Report No. 496

epa Research

INHALE: Irish Nationwide Health and Air Quality Linkage

Authors: Éilis J. O'Reilly, Claire Buckley, Christina Dillon and Stig Hellebust

Lead organisation: University College Cork

www.epa.ie

Rialtas na hÉireann
Government of Ireland

Environmental Protection Agency

The EPA is responsible for protecting and improving the environment as a valuable asset for the people of Ireland. We are committed to protecting people and the environment from the harmful effects of radiation and pollution.

The work of the EPA can be divided into three main areas:

Regulation: Implementing regulation and environmental compliance systems to deliver good environmental outcomes and target those who don't comply.

Knowledge: Providing high quality, targeted and timely environmental data, information and assessment to inform decision making.

Advocacy: Working with others to advocate for a clean, productive and well protected environment and for sustainable environmental practices.

Our Responsibilities Include:

Licensing

- Large-scale industrial, waste and petrol storage activities;
- Urban waste water discharges;
- The contained use and controlled release of Genetically Modified Organisms;
- Sources of ionising radiation;
- Greenhouse gas emissions from industry and aviation through the EU Emissions Trading Scheme.

National Environmental Enforcement

- Audit and inspection of EPA licensed facilities;
- Drive the implementation of best practice in regulated activities and facilities;
- Oversee local authority responsibilities for environmental protection;
- Regulate the quality of public drinking water and enforce urban waste water discharge authorisations;
- Assess and report on public and private drinking water quality;
- Coordinate a network of public service organisations to support action against environmental crime;
- Prosecute those who flout environmental law and damage the environment.

Waste Management and Chemicals in the Environment

- Implement and enforce waste regulations including national enforcement issues;
- Prepare and publish national waste statistics and the National Hazardous Waste Management Plan;
- Develop and implement the National Waste Prevention Programme;
- Implement and report on legislation on the control of chemicals in the environment.

Water Management

- Engage with national and regional governance and operational structures to implement the Water Framework Directive;
- Monitor, assess and report on the quality of rivers, lakes, transitional and coastal waters, bathing waters and groundwaters, and measurement of water levels and river flows.

Climate Science & Climate Change

- Publish Ireland's greenhouse gas emission inventories and projections;
- Provide the Secretariat to the Climate Change Advisory Council and support to the National Dialogue on Climate Action;
- Support National, EU and UN Climate Science and Policy development activities.

Environmental Monitoring & Assessment

- Design and implement national environmental monitoring systems: technology, data management, analysis and forecasting;
- Produce the State of Ireland's Environment and Indicator Reports;
- Monitor air quality and implement the EU Clean Air for Europe Directive, the Convention on Long Range Transboundary Air Pollution, and the National Emissions Ceiling Directive;
- Oversee the implementation of the Environmental Noise Directive;
- Assess the impact of proposed plans and programmes on the Irish environment.

Environmental Research and Development

- Coordinate and fund national environmental research activity to identify pressures, inform policy and provide solutions;
- Collaborate with national and EU environmental research activity.

Radiological Protection

- Monitoring radiation levels and assess public exposure to ionising radiation and electromagnetic fields;
- Assist in developing national plans for emergencies arising from nuclear accidents;
- Monitor developments abroad relating to nuclear installations and radiological safety;
- Provide, or oversee the provision of, specialist radiation protection services.

Guidance, Awareness Raising, and Accessible Information

- Provide independent evidence-based reporting, advice and guidance to Government, industry and the public on environmental and radiological protection topics;
- Promote the link between health and wellbeing, the economy and a clean environment;
- Promote environmental awareness including supporting behaviours for resource efficiency and climate transition;
- Promote radon testing in homes and workplaces and encourage remediation where necessary.

Partnership and Networking

- Work with international and national agencies, regional and local authorities, non-governmental organisations, representative bodies and government departments to deliver environmental and radiological protection, research coordination and science-based decision making.

Management and Structure of the EPA

The EPA is managed by a full time Board, consisting of a Director General and five Directors. The work is carried out across five Offices:

1. Office of Environmental Sustainability
2. Office of Environmental Enforcement
3. Office of Evidence and Assessment
4. Office of Radiation Protection and Environmental Monitoring
5. Office of Communications and Corporate Services

The EPA is assisted by advisory committees who meet regularly to discuss issues of concern and provide advice to the Board.

INHALE: Irish Nationwide Health and Air Quality Linkage

Authors: Éilis J. O'Reilly, Claire Buckley, Christina Dillon and Stig Hellebust

Lead organisation: University College Cork

What did this research aim to address?

Every day, every hour, air quality is monitored across Ireland. At the same time, thousands of routine health and social care interactions take place, each generating data that could be used to advance a programme of air quality and health research in Ireland. Safeguarding patient and service user privacy is the foremost concern, but barriers to accessing data prevent meaningful research on the health effects of pollution.

What this means in practice is that researchers with air quality data, curated and made accessible by the EPA, cannot link the relevant pollution exposure to specific individuals in health data. To protect privacy, health data are made available on aggregated units (hospital attended or county; monthly counts). Air quality varies substantially within these aggregations; therefore, this level of aggregation does not let us establish who is exposed to higher levels of air pollution.

The Irish Nationwide Health and Air Quality Linkage (INHALE) project reviewed the methods of linking air quality and health data in other jurisdictions and recommended ways to create capacity for world-leading environmental research in the Irish population.

What did this research find?

The INHALE project identified best-practice and novel approaches to data linkage to support research and policy-driven change. The report details two main methods for safely linking air quality and health data. The first recommendation is investment in trusted research environment (TRE) infrastructure that can fulfil requests to create linked, de-identified individual-level data for researchers or government departments through a secure platform.

In the absence of a TRE, the INHALE project recommends a second researcher-led solution where data custodians provide aggregated data on a unit relevