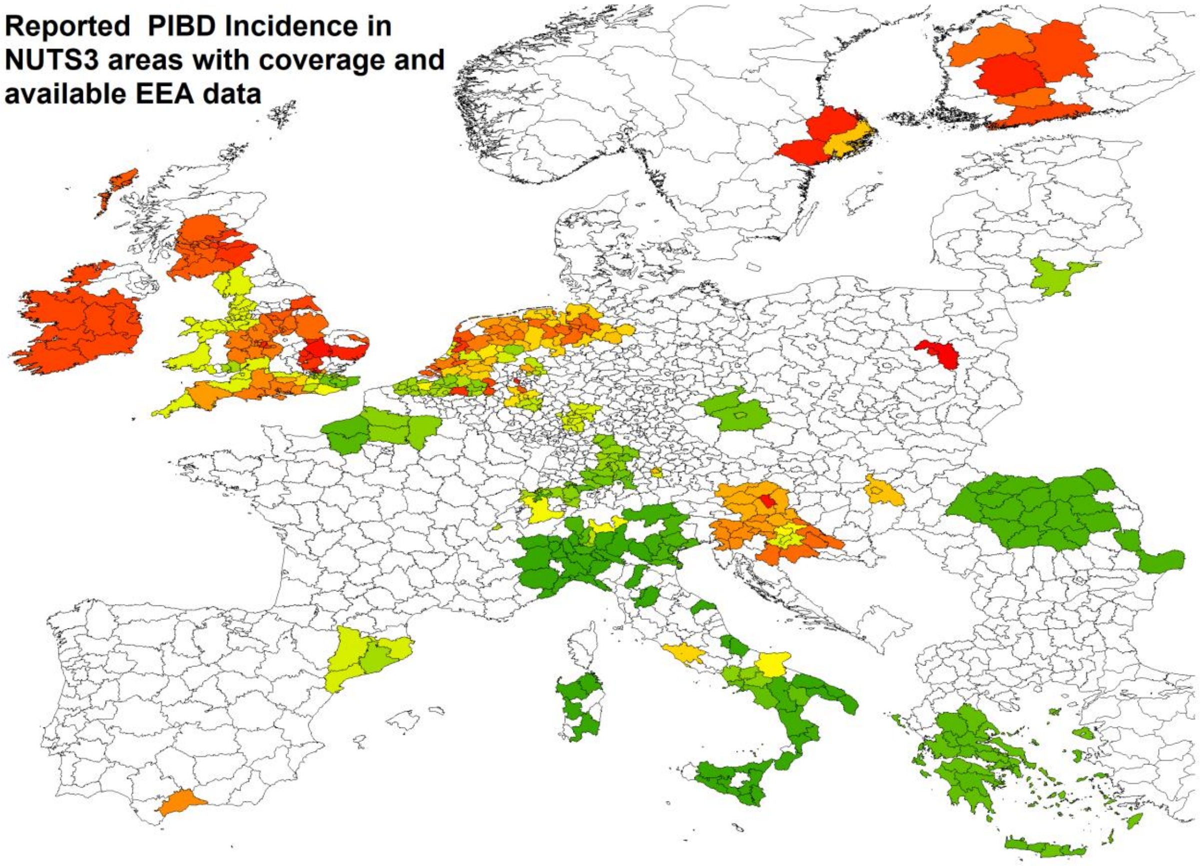


Figure 65. The reported incidence in Europe was mapped using the NUTS3 regions with available EEA and Eurostat data that were included in the geostatistical analysis.

Areas in green reported lower IBD incidence rates compared to the areas in red. The incidence was also reported in additional areas in Europe and in other continents which were excluded due to the data availability restrictions.

With the application of spatial empirical Bayesian smoothing, we were able to adjust and smoothen the spatial variable of incidence thus eliminating the influence of potentially small samples and extreme values. The map smoothing reduces unreliable information and provides a more stable incidence estimate. In the following **Figure 66**, the smoothened and final PIBD incidence map of the NUTS3 areas containing evaluable information for the geostatistical analysis is presented.

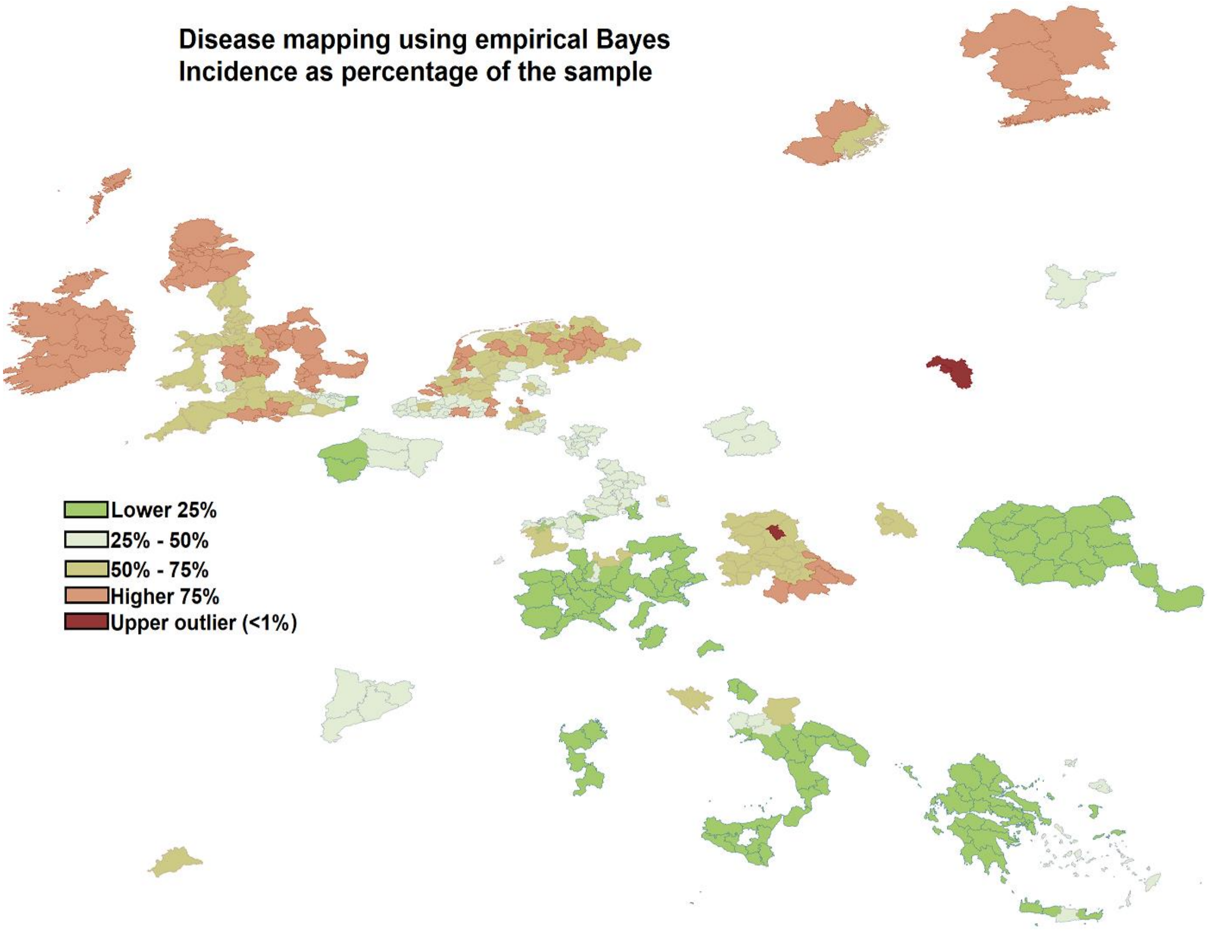

Disease mapping using empirical Bayes
Incidence as percentage of the sample
🟩 Lower 25%
▭ 25% - 50%
▨ 50% - 75%
🟧 Higher 75%
🟥 Upper outlier (<1%)

Figure 66. The map of Europe has been filtered for the areas included in the geostatistical analysis.

The choropleth map splits the NUTS3 regions into 4 quartile groups based on the reported incidence. Two regions are marked as outliers (highlighted in red). Italy, Greece, Romania, Finland, Croatia, Ireland and Slovenia show consistent results within their regions. Scotland stands out in the United Kingdom with the latter presenting significant variation overall. The Belgium, Netherlands and North Germany areas also present significant variation.

Using Moran's I, we measured the extent of spatial autocorrelation in the estimated outcome of incidence. The Moran's I was 0.718 suggesting the presence of significant autocorrelation in the observations. As shown in the following **Figures 67 & 68** the incidence observations are spatially lagged and tend to be positively correlated in closer proximity.

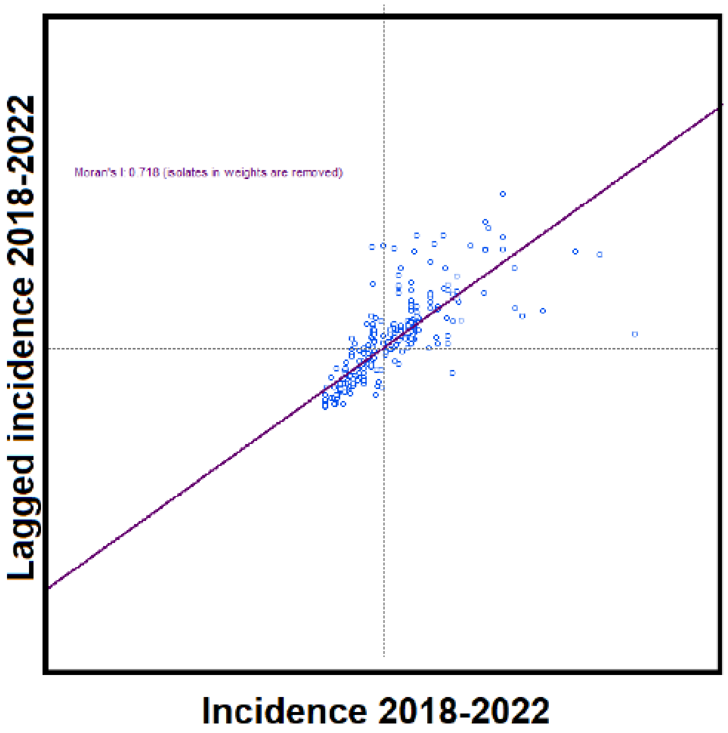


Figure 67. Correlogram of the incidence in the 2018 – 2022 period.

The summary of the autocorrelation function. The x-axis of a correlogram represents the time lags, or the number of time units between the observation and its lagged value. The y-axis represents the magnitude of the correlation between the observation and its lagged value.

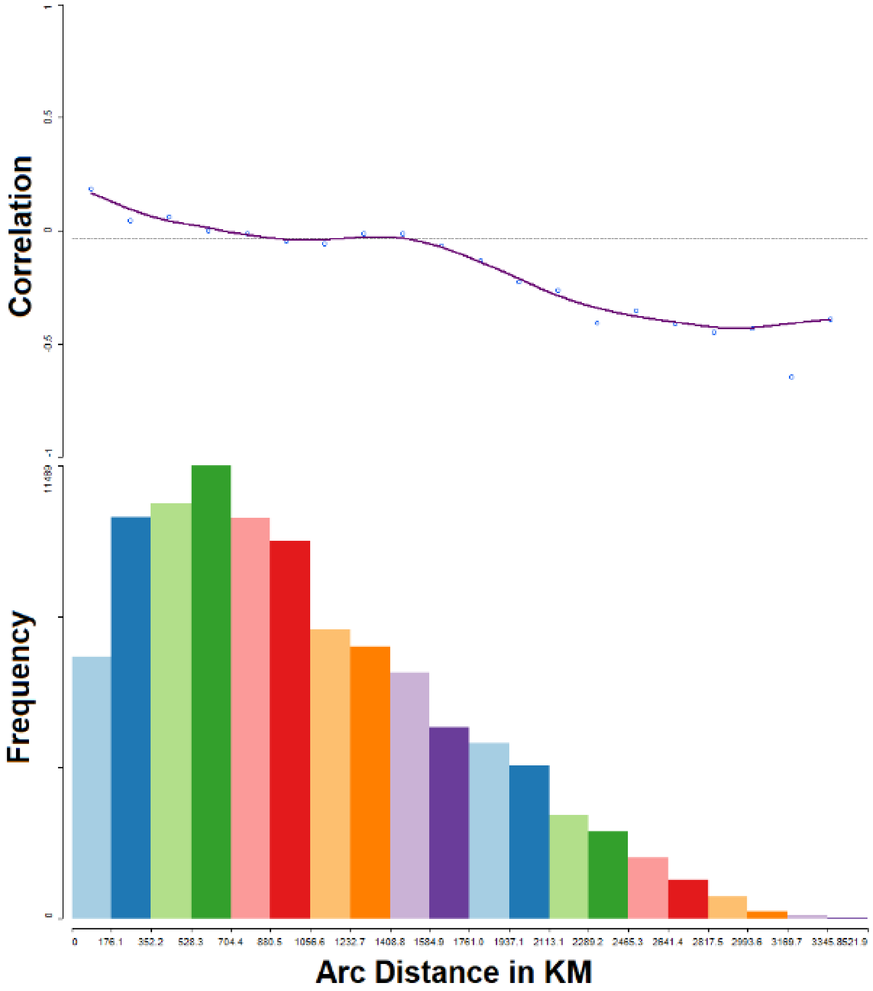


Figure 68. The autocorrelation of the reported incidence.

The correlation of the disease incidence between the NUTS3 areas at different distances and the distribution of the distance between all the NUTS3 areas in our dataset.

A local indicator of spatial association analysis suggests that the correlation is, as expected, influenced by the design of the study where neighbouring regions are more likely to be covered by the same clinic and therefore are assigned with similar incidence values. This demonstrates that the ID of the unit that claims coverage and geographical location of the NUTS3 areas are highly associated with the observed clusters. However, as shown in **Figure 69,** the observed clusters of low and high reported incidence appear to be separated based on their latitude.

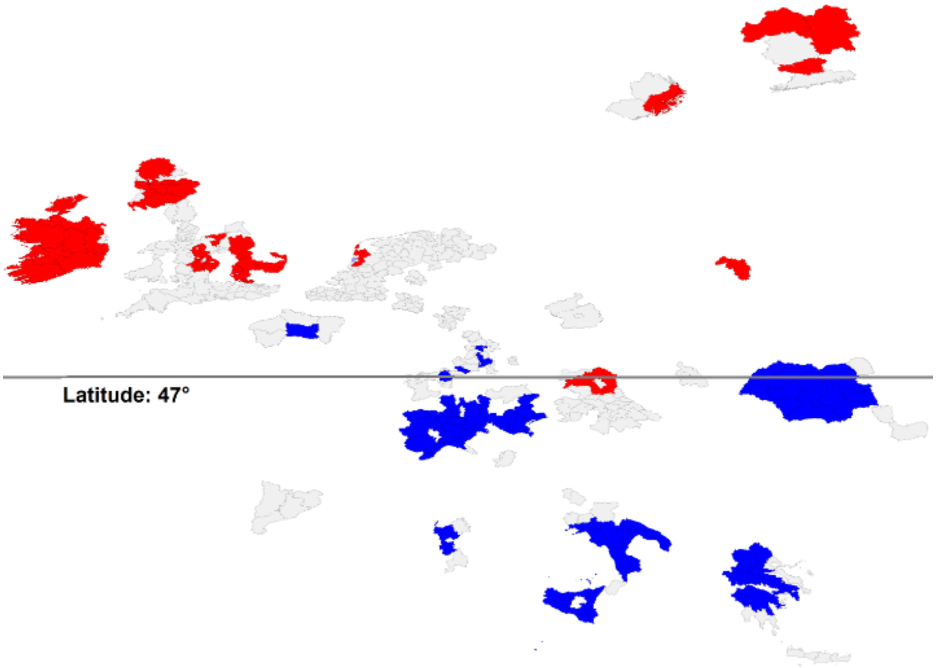


Figure 69. Cluster of higher and lower incidence.

Our data show that almost all NUTS3 clusters in the north appear to be clusters of high incidences, in contrast with the clusters in the south, which appear to be clusters of low incidence rates. The 47^{o}N can separate these clusters almost perfectly.

6.2. Predictor data

In the analysis of the European countries, we processed several sets of environmental pollutants as discussed in the methods. The lists of the emission type for the ten most frequent release's locations are summarised in the tables below for the air and water, respectively.

Table 18 The most common EEA air pollutants used in the geospatial analysis.

The most frequent air releases are not common in water bodies.

Pollutant Name	Sites - Total	Sites -Air	Sites -Water	Sites - Soil
Ammonia (NH3)	84,970	84,944	26	0
Nitrogen oxides (NOx/NO2)	47,570	47,569	1	0
Carbon dioxide (CO2)	41,185	41,185	0	0
Sulphur oxides (SOx/SO2)	24,812	24,799	13	0
Methane (CH4)	24,388	24,388	0	0
Non-methane VOC	16,803	16,803	0	0
Carbon monoxide (CO)	12,365	12,364	1	0
Nitrous oxide (N2O)	11,802	11,802	0	0
Particulate matter (PM10)	11,672	11,671	1	0

Table 19 The most common EEA pollutants released in water used in the geospatial analysis.

The most frequently released pollutants in water bodies are also released in the air.

Pollutant Name	Sites - Total	Sites - Water	Sites -Air	Sites - Soil
Zinc and compounds (as Zn)	38,948	29,734	8,634	580
Total organic carbon (TOC)	29,815	29,491	271	53
Total nitrogen	23,368	23,299	6	63
Total phosphorus	21,251	21,135	0	116
Nickel and compounds (as Ni)	29,063	19,905	8,735	423
Copper and compounds (as Cu)	19,726	15,169	4,069	488
Arsenic and compounds (as As)	17,136	12,104	4,881	151
Chlorides (as total Cl)	11,550	11,466	4	80
Lead and compounds (as Pb)	15,751	10,555	4,810	386

In the following **Figure 70**, the spatial distribution of four of the interpolated pollutants is shown as an example of all the processed pollutants. These pollutants have been interpolated using the EEA E-PRTR pollutant releases data as described in the methods.

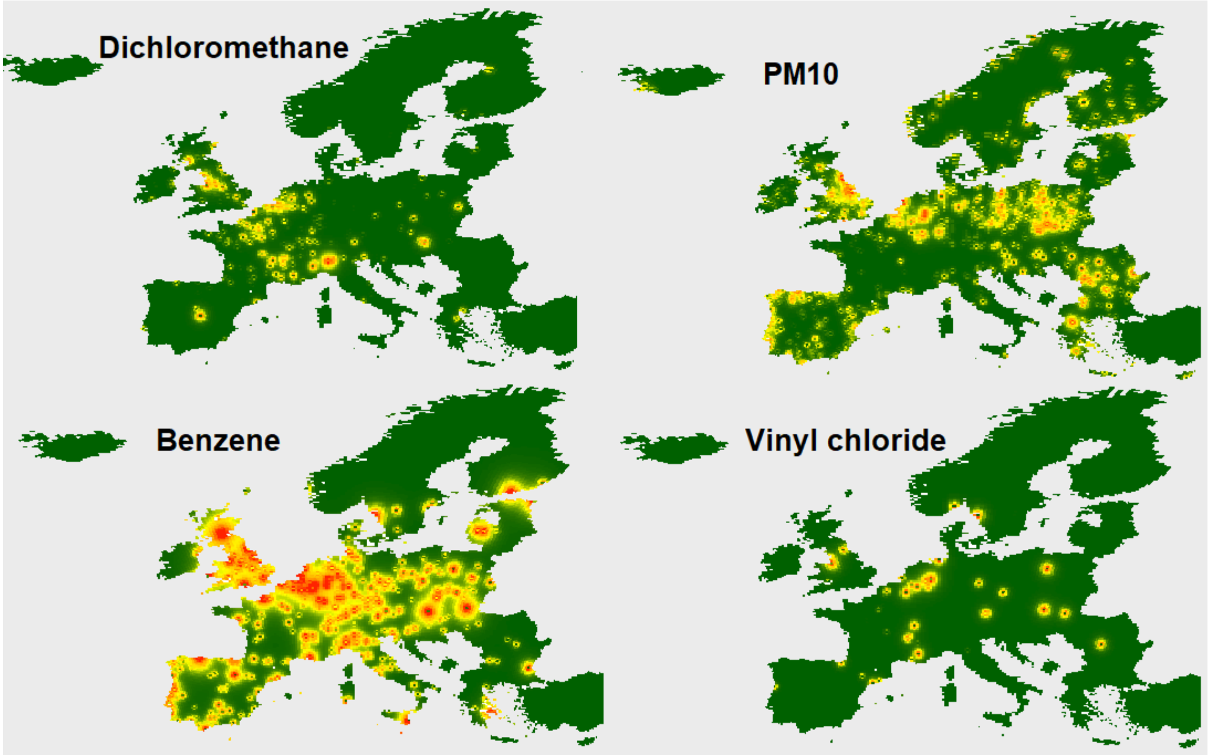


Figure 70. Example of four interpolated pollutants.

The pollutants that were used as the predictors in our analysis were interpolated as shown in this figure. This allowed us to merge them with the NUTS3 areas and prepare them for the *geostatistical analysis. The pollutants in this example were provided by the European Environment Agency's Pollutant Release and Transfer Register.*

The initial step before performing statistical modelling on the incidence data also involved investigating the correlation between the pollutant variables. To determine the level of independence among these variables, the distribution of each pollutant in the studied regions (NUTS3 regions with incidence data) was examined. As illustrated in **Figure 71**, the results revealed that about 25% of the pollutants exhibit a high correlation. This is a finding that needs to be considered before finalising our models describing the relationship between the pollutants and PIBD incidence since it suggests that some of the pollutants may be used interchangeably in the models.

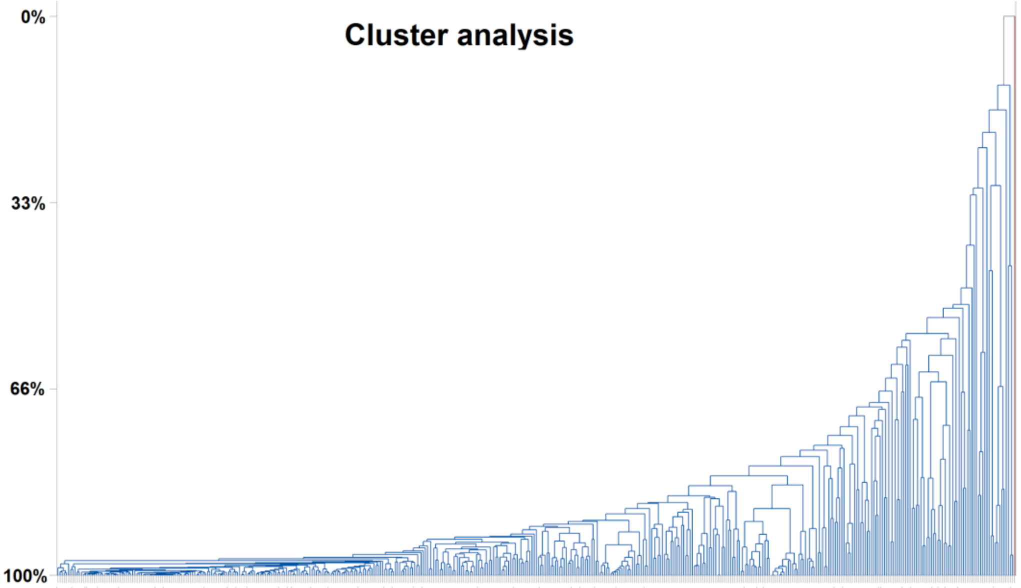


Figure 71. Cluster analysis of the pollutants in our dataset for the covered NUTS3 regions.

Note that this is not an analysis of spatial clusters but rather an analysis of their similarity based on the Euclidean distance between the pollutants.

The pollutants were also found to deviate significantly from the normal distribution. The Shapiro–Wilk test confirms that all pollutants are not normally distributed, with Methane being the closest one to normality and yet returning significant results in the Shapiro–Wilk test ($p<0.00001$). The following **Figure 72** shows four examples of the pollutant frequency distributions.

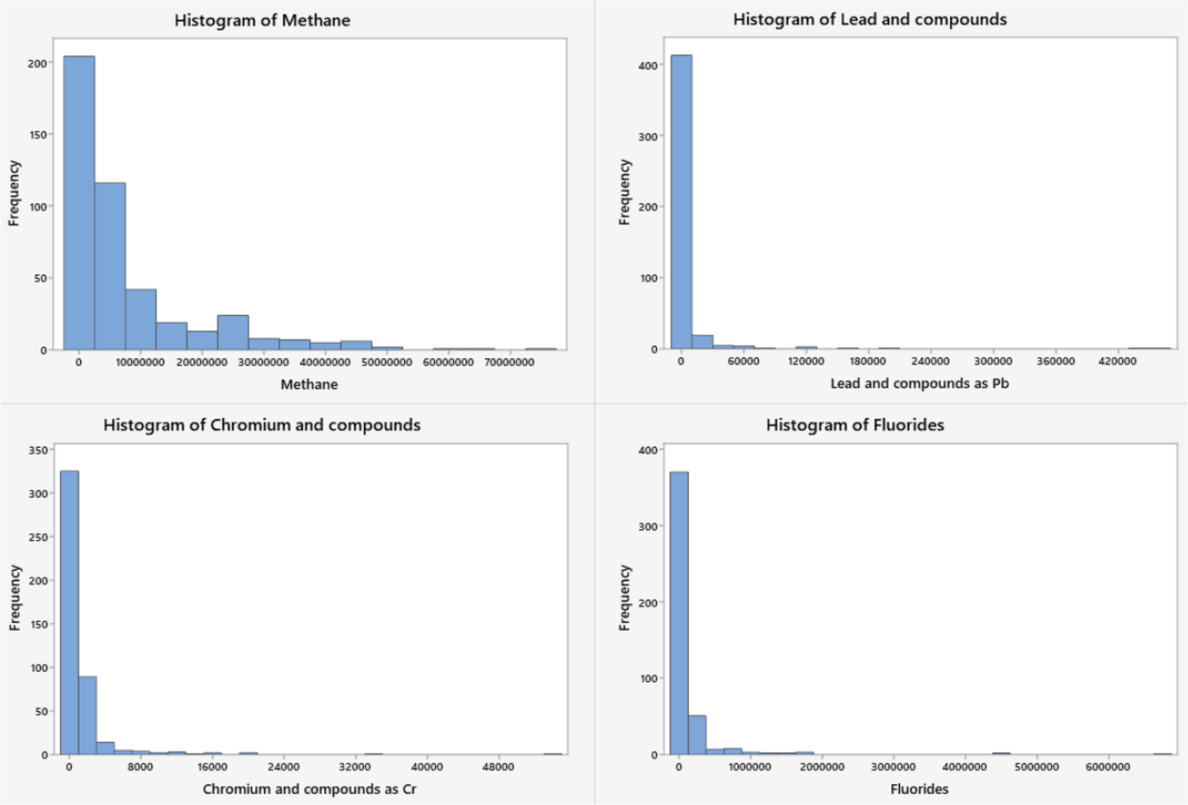


Figure 72. Frequency distribution of the pollutants.

The histograms of the four pollutants that were interpolated in the previous example in Figure 70. These histograms show that a transformation could improve the distribution of the pollutants.

Considering that the predictor data are skewed to the right and highly clustered at lower values, it becomes clear that a transformation would improve the spread of the data and contribute to a more robust fit in the subsequent modelling. In cases of datasets with clustered observations and a few extreme values, the latter characteristic may be particularly influential, affecting the model fit disproportionally compared to the other observations. Taking into account that the outcome of the Shapiro–Wilk test statistic *W* spans from 0 to 1, with 1 being a perfect match to a normal distribution, we hope to perform a transformation that improves *W*. After performing a logarithmic transformation (base 10 logarithm) to the predictor data, the average *W* increased from 0.191 to 0.96. In the figure below the distributions of the four pollutants of the previous example are shown after the transformation (**Figure 73**).

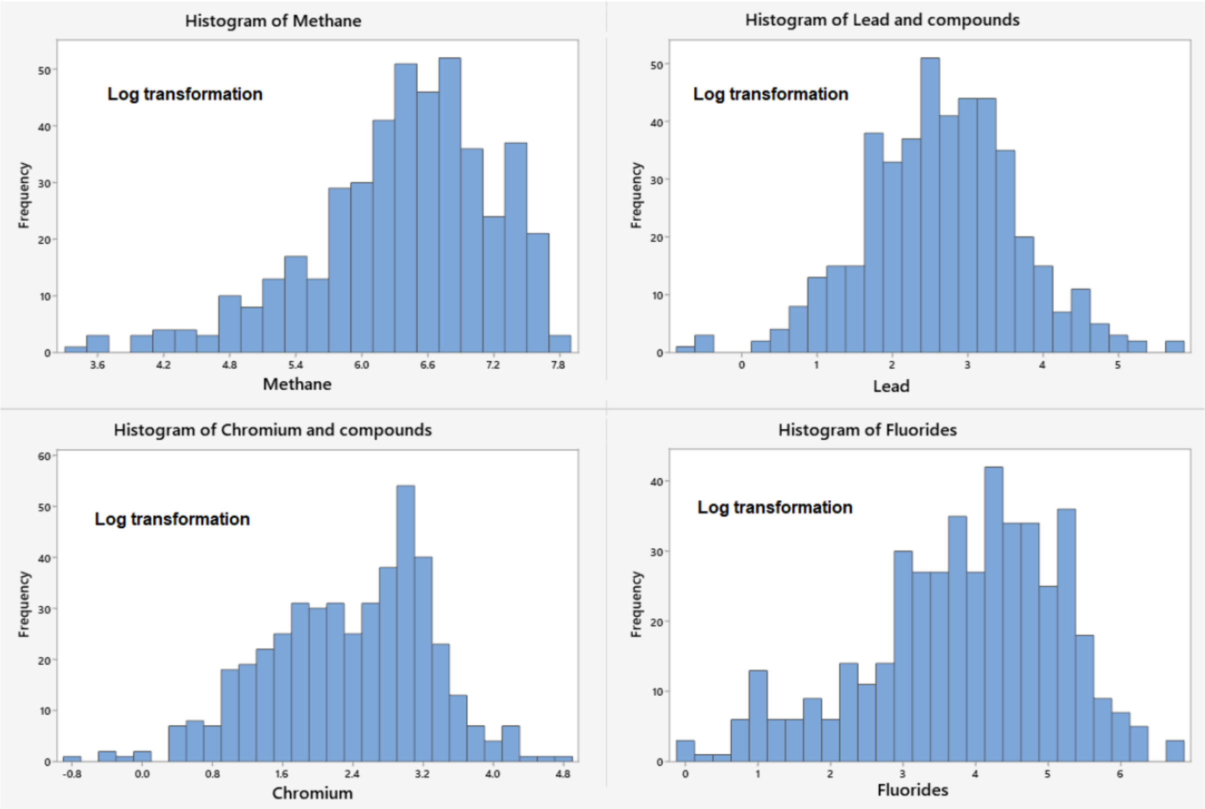


Figure 73. Frequency distribution of the log-transformed pollutants.

The histograms of the four pollutants in the previous example [Figure 70 and Figure 72]. The data transformation improved the distribution of the pollutants significantly.

6.3. Model outliers

In this section we discuss the presence, effects and handling of outliers in the outcome variable of incidence, as these observations may invalidate the analysis when they become highly influential.

As discussed previously, the dataset contains two countries that could be considered as outliers, namely Austria and Poland, with Finland being a less probable outlier. In addition to the outlier analysis, further investigation into the within-country reporting revealed that Austria and Poland have each submitted one extreme measurement at one data collection year which accounts for only a 0.2% of the total number of the incidence reports. However, despite their small contribution to the overall data, these outliers can have an outsized influence on the results. For example, in a simple linear regression using the interpolated polycyclic aromatic hydrocarbon pollution to explain the incidence, the inclusion of these two outliers can inflate the model's coefficient by 18%, which is more than 30 times greater than the next most influential case. This effect is illustrated in **Figure 74** which shows that the Cook's D values for these outliers are significantly higher than the rest of the observations. High Cook's D values indicate that an observation has both high leverage and high residual values, which can have a detrimental effect on the fit of the model.

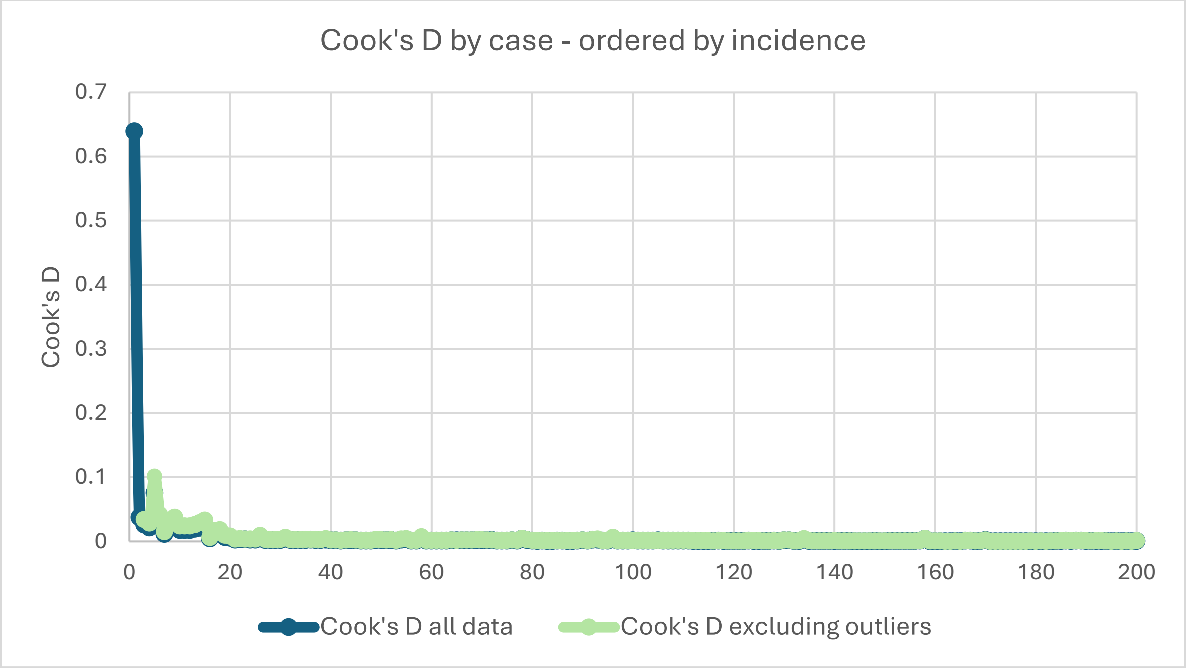


Figure 74. The Cook's distance metric is used to identify influential data.

The Cook's D values for all cases in our test model using polycyclic aromatic hydrocarbon pollution to explain the disease incidence with and without the two outliers (Austria and Poland). This randomly selected pollutant shows the importance of using model fit diagnostics and the need for appropriate outlier management.

Based on these findings and in conjunction with the findings of the data quality and outlier assessment metrics, these outliers were removed from the dataset prior to the risk factor analysis.

6.4. Significant findings

In this section, we outline the process of developing models to explain disease incidence using our predictor dataset and highlight the key variables that are likely to have a significant impact on the incidence of the disease. Our analysis is based on two approaches that consider the presence of autocorrelation in our data, namely spatial regression, and mixed-effects linear models. The latter approach was also expanded to the analysis of each PIBD phenotype separately in 6.5.

Pearson's and Spearman's correlation tests between the incidence and the pollutant levels revealed that 58% and 78% of the variables, respectively, were positively correlated with the outcome. Focusing on the Pearson coefficient, after adjusting the critical level of significance to address the inflation of the family-wise error rate, 27 pollutants were selected due to their significant correlation with the PIBD incidence. Within this group, only 2 pollutants showed a negative correlation meaning that 93% of the pollutants with the highest correlation were positively correlated with the incidence. In addition to the 27 selected variables, another 19 pollutants that presented an absolute Spearman Rho over 0.35, were selected for further investigation using linear modelling and spatial regression.

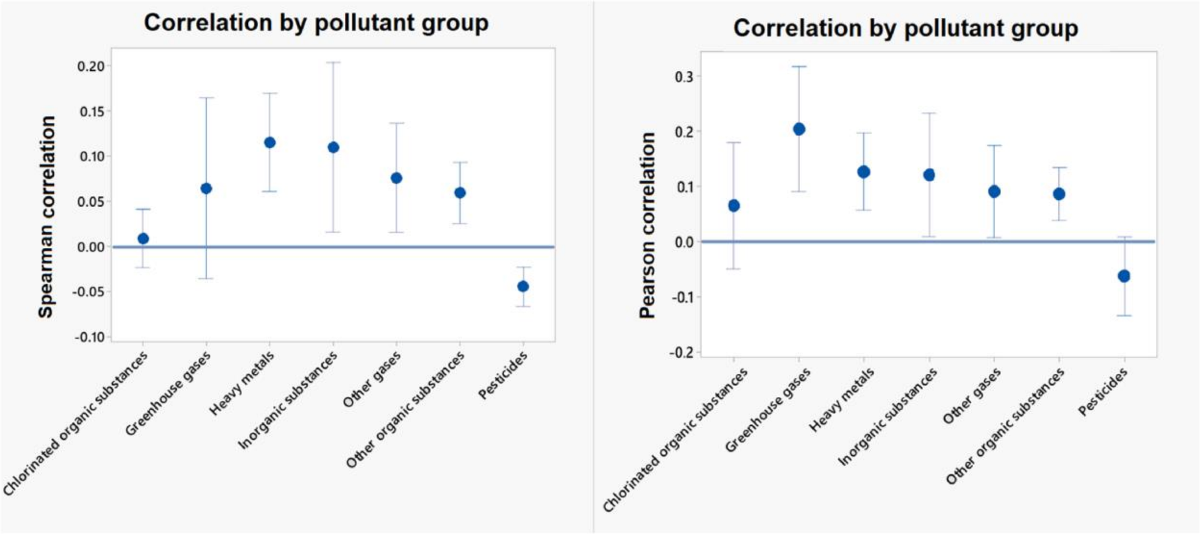


Figure 75. The average correlation of pollutant groups with the incidence.

The average Spearman and Pearson correlation of the pollutants with the disease incidence, grouped by broad pollution category.

6.5. Linear modelling of PIBD incidence

The desired final model must satisfy several requirements as listed below:

- Account for the spatial auto-correlation and clustering of the results reported by the same country

- Account for the non-independence of the incidence reports from the same NUTS3 over the multiple collection rounds.
- Weigh the results based on the population of the NUTS3 regions.
- Estimate the effects of the pollutants on the incidence
- Incorporate the effects of time on the incidence

The model that satisfies all the listed requirements is a mixed effects linear regression (LMM). The variables of NUTS3 and country will be introduced with random intercepts where the NUTS3 areas are nested under the country. The variables of time and investigated pollutants will be introduced as a fixed effects in the model. Random slopes were not introduced in the models as there is no assumption that patients in different areas are expected to be affected in variable manner by the same pollution effects over time. Furthermore, in practice, adding random slopes did not improve the model fit. Lastly, by introducing the weights in our model, we are increasing the contribution of observations derived from large populations to the log-likelihood as follows: $\log(L(\theta)) \sum_{i=1}^{n} w_i \log(P(y_i|x_i, \theta))$. Considering the standardised NUTS3 size shown in **Table 1** the weights will improve the model fit but are not expected to have a great effect.

Therefore, the model is built as follows:

$$Incidence_{ijk} = \beta_0 + \beta_{0j} + \beta_{i(j)} + \beta_1 \cdot time_{ij} + \beta_2 \cdot pollutant1_{ij} + \beta_3 \cdot pollutant2_{ij} + e_{ijk}$$

Where:

- $Incidence_{ijk}$ is the observed incidence for the *ith* NUTS3 in the *jth* country at the *kth* data collection round
- β_0 is the fixed intercept
- β_{0j} is the random intercept for the *jth* Country
- $\beta_{j(i)}$ is the random intercept for the *ith* NUTS3 within the *jth* Country (nested)
- β_1 is the fixed effect coefficient for the time, representing the fixed effect of time on the incidence
- $\beta_2$and β_3 are the fixed effects coefficients for the pollutants 1 and 2, representing their fixed effects on the incidence

- e_{ij} is the residual error, which represents the variability in incidence that is not accounted for the time or random effects

Based on the results, it appears that only a small proportion (13%) of the variables that were significant in the initial correlation analysis remained significant in the univariate mixed effects model (LMM) analysis. The most important risk factors are summarised in **Table 20**. The LMM models were fit using single predictors, predictors combined with the variable of time, and predictors paired with an interaction term. The models were evaluated based on the Akaike information criterion (AIC) and the t-value estimate of each term per tested model. The latter refers to the t-statistic associated with each fixed effect in our linear mixed effects model, and it measures the magnitude of the estimated fixed effect in relation to its standard error. This value is calculated by dividing the estimated fixed effect by its standard error, and therefore, larger absolute t-values provide more substantial evidence against the null hypothesis of the fixed effect being equal to zero. In our analysis any t-values with an absolute value greater than 1.96 were considered statistically significant at a significance level of 0.05.

The variable with the strongest effect on the outcome was particulate matter 10 emissions (PM10), which was found to be a significant risk factor in all models followed by Carbon monoxide (CO), Carbon dioxide (CO_2) and Chlorine with inorganic compounds (HCl) as shown in **Table 21**. None of the combinations of pollutants returned significant results, but the variable "Other gases" had an important effect in the opposite direction, suggesting that it may be a protective factor (or being correlated with one). Overall, these findings suggest that the initial correlation analysis may have overestimated the importance of some variables and that using the LMM model can help to identify the most important risk factors more accurately for the outcome of interest. The results also highlight the importance of considering individual predictors in combination with other variables, such as time and interaction terms, when assessing the effects of multiple risk factors on an outcome.

Table 20 Summary of significant PIBD incidence risk factors based on the LMM analysis.

Variable Name	**AIC**	**t- value**
Particulate matter (PM10)	5895.45	3.42
Carbon monoxide (CO)	5895.63	3.40
Carbon dioxide (CO_2)	5901.13	2.42
Chlorine and inorganic compounds (HCl)	5902.24	2.18

Although, all four variables shown in **Table 20** were significant when combined with the fixed effect of time in the LMM model, in the following step we present the fit diagnostics for PM10 which had the strongest fit.

Table 21 The model with the best fit and lowest AIC score.

Model	**AIC**	**Fixed effect term**	**Term t-value**
Incidence ~ PM10 + Time	5895.45	PM10	3.42
		Time	5.588

When fitting the following model from **Table 21**:

$$Incidence_{ijk} = \beta_{0j} + \beta_{0i} + \beta_1 \cdot time_{ij} + \beta_2 \cdot PM10_{ij} + e_{ijk}$$

we observe that the residuals deviate significantly from the normal distribution and present high heteroscedasticity, while the random effects intercept estimates also deviate significantly from normality (**Figure 76**).

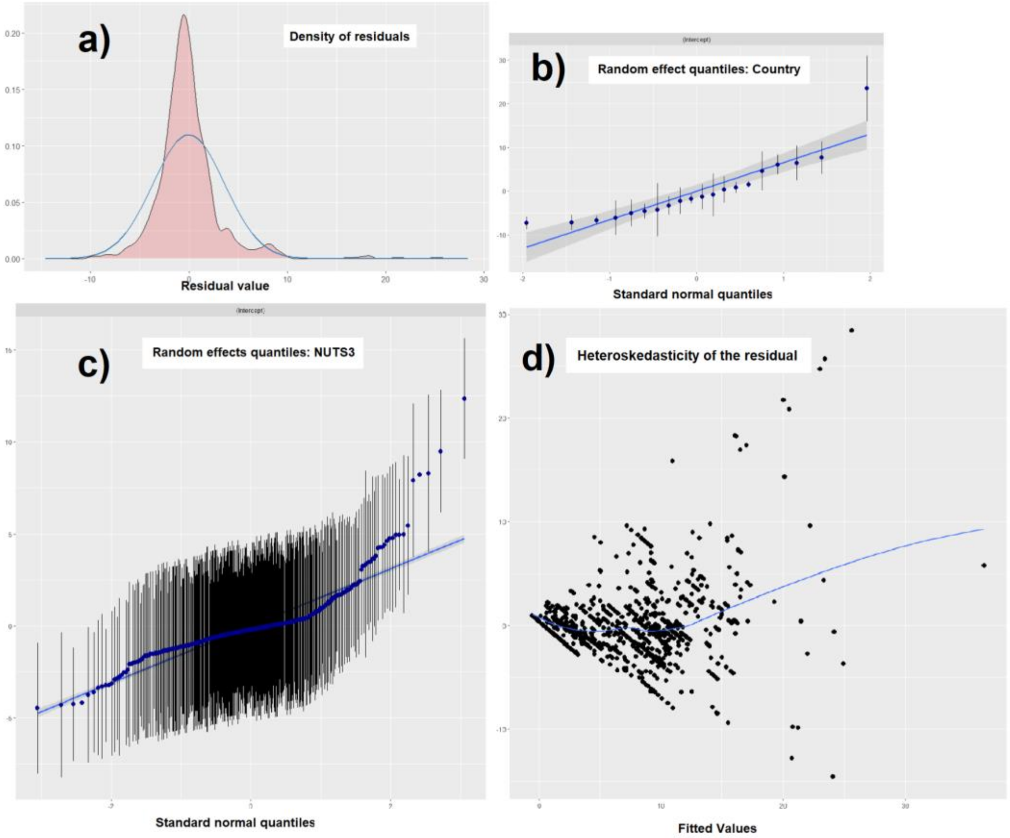


Figure 76. LMM fit summaries of the PM10 + Time model.

The panels above a) indicate a significant deviation of the residuals from the normal distribution, b) and c) show that the random intercepts are also deviating from the normal distribution due to a few outliers, which may be problematic for the accurate estimation of the variance and covariance structure of the random effects and d) showing the high and heterogeneous heteroscedasticity.

By removing Poland as a major outlier in our dataset and using a square root transformation on the incidence, we were able to significantly improve the model's fit as shown in **Figures 76 & 77**. When the model was fitted with the transformed incidence, we observed that the residuals were closer to a normal distribution and the intercept estimates of the random effects for the country improved significantly (**Figure 77**). In addition, we almost fully eliminated the heteroscedasticity, and the residual moving average term was reduced by a factor of 10 (**Figure 77**). A final remark about the selected LMM is that it shows no signs of multicollinearity, as the correlation between its fixed effects is minimal at 0.025. Although only the diagnostics of the PM10 + time model are presented here both the transformation of the incidence, and the removal of the outliers improved the fit of all significant pollutants listed in **Table 20**.

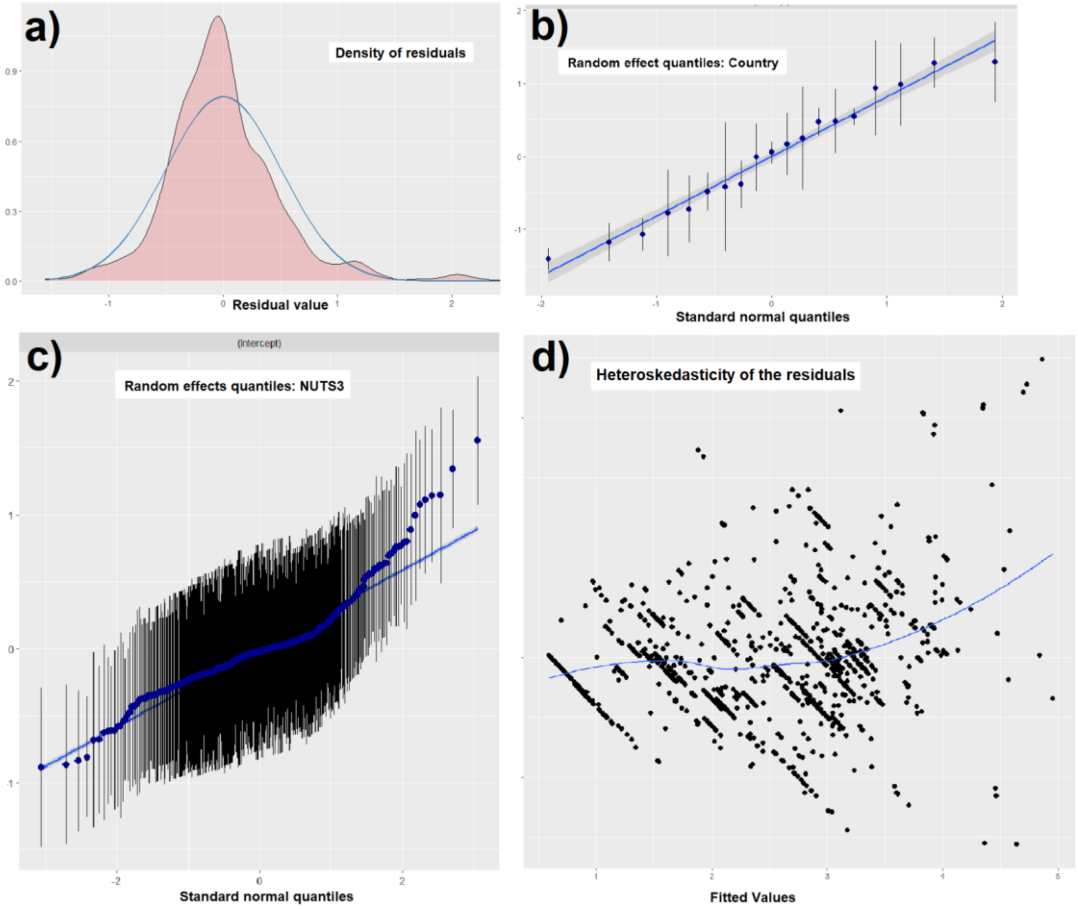


Figure 77. LMM fit summaries of the PM10 + Time model with the transformed outcome.

The panels above a) indicate a significant improvement in the distribution of the residuals, b) and c) show a significant improvement in the distribution of the random intercepts for both the regions and countries and d) show the low and significantly more homogenous heteroscedasticity. Please note that the scale of the y axes in Figure 77 were adjusted due to the great improvement goodness of fit of the latter optimise model.

Table 22 Improvement of the final model with the lowest AIC score.

The updated model excluded 2 outlying observations and used the transformed incidence.

Model	AIC	Fixed effect term	Term t-value
Incidence ~ PM10 + Time	5895.45	PM10	3.42
		Time	5.588
Incidence (transformed) ~ PM10 + Time (exc. outliers)	2174.4	PM10	3.243
		Time	5.801

Acknowledging that several actions were undertaken to refine the mixed effects linear models is important. While these efforts markedly enhanced the models' fit, it is pertinent to note that the optimisation process was not comprehensive. As shown in the preceding sections, the transformation of pollutant variables was imperative due to their significantly skewed distribution. Similarly, modifying the outcome variable enhanced model fit, albeit without substantially altering the analytical results. However, further refinements, including expanding the regressors and attempting to add terms of higher order, were not explored, delineating an avenue for future research.

6.6. Spatial regression

Following the LMM approach, we have also fit a spatial regression model using the mapped incidence for each NUT3 area as shown in **Figure 65** and the predictor dataset (example of predictors used in **Figure 70**). The advantage of this approach is that it can account for spatial autocorrelation and address more complex spatial patterns that may not be captured by LMM models. To fit the spatial regression, the results collected at different times were averaged per NUTS3 region. Only a single case was removed from this dataset, the discussed NUTS3 region of Poland. For the spatial regression, the Rook contiguity weights were calculated (based on our findings in 2.4). As shown in **Figure 78**, a small number of regions were isolated and in no contact with other areas meaning that they did not share any spatial information in the spatial analysis. However, the majority of NUTS3 regions shared a border and, therefore, spatial weights with multiple other regions.

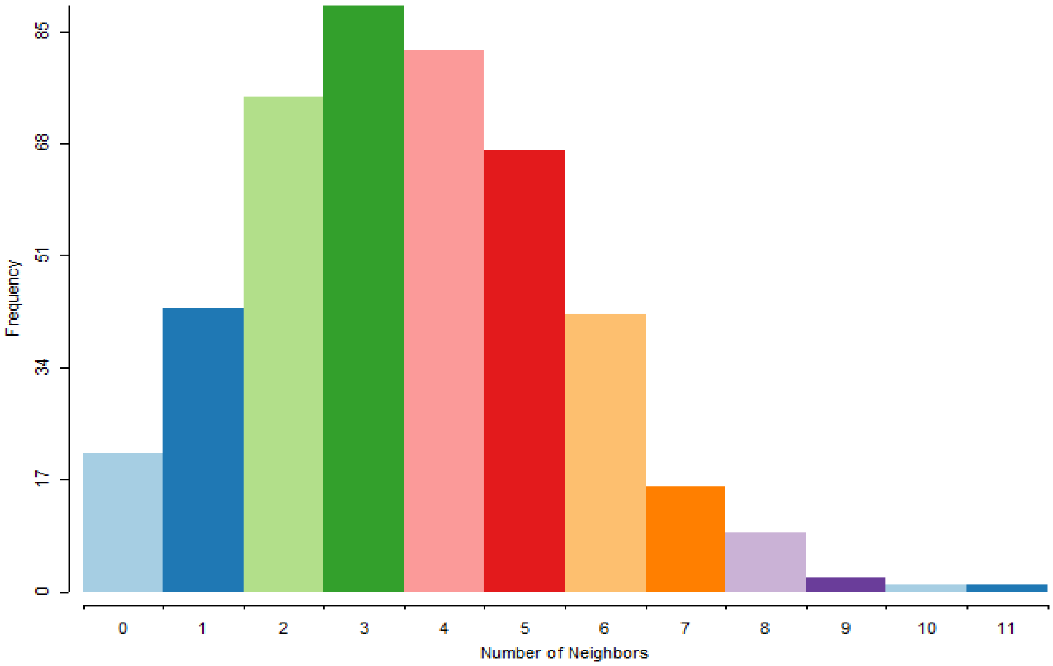


Figure 78. Number of NUTS3 areas with spatial relationships

The counts of the NUTS3 areas in our dataset (y axis) by the number of other regions (x axis) they share spatial with.

The spatial regression returned several predictors as potentially significant variables and explained a great proportion of the observed variance based on the spatial auto correlation and clustering effects. As discussed in the methods (2.3.3), the spatial error regression model was the most appropriate for our data due to the presence of clusters. The model with the best fit reached an R squared of 65% and included two significant predictors: the particular matter 10 and Chlorine with inorganic compounds (HCl). As shown in the following table, the model returns an increased Lambda suggesting a high degree of spatial autocorrelation (**Table 23**).

Table 23 The 3 significant predictors in the incidence spatial regression model.

Variable	Coefficient	Std. Error	z-value	P value
Constant	-0.10	1.41	-0.07	0.943
Particular matter	0.33	0.15	2.24	0.025
Chlorine and inorganic compounds (HCl)	0.36	0.13	2.88	<0.004
Lamda (autocorrelation parameter)	0.70	0.03	20.59	<< 0.000

Using the spatial weights and coefficients of the spatial model, as summarised in **Table 23,** we can obtain the predicted values and residuals and investigate the model fit. As shown below in **Figure 79**, the predicted map presents very similar patterns and disease distribution compared to the observed map in **Figure 66,** suggesting a good fit for the spatial regression.

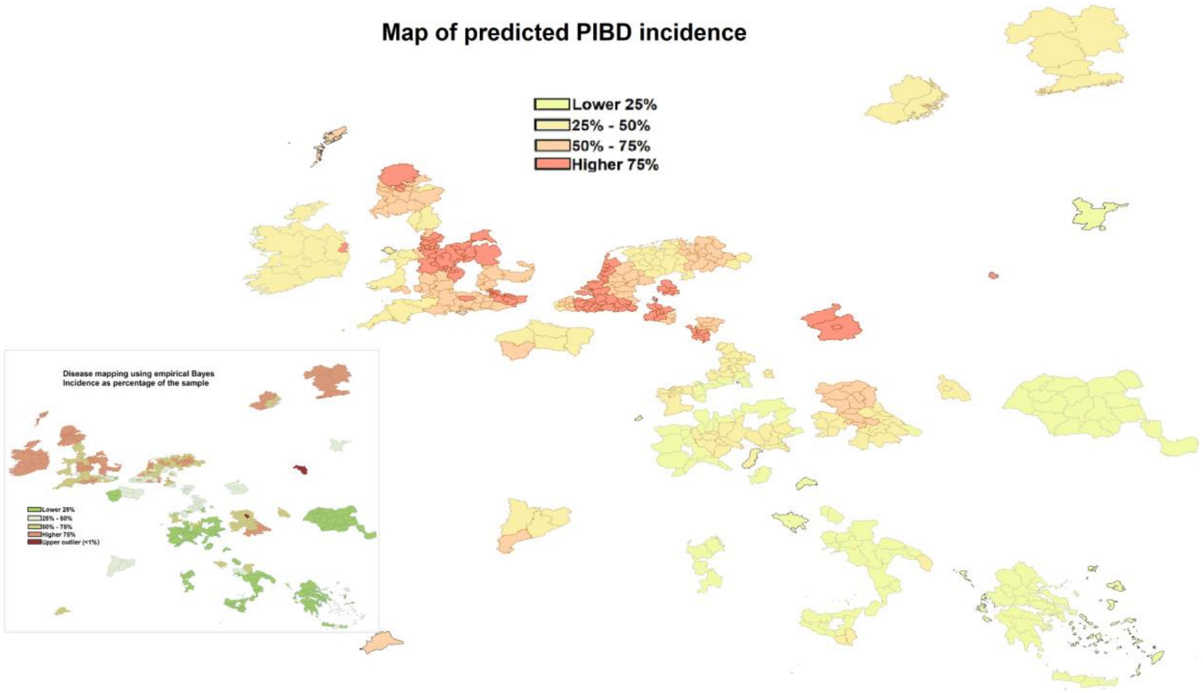

Map of predicted PIBD incidence
Lower 25%
25% - 50%
50% - 75%
Higher 75%
Disease mapping using empirical Bayes
Incidence as percentage of the sample
Lower 25%
25% - 50%
50% - 75%
Higher 75%
Upper outlier (<1%)

Figure 79. Spatial empirical Bayes mapping of the predicted incidence

The map of the predicted incidence is based on the spatial regression model (main map) compared with the observed incidence map from Figure 66 (additional map). Although the colour scale used varies between the maps, the values present high concordance. The predicted values on the main map were estimated using the spatial weights and coefficients of the final spatial error regression model.

The study of the residuals, as shown in **Figure 80**, confirms that most areas have a good fit with marginal deviations between the observed and expected incidence. The only extreme residuals were found in Finland, suggesting that according to our model, the pollution and spatial relationships cannot explain sufficiently the high reported incidence in that region.

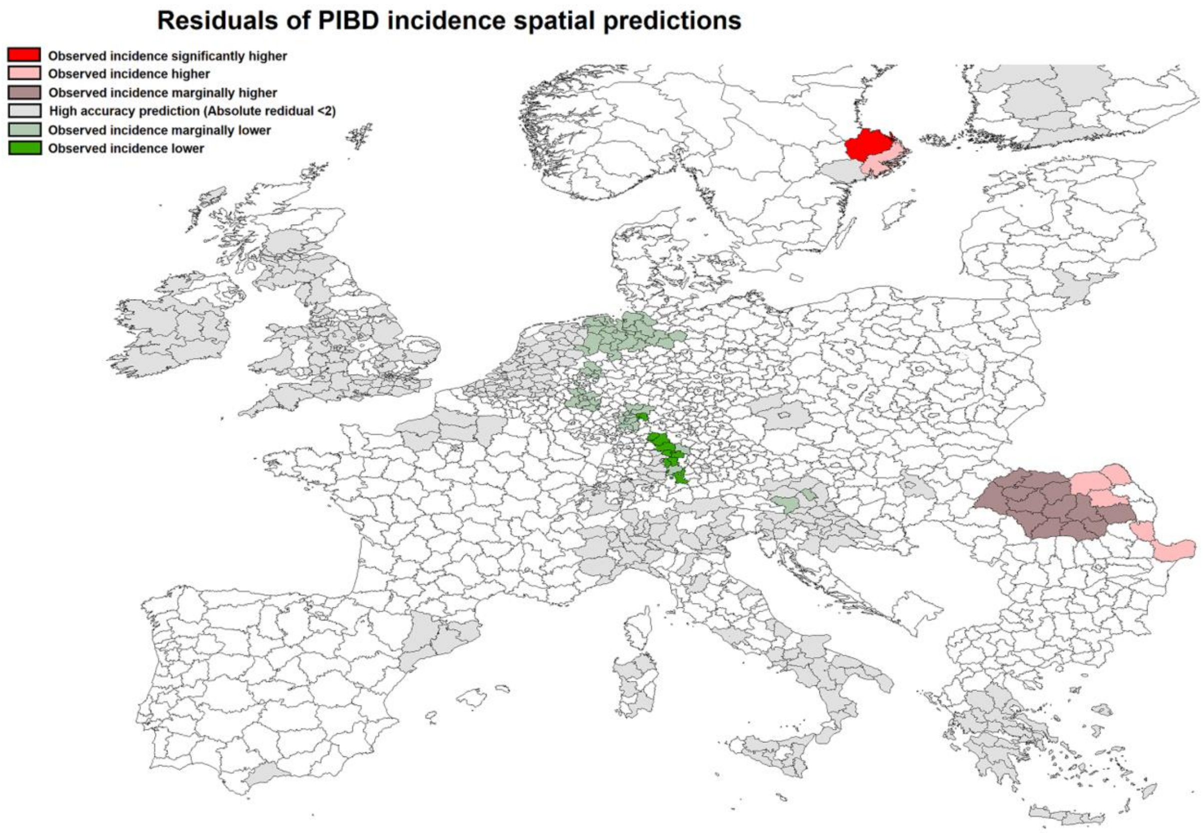

Residuals of PIBD incidence spatial predictions
🟥 Observed incidence significantly higher
Observed incidence higher
Observed incidence marginally higher
High accuracy prediction (Absolute redidual <2)
Observed incidence marginally lower
🟩 Observed incidence lower

Figure 80. Residual analysis of the spatial regression.

The residuals of the spatial regression model indicate a good fit in most countries, with only a few minor discrepancies. Specifically, the fitted incidence was slightly higher for several reporting areas in Germany and slightly lower in all reporting areas in. However, the model showed a significant outlier in Finland, where the predicted incidence was substantially lower than the observed values.

In the final step of the spatial analysis, we utilised the coefficients from the spatial regression and combined the average values of PM10 and Chlorine with inorganic compounds for each NUTS3 region (**Figure 81**) to estimate the expected PIBD incidence for all NUTS3 regions in Europe (**Figure 82**). It is important to note that these estimates, as displayed in **Figure 82,** do not incorporate any spatial information, and are solely based on the pollution measures.

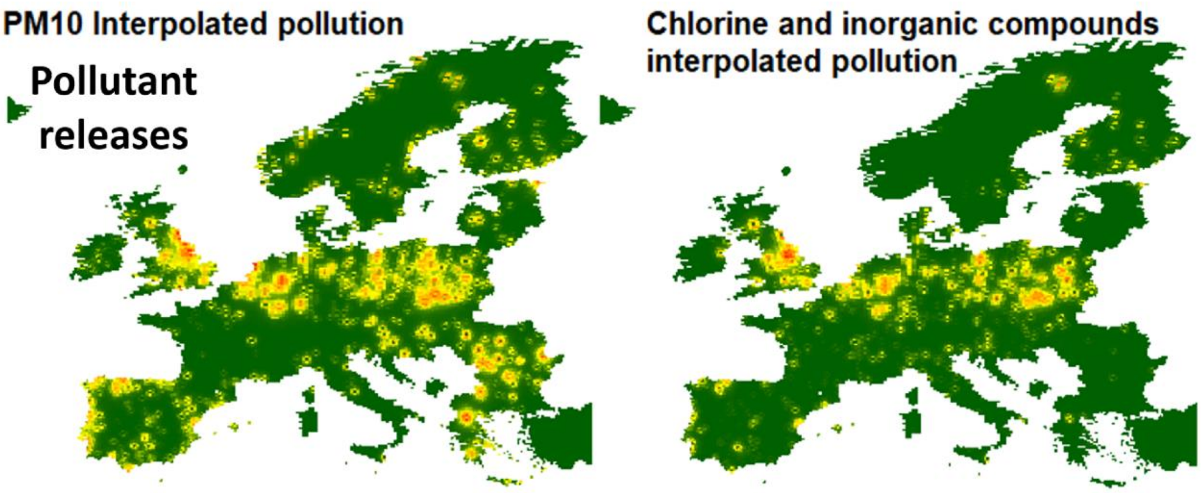

PM10 Interpolated pollution
Pollutant releases
Chlorine and inorganic compounds interpolated pollution

Figure 81. The interpolated map of the two most significant findings is based on spatial modelling.

The raster map of the interpolated PM10 and Chlorine with inorganic compounds was used to estimate the predicted PIBD incidence in Europe [Figure 82] based on the coefficients of our spatial regression model.

Although the prediction map in **Figure 82** provides useful information, it should be interpreted with caution and not compared directly to the reported incidence map. This is because the prediction map is solely based on pollution measures and does not consider some additional factors that could affect the reported mapped incidence of PIBD. For example, although areas such as Scotland may appear primarily green on the prediction map which indicates low expected incidence, this could be misleading. For Scotland the large areas in green include only a small fraction of the population and the overall incidence estimates are influenced mostly by smaller areas such as Edinburgh and Glasgow, which are predicted by the model to have a high incidence. Therefore, it is essential to consider the importance of aggregation methods when interpreting the prediction map. (This consideration was also discussed in 2.3.1).

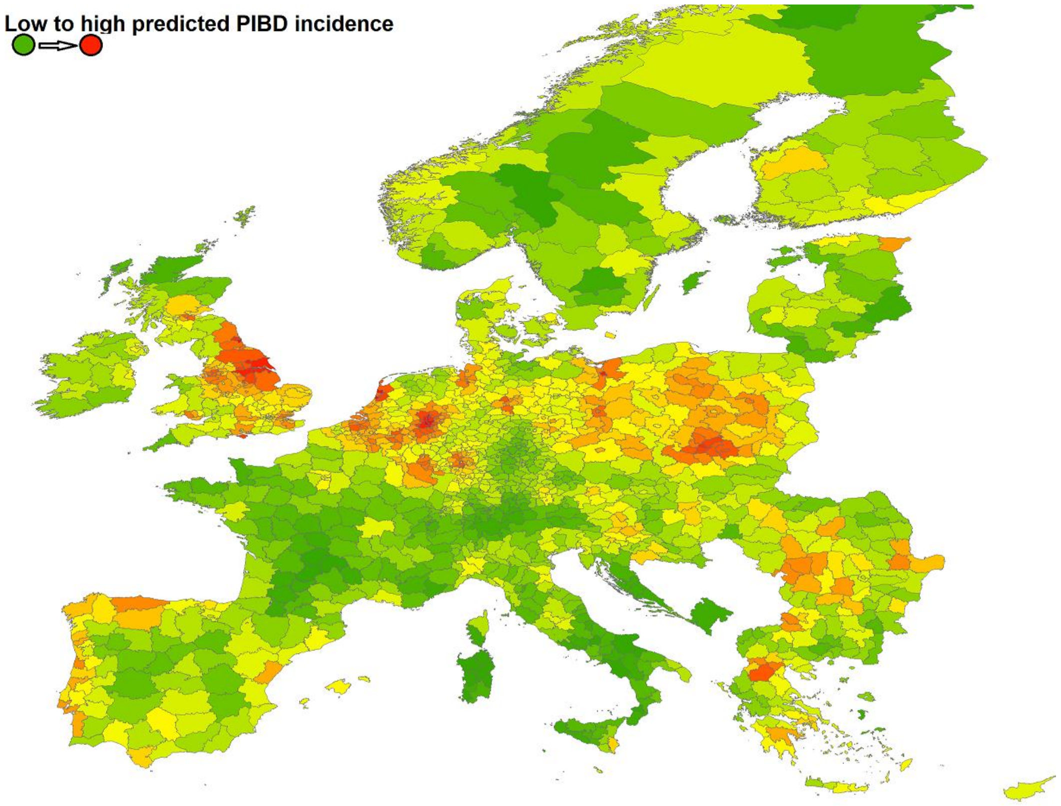


Figure 82. The predicted PIBD incidence in Europe based on the spatial regression model.

The predicted PIBD incidence was calculated for all NUTS3 regions based on the spatial regression model coefficients and the EEA interpolated maps of the pollutants that were found to be significant in our analysis.

6.7. Linear modelling of CD and UC with IBDU incidence

As discussed in the introduction, certain risk factors may have different effects on the CD and UC phenotypes. Thus, we have repeated the LMM analysis, testing the fixed effects of time and pollutants, with the outcome of PIBD incidence split by the two disease subtypes. The two investigated outcomes were the CD and UC/IBDU incidence. As shown in **Figure 83**, the four pollutants that were found to be significant, combined with time, in the PIBD analysis were also significant in the analysis of the individual phenotypes. Furthermore, five additional pollutants that did not exceed the significance threshold for PIBD were significant when analysed for the individual phenotypes. A surprising observation in the results is that certain pollutants exceeded the significance threshold of both phenotypes individually but not when combined into the PIBD incidence outcome. Further investigation of this occurrence revealed that the discrepancy is caused by a higher intercept estimate of the PIBD model compared to the individual CD and UC/IBDU models.

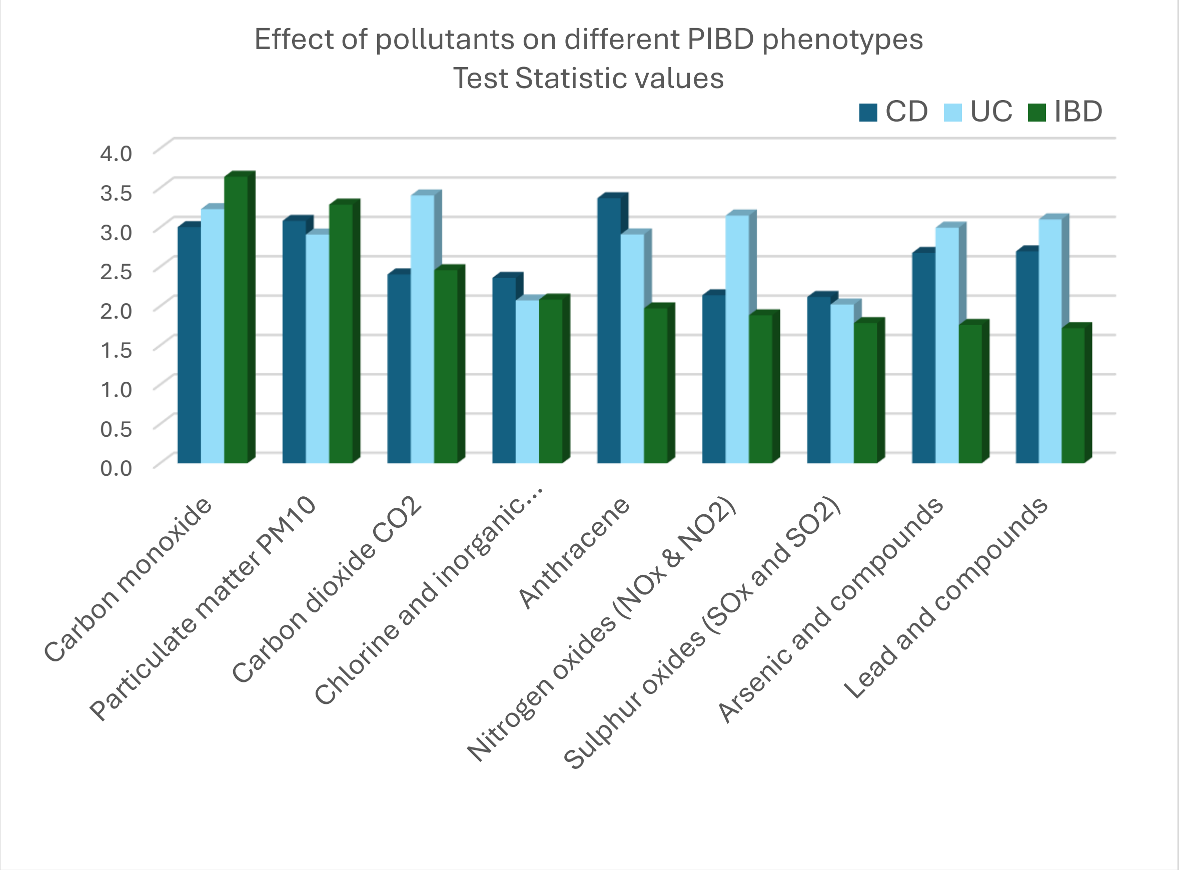


Figure 83. Summary of the significant effects of pollutants on the CD, UC/IBDU and PIBD incidence.

The t-test values of the fixed effect show the magnitude of the effects per pollutant on the PIBD, CD and UC/IBDU incidence. These results are also summarised in the following ***table 24****. Please note that any values above 1.96 are equivalent to a p-value <0.05. Time was included as a fixed effect in all models.*

As shown in **Figure 83** and **Table 24** below, the pollutants are ordered by their significance in the univariate analysis with PIBD as the outcome.

Table 24 Summary of all pollutants that exceeded the significance threshold for at least one phenotype.

This table summarises the t-values of all pollutants that exceeded the significance threshold for any phenotype. The fixed effect of time was significant in all models. The t-value for time is included in the table as an example of the PM10 model.

t-values of the pollutant fixed effects in the LMM models			
Pollutant	CD	UC	PIBD
Carbon monoxide	3.0	3.2	3.6
Particulate matter PM10	3.1	2.9	3.3
Carbon dioxide CO2	2.4	3.4	2.5
Chlorine and inorganic compounds	2.4	2.1	2.1
Anthracene	3.4	2.9	2.0
Nitrogen oxides (NOx and NO2)	2.1	3.1	1.9
Sulphur oxides (SOx and SO2)	2.1	2.0	1.8
Arsenic and compounds	2.7	3.0	1.8
Lead and compounds	2.7	3.1	1.7
Time (for PM10 model)	**5.4**	**6.3**	**5.38**

6.8. Additional findings, sun irradiance and population density

In the following paragraphs, we will analyse the predictors of population density and sun exposure separately and in addition to the previous models due to their distinct characteristics. Population density is expected to be correlated with several predictors, while sun irradiance is anticipated to have a strong correlation with latitude, which has a significant and well-established link with the disease incidence. The primary focus of our investigation will be to determine if these predictors can be incorporated into the existing models of our research. We will examine if they are essential and whether they will replace existing terms or explain additional portions of the observed disease incidence variance. In the following three analyses we update: i) the LMM PIBD incidence model, ii) the spatial regression PIBD model and iii) the CD - UC/IBDU LMM model.

6.8.1. Updated Linear modelling of PIBD Incidence

The same mixed effects analysis was repeated based on the following model, as described in 6.5:

$$Incidence_{ijk} = \beta_0 + \beta_{0j} + \beta_{i(j)} + \beta_1 \cdot time_{ij} + \beta_2 \cdot Exposure_{ij} + e_{ij}$$

The sun exposure was found to be a significant protective factor for PIBD while in contrast, the population density shows a less significant but positive correlation with the disease incidence. The t-value of the univariate model for solar exposure and population density is shown below in comparison with the four significant findings from the LMM pollution analysis (**Table 25**).

Table 25 Summary of significant risk factors for solar exposure.

Variable Name	**AIC**	**t- value**
Particulate matter (PM10)	5895.45	3.42
Carbon monoxide (CO)	5895.63	3.40
Carbon dioxide (CO2)	5901.13	2.42
Chlorine and inorganic compounds (HCl)	5902.24	2.18
Sun irradiation	5897.54	-2.62
Population density	5900.32	2.02

Furthermore, the addition of the solar exposure into the existing PIBD incidence model from section 6.5 of this thesis, is reducing the PM10 significance marginally while improving the overall fit significantly as shown in **Table 26**. A similar performance is observed for the rest of the significant pollutants. However, the population density was not added in these models as it marginally improved the fit when the sun irradiance and pollution terms were already included.

Table 26 The updated LMM model with solar irradiation, PM10 and time.

Model	AIC	Fixed effect term	Term t-value
Incidence ~ PM10 + Time	2174.4	PM10	3.243
		Time	5.80
Incidence ~ PM10 + Time + Sun irradiation	2137.7	PM10	2.65
		Time	5.38
		Sun irradiation	-2.57

Therefore, we can conclude that sun irradiation appears to be a strong protective factor in the PIBD risk.

6.8.2. Updated spatial regression for PIBD incidence

Similarly to the previous PIBD incidence analysis that was based on pollution exposures, the LMM model fit is followed by a spatial regression analysis. The results of the spatial regression model with the addition of sun irradiation as a predictor are shown in **Table 27** and are compared against the model that does not include sun exposure.

Table 27 Results of the PIBD incidence spatial regression model based on the sun and pollution exposures.

The results from the spatial regression model that did not include the solar exposure are shown in brackets for comparison purposes.

Variable	Coefficient	Std. Error	z-value	P value
Constant	-1.17 (-0.1)	1.52	-0.77	0.0243 (0.943)
Particular matter 10	0.32 (0.33)	0.15	2.11	0.0353 (0.025)
Chlorine and inorganic compounds (HCl)	0.46 (0.36)	0.13	4.11	0.0008 (0.004)
Sun Exposure	-0.037 (NA)	0.009	-4.07	0.0001 (NA)
Lamda (autocorrelation parameter)	0.67 (0.7)	0.03	18.79	<< 0.00 (<< 0.00)

Table 27, shows that the sun exposure has a significant negative effect on the PIBD incidence. Adding the sun exposure predictor in the spatial regression did not change the importance of the PM10 and Chlorine with inorganic compounds (HCl).

6.8.3. Updated linear modelling of CD and UC with IBDU incidence

In a separate LMM analysis for CD and UC/IBDU, solar exposure was strongly associated with lower CD incidence rates, while the population density showed a significant positive effect with CD. In contrast, for UC/IBDU, the solar exposure showed a negative effect that was not significant, while the population density presented a very strong positive effect. These findings are summarised in the following **Table 28**.

Table 28 Summary of the fixed effects of population density and sun exposure

The summary of the fixed effects of population density and sun exposure on the two PIBD subtypes is based on univariate LMM models with time.

Phenotype	Predictor	t- value
UC with IBD-U	Population Density	3.75
UC with IBD-U	Sun exposure	-1.53
CD	Population Density	2.28
CD	Sun exposure	-2.48

The introduction of population density and solar exposure reduced the number of predictors that remained significant in the LMM modelling. The best fitted LMM models for the CD and UC/IBDU incidence are summarised below (Table 29). For CD, the pollutant with the highest t-value and best fit was again PM10 (based on AIC). However, replacing it with the chlorine inorganic compounds also returned a similar fit (marginally inferior). For CD, carbon monoxide was the pollutant with the highest t-value and best fit. However, replacing it with Anthracene also returned a similar fit (inferior). The best fitted mixed models for the CD and UC incidence are summarised in the following **Table 29**.

Table 29 Final LMM models with the best fit for the UC/IBDU and CD incidence

UC/IBDU Incidence LMM model		**CD Incidence LMM model**	
Term	**t-value**	**Term**	**t-value**
Year (Time)	5.401	Year (Time)	4.918
Population Density	2.015	Sun Exposure	-2.662
Carbon monoxide (CO)	2.488	PM10	2.218
		Population Density	2.176

These results suggest that sun exposure is a strong protective factor while PM10, Carbon Monoxide, Chlorine with inorganic compounds remain important findings. The population density may also be important as it improves the overall fit when UC and CD are assessed separately. Lastly, the pollutant of anthracene also emerged as a finding that requires further attention. However, further investigation revealed that this pollutant is extremely clustered, localised and essentially present in two regions with high CD incidence. Therefore, it appeared to be a potential finding although it explained only a very small fraction of the disease variance.

6.8.4. Discussion

In this subchapter, we have mapped the disease incidence and predictors and investigated important properties of our data, such as the distribution characteristics, presence of clusters, outliers, autocorrelation, and other important properties. This work, in conjunction with model fitness diagnostics, allowed us to prepare our data and use appropriate methodologies for the geostatistical analysis. The disease incidence was studied for the PIBD cases and for the two disease phenotypes separately. Overall, from hundreds of predictors examined, we have identified sun exposure, population density, particulate matter 10, the year of data collection, carbon monoxide/dioxide and Chlorine with inorganic compounds (HCl) as the most important factors that were strongly associated with the observed incidence of the disease and its individual subtypes.

The mapping of the disease incidence reveals a very strong spatial autocorrelation (AR) in the reported incidence. The main AR sources are adjacency, the latitude and the data collection methods, as all regions covered by the same clinic were assigned the same incidence values. The presence of spatial autocorrelation in the data also suggests that the incidence of PIBD and its subtypes is likely influenced by local environmental factors. This, combined with the repeated measures aspects of our data, introduces a multidimensional source of autocorrelation which requires the use of specific models for the analysis of the incidence. Our analyses were based on two types of models, linear mixed effects models and spatial regression. Very importantly, our results underline the importance of tailoring the analysis methodologies to the characteristics of our data. Although many pollutants were correlated with the disease incidence, we were able to reject several variables that were not significant in the final analyses.

As this is the first and largest epidemiological study of its kind, focusing on the effects of the environment and specific pollutants on PIBD while considering hundreds of possible risk factors, it is crucial to validate our findings by identifying common points with similar previous studies. Specifically, our main protective factor finding, sun exposure, has also been reported as a finding in the literature. Several articles have reported sun exposure and vitamin D to be protective against IBD, especially Crohn’s (Jørgensen et al., 2010; Kappelman et al., 2007). This is aligned with our findings as we have identified the sun as a protective factor in PIBD and primarily for Crohn’s disease. Furthermore, according to our results, PM10 was identified as a major risk factor in PIBD affecting both the UC and CD phenotypes, which as a finding is also in agreement with the literature (Ding et al., 2022; Kaplan et al., 2010b). According to our findings, population density is also a risk factor in PIBD. We were able to identify only one study with similar findings (Ng et al., 2019). However, considering the high correlation between population density and the great number of environmental exposures in highly urbanised settings, this finding was not unexpected. Moreover, carbon monoxide and dioxide were also identified as significant risk factors for PIBD. This, again, has also been suggested previously, but the evidence is very limited (Ding et al., 2022). However, the one novel association in our study is the link between Chlorine and inorganic compounds with the PIBD incidence. Chlorine and inorganic compounds such as hydrogen chloride (HCl) are widely used in various industries today, while this chemical is found in plastics, solvents, textiles, and pharmaceuticals. It is also commonly used as a disinfectant in water treatment to kill bacteria and other pathogens. The uses of this chemical are very closely aligned with the "westernised lifestyle" and urbanisation, making it a promising risk factor that must be investigated further. In addition, this risk factor was processed and interpolated with a very high confidence, as close to 9,000 observations across Europe were used to determine its spatial distribution. As outlined in the validation section (2.4), precise results are greatly dependent on having a high level of confidence in the interpolation of the risk factor.

7. Analysis of the disease phenotype in the Inception Cohort population

The aim of this subchapter is to map the patients recruited by the several centres participating in the Inception Cohort and understand the spatial distribution of the disease phenotype. This patient level analysis does not include healthy control data and therefore, the disease incidence is not analysed. The studied outcome in this chapter is the variation of the disease phenotype depending on the location, exposures and exposome of the individual patients. Similarly, to the geostatistical work in the Safety Registry, this analysis also includes a large number of suspected risk factors. Although the sample size of this dataset is smaller compared to the Safety Registry, it is available on the patient level and has been validated and submitted by trained stuff.

7.1. Available data

The Inception Cohort dataset includes 432 prospectively recruited patients with available information from a detailed environmental questionnaire and follow-up data over one to five years depending on the time of the recruitment. The specific external exposome (1.3.4.1) analysis included four continuous variables and 230 factors. This specific external exposome dataset included various types of exposures such as medications received, dietary habits, pets, water source, type of dwelling and more. Within this dataset, a subgroup of patients provided postcode information allowing the calculation of the approximate residence coordinates for 432 individuals that were living in the UK and the Netherlands at the time of the study as shown in **Figure 84** below. Based on the geographical information of the patients, individual patient exposure data were extracted successfully for 113 types of environmental exposures.

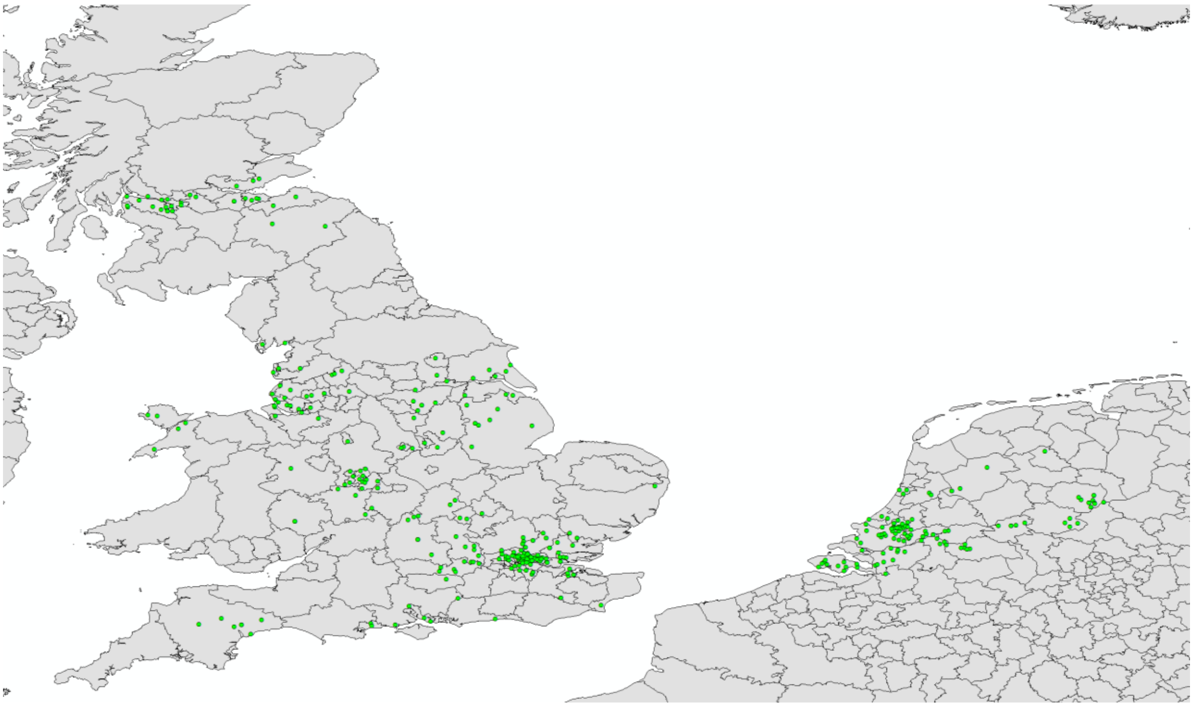


Figure 84. The locations of patients recruited in the UK and Netherlands were included in the patient-level data phenotype analysis.

The patients tend to cluster in specific locations which are closer to the Inception Cohort recruitment sites.

7.2. Findings

The aim of this subchapter is to perform a patient-level analysis and report factors that may have significant effects on the disease phenotype. It must be pointed out that variables with reports of significant variations in the CD/IBD ratio should not be interpreted as risk or protective factors. In the following analysis we report variables that appear to have significantly different effects on each disease phenotype but without control data we cannot determine whether they increase or reduce the incidence for any phenotype in a disproportional manner, or they have opposite effects on CD and UC/IBDU. The investigated predictors were examined univariately as terms in a logistic regression model and in combinations of two including their interaction terms. Considering the number of predictors and patients in our study, any analysis with additional interaction terms or single 3-way interaction terms will result to an extreme number of several thousand examined combinations.

In the broader dataset with the 432 patients, 7 variables were found to be significant predictors of the disease phenotype and are summarised in the **Table 30** below.

Table 30 The results of the univariate logistic regression analysis of the disease phenotype.

The exposome variables were studied to assess if they could be used to predict the CD diagnosis over the UC/IBDU. The most significant predictors are included in this table with their respective p-values.

Variable	Format	p value (logit)
Subject's Age at Diagnosis	Continuous	0.0006
Owing "Other" Pets/animals were	Factor	0.0041
Biological Father's Ethnic Background	Factor	0.0159
Varicella Vaccine	Factor	0.0218
Consecutive years did you take multivitamins?	Factor	0.0259
BCG Vaccine	Factor	0.0317
Water supply	Factor	0.0456

In our sample, older age at diagnosis is favouring CD phenotypes against the UC and IBD-U. As shown in **Figures 85 & 86** below, the CD diagnoses in the early years were marginally less common but increased rapidly after the age of 10. It should be noted that 36.7% of the youngest 30 patients in our study presented CD, while the percentage increases by age reaching 66.7% for the oldest 30 patients recruited. Very importantly, because the phenotype was split into two different outcomes, an important interaction was not captured by the logistic model. As shown in **Figure 85**, the effect of age on the phenotype depends on the sex of the patient.

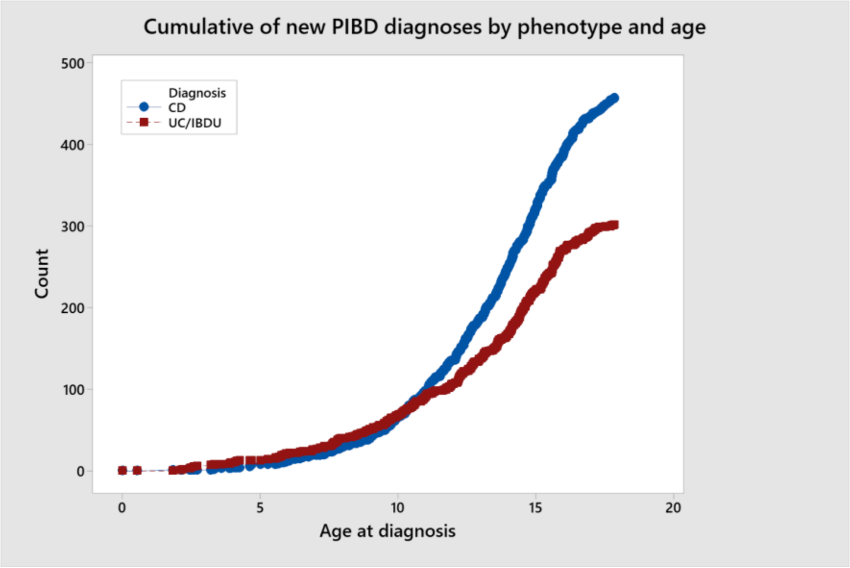


Figure 85. The cumulative number of new CD and UC/IBDU cases in the Inception Cohort by age.

Younger age favours the UC/IBDU phenotype marginally up to the age of 10 where the CD phenotype becomes more prevalent. Both phenotypes show an increase of PIBD diagnoses after the age of 10.

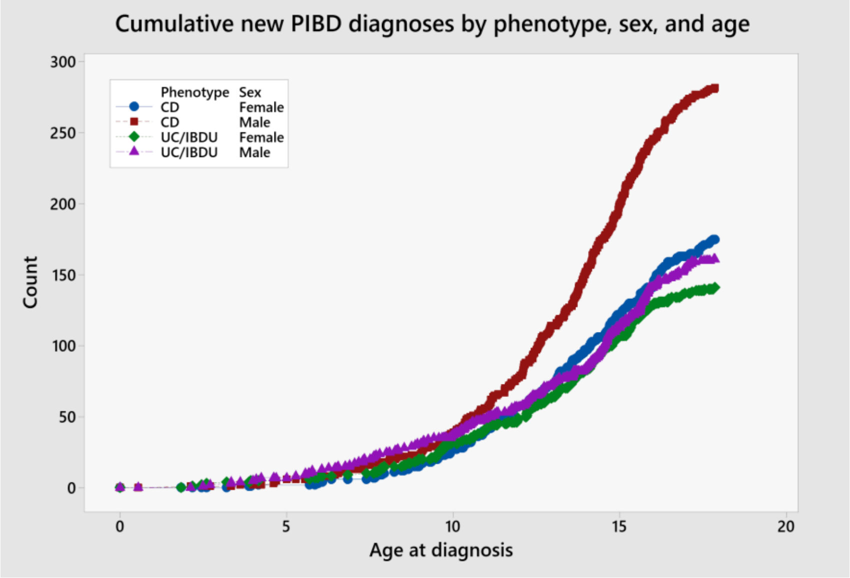


Figure 86. The cumulative number of new CD and UC/IBDU diagnoses in the Inception Cohort by sex and age.

The CD rates were found to be higher for both females and males. However, the main driver of the increase in PIBD is the steep increase in CD diagnosis in males older than 10 years.

As shown in **Table 30**, exposure to pets is also an important predictor. When this question is answered by the patients as “other”, it also favours the CD phenotypes in the Inception Cohort population. While the overall CD rate in our sample is 58.1%, this percentage decreases to

56.3% for the patients who did not own a pet from the “other” category and increases to 73.2% for those who did (60 out of the 83 patients, 2-proportions test p = 0.002). The “other” category included primarily rodents and rabbits. It should be noted that from the 67 households with rodents and rabbits, 78% of the patients presented CD.

Furthermore, black ethnic backgrounds of the biological father seem to favour the CD phenotype. The results were similar for the black ethnic background of the biological mother too, but the effect size was marginally smaller hence increasing the p value of the test. 75% and 90% of the patients with a black mother and black father respectively were diagnosed with Crohn’s disease, while all non-white ethnic groups presented higher CD rates compared to the white ethnic background patients. Repeating the analysis for patients with at least one biological parent with black ethnic background showed that this group’s phenotype ratio changes by 27.3% (95% C.I:13.7, 41), favouring the CD diagnosis (2-proportions test p= 0.00009, exact test p= 0.006).

The variable of consecutive years of multivitamins use was rejected upon further investigation. The results suggest that receiving multivitamins in the year before the diagnosis reduces the percentage of CD compared to UC and IBD-U. However, the CD rate increases again significantly for all groups receiving multivitamins for both longer and a shorter time periods.

The two types of vaccines that appeared to be significant in the analysis were the Varicella and BCG vaccines (**Table 30**). The latter, when received at least at one dose was associated with lower CD rates, while the Varicella vaccine was associated with higher CD rates (two proportion tests p-value: 0.045 and 0.031, respectively). It is essential to clarify that this association is not significant after factoring in the required adjustment of the significance level, to account for the multiple comparisons. Furthermore, as stated in the introduction, the reported effects may be increase or decrease the risk for both phenotypes resulting to change in the phenotype ratio.

Lastly, upon further investigation, the water supply finding reveals a strong association between the consumption of bottled/non-bottled water and the rate of CD. A two-proportion test comparison of the group consuming bottled water against the groups consuming water from the main supply (filtered or unfiltered) shows a 19.2% higher rate of CD, in the latter group with a p-value of 0.0029.

7.2.1. Spatial regression phenotype analysis of the Inception Cohort patients

By incorporating the geographical information provided by 423 patients, it becomes possible to identify the exposures of each patient to the interpolated risk factors similarly to the previous subchapter. This expanded analysis involves adding the values of all risk factors at the participant's location and incorporating their corresponding values associated with that location to the patient data. The univariate logistic regression analysis that was used for the specific exposome analysis in the previous paragraphs, was used to identify environmental pollutants that warranted further investigation. A spatial regression with the selected pollutants returned three variables of interest with positive and negative associations to the disease phenotype ratio. These variables are summarised in the following **Table 31**.

Table 31 The environmental pollutants that may affect the PIBD phenotype ratio.

Variable	**Effect on CD/UC-IBDU ratio**	**Format**	**p-value (logit)**	**Logistic model AIC**
Hydrofluorocarbons (HFCs)	Increase	Continuous	0.029	534.1
Chlordane	Decrease	Continuous	0.036	535.4
Tetrachloroethane	Increase	Continuous	0.037	533.8

Upon further investigation of potential interactions among the predictors in our dataset, no associations were found to be significant after accounting for a large number of combinations of factors and levels (>1000). Multiple testing increases the likelihood of false positive findings since each additional combination of factors included in the model raises the probability of such findings. As shown in **Figure 87** the dataset was highly clustered since the patient location were most frequently close to the clinic that recruited the patients. Although we can account for the clustering effect this reduces the variability and area coverage required to identify true effects. In contrast to the lattice data analysis in the previous chapters, the sample size becomes irrelevant in point locations dataset when they are all found in the same areas.

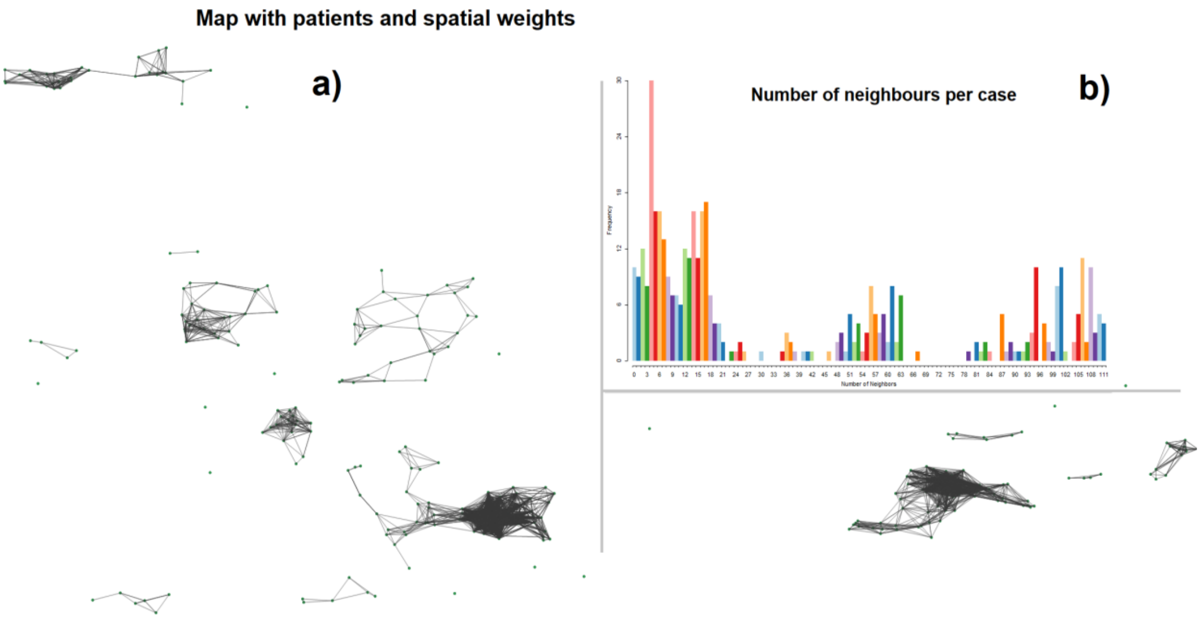


Figure 87. The patient spatial weights and the distribution of spatial dependencies count per the patient.

Panel a) shows the spatial relationships among all patients with neighbouring observations, while panel b) displays the number of spatial dependencies per patient.

In summary, the Inception Cohort analysis revealed strong associations between the disease phenotype and certain variables, such as age at diagnosis, specific types of pets, parents' ethnic background, and water source. Additionally, five predictors were identified as potentially significant, namely, BCG and Varicella vaccines, as well as Hydrofluorocarbons (HFCs), Chlordane, and Tetrachloroethane, with the latter three to be revealed through geostatistical analysis.

7.3. Discussion

In this subchapter, we have investigated the ratio of PIBD subtypes using patient level data from the Inception Cohort and mapped patients recruited in the UK and the Netherlands. Although both CD and UC phenotypes involve chronic inflammation of the gastrointestinal tract and fall under the IBD category, their molecular mechanisms underlying their pathogenesis differ (Podolsky, 2002). Subsequently and as discussed in the introduction, certain IBD risk factors may be different for the CD and UC phenotypes. Therefore, these factors can be studied in our dataset when the CD cases are essentially used as control data to study the risk factors in UC and vice versa for the CD risk factors. To understand further our findings from the Inception Cohort phenotype analysis, in the following paragraphs we will study the relevance of our results with the current literature. Similarly to the incidence, prevalence analysis, this is important for the validation of the results and the identification of potentially novel findings.

In this subchapter we showed that in the studied population, age does not only increase the overall risk of IBD but disproportionately increases the Crohn's disease risk for males over 10 years old. As discussed in the introduction, several studies have reported that the disease incidence and the ratio of the diagnoses between males and females changes with age (Ludvigsson et al., 2017; Urlep et al., 2015). Several studies have reported that in the paediatric population, the CD phenotype is more common in males, while the UC phenotype is prevalent in females (Sýkora et al., 2018 ; Ha et al., 2010). However, the difference observed in our data significantly exceeds any results reported in the literature. This may suggest a recent change related to the specific characteristics of the population recruited in the Inception Cohort.

The exposure to certain types of pets, such as rabbits and rodents, was also associated with a higher incidence of CD in our study. There have been several studies investigating the association between exposure to pets and the incidence of Crohn's disease and ulcerative colitis (Cholapranee and Ananthakrishnan, 2016). While the evidence is not consistent and the strength of the association remains unclear, a large review and meta-analysis reported that the exposure to pets and farm animals at an early age reduces the overall risk of IBD significantly (Cholapranee and Ananthakrishnan, 2016). Cholapranee, also reported that this effect was marginally greater for UC compared to CD. Our results could possibly reflect the same effects but with specific types of pets having an even greater protective effect against the UC type. Alternatively, it is also possible that rodents as pets could be a risk factor for paediatric IBD, but this has not yet been identified in the literature, as previous studies have grouped animals into broad categories.

Our study also identified an association between black ethnic background and a higher incidence of CD compared to UC and IBDU. The effects of race on the IBD phenotype varies significantly depending on several factors including the location, immigration patterns and the methods of the conducted study. According to current literature, it is more likely for the white population to present a higher CD to UC ratio compared to the black population(Aniwan et al., 2019; Barnes et al., 2021; Misra et al., 2019), making this the first finding of our study that contradicts the literature. However, considering the fast-changing incidence of IBD, it should be noted that there are no recent relevant published data. Therefore, we cannot formally assess the effects of race on the diagnosis phenotype in the paediatric population based on previously published information.

Additionally, our results suggested that the source of water consumed by patients may also play a role in the development of IBD. Specifically, we found that consumption of bottled water was associated with reduced risk of CD compared to UC. Studies that analysed data from thousands of subjects have also revealed associations between water source and water quality with the presentation of PIBD. Vanhaecke et al., 2022 reported that the consumption of bottled water was associated with smaller IBD rates compared to the other sources. However, the reported effect size was rather small and not significant. Holik et al., 2020 conducted a very relevant study investigating specifically the drinking water and its effects on patients with IBD. Holik reported that the rural and well water supply increased the CD - UC ratio by a small percentage. He also reported that the rate of CD to UC was 37% higher in patients consuming low quality water, although this was reported as not statistically significant. The effects of drinking water on IBD have been studied by others (Aamodt et al., 2008; Hermon-Taylor, 1993; van Kruiningen and Freda, 2001) and given the current literature, we cannot verify that there is an effect on IBD overall. However, our findings combined with the literature suggest that water quality may affect more the CD phenotype compared to the UC and due to the small effect size, the detection of this is challenging.

Finally, our study identified five suspected factors that may influence the disease phenotype. The suspected pollutants, Hydrofluorocarbons (HFCs), Tetrachloroethane and Chlordane, are not reported in the literature in relation to the IBD. It is important to underline that the spatial analysis of the Inception Cohort was considerably limited compared to the Safety Registry analysis. This is related to the geographic coverage of the Inception Cohort which is equal to a small fraction of the Safety Registry. The patients included in this analysis were based in certain parts of the Netherlands, Greater London, Midlands and the areas close to Edinburgh and Glasgow in Scotland. Furthermore, it seems that the patient location is clustered compared to the general population, reducing the variability of exposures dramatically. Therefore, the findings of the spatial analysis will not be investigated further. Similarly, the final two suspected findings of BCG and varicella vaccinations were borderline significant findings that have not been reported to the literature previously. Therefore, it is very likely that they are not influencing the incidence of IBD and its phenotype ratio.

8. DISCUSSION

8.1. Summary of findings

In this chapter, I have outlined and summarised the study findings parallel to the relevant literature.

8.1.1. Incidence and prevalence findings

Overall, the results of this study were consistent with established research findings, which supports our methods for data collection and incidence analysis. In terms of the geographical trends of PIBD, as previously discussed, countries located closer to northern latitudes tend to have higher incidence rates compared to those situated further South. Our results indicate a strong increasing latitude trend in PIBD incidence rates, which agrees with the previous literature. In terms of the country-specific comparisons, as shown in the final incidence map in **Figure 66**, common patterns reported in large systematic reviews emerge. For instance, the higher incidence in Scotland compared to the rest of the UK emerges while the low incidence in Greece and Italy is also confirmed. Similarly, the increased reported incidence in Finland is also in agreement with the literature. Lastly, data from Austria and Poland were the main discrepancies in our data, but as discussed, this is related to the reporting error of the catchment areas. In our future work, we will have the option to rectify this retrospectively and prospectively.

Additionally, as discussed in the introduction, the incidence of PIBD has been rising in all European countries, apart from countries in Eastern and Southeast Europe. Our findings reveal a marginal decrease in the disease incidence in countries such as Serbia, Lithuania, Hungary, and Greece. Temporally, although our results are generally consistent with current literature, we observe a steep increase in PIBD incidence rates from 2020 onwards, which may be related to the COVID-19 pandemic. The role of viruses in IBD risk and development has been studied by several research groups reporting that there are viral infections that can influence the development of IBD because of their interaction with the patient’s microbiota (Tarris et al., 2021; Ungaro et al., 2019). Other studies have investigated the role of norovirus, rotaviruses and cytomegalovirus, with the latter being confirmed as a risk factor in IBD (de Hertogh and Geboes, 2004; Mavropoulou et al., 2019). The cytomegalovirus (CMV) infection increases the risk for IBD through several possible mechanisms, including increasing gut permeability which is a known mechanism in the CMV pathogenesis, expression of vascular cell adhesion molecule-1, or increased interleukin-6. Recent studies have also shown that COVID-19 introduces microbiota alterations (Delgado-Gonzalez et al., 2021; Ungaro et al., 2022). Studies in murine models demonstrated how the virus could increase susceptibility to inflammation and colitis development due to decreased antimicrobial peptides and the alteration of gut microbiota (Delgado-Gonzalez et al., 2021). Furthermore, other studies also reported that COVID19 increases the gastro-intestinal symptoms of IBD patients (Ungaro et al., 2022). In summary, the literature provides examples of mechanisms where COVID-19 may increase the IBD risk. However, the rate of increase in our study was too high to be explained by COVID-19 alone. Therefore, we should consider the possibility of the impacts on the lifestyle as discussed in 3.2.5. or major reporting and patient management changes after the pandemic. However, the latter seems unlikely considering that the increased incidence was reported in both 2020-2021 and 2021-2022 at an essentially identical rate.

Geostatistical analysis findings

Although IBD has a genetic component, environmental and lifestyle factors play a more significant role in its development. As discussed in the introduction, several studies suggest that a low percentage of cases are due to genetics alone. The environmental effects are significant, as evidenced by differences in gene expression between homozygotic twins and other studies. Hence the need to identify the underlying environmental factors is a key epidemiological aim. As mentioned in the introduction, although the current literature on the topic is developing, a few environmental factors have been reported repeatedly with high confidence. Sun exposure/vitamin D is one of the most representative examples of such factors. This was also a finding in our study that we can report with high confidence. Furthermore, PM10, COx and Chlorine with inorganic compounds were important findings in our analysis. PM10 is an important predictor for various chronic diseases and recently has been associated with IBD as well. PM10 is often a proxy for PM2.5 which due to its size is more difficult to measure. PM10 pollution refers to the presence of particulate matter with a diameter of 10 micrometres or less in the air. These particles are small enough to be inhaled deep into the lungs and can cause respiratory and cardiovascular problems, especially in vulnerable populations such as children, elderly, and people with pre-existing health conditions. Sources of PM10 pollution can include traffic, industry, construction, and natural phenomena such as dust storms or wildfires. Therefore, PM10 seems to be a relatively broad indication of pollution which may explain the strong positive effect that it presented on the incidence of PIBD. Carbon oxides, although they are specific chemicals, they are also indicators of overall pollution. Carbon monoxide in particular, has been associated with IBD and this is an additional finding that we can confirm. The Chlorine with Inorganic compounds finding however, is novel. Inorganic compounds and halogens in particular, have not been associated with the development of IBD thus far. The EEA provides these data grouped which complicates the identification of the exact compound that is responsible for the observed effects. The E-PRTR (European Pollutant Release and Transfer Register Regulation) provides additional information on every pollutant source that we included in our dataset and analysis. Therefore, future work is required to identify which of the 9000 observations in the Chlorine with Inorganic compounds variable are associated with the disease incidence and what are their common characteristics. Lastly, the population density was also found to be significant in our study, again in agreement with some articles in the literature. However, this should not be accepted as a finding itself since population density alone cannot influence the disease.

This is a proxy finding for the true underlying effect. Urban areas are frequently more polluted or may have a higher risk of infection, limited sun exposure and other differences compared to rural areas. This is possibly similar to the protective effect we see for the group of pesticides on PIBD as shown in **Figure 75**. It is unlikely that high level of pesticides in an area will reduce the rate of PIBD but it is possible that that low levels of pesticides are linked to urban settings and specific kinds of pollution that are also found in urban settings. Perhaps this may also be related to lifestyle choices that can vary between urban and rural settings. These are examples of proxy predictors and results that we need to consider before presenting a finding as significant.

Phenotype ratio findings

As previously reported in the literature, the ratio of CD to UC tends to decrease as we move from Western to the Eastern Europe (Goldiș et al., 2019; Khalif and Shapina, 2017). Again, in agreement with the literature, we have also identified this spatial trend in our data. Specifically, we have observed a strong West-East reduction of the CD rate. Furthermore, we have also identified a less significant reduction from the North to the South. The patient-level analysis also returned two important factors that warrant further investigation, drinking water and having pets. The finding of pets contradicts the literature, as contact with animals, especially at a young age, seems to have a protective effect. However, there are no studies investigating specific types of pets, so we cannot exclude this as a possibility. Perhaps this might not be related to the pets but the type of food they consume or another indirectly linked to them predictor. The water source is essentially related to the water quality and therefore is not unexpected that the studies in different areas reported different results. In future research, in order to verify this finding from the Inception Cohort data, we should investigate whether this effect (bottled water protective for CD) is seen in patients from all participating countries or from specific countries only.

8.2. Novelty of the study

In terms of its design, this study can collect epidemiological information prospectively from hundreds of clinics in over 30 countries simultaneously, using the same methodology. This methodology is a vital and novel element, as we present here the most extensive study in scale in the field of PIBD epidemiology and reduce the heterogeneity that other studies might have, when combining data from different sources. Firstly, it is important to note that when combining information from different regions, various methods were used for data collection and analysis. Secondly, the observed differences might not reflect any underlying effects, but variations that were introduced from the methodological inconsistencies. To underline the importance of using harmonised methods, we can hypothesise a scenario where our methods introduce bias, such as underestimating the incidence of PIBD. Given the consistency of the methods used in our study, this bias is expected to be consistent across all regions. Consequently, the geostatistical analysis will remain unaffected, as it is based on the differences between the studied Nomenclature of Territorial Units for Statistics (NUTS3) areas instead of the actual values. Also, as we have seen from the literature review in the introduction and our results, the PIBD incidence is changing rapidly over time. This requires that the data collection from different regions occurs within a small-time margin; otherwise, any comparison will be invalid.

As mentioned, the coverage and scale of this study also makes it novel in the field of PIBD. As our network keeps expanding, we have exceeded 30 million general paediatric person-years and 10,000 PIBD patients-years in coverage in Europe alone. The data collected from 23 European countries also provide significant geostatistical variation to capture the environmental effects on the disease incidence and prevalence. Therefore, this study can produce results that would traditionally require several studies combined with a systematic review and meta-analysis.

The prospective nature of our study also makes it a unique project that can provide information on such a large PIBD population prospectively and with frequent updates. This is the only study in the field of IBD that is able to report the incidence and prevalence rates for each participating country annually, with the coverage in some countries such as the UK, Israel and the Netherlands reaching 100%. This makes our study very sensitive in detecting phenomena in an almost real-time manner, making it a novel characteristic, considering that alternative research methods would take require several years to detect such events. The suspected impact of COVID-19 on the incidence is an indicator of this potential.

Regarding the methods, this was also the first study in the field of PIBD, and generally IBD, where the statistical analysis incorporated and adjusted for the effects of clustering and autocorrelation (that emerge from the spatial and temporal nature of our data). In total, we were able to identify three studies in IBD that have considered the autocorrelation, but these studies focused on mapping and not the analysis of the disease risk factors. Therefore, this is the first project in IBD employing such methods, including linear mixed effects models and especially spatial regression analysis.

One of the main methodological novelties of this project lies in the management of multiple spatial datasets using different spatial support and formats. In this study, we have collected the incidence data using the NUTS3 that had misalignment issues due to nomenclature updates over different years. The pollution and environmental datasets were also misaligned as they were available in various formats, primarily in point data using several geodetic systems and random locations. Using conversions of geodetic reference systems, population and polygon centroids, interpolation, aggregation, and other methods, we overcame these misalignment challenges. After rigorous testing and several simulations, we have determined that Kriging interpolation algorithms and EBK in particular, can be powerful tools depending on the extent and type of autocorrelation of the target, while IDW with small cell size and appropriate parameter settings that reflect the exposure mechanism can perform exceptionally well for certain pollutants.

Finally, we can also make novelty remarks regarding the findings of our study. Upon investigating the factors and covariates that may drive the incidence of the disease, we identified specific novel associations. Surprisingly, a steep rise in disease incidence was detected from 2020 onwards, a fact that has yet to be reported by other studies. In addition, chlorine and inorganic compounds pollution is also a novel finding regarding the chemicals and environmental exposures. Lastly, from the list of lifestyle risk factors, the water source and rodents as pets were also novel associations. However, our data do not provide evidence that they are risk factors for IBD; rather that they may affect one of the two disease phenotypes or both, but in a disproportional manner.

8.3. Limitations of this research

In clinical research, it is important to acknowledge and address the limitations of the study design, data collection and analysis, as these limitations can affect the interpretation and generalisability of the findings and influence the conclusions drawn. Therefore, a thorough discussion of the limitations is a crucial component of our project. Identifying and addressing the limitations of our study is also an essential step in designing future research where these limitations are minimised.

The first limitation is related to the Safety Registry, where the submitted data cannot be validated against their source. For instance, in the Inception Cohort, being a clinical study, when we suspect possible discrepancies in the entered data, we can crosscheck the submitted information against the clinical records of the patient after reaching out to the recruitment sites. However, in the safety registry, the information from the participating PIBD experts cannot be validated in a similar manner as frequently. In some cases, this has been possible, but it is not always expected as the safety registry relies on voluntary participation. As discussed in the methods and results chapters, we have used alternative approaches based on metrics focusing on outliers and the consistency across and within different centres to address these issues.

However, although this approach is likely to identify mistakes, it may also be biased against extreme results and sudden shifts over time that may seem inaccurate when applying our methodology. We have used a stringent criterion to remove observations with outlying metrics results to minimise this risk. Another disadvantage of these metrics is that they can only detect extreme reports and therefore, may overlook inaccurate reports that randomly fall closer to the expected values. Such a discrepancy can only be detected in centres that reported over multiple years, giving us the necessary information to verify the consistency of their reports.

In addition to the difficulties encountered in verifying the results reported in the safety registry, another significant limitation identified is the inconsistencies in the catchment areas reported. The findings highlighted countries such as Poland, Austria, and Finland exhibiting notably high results that appeared implausible. Despite this, the validation metrics indicated that the figures submitted were consistent across different years and in relation to both the incidence-prevalence ratio and the CD-UC/IBDU ratio. This suggests that the inaccuracies in the incidence and prevalence rates reported by these countries are likely attributed to issues with defining the catchment areas accurately, rather than inaccuracies in the number of new and existing PIBD cases managed by the reporting centres. The challenge lies in the reporting centres’ ability to precisely identify all the regions their patients originate from, leading to an underestimation of the clinic's coverage. As a result, a smaller denominator of person-years is applied to these centres, thereby artificially elevating the estimated incidence and prevalence rates. In future work, to address this limitation, we might consider obtaining patient-level data directly from the reporting centres to accurately define the catchment areas ourselves. Alternatively, we could add fields to the electronic reporting forms that allow reporting experts to indicate their confidence level regarding their defined catchment areas. This approach would enable us to filter the data, focusing only on areas where the confidence in the denominator data is high, thereby enhancing the accuracy of our incidence and prevalence estimates.

Another limitation related to the reporting is the overlap between different reporting centres. As shown in **Table 13**, the overlap is limited in our study but nevertheless present. In these cases, we have assumed that the coverage is equally split between the centres claiming the same NUTS3 regions. However, we are in no position to know whether this is true, since it is likely for one centre to cover a larger part of the overlapped area. When this occurs, it will inflate the incidence for one centre and decrease it for the other one, as the denominator of that region should not be split as 1:1 but in an imbalanced manner. In practice, considering that the reporting units are covering multiple regions, this has a small impact on the incidence estimate (~5%).

We have also identified limitations in our geostatistical analysis, with the size of the mapped NUTS3 areas being one of them. Using the NUTS3 database was essentially our only option since it provided a platform of harmonised territorial units that we used for data collection, disease mapping and calculation of the disease incidence. The demographics, several predictors and denominator population were available for all countries by Eurostat on the NUTS3 level. The population determines the number of these territories and not the area size, meaning that low population density regions have a small number of geographically large NUTS3 territories. In practice, this means that in areas of low population density, we may overestimate the population exposure to certain risk factors. This is because these areas contain large NUTS3 territories that are more likely to include locations with high pollution due to their size. Given their size, it is also possible that the population in these NUTS3 territories is far from the point of the location, leading to the overestimation of the exposure. In addition, in the rare cases where a large NUTS3 territory contains two or more population clusters that are far apart, it is challenging to produce an objective estimate for the average exposure of the total population in that area. This limitation falls under the general limitation of using aggregate data, a common practice in epidemiology. Patient-level data that we could use for the geostatistical analysis of the incidence would contribute to a significantly higher statistical power, as we discuss in the following section.

An additional limitation is related to the assumption that newly diagnosed patients have not relocated outside the areas covered by each reporting unit. Assuming that for a specific exposure to influence the disease incidence, a certain amount of time, the accuracy of our results will be limited by a large number of patients who have relocated recently outside the regions covered by the reported clinic. Although we cannot control or adjust our data for this effect, studying the paediatric population exclusively reduces this effect significantly as this population relocates less frequently.

Regarding the geostatistical analysis, it must be acknowledged that there is a potential bias in reporting the environmental factors, specifically the pollutants. Suppose any specific contaminants are underreported in areas where lower PIBD incidence has been detected. In that case, a false association between the areas of low incidence and low pollution levels will emerge. This issue arises from the assumption that areas with no emissions and reported pollutant releases are low-pollution areas. However, it should also be noted that no evidence suggests that the European Environment Agency dataset is incomplete. A similar bias may also be present in the phenotype analysis, which uses questionnaire responses from the inception cohort patients. It is possible that for risk factors such as smoking and vaping, not all paediatric participants will be willing to share that kind of information with the site staff during the data collection. Therefore, this very likely to skew the relevant data making this risk factor potentially invalid for the analysis.

8.4. Translational potential and future research

The future goals in this work are related to both the Safety Registry and Inception Cohort studies. Although several centres in their Inception Cohort have reached their targets and stopped recruiting, most patients have many years of follow-up available and a proportion of them are still followed-up. Therefore, we would like to process additional data from these patients and investigate whether their geographical location, lifestyle choices and environmental exposures have affected the development of their disease and, subsequently, their quality of life. During my PhD, I have studied the phenotype ratio using this patient-level dataset, but future research must be expanded to the numerous clinical outcomes available from the Inception Cohort. Another significant finding that emerged from the Inception Cohort data which warrants further investigation is the sudden increase of CD in males older than 10. Based on the follow-up data of the Inception Cohort it may be possible identify additional explanatory factors, possibly associated to the patients' sex that explain this increase.

Regarding the Safety Registry, the immediate goal is to proceed with an additional collection round for 2023-2024 period and validate our findings prospectively. This would require setting specific hypotheses based on our current findings to confirm the significant predictors of PIBD incidence. This will include the four pollutants discussed in the previous paragraphs, solar exposure, population density and the potential effects of the COVID-19 pandemic. If the expansion of the Safety Registry proves to be too challenging, as an alternative, we will consider expanding in targeted areas only, where according to our results and subsequent modelling, are expected to present the highest or lowest PIBD incidence in Europe. Continuing the data collection and ensuring that the database remains operational is a crucial task itself, as this would provide insights on the incidence trends almost in real-time, as discussed in the previous paragraphs. This may be of great value in detecting important factors that influence the incidence in different regions over time. The disease prevalence data collection is also important as it supports the calculations of the incidence of rare and severe complications in PIBD, which is the primary function of the Safety Registry. This is, however, out of scope for this project. Another future goal is to expand the Safety Registry further and increase the number of active participants. Adding more regions in the geostatistical analysis will increase the statistical power of our analysis and may subsequently help us identify additional important factors. Finally, two important additional future goals, related to the data and methodology of the Safety Registry, are the collection of patient-level data and the expansion of the predictor data over different time periods. The latter would allow us to also incorporate the "lag effect" from the time of exposure until the time of disease presentation. It will also allow us to quantify the exposure needed before the development of the disease, assuming that this fits the underlying disease mechanism (i.e., three years or longer of exposure to high PM10 increases the incidence by 15%). Although increasing the temporal granularity is important, an increase of the spatial granularity with the use of patient-level data would be the single most important improvement in the accuracy and precision of the study. In our last data collection round, most participants responded positively to our request for patient-level data with location information in the future. Such a task would be challenging as it would require local ethics approvals, but it would also provide high-detail environmental exposure data for thousands of patients across several European regions. This would eliminate a lot of the uncertainty related to the use of aggregate data and further increase the validity of our results.

Successful epidemiological and clinical projects in general, should have a translational potential. This is the potential for our research findings to be translated into practical applications that can improve patient outcomes and inform clinical practice. The Safety Registry has already supported our work that may influence the clinical practice related to the management and prevention of the rare complications in PIBD. In the context of my PhD, the translational potential may refer to the application of research findings to improve the diagnosis, treatment, and management of PIBD in clinical settings. For example, findings on the environmental and lifestyle factors associated with PIBD incidence and prevalence may inform targeted interventions aimed at reducing exposure to these factors in affected populations. Similarly, geostatistical analysis of PIBD incidence and prevalence may help identify high-risk regions, inform the allocation of healthcare resources, and develop targeted interventions in these areas. Most importantly, identifying the risk factors of PIBD, a disease with a major environmental component, can provide targets for basic research to help us understand the underlying mechanism of IBD. This would ultimately improve patient outcomes and inform clinical decision-making in preventing, diagnosing, and treating PIBD.

CITATIONS

Aamodt, G., Bukholm, G., Jahnsen, J., Moum, B., Vatn, M.H., 2008. The Association Between Water Supply and Inflammatory Bowel Disease Based on a 1990-1993 Cohort Study in Southeastern Norway. Am J Epidemiol 168, 1065–1072. https://doi.org/10.1093/aje/kwn218

Aardoom, M.A., Klomberg, R.C.W., Kemos, P., Ruemmele, F.M., Fagbemi, A., Kiparissi, F., Schweizer, J.J., Sebastian, S., Russell, R.K., Torrente, F., van Mill, M., de Ridder, L., Croft, N.M., Tempia-Caliera, M., Lee, W.S., Pigott, A.J., Classen, M., Morris, M.-A., Muhammed, R., Hussey, S., Cananzi, M., Menz, T.J., Wahbeh, G.T., van Ommen, C.H.H., de Ridder, L., Croft, N.M., Turner, D., Focht, G., Croft, N., de Ridder, L., Samsom, J., Veereman, G., Neyt, M., Kemos, P., Koletzko, S., Brückner, A., Levine, A., Russell, R., Levine, A., Weiner, D., Griffiths, A., Aloi, M., Raes, J., Christiaens, A., Walters, T., Walker, M., Ruemelle, F., Demange, C.N., Bigot, L., 2022. The Incidence and Characteristics of Venous Thromboembolisms in Paediatric-Onset Inflammatory Bowel Disease: A Prospective International Cohort Study Based on the PIBD-SETQuality Safety Registry. J Crohns Colitis 16. https://doi.org/10.1093/ecco-jcc/jjab171

Abegunde, A.T., Muhammad, B.H., Bhatti, O., Ali, T., 2016. Environmental risk factors for inflammatory bowel diseases: Evidence based literature review. World J Gastroenterol 22, 6296. https://doi.org/10.3748/wjg.v22.i27.6296

Adamiak, T., Walkiewicz-Jedrzejczak, D., Fish, D., Brown, C., Tung, J., Khan, K., Faubion Jr, W., Park, R., Heikenen, J., Yaffee, M., Rivera-Bennett, M.T., Wiedkamp, M., Stephens, M., Noel, R., Nugent, M., Nebel, J., Simpson, P., Kappelman, M.D., Kugathasan, S., 2013. Incidence, clinical characteristics, and natural history of pediatric IBD in Wisconsin: a population-based epidemiological study. Inflamm Bowel Dis 19, 1218–1223. https://doi.org/10.1097/MIB.0b013e318280b13e

Allocca, M., Fiorino, G., Zallot, C., Furfaro, F., Gilardi, D., Radice, S., Danese, S., Peyrin-Biroulet, L., 2020. Incidence and Patterns of COVID-19 Among Inflammatory Bowel Disease Patients From the Nancy and Milan Cohorts. Clinical Gastroenterology and Hepatology 18, 2134–2135. https://doi.org/10.1016/j.cgh.2020.04.071

Ananthakrishnan, A.N., McGinley, E.L., Binion, D.G., Saeian, K., 2011. Ambient air pollution correlates with hospitalizations for inflammatory bowel disease. Inflamm Bowel Dis 17, 1138–1145. https://doi.org/10.1002/ibd.21455

Aniwan, S., Harmsen, W.S., Tremaine, W.J., Loftus, E. v., 2019. Incidence of inflammatory bowel disease by race and ethnicity in a population-based Inception Cohort from 1970 through 2010. Therap Adv Gastroenterol 12, 175628481982769. https://doi.org/10.1177/1756284819827692

Anselin, L., Syabri, I., Kho, Y., 2006. GeoDa: An Introduction to Spatial Data Analysis. Geogr Anal 38, 5–22. https://doi.org/10.1111/j.0016-7363.2005.00671.x

Archimandritis, A.J., Kourtesas, D., Sougioultzis, S., Giontzis, A., Grigoriadis, P., Davaris, P., Tzivras, M., 2002. Inflammatory bowel disease in Greece - A hospital-based clinical study of 172 consecutive patients. Medical Science Monitor 8.

Ashton, J.J., Wiskin, A.E., Ennis, S., Batra, A., Afzal, N.A., Beattie, R.M., 2014. Rising incidence of paediatric inflammatory bowel disease (PIBD) in Wessex, Southern England. Arch Dis Child 99, 659–664. https://doi.org/10.1136/archdischild-2013-305419

Barnes, E.L., Loftus, E. v., Kappelman, M.D., 2021. Effects of Race and Ethnicity on Diagnosis and Management of Inflammatory Bowel Diseases. Gastroenterology 160, 677–689. https://doi.org/10.1053/j.gastro.2020.08.064

Beamish, L.A., Osornio-Vargas, A.R., Wine, E., 2011. Air pollution: An environmental factor contributing to intestinal disease. J Crohns Colitis 5, 279–286. https://doi.org/10.1016/j.crohns.2011.02.017

Beaugerie, L., Langholz, E., Nyboe-Andersen, N., Pigneur, B., Sokol, H., 2018. Differences in epidemiological features between ulcerative colitis and Crohn's disease: The early life-programmed versus late dysbiosis hypothesis. Med Hypotheses 115, 19–21. https://doi.org/10.1016/j.mehy.2018.03.009

Benchimol, E.I., Fortinsky, K.J., Gozdyra, P., van den Heuvel, M., van Limbergen, J., Griffiths, A.M., 2011a. Epidemiology of pediatric inflammatory bowel disease: A systematic review of international trends. Inflamm Bowel Dis 17, 423–439. https://doi.org/10.1002/ibd.21349

Benchimol, E.I., Guttmann, A., Griffiths, A.M., Rabeneck, L., Mack, D.R., Brill, H., Howard, J., Guan, J., To, T., 2009. Increasing incidence of paediatric inflammatory bowel disease in Ontario, Canada: evidence from health administrative data. Gut 58, 1490–1497. https://doi.org/10.1136/gut.2009.188383

Benchimol, E.I., Mack, D.R., Nguyen, G.C., Snapper, S.B., Li, W., Mojaverian, N., Quach, P., Muise, A.M., 2014. Incidence, Outcomes, and Health Services Burden of Very Early Onset Inflammatory Bowel Disease. Gastroenterology 147, 803-813.e7. https://doi.org/10.1053/j.gastro.2014.06.023

Benchimol, E.I., To, T., Griffiths, A.M., Rabeneck, L., Guttmann, A., 2011b. Outcomes of Pediatric Inflammatory Bowel Disease: Socioeconomic Status Disparity in a Universal-Access Healthcare System. J Pediatr 158, 960-967.e4. https://doi.org/10.1016/j.jpeds.2010.11.039

Bequet, E., Sarter, H., Fumery, M., Vasseur, F., Armengol-Debeir, L., Pariente, B., Ley, D., Spyckerelle, C., Coevoet, H., Laberenne, J.E., Peyrin-Biroulet, L., Savoye, G., Turck, D., Gower-Rousseau, C., 2017. Incidence and Phenotype at Diagnosis of Very-early-onset Compared with Later-onset Paediatric Inflammatory Bowel Disease: A Population-based Study [1988-2011. J Crohns Colitis 11. https://doi.org/10.1093/ecco-jcc/jjw194

Bernstein, C.N., Kraut, A., Blanchard, J.F., Rawsthorne, P., Yu, N., Walld, R., 2001. The relationship between inflammatory bowel disease and socioeconomic variables. American Journal of Gastroenterology 96, 2117–2125. https://doi.org/10.1111/j.1572-0241.2001.03946.x

Bian, J., Li, D., Bai, Z., Li, Q., Lyu, D., Zhou, X., 2020. Transport of Asian surface pollutants to the global stratosphere from the Tibetan Plateau region during the Asian summer monsoon. Natl Sci Rev 7, 516–533. https://doi.org/10.1093/nsr/nwaa005

Blanchard, J.F., 2001. Small-area Variations and Sociodemographic Correlates for the Incidence of Crohn's Disease and Ulcerative Colitis. Am J Epidemiol 154, 328–335. https://doi.org/10.1093/aje/154.4.328

Brook, R.D., Rajagopalan, S., Pope, C.A., Brook, J.R., Bhatnagar, A., Diez-Roux, A. v., Holguin, F., Hong, Y., Luepker, R. v., Mittleman, M.A., Peters, A., Siscovick, D., Smith, S.C., Whitsel, L., Kaufman, J.D., 2010. Particulate Matter Air Pollution and Cardiovascular Disease. Circulation 121, 2331–2378. https://doi.org/10.1161/CIR.0b013e3181dbece1

Burghardt, K., Guo, S., Lerman, K., 2022. Unequal impact and spatial aggregation distort COVID-19 growth rates. Philosophical Transactions of the Royal Society A:

Mathematical, Physical and Engineering Sciences 380. https://doi.org/10.1098/rsta.2021.0122

Burisch, J., Pedersen, N., Čuković-Čavka, S., Brinar, M., Kaimakliotis, I., Duricova, D., Shonová, O., Vind, I., Avnstrøm, S., Thorsgaard, N., Andersen, V., Krabbe, S., Dahlerup, J.F., Salupere, R., Nielsen, K.R., Olsen, J., Manninen, P., Collin, P., Tsianos, E. v, Katsanos, K.H., Ladefoged, K., Lakatos, L., Björnsson, E., Ragnarsson, G., Bailey, Y., Odes, S., Schwartz, D., Martinato, M., Lupinacci, G., Milla, M., de Padova, A., D'Incà, R., Beltrami, M., Kupcinskas, L., Kiudelis, G., Turcan, S., Tighineanu, O., Mihu, I., Magro, F., Barros, L.F., Goldis, A., Lazar, D., Belousova, E., Nikulina, I., Hernandez, V., Martinez-Ares, D., Almer, S., Zhulina, Y., Halfvarson, J., Arebi, N., Sebastian, S., Lakatos, P.L., Langholz, E., Munkholm, P., 2014a. East–West gradient in the incidence of inflammatory bowel disease in Europe: the ECCO-EpiCom Inception Cohort. Gut 63, 588–597. https://doi.org/10.1136/gutjnl-2013-304636

Burisch, J., Pedersen, N., Cukovic-Cavka, S., Turk, N., Kaimakliotis, I., Duricova, D., Bortlik, M., Shonová, O., Vind, I., Avnstrøm, S., Thorsgaard, N., Krabbe, S., Andersen, V., Dahlerup, J.F., Kjeldsen, J., Salupere, R., Olsen, J., Nielsen, K.R., Manninen, P., Collin, P., Katsanos, K.H., Tsianos, E. v, Ladefoged, K., Lakatos, L., Ragnarsson, G., Björnsson, E., Bailey, Y., O'Morain, C., Schwartz, D., Odes, S., Giannotta, M., Girardin, G., Kiudelis, G., Kupcinskas, L., Turcan, S., Barros, L., Magro, F., Lazar, D., Goldis, A., Nikulina, I., Belousova, E., Martinez-Ares, D., Hernandez, V., Almer, S., Zhulina, Y., Halfvarson, J., Arebi, N., Tsai, H.H., Sebastian, S., Lakatos, P.L., Langholz, E., Munkholm, P., 2014b. Environmental factors in a population-based Inception Cohort of inflammatory bowel disease patients in Europe — An ECCO-EpiCom study. J Crohns Colitis 8, 607–616. https://doi.org/10.1016/j.crohns.2013.11.021

Carbonnel, F., Jantchou, P., Monnet, E., Cosnes, J., 2009. Environmental risk factors in Crohn's disease and ulcerative colitis: an update. Gastroenterol Clin Biol 33, S145–S157. https://doi.org/10.1016/s0399-8320(09)73150-1

Castro, M., Papadatou, B., Baldassare, M., Balli, F., Barabino, A., Barbera, C., Barca, S., Barera, G., Bascietto, F., Canani, B.R., Calacoci, M., Campanozzi, A., Castellucci, G., Catassi, C., Colombo, M., Covoni, M.R., Cucchiara, S., D'Altilia, M.R., de Angelis, G.L., de Virgilis, S., di Ciommo, V., Fontana, M., Guariso, G., Knafelz, D., Lambertini, A., Licciardi, S., Lionetti, P., Liotta, L., Lombardi, G., Maestri, L., Martelossi, S., Mastella, G., Oderda, G., Perini, R., Pesce, F., Ravelli, A., Roggero, P., Romano, C., Rotolo, N., Rutigliano, V., Scotta, S., Sferlazzas, C., Staiano, A., Ventura, A., Zaniboni, M.G., 2008. Inflammatory bowel disease in children and adolescents in Italy: Data from the pediatric national IBD register (1996–2003). Inflamm Bowel Dis 14, 1246–1252. https://doi.org/10.1002/ibd.20470

Chan, S.S.M., Luben, R., Bergmann, M.M., Boeing, H., Olsen, A., Tjonneland, A., Overvad, K., Kaaks, R., Kennedy, H., Khaw, K.-T., Riboli, E., Hart, A.R., 2011. Aspirin in the aetiology of Crohn's disease and ulcerative colitis: a European prospective cohort study. Alimentary Pharmacology &amp; Therapeutics 34, 649–655. https://doi.org/10.1111/j.1365-2036.2011.04784.x

Chen, S.-Y., Feng, Z., Yi, X., 2017. A general introduction to adjustment for multiple comparisons. J Thorac Dis 9, 1725–1729. https://doi.org/10.21037/jtd.2017.05.34

Chetcuti Zammit, S., Ellul, P., Girardin, G., Valpiani, D., Nielsen, K.R., Olsen, J., Goldis, A., Lazar, D., Shonová, O., Nováková, M., Sebastian, S., Whitehead, E., Carmona, A., Martinez-Cadilla, J., Dahlerup, J.F., Kievit, A.L.H., Thorsgaard, N., Katsanos, K.H., Christodoulou, D.K., Magro, F., Salupere, R., Pedersen, N., Kjeldsen, J., Carlsen, K., Ioannis, K., Bergemalm, D., Halfvarson, J., Duricova, D., Bortlik, M., Collin, P., Oksanen, P., Kiudelis, G., Kupcinskas, L., Kudsk, K., Andersen, V., O'Morain, C., Bailey, Y., Doron, S., Shmuel, O., Almer, S., Arebi, N., Misra, R., Čuković-Čavka, S., Brinar, M., Munkholm, P., Vegh, Z., Burisch, J., 2018. Vitamin D deficiency in a European inflammatory bowel disease Inception Cohort: an Epi-IBD study. European Journal of Gastroenterology &amp; Hepatology 30, 1297–1303. https://doi.org/10.1097/meg.0000000000001238

Cholapranee, A., Ananthakrishnan, A.N., 2016. Environmental Hygiene and Risk of Inflammatory Bowel Diseases. Inflamm Bowel Dis 22, 2191–2199. https://doi.org/10.1097/MIB.0000000000000852

Chowers, Y., Odes, S., Bujanover, Y., Eliakim, R., Bar Meir, S., Avidan, B., 2004. The Month of Birth is Linked to the Risk of Crohn's Disease in the Israeli Population. Am J Gastroenterol 99, 1974–1976. https://doi.org/10.1111/j.1572-0241.2004.40058.x

Coughlan, A., Wylde, R., Lafferty, L., Quinn, S., Broderick, A., Bourke, B., Hussey, S., 2017. A rising incidence and poorer male outcomes characterise early onset paediatric inflammatory bowel disease. Aliment Pharmacol Ther 45, 1534–1541. https://doi.org/10.1111/apt.14070

Crumeyrolle, S., Augustin, P., Rivellini, L.-H., Choël, M., Riffault, V., Deboudt, K., Fourmentin, M., Dieudonné, E., Delbarre, H., Derimian, Y., Chiapello, I., 2019. Aerosol variability induced by atmospheric dynamics in a coastal area of Senegal, North-Western Africa. Atmos Environ 203, 228–241. https://doi.org/10.1016/j.atmosenv.2019.01.041

Day, A.S., Lemberg, D.A., Gearry, R.B., 2014. Inflammatory bowel disease in australasian children and adolescents. Gastroenterol Res Pract 2014, 703890. https://doi.org/10.1155/2014/703890

de Hertogh, G., Geboes, K., 2004. Crohn's disease and infections: a complex relationship. MedGenMed 6, 14.

de Silva, P.S., Yang, X., Korzenik, J.R., Goldman, R.H., Arheart, K.L., Caban-Martinez, A.J., 2017. Association of urinary phenolic compounds, inflammatory bowel disease and chronic diarrheal symptoms: Evidence from the National Health and Nutrition Examination Survey. Environmental Pollution 229, 621–626. https://doi.org/10.1016/j.envpol.2017.06.023

DeBord, D.G., Carreón, T., Lentz, T.J., Middendorf, P.J., Hoover, M.D., Schulte, P.A., 2016. Use of the "Exposome" in the Practice of Epidemiology: A Primer on -Omic Technologies. Am J Epidemiol 184, 302–314. https://doi.org/10.1093/aje/kwv325

Dimakou, K., Pachoula, I., Chouliaras, G., Panayotou, I., Orfanou, I., Lagona, E., Roma-Giannikou, E., 2012. P170 Paediatric inflammatory bowel disease in Greece: 30 years experience of a single center. J Crohns Colitis 6, S77. https://doi.org/10.1016/s1873-9946(12)60190-1

Economou, M., Filis, G., Tsianou, Z., Alamanos, J., Kogevinas, A., Masalas, K., Petrou, A., Tsianos, E. v, 2007. Crohn's disease incidence evolution in North-western Greece is not associated with alteration of NOD2/CARD15 variants. World J Gastroenterol 13, 5116–5120. https://doi.org/10.3748/wjg.v13.i38.5116

el Mouzan, M.I., Saadah, O., Al-Saleem, K., al Edreesi, M., Hasosah, M., Alanazi, A., al Mofarreh, M., Asery, A., al Qourain, A., Nouli, K., al Hussaini, A., Telmesani, A., AlReheili, K., Alghamdi, S., Alrobiaa, N., Alzaben, A., Mehmadi, A., al Hebbi, H., al Sarkhy, A., al Mehaidib, A., al Saleem, B., Assiri, A., Wali, S., 2014. Incidence of Pediatric Inflammatory Bowel Disease in Saudi Arabia. Inflamm Bowel Dis 1. https://doi.org/10.1097/mib.0000000000000048

ESRI 2011. ArcGIS Desktop: Release 10. Redlands, CA: Environmental Systems Research Institute, n.d.

Farrokhyar, F., Swarbrick, E.T., Irvine, E.J., 2001. A Critical Review of Epidemiological Studies in Inflammatory Bowel Disease. Scand J Gastroenterol 36, 2–15. https://doi.org/10.1080/00365520150218002

Felder, J.B., Korelitz, B.I., Rajapakse, R., Schwarz, S., Horatagis, A.P., Gleim, G., 2000. Effects of nonsteroidal antiinflammatory drugs on inflammatory bowel disease: a case-control study. Am J Gastroenterol 95, 1949–1954. https://doi.org/10.1111/j.1572-0241.2000.02262.x

Fernández, A., Hernández, V., Martínez-Ares, D., Sanromán, L., de Castro, M.L., Pineda, J.R., Carmona, A., González-Portela, C., Salgado, C., Martínez-Cadilla, J., Pereira, S., García-Burriel, J.I., Vázquez, S., Rodríguez-Prada, I., 2015. Incidence and phenotype at diagnosis of inflammatory bowel disease. Results in Spain of the EpiCom study. Gastroenterol Hepatol 38, 534–540. https://doi.org/10.1016/j.gastrohep.2015.03.001

Fletcher, J., Cooper, S.C., Ghosh, S., Hewison, M., 2019. The Role of Vitamin D in Inflammatory Bowel Disease: Mechanism to Management. Nutrients 11, 1019. https://doi.org/10.3390/nu11051019

Fuentes, M., Song, H.-R., Ghosh, S.K., Holland, D.M., Davis, J.M., 2006. Spatial Association between Speciated Fine Particles and Mortality. Biometrics 62, 855–863. https://doi.org/10.1111/j.1541-0420.2006.00526.x

García, M.Á., Sánchez, M.L., Pérez, I.A., Torre, B. de, 2008. Continuous Carbon Dioxide Measurements in a Rural Area in the Upper Spanish Plateau. J Air Waste Manage Assoc 58, 940–946. https://doi.org/10.3155/1047-3289.58.7.940

Gearry, R.B., Dodgshun, A.J., 2012. The “hygiene hypothesis” in IBD. J Crohns Colitis 6, 869. https://doi.org/10.1016/j.crohns.2012.04.010

Gearry, R.B., Richardson, A.K., Frampton, C.M., Dodgshun, A.J., Barclay, M.L., 2010. Population-based cases control study of inflammatory bowel disease risk factors. J Gastroenterol Hepatol 25, 325–333. https://doi.org/10.1111/j.1440-1746.2009.06140.x

Ghaly, S., Hart, P.H., Lawrance, I.C., 2019. Inflammatory bowel diseases: interrelationships between dietary vitamin D, exposure to UV radiation and the fecal microbiome. Expert Review of Gastroenterology &amp; Hepatology 13, 1039–1048. https://doi.org/10.1080/17474124.2019.1685874

Gilat, T., Hacohen, D., Lilos, P., Langman, M.J.S., 1987. Childhood Factors in Ulcerative Colitis and Crohn’s Disease: An International Cooperative Study. Scand J Gastroenterol 22, 1009–1024. https://doi.org/10.3109/00365528708991950

Gleeson, M.H., Davis, A.J.M., 2003. Non-steroidal anti-inflammatory drugs, aspirin and newly diagnosed colitis: a case-control study. Alimentary Pharmacology &amp; Therapeutics 17, 817–825. https://doi.org/10.1046/j.1365-2036.2003.01519.x

Goldiș, A., Lupușoru, R., Gheorghe, L., Gheorghe, C., Trifan, A., Dobru, D., Cijevschi, C., Tanțău, A., Constantinescu, G., Iacob, R., Goldiș, R., Diculescu, M., 2019. Geographic Distribution, Phenotype and Epidemiological Tendency in Inflammatory Bowel Disease Patients in Romania. Medicina (Kaunas) 55, 704. https://doi.org/10.3390/medicina55100704

Goshua, A., Akdis, C.A., Nadeau, K.C., 2022. World Health Organization global air quality guideline recommendations: Executive summary. Allergy 77, 1955–1960. https://doi.org/10.1111/all.15224

Grieci, T., Bütter, A., 2009. The incidence of inflammatory bowel disease in the pediatric population of Southwestern Ontario. J Pediatr Surg 44, 977–980. https://doi.org/10.1016/j.jpedsurg.2009.01.038

Gryparis, A., Paciorek, C.J., Zeka, A., Schwartz, J., Coull, B.A., 2009. Measurement error caused by spatial misalignment in environmental epidemiology. Biostatistics 10, 258–274. https://doi.org/10.1093/biostatistics/kxn033

Ha, C.Y., Newberry, R.D., Stone, C.D., Ciorba, M.A., 2010. Patients With Late-Adult-Onset Ulcerative Colitis Have Better Outcomes Than Those With Early Onset Disease. Clinical Gastroenterology and Hepatology 8, 682-687.e1. https://doi.org/10.1016/j.cgh.2010.03.022

Halme, L., 2006. Family and twin studies in inflammatory bowel disease. World J Gastroenterol 12, 3668. https://doi.org/10.3748/wjg.v12.i23.3668

Hammer, T., Nielsen, K.R., Munkholm, P., Burisch, J., Lynge, E., 2016. The Faroese IBD Study: Incidence of Inflammatory Bowel Diseases Across 54 Years of Population-based Data. J Crohns Colitis 10, 934–942. https://doi.org/10.1093/ecco-jcc/jjw050

Hansen, T.S., Jess, T., Vind, I., Elkjaer, M., Nielsen, M.F., Gamborg, M., Munkholm, P., 2011. Environmental factors in inflammatory bowel disease: A case-control study based on a Danish Inception Cohort. J Crohns Colitis 5, 577–584. https://doi.org/10.1016/j.crohns.2011.05.010

Henderson, P., Hansen, R., Cameron, F.L., Gerasimidis, K., Rogers, P., Bisset, M.W., Reynish, E.L., Drummond, H.E., Anderson, N.H., van Limbergen, J., Russell, R.K., Satsangi, J., Wilson, D.C., 2012. Rising incidence of pediatric inflammatory bowel disease in Scotland*. Inflamm Bowel Dis 18, 999–1005. https://doi.org/10.1002/ibd.21797

Hermon-Taylor, J., 1993. Causation of Crohn’s disease: the impact of clusters. Gastroenterology 104, 643–646. https://doi.org/10.1016/0016-5085(93)90438-i

Hodges, P., Kelly, P., 2020. Inflammatory bowel disease in Africa: what is the current state of knowledge? Int Health 12, 222–230. https://doi.org/10.1093/inthealth/ihaa005

Holik, D., Bezdan, A., Marković, M., Orkić, Ž., Milostić-Srb, A., Mikšić, Š., Včev, A., 2020. The Association between Drinking Water Quality and Inflammatory Bowel Disease—A Study in Eastern Croatia. Int J Environ Res Public Health 17, 8495. https://doi.org/10.3390/ijerph17228495

Holmes, E.A., Ponsonby, A.-L., Pezic, A., Ellis, J.A., Kirkwood, C.D., Lucas, R.M., consortium, P., 2019. Higher Sun Exposure is Associated With Lower Risk of Pediatric Inflammatory Bowel Disease: A Matched Case-control Study. J Pediatr Gastroenterol Nutr 69, 182–188. https://doi.org/10.1097/MPG.0000000000002390

Holmes, E.A., Rodney Harris, R.M., Lucas, R.M., 2018. Low Sun Exposure and Vitamin D Deficiency as Risk Factors for Inflammatory Bowel Disease, With a Focus on Childhood Onset. Photochem Photobiol 95, 105–118. https://doi.org/10.1111/php.13007

Hong, S.J., Cho, S.M., Choe, B.-H., Jang, H.J., Choi, K.H., Kang, B., Kim, J.E., Hwang, J.H., 2018. Characteristics and Incidence Trends for Pediatric Inflammatory Bowel Disease in Daegu-Kyungpook Province in Korea: a Multi-Center Study. J Korean Med Sci 33, e132–e132. https://doi.org/10.3346/jkms.2018.33.e132

Hope, B., Shahdadpuri, R., Dunne, C., Broderick, A.M., Grant, T., Hamzawi, M., O’Driscoll, K., Quinn, S., Hussey, S., Bourke, B., 2012. Rapid rise in incidence of Irish paediatric inflammatory bowel disease. Arch Dis Child 97, 590–594. https://doi.org/10.1136/archdischild-2011-300651

Huang, J.G., Aw, M.M., 2020. Pediatric Inflammatory Bowel Disease in Asia: Epidemiology and natural history. Pediatrics &amp; Neonatology 61, 263–271. https://doi.org/10.1016/j.pedneo.2019.12.008

Hund, L., Chen, J.T., Krieger, N., Coull, B.A., 2012. A geostatistical approach to large-scale disease mapping with temporal misalignment. Biometrics 68, 849–858. https://doi.org/10.1111/j.1541-0420.2011.01721.x

Iwata, A., Fujioka, K., Yonemichi, T., Fukagata, K., Kurosawa, K., Tabata, R., Kitagawa, M., Takashima, T., Okuda, T., 2019. Seasonal variation in atmospheric particle electrostatic charging states determined using a parallel electrode plate device. Atmos Environ 203, 62–69. https://doi.org/10.1016/j.atmosenv.2019.01.040

Jabandziev, P., Pinkasova, T., Kunovsky, L., Papez, J., Jouza, M., Karlinova, B., Novackova, M., Urik, M., Aulicka, S., Slaby, O., Bohosova, J., Bajerova, K., Bajer, M., Goel, A., 2020. Regional Incidence of Inflammatory Bowel Disease in a Czech Pediatric Population: 16 Years of Experience (2002-2017). J Pediatr Gastroenterol Nutr 70, 586–592. https://doi.org/10.1097/MPG.0000000000002660

Jakobsen, C., Paerregaard, A., Munkholm, P., Faerk, J., Lange, A., Andersen, J., Jakobsen, M., Kramer, I., Czernia-Mazurkiewicz, J., Wewer, V., 2011. Pediatric inflammatory bowel disease: Increasing incidence, decreasing surgery rate, and compromised nutritional status: A prospective population-based cohort study 2007–2009. Inflamm Bowel Dis 17, 2541–2550. https://doi.org/10.1002/ibd.21654

Jandarov, R.A., Sheppard, L.A., Sampson, P.D., Szpiro, A.A., 2017. A novel principal component analysis for spatially misaligned multivariate air pollution data. J R Stat Soc Ser C Appl Stat 66, 3–28. https://doi.org/10.1111/rssc.12148

Jeffery, C., Ozonoff, A., Pagano, M., 2014. The effect of spatial aggregation on performance when mapping a risk of disease. Int J Health Geogr 13, 9. https://doi.org/10.1186/1476-072X-13-9

Johnston, R.D., Logan, R.F.A., 2008. What is the peak age for onset of IBD? Inflamm Bowel Dis 14, S4–S5. https://doi.org/10.1002/ibd.20545

Joossens, M., Joossens, S., van Steen, K., Pierik, M., Vermeire, S., Rutgeerts, P., van Ranst, M., 2005. Crohn’s Disease and Month of Birth. Inflamm Bowel Dis 11, 597–599. https://doi.org/10.1097/01.mib.0000163697.34592.d4

Kaplan, G.G., Bernstein, C.N., Coward, S., Bitton, A., Murthy, S.K., Nguyen, G.C., Lee, K., Cooke-Lauder, J., Benchimol, E.I., 2019. The Impact of Inflammatory Bowel Disease in Canada 2018: Epidemiology. J Can Assoc Gastroenterol 2, S6–S16. https://doi.org/10.1093/jcag/gwy054

Kaplan, G.G., Hubbard, J., Korzenik, J., Sands, B.E., Panaccione, R., Ghosh, S., Wheeler, A.J., Villeneuve, P.J., 2010. The inflammatory bowel diseases and ambient air pollution: a novel association. Am J Gastroenterol 105, 2412–2419. https://doi.org/10.1038/ajg.2010.252

Kaplan, G.G., Windsor, J.W., 2021. The four epidemiological stages in the global evolution of inflammatory bowel disease. Nat Rev Gastroenterol Hepatol 18, 56–66. https://doi.org/10.1038/s41575-020-00360-x

Karolewska-Bochenek, K., Lazowska-Przeorek, I., Albrecht, P., Grzybowska, K., Ryzko, J., Szamotulska, K., Radzikowski, A., Landowski, P., Krzesiek, E., Ignys, I., Fyderek, K., Czerwionka-Szaflarska, M., Jarocka-Cyrta, E., 2009. Epidemiology of Inflammatory Bowel Disease among Children in Poland. Digestion 79, 121–129. https://doi.org/10.1159/000209382

Khalif, I.L., Shapina, M. v, 2017. Inflammatory bowel disease treatment in Eastern Europe. Curr Opin Gastroenterol 33, 230–233. https://doi.org/10.1097/mog.0000000000000370

Koloski, N.-A., Bret, L., Radford-Smith, G., 2008. Hygiene hypothesis in inflammatory bowel disease: a critical review of the literature. World J Gastroenterol 14, 165–173. https://doi.org/10.3748/wjg.14.165

Lehtinen, P., Ashorn, M., Iltanen, S., Jauhola, R., Jauhonen, P., Kolho, K.-L., Auvinen, A., 2011. Incidence trends of pediatric inflammatory bowel disease in Finland, 1987–2003, a nationwide study. Inflamm Bowel Dis 17, 1778–1783. https://doi.org/10.1002/ibd.21550

Lehtinen, P., Pasanen, K., Kolho, K.-L., Auvinen, A., 2016. Incidence of Pediatric Inflammatory Bowel Disease in Finland. Journal of Pediatric Gastroenterology &amp; Nutrition 63, 65–70. https://doi.org/10.1097/mpg.0000000000001050

Li, S., Lv, Z., 2021. Do spatial spillovers matter? Estimating the impact of tourism development on CO2 emissions. Environmental Science and Pollution Research 28, 32777–32794. https://doi.org/10.1007/s11356-021-12988-6

Liang, D., Kumar, N., 2013. Time-space Kriging to address the spatiotemporal misalignment in the large datasets. Atmos Environ (1994) 72, 60–69. https://doi.org/10.1016/j.atmosenv.2013.02.034

Limketkai, B.N., Bayless, T.M., Brant, S.R., Hutfless, S.M., 2014. Lower regional and temporal ultraviolet exposure is associated with increased rates and severity of inflammatory bowel disease hospitalisation. Alimentary Pharmacology &amp; Therapeutics n/a-n/a. https://doi.org/10.1111/apt.12845

Loddo, I., Romano, C., 2015. Inflammatory Bowel Disease: Genetics, Epigenetics, and Pathogenesis. Front Immunol 6. https://doi.org/10.3389/fimmu.2015.00551

Lopez, R.N., Appleton, L., Gearry, R.B., Day, A.S., 2018. Rising Incidence of Paediatric Inflammatory Bowel Disease in Canterbury, New Zealand, 1996–2015. Journal of Pediatric Gastroenterology &amp; Nutrition 66, e45–e50. https://doi.org/10.1097/mpg.0000000000001688

Lovasz, B.D., Lakatos, L., Horvath, A., Pandur, T., Erdelyi, Z., Balogh, M., Szipocs, I., Vegh, Z., Veres, G., Müller, K.E., Golovics, P.A., Kiss, L.S., Mandel, M.D., Lakatos, P.L., 2014. Incidence rates and disease course of paediatric inflammatory bowel diseases in Western Hungary between 1977 and 2011. Digestive and Liver Disease 46, 405–411. https://doi.org/10.1016/j.dld.2013.12.013

Lu, C., Yang, J., Yu, W., Li, D., Xiang, Z., Lin, Y., Yu, C., 2015. Association between 25(OH)D Level, Ultraviolet Exposure, Geographical Location, and Inflammatory Bowel Disease Activity: A Systematic Review and Meta-Analysis. PLoS One 10, e0132036–e0132036. https://doi.org/10.1371/journal.pone.0132036

Ludvigsson, J.F., Büsch, K., Olén, O., Askling, J., Smedby, K.E., Ekbom, A., Lindberg, E., Neovius, M., 2017. Prevalence of paediatric inflammatory bowel disease in Sweden: a nationwide population-based register study. BMC Gastroenterol 17, 23. https://doi.org/10.1186/s12876-017-0578-9

Lv, W., Wu, Y., Zang, J., 2021. A Review on the Dispersion and Distribution Characteristics of Pollutants in Street Canyons and Improvement Measures. Energies (Basel) 14, 6155. https://doi.org/10.3390/en14196155

Maaser, C., Langholz, E., Gordon, H., Burisch, J., Ellul, P., Hernández Ramirez, V., Karakan, T., Katsanos, K.H., Krustins, E., Levine, A., Mantzaris, G.J., O’Morain, C., Saritas Yuksel, E., Strid, H., Annese, V., 2016. European Crohn’s and Colitis Organisation Topical Review on environmental factors in IBD. J Crohns Colitis jjw223. https://doi.org/10.1093/ecco-jcc/jjw223

Malmborg, P., Grahnquist, L., Lindholm, J., Montgomery, S., Hildebrand, H., 2013. Increasing Incidence of Paediatric Inflammatory Bowel Disease in Northern Stockholm County, 2002–2007. Journal of Pediatric Gastroenterology &amp; Nutrition 57, 29–34. https://doi.org/10.1097/mpg.0b013e31828f21b4

Martín-de-Carpi, J., Rodríguez, A., Ramos, E., Jiménez, S., Martínez-Gómez, M.J., Medina, E., 2013. Increasing Incidence of Pediatric Inflammatory Bowel Disease in Spain (1996–2009). Inflamm Bowel Dis 19, 73–80. https://doi.org/10.1002/ibd.22980

Martín-de-Carpi, J., Rodríguez, A., Ramos, E., Jiménez, S., Martínez-Gómez, M.J., Medina, E., Navas-López, V.M., 2014. The complete picture of changing pediatric inflammatory bowel disease incidence in Spain in 25years (1985–2009): The EXPERIENCE registry. J Crohns Colitis 8, 763–769. https://doi.org/10.1016/j.crohns.2014.01.005

Mavropoulou, E., Ternes, K., Mechie, N.-C., Bremer, S.C.B., Kunsch, S., Ellenrieder, V., Neesse, A., Amanzada, A., 2019. Cytomegalovirus colitis in inflammatory bowel disease and after haematopoietic stem cell transplantation: diagnostic accuracy, predictors, risk factors and disease outcome. BMJ Open Gastroenterol 6, e000258. https://doi.org/10.1136/bmjgast-2018-000258

Minitab, LLC, 2021. Minitab, Available at: https://www.minitab.com., n.d.

Misra, R., Limdi, J., Cooney, R., Sakuma, S., Brookes, M., Fogden, E., Pattni, S., Sharma, N., Iqbal, T., Munkholm, P., Burisch, J., Arebi, N., 2019. Ethnic differences in inflammatory bowel disease: Results from the United Kingdom inception epidemiology study. World J Gastroenterol 25, 6145–6157. https://doi.org/10.3748/wjg.v25.i40.6145

Nerich, V., Jantchou, P., Boutron-Ruault, M.-C., Monnet, E., Weill, A., Vanbockstael, V., Auleley, G.-R., Balaire, C., Dubost, P., Rican, S., Allemand, H., Carbonnel, F., 2011. Low exposure to sunlight is a risk factor for Crohn’s disease. Alimentary Pharmacology &amp; Therapeutics 33, 940–945. https://doi.org/10.1111/j.1365-2036.2011.04601.x

Ng, S.C., Shi, H.Y., Hamidi, N., Underwood, F.E., Tang, W., Benchimol, E.I., Panaccione, R., Ghosh, S., Wu, J.C.Y., Chan, F.K.L., Sung, J.J.Y., Kaplan, G.G., 2017. Worldwide incidence and prevalence of inflammatory bowel disease in the 21st century: a systematic review of population-based studies. The Lancet 390, 2769–2778. https://doi.org/10.1016/s0140-6736(17)32448-0

Ntirampeba, D., Neema, I., Kazembe, L., 2018. Modelling spatio-temporal patterns of disease for spatially misaligned data: An application on measles incidence data in Namibia from 2005-2014. PLoS One 13, e0201700–e0201700. https://doi.org/10.1371/journal.pone.0201700

Olmedo-Martín, R.V., González-Molero, I., Olveira, G., Amo-Trillo, V., Jiménez-Pérez, M., 2019. Sunlight exposure in inflammatory bowel disease outpatients: predictive factors and correlation with serum vitamin D. Gastroenterología y Hepatología (English Edition) 42, 604–613. https://doi.org/10.1016/j.gastre.2019.07.002

Ong, C., Aw, M.M., Liwanag, M.J., Quak, S.H., Phua, K.B., 2018. Rapid rise in the incidence and clinical characteristics of pediatric inflammatory bowel disease in a South-East Asian cohort in Singapore, 1994-2015. J Dig Dis 19, 395–403. https://doi.org/10.1111/1751-2980.12641

Opstelten, J.L., Beelen, R.M.J., Leenders, M., Hoek, G., Brunekreef, B., van Schaik, F.D.M., Siersema, P.D., Eriksen, K.T., Raaschou-Nielsen, O., Tjønneland, A., Overvad, K., Boutron-Ruault, M.-C., Carbonnel, F., de Hoogh, K., Key, T.J., Luben, R., Chan, S.S.M., Hart, A.R., Bueno-de-Mesquita, H.B., Oldenburg, B., 2016. Exposure to Ambient Air Pollution and the Risk of Inflammatory Bowel Disease: A European Nested Case-Control Study. Dig Dis Sci 61, 2963–2971. https://doi.org/10.1007/s10620-016-4249-4

Orel, R., Kamhi, T., Vidmar, G., Mamula, P., 2009. Epidemiology of Pediatric Chronic Inflammatory Bowel Disease in Central and Western Slovenia, 1994–2005. Journal of Pediatric Gastroenterology &amp; Nutrition 48, 579–586. https://doi.org/10.1097/mpg.0b013e318164d903

Pacifici, K., Reich, B.J., Miller, D.A.W., Pease, B.S., 2019. Resolving misaligned spatial data with integrated species distribution models. Ecology 100, e02709–e02709. https://doi.org/10.1002/ecy.2709

Piovani, D., Danese, S., Peyrin-Biroulet, L., Bonovas, S., 2019. Environmental, Nutritional, and Socioeconomic Determinants of IBD Incidence: A Global Ecological Study. J Crohns Colitis 14, 323–331. https://doi.org/10.1093/ecco-jcc/jjz150

Podolsky, D.K., 2002. Inflammatory Bowel Disease. New England Journal of Medicine 347, 417–429. https://doi.org/10.1056/NEJMra020831

QGIS.org, %Y. QGIS Geographic Information System. QGIS Association. http://www.qgis.org, n.d.

Qin, X., 2011. What made Canada become a country with the highest incidence of inflammatory bowel disease: could sucralose be the culprit? Can J Gastroenterol 25, 511. https://doi.org/10.1155/2011/451036

Quigley, E.M., 2012. Epigenetics: filling in the “heritability gap” and identifying gene-environment interactions in ulcerative colitis. Genome Med 4, 72. https://doi.org/10.1186/gm373

R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.). [WWW Document], n.d.

Roquette, R., Nunes, B., Painho, M., 2018. The relevance of spatial aggregation level and of applied methods in the analysis of geographical distribution of cancer mortality in mainland Portugal (2009–2013). Popul Health Metr 16, 6. https://doi.org/10.1186/s12963-018-0164-6

Saidel-Odes, L., Odes, S., 2014. Hygiene hypothesis in inflammatory bowel disease, Annals of Gastroenterology.

Salim, S.Y., Kaplan, G.G., Madsen, K.L., 2014. Air pollution effects on the gut microbiota: a link between exposure and inflammatory disease. Gut Microbes 5, 215–219. https://doi.org/10.4161/gmic.27251

Salupere, R., 2001. Inflammatory bowel disease in Estonia: a prospective epidemiologic study 1993-1998. World J Gastroenterol 7, 387–388. https://doi.org/10.3748/wjg.v7.i3.387

Schwarz, J., Sýkora, J., Cvalínová, D., Pomahačová, R., Klečková, J., Kryl, M., Včelák, P., 2017. Inflammatory bowel disease incidence in Czech children: A regional prospective study, 2000-2015. World J Gastroenterol 23, 4090–4101. https://doi.org/10.3748/wjg.v23.i22.4090

Selvaratnam, S., Gullino, S., Shim, L., Lee, E., Lee, A., Paramsothy, S., Leong, R.W., 2019. Epidemiology of inflammatory bowel disease in South America: A systematic review. World J Gastroenterol 25, 6866–6875. https://doi.org/10.3748/wjg.v25.i47.6866

Shaw, S.Y., Blanchard, J.F., Bernstein, C.N., 2015. Early Childhood Measles Vaccinations are not Associated with Paediatric IBD: A Population-based Analysis. J Crohns Colitis 9, 334–338. https://doi.org/10.1093/ecco-jcc/jjv029

Shaw, S.Y., Blanchard, J.F., Bernstein, C.N., 2011. Association Between the Use of Antibiotics and New Diagnoses of Crohn’s Disease and Ulcerative Colitis. American Journal of Gastroenterology 106, 2133–2142. https://doi.org/10.1038/ajg.2011.304

Sobotka, T., Skirbekk, V., Philipov, D., 2011. Economic Recession and Fertility in the Developed World. Popul Dev Rev 37, 267–306. https://doi.org/10.1111/j.1728-4457.2011.00411.x

Soon, I.S., Molodecky, N.A., Rabi, D.M., Ghali, W.A., Barkema, H.W., Kaplan, G.G., 2012. The relationship between urban environment and the inflammatory bowel diseases: a systematic review and meta-analysis. BMC Gastroenterol 12, 51. https://doi.org/10.1186/1471-230X-12-51

Sumetsky, N., Mair, C., Anderson, S., Gruenewald, P.J., 2020. A spatial partial differential equation approach to addressing unit misalignments in Bayesian poisson space-time models. Spat Spatiotemporal Epidemiol 33, 100337. https://doi.org/10.1016/j.sste.2020.100337

Sýkora, J., Pomahačová, R., Kreslová, M., Cvalínová, D., Štych, P., Schwarz, J., 2018. Current global trends in the incidence of pediatric-onset inflammatory bowel disease. World J Gastroenterol 24, 2741–2763. https://doi.org/10.3748/wjg.v24.i25.2741

Tarris, G., de Rougemont, A., Charkaoui, M., Michiels, C., Martin, L., Belliot, G., 2021. Enteric Viruses and Inflammatory Bowel Disease. Viruses 13. https://doi.org/10.3390/v13010104

Thia, K.T., Loftus Edward V., J., Sandborn, W.J., Yang, S.-K., 2008. An Update on the Epidemiology of Inflammatory Bowel Disease in Asia. Am J Gastroenterol 103, 3167–3182. https://doi.org/10.1111/j.1572-0241.2008.02158.x

Tobler, W., 2004. On the First Law of Geography: A Reply. Annals of the Association of American Geographers 94, 304–310. https://doi.org/10.1111/j.1467-8306.2004.09402009.x

Tobler, W.R., 1959. Automation and Cartography. Geogr Rev 49, 526. https://doi.org/10.2307/212211

Tsianos, E. v, Masalas, C.N., Merkouropoulos, M., Dalekos, G.N., Logan, R.F., 1994. Incidence of inflammatory bowel disease in north west Greece: rarity of Crohn’s disease in an area where ulcerative colitis is common. Gut 35, 369–372. https://doi.org/10.1136/gut.35.3.369

Ungaro, F., Massimino, L., D’Alessio, S., Danese, S., 2019. The gut virome in inflammatory bowel disease pathogenesis: From metagenomics to novel therapeutic approaches. United European Gastroenterol J 7, 999–1007. https://doi.org/10.1177/2050640619876787

Ungaro, R., Bernstein, C.N., Gearry, R., Hviid, A., Kolho, K.-L., Kronman, M.P., Shaw, S., van Kruiningen, H., Colombel, J.-F., Atreja, A., 2014. Antibiotics Associated With Increased Risk of New-Onset Crohn’s Disease But Not Ulcerative Colitis: A Meta-Analysis. American Journal of Gastroenterology 109, 1728–1738. https://doi.org/10.1038/ajg.2014.246

Ungaro, R.C., Agrawal, M., Brenner, E.J., Zhang, X., Colombel, J.-F., Kappelman, M.D., Reinisch, W., 2022. New Gastrointestinal Symptoms Are Common in Inflammatory Bowel Disease Patients With COVID-19: Data From an International Registry. Inflamm Bowel Dis 28, 314–317. https://doi.org/10.1093/ibd/izab184

Ungaro, R.C., Kappelman, M.D., Rubin, D.T., Colombel, J.-F., 2021. COVID-19 and Inflammatory Bowel Disease: Lessons Learned, Practical Recommendations, and Unanswered Questions. Gastroenterology 160, 1447–1451. https://doi.org/10.1053/j.gastro.2020.12.042

Urlep, D., Blagus, R., Orel, R., 2015. Incidence Trends and Geographical Variability of Pediatric Inflammatory Bowel Disease in Slovenia: A Nationwide Study. Biomed Res Int 2015, 921730. https://doi.org/10.1155/2015/921730

Urlep, D., Trop, T.K., Blagus, R., Orel, R., 2014. Incidence and Phenotypic Characteristics of Pediatric IBD in Northeastern Slovenia, 2002–2010. Journal of Pediatric Gastroenterology &amp; Nutrition 58, 325–332. https://doi.org/10.1097/mpg.0000000000000207

Utazi, C.E., Thorley, J., Alegana, V.A., Ferrari, M.J., Nilsen, K., Takahashi, S., Metcalf, C., Lessler, J., Tatem, A.J., 2019. A spatial regression model for the disaggregation of areal unit based data to high-resolution grids with application to vaccination coverage mapping. Stat Methods Med Res 28, 3226–3241. https://doi.org/10.1177/0962280218797362

van der Zaag-Loonen, H.J., Casparie, M., Taminiau, J.A.J.M., Escher, J.C., Pereira, R.R., Derkx, H.H.F., 2004. The Incidence of Pediatric Inflammatory Bowel Disease in the Netherlands: 1999–2001. J Pediatr Gastroenterol Nutr 38, 302–307. https://doi.org/10.1097/00005176-200403000-00014

van Kruiningen, H.J., Freda, B.J., 2001. A Clustering of Crohn’s Disease in Mankato, Minnesota. Inflammatory Bowel Disease 7, 27–33. https://doi.org/10.1097/00054725-200102000-00004

Vanhaecke, T., Bretin, O., Poirel, M., Tap, J., 2022. Drinking Water Source and Intake Are Associated with Distinct Gut Microbiota Signatures in US and UK Populations. J Nutr 152, 171–182. https://doi.org/10.1093/jn/nxab312

Virta, L., Auvinen, A., Helenius, H., Huovinen, P., Kolho, K.-L., 2012. Association of Repeated Exposure to Antibiotics With the Development of Pediatric Crohn’s Disease--A Nationwide, Register-based Finnish Case-Control Study. Am J Epidemiol 175, 775–784. https://doi.org/10.1093/aje/kwr400

Virta, L.J., Saarinen, M.M., Kolho, K.-L., 2016a. Inflammatory Bowel Disease Incidence is on the Continuous Rise Among All Paediatric Patients Except for the Very Young: A Nationwide Registry-based Study on 28-Year Follow-up. J Crohns Colitis 11, 150–156. https://doi.org/10.1093/ecco-jcc/jjw148

Virta, L.J., Saarinen, M.M., Kolho, K.-L., 2016b. Inflammatory Bowel Disease Incidence is on the Continuous Rise Among All Paediatric Patients Except for the Very Young: A Nationwide Registry-based Study on 28-Year Follow-up. J Crohns Colitis 11, 150–156. https://doi.org/10.1093/ecco-jcc/jjw148

Vrijheid, M., 2014. The exposome: a new paradigm to study the impact of environment on health. Thorax 69, 876–878. https://doi.org/10.1136/thoraxjnl-2013-204949

Wild, C.P., 2012. The exposome: from concept to utility. Int J Epidemiol 41, 24–32. https://doi.org/10.1093/ije/dyr236

Wittig, R., Albers, L., Koletzko, S., Saam, J., von Kries, R., 2019. Pediatric Chronic Inflammatory Bowel Disease in a German Statutory Health INSURANCE—Incidence Rates From 2009 to 2012. Journal of Pediatric Gastroenterology &amp; Nutrition 68, 244–250. https://doi.org/10.1097/mpg.0000000000002162

Zhang, L., Wang, Y., Liu, Y., Li, Z., Li, X., 2019. Variation of platinum group elements (PGE) in airborne particulate matter (PM2.5) in the Beijing urban area, China: A case study of the 2014 APEC summit. Atmos Environ 198, 70–76. https://doi.org/10.1016/j.atmosenv.2018.10.044

Zhang, X.-S., 2015. Strain Interactions as a Mechanism for Dominant Strain Alternation and Incidence Oscillation in Infectious Diseases: Seasonal Influenza as a Case Study. PLoS One 10, e0142170–e0142170. https://doi.org/10.1371/journal.pone.0142170

Zhang, Z., Manjourides, J., Cohen, T., Hu, Y., Jiang, Q., 2016. Spatial measurement errors in the field of spatial epidemiology. Int J Health Geogr 15, 21. https://doi.org/10.1186/s12942-016-0049-5

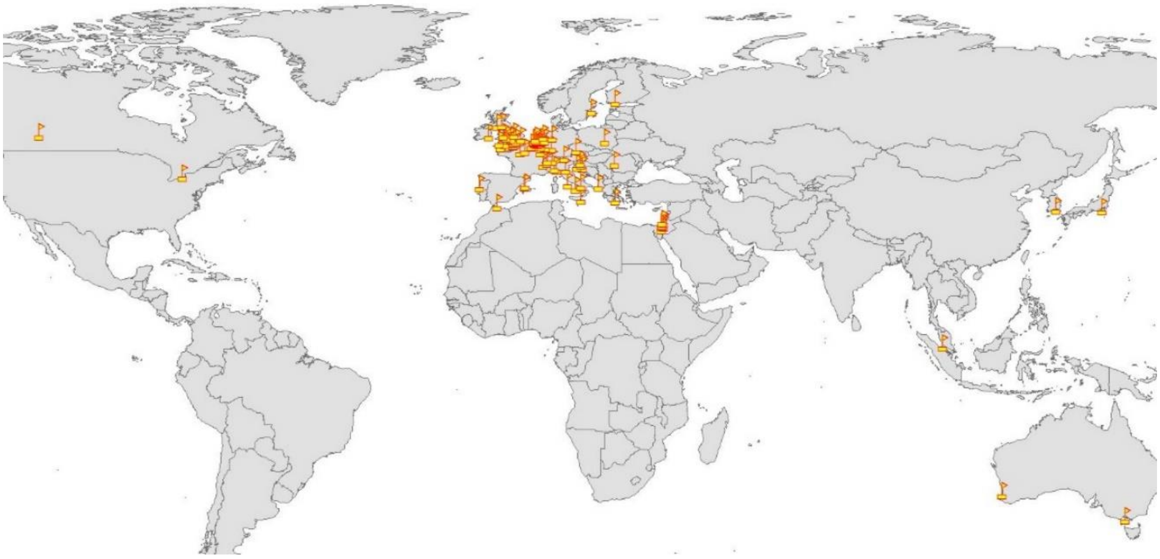
Jairath, V., & Feagan, B. G., 2020. Global burden of inflammatory bowel disease The Lancet Gastroenterology &amp; Hepatology 5, 1. https://doi.org/10.1016/s2468-1253(19)30358-9

Appendix

The PIBD safety registry relies on a network of 150+ currently active participants with specialisation in paediatric Gastroenterology. We have established an online registry of rare and serious complications of paediatric IBD by sending a monthly E-card to experts like yourself and collating all responses to a dedicated database.

Safety registry additional information

Appendix Figure 1



A world map featuring the locations of all reporting PIBD experts illustrates the expansive reach of our safety registry. Initially, the registry aimed to gather data on the incidence of rare and severe complications associated with PIBD. To estimate this incidence, it was necessary to collect information on the number of patients each PIBD expert was managing at the time they completed the electronic surveys; this figure served as the denominator for calculating the incidence of these rare complications. Building upon this objective, we extended our data collection to include information about the catchment areas covered by the PIBD experts. This additional data helps us estimate the total population under 18 years of age that each reporting centre is responsible for, which in turn serves as the denominator for calculating both the incidence and prevalence of PIBD.

Within the registry, whenever a complication is reported, we also gather follow-up data that aids in identifying significant risk factors and potentially refining the management of these cases. Although we have documented 150 complications to date, our goal is to collect even more comprehensive data, enabling us to conduct detailed analyses of each complication type.

Inception cohort protocol synopsis

PHASE:	**IV WITH NO DIRECT BENEFIT**
DESCRIPTION:	**European prospective inception cohort and safety registry:** • a registry will be specifically designed to analyse effectiveness and safety signals of current treatment strategies in routine practice and to correlate them to individual risk factors • in combination with a safety registry, incidence and prevalence of severe and rare complications of the disease will be estimated • this will ultimately lead to improvement of treatment algorithms and paediatric IBD patient outcomes
STUDY POPULATION :	Children with newly diagnosed IBD (age 0-17 years). Patients included in the inception cohort will be followed up until 20 years from inclusion. To identify patients with rare and serious complications of IBD the safety registry will prospectively identify these conditions through both the inception cohort and the wider European networks of paediatric gastroenterologists (PIBD-NET and PEDDCReN).
INCLUSION:	Inception cohort: – New patients, 0-17 years of age, with a confirmed diagnosis of IBD (Crohn's disease, Ulcerative colitis, IBD-Unclassified) within 2 months of inclusion based on history, physical examination, laboratory, endoscopic, radiological and histological features according to the revised Porto criteria – Informed consent of patient (if indicated) and parents has been obtained – Concerning the patients of whom biological specimens will be included: patients should nor have started IBD treatment yet Safety registry: Any child with IBD <19 years old with complications as detailed in the safety monitoring list (or future updates of the list of conditions) can be reported. For more detailed phenotyping including patient identifiable information and for collection and analysis of biological specimens such as DNA consent will be required.

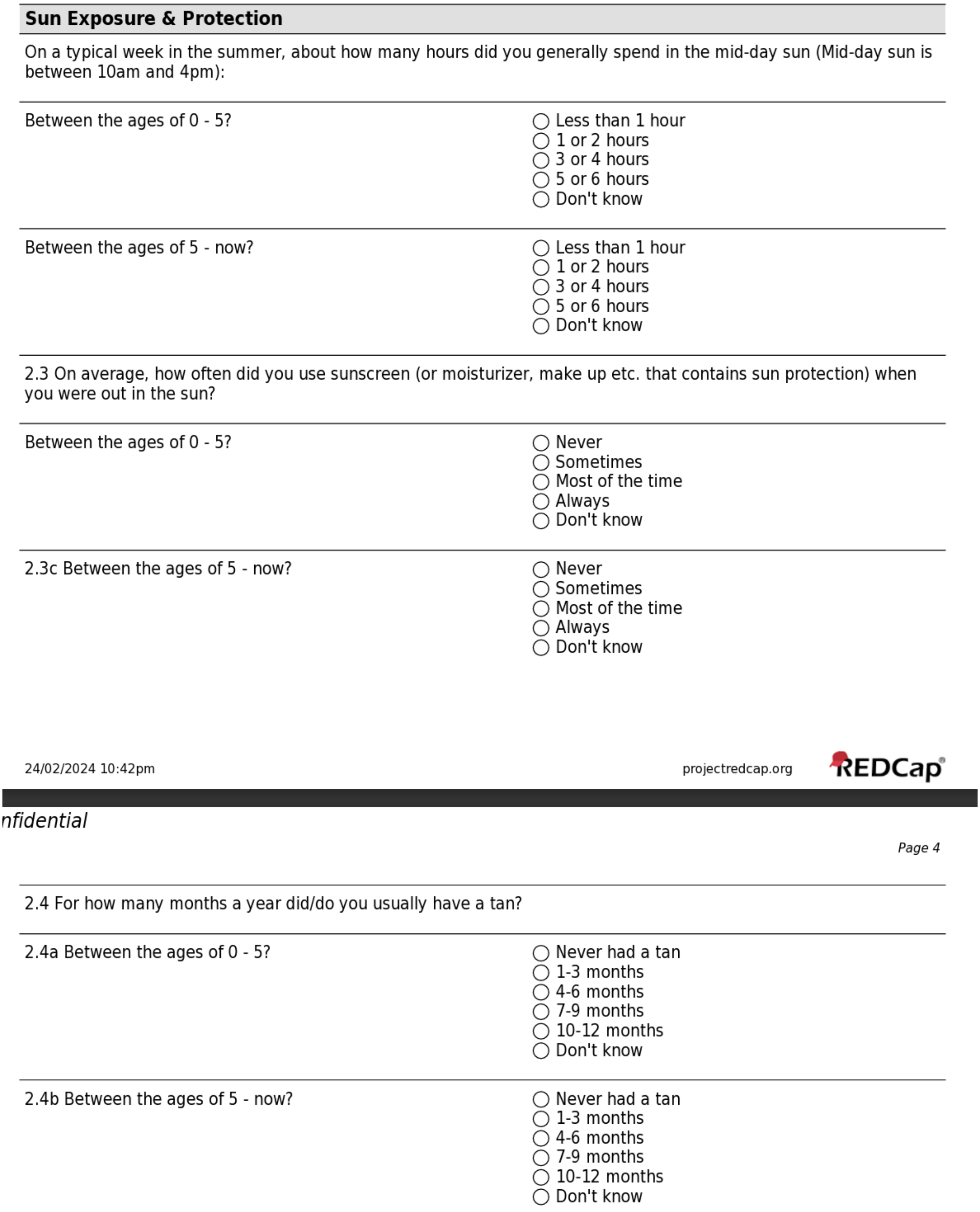
EXCLUSION	– Inability to read and understand the patient and family information sheets – Informed consent of patient or parents has not been obtained when required – Patients on similar treatments as for IBD but for other conditions, or known with conditions directly affecting the IBD

PRIMARY OBJECTIVE	The primary objective of the PIBD-NET inception cohort is to search for predictive factors for outcome, specific serious adverse events (SAEs) and for predictors factors for therapy outcomes.
SECONDARY OBJECTIVES	The secondary objective is the identification of rare complications of disease or treatment in paediatric IBD patients. This will be performed by: – Establishment a pan-European monitoring system for identifying patients with rare and serious complications of PIBD and its treatments – Identifying and characterising patient's clinical phenotype with these complications with the aim of being able to better predict and therefore prevent the complications – looking for immunological and/or genetic predictors/risk factors of these complications.

STUDY DESIGN:	The PIBD-Net inception cohort and safety registry (WP7) is an observational study supported by the European H2020 program. Inception cohort: An observational registry including a subcohort of patients, in which biological specimen will be collected, will be set up and collection of safety signaling on a wide scale will be performed. A robust and highly secured prospective multicenter long term database tool for PIBD will be created. A total of 1000 patients Children (age 0-18 years) with new-onset IBD will be included during a 3-year period. Per year 200 CD patients and 100 UC patients will be included. Moreover, within these three years, in specific centers able to perform these immunological techniques, 150 children (age 0-17 years), with new- onset IBD will be included for collection of biological specimens. Patients will be closely monitored for disease progression during preferably twenty years of follow up to examine effectiveness, identify treatment or disease related risks as well as complications related to disease progression. It will also allow to collect longitudinal data on psychosocial outcomes and health-related costs. Safety registry: A pan-European safety registry of rare complications of drugs and the disease will be created. Investigators will monthly be requested to identify patients with rare and serious complications according to the safety monitoring list. **Nature and extent of the burden and risks associated with participation, benefit and group relatedness:** Both the inception cohort study and the safety registry do not bring any risks for the patients. The burden is considered minimal. Since disease phenotype, course of the disease and benefits and risks of treatment differ between children and adults, this study cannot be performed in adult patients.
RECRUITEMENT	**36 Months**
STUDY DURATION	**Total Duration of the study: 4 years** **20 years of follow up**

Environmental questionnaire examples

Appendix Figure 2 Inception cohort environmental questionnaire samples



Sun Exposure & Protection

On a typical week in the summer, about how many hours did you generally spend in the mid-day sun (Mid-day sun is between 10am and 4pm):

Between the ages of 0 - 5?

○ Less than 1 hour
○ 1 or 2 hours
○ 3 or 4 hours
○ 5 or 6 hours
○ Don't know

Between the ages of 5 - now?

○ Less than 1 hour
○ 1 or 2 hours
○ 3 or 4 hours
○ 5 or 6 hours
○ Don't know

2.3 On average, how often did you use sunscreen (or moisturizer, make up etc. that contains sun protection) when you were out in the sun?

Between the ages of 0 - 5?

○ Never
○ Sometimes
○ Most of the time
○ Always
○ Don't know

2.3c Between the ages of 5 - now?

○ Never
○ Sometimes
○ Most of the time
○ Always
○ Don't know

24/02/2024 10:42pm projectredcap.org REDCap

nfidential

Page 4

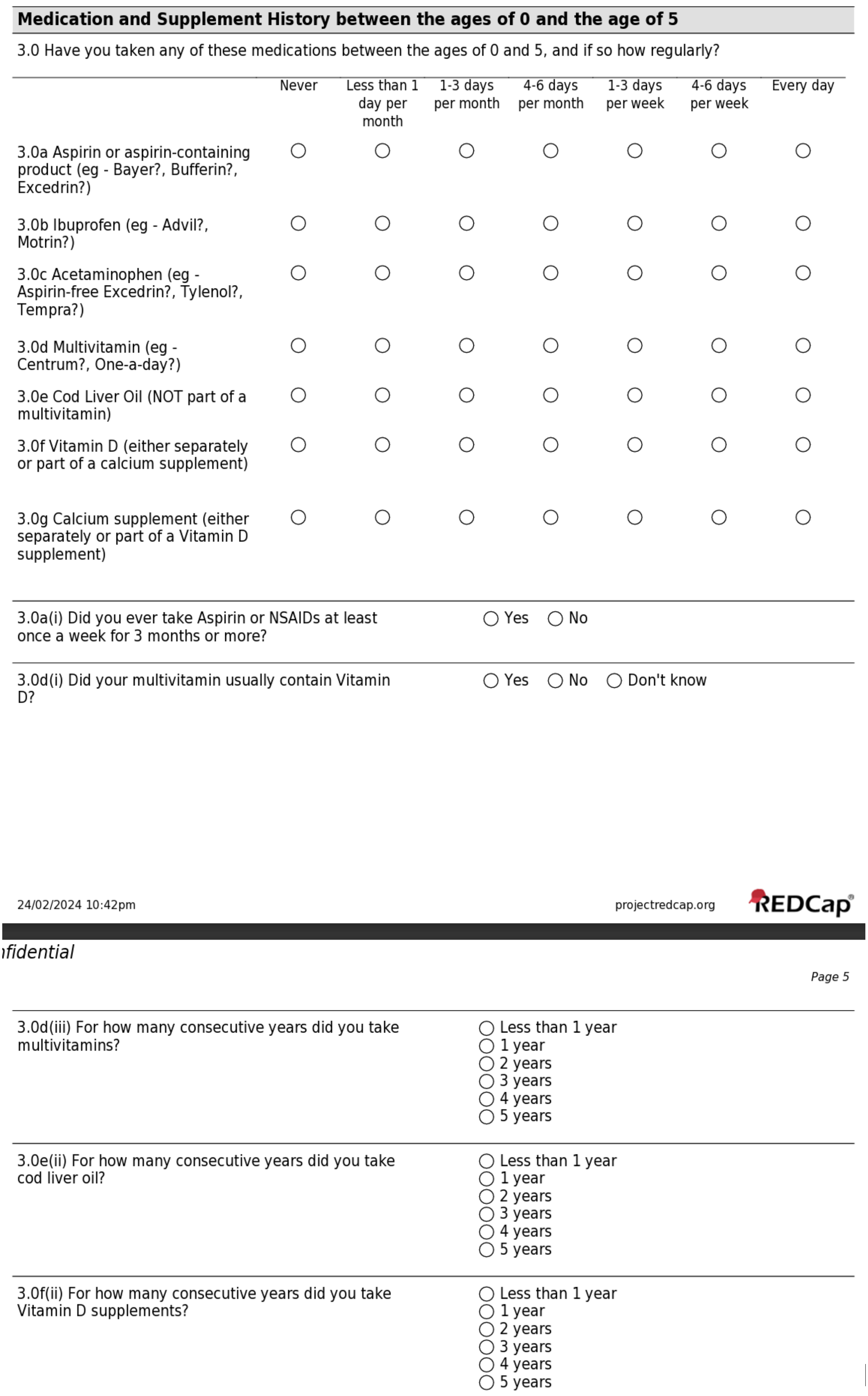
2.4 For how many months a year did/do you usually have a tan?

2.4a Between the ages of 0 - 5?

○ Never had a tan
○ 1-3 months
○ 4-6 months
○ 7-9 months
○ 10-12 months
○ Don't know

2.4b Between the ages of 5 - now?

○ Never had a tan
○ 1-3 months
○ 4-6 months
○ 7-9 months
○ 10-12 months
○ Don't know



Medication and Supplement History between the ages of 0 and the age of 5

3.0 Have you taken any of these medications between the ages of 0 and 5, and if so how regularly?

	Never	Less than 1 day per month	1-3 days per month	4-6 days per month	1-3 days per week	4-6 days per week	Every day
3.0a Aspirin or aspirin-containing product (eg - Bayer?, Bufferin?, Excedrin?)	○	○	○	○	○	○	○
3.0b Ibuprofen (eg - Advil?, Motrin?)	○	○	○	○	○	○	○
3.0c Acetaminophen (eg - Aspirin-free Excedrin?, Tylenol?, Tempra?)	○	○	○	○	○	○	○
3.0d Multivitamin (eg - Centrum?, One-a-day?)	○	○	○	○	○	○	○
3.0e Cod Liver Oil (NOT part of a multivitamin)	○	○	○	○	○	○	○
3.0f Vitamin D (either separately or part of a calcium supplement)	○	○	○	○	○	○	○
3.0g Calcium supplement (either separately or part of a Vitamin D supplement)	○	○	○	○	○	○	○

3.0a(i) Did you ever take Aspirin or NSAIDs at least once a week for 3 months or more? ○ Yes ○ No

3.0d(i) Did your multivitamin usually contain Vitamin D? ○ Yes ○ No ○ Don't know

24/02/2024 10:42pm projectredcap.org

nfidential

Page 5

3.0d(iii) For how many consecutive years did you take multivitamins?

○ Less than 1 year
○ 1 year
○ 2 years
○ 3 years
○ 4 years
○ 5 years

3.0e(ii) For how many consecutive years did you take cod liver oil?

○ Less than 1 year
○ 1 year
○ 2 years
○ 3 years
○ 4 years
○ 5 years

3.0f(ii) For how many consecutive years did you take Vitamin D supplements?

○ Less than 1 year
○ 1 year
○ 2 years
○ 3 years
○ 4 years
○ 5 years

4.1g Currently, the mother smokes cigarettes every day, some days, or not at all?

○ Every day
○ Some days
○ Not at all
○ Don't know
○ Refused

4.1h Has the mother EVER smoked cigarettes EVERY DAY for at least 6 months?

○ Yes ○ No ○ Don't know
○ Refused

4.1i How old was the mother when first started smoking cigarettes every day?

____
(answer in years)

4.1k On the average, about how many cigarettes does the mother currently smoke each day?

○ 0 - 5
○ 5 - 10
○ 10 - 20
○ >20
○ Don't know
○ Refused

4.1l On how many of the past 30 days did the mother smoke cigarettes?

○ 0
○ 1
○ 2
○ 3
○ 4
○ 5
○ 6
○ 7
○ 8
○ 9
○ 10
○ 11
○ 12
○ 13
○ 14
○ 15
○ 16
○ 17
○ 18
○ 19
○ 20
○ 21
○ 22
○ 23
○ 24
○ 25
○ 26
○ 27
○ 28
○ 29
○ 30
○ Don't know
○ Refused

4.1m On the average, on those days, how many cigarettes did the mother usually smoke each day?

○ 0 - 5
○ 5 - 10
○ 10 - 20
○ >20
○ Don't know
○ Refused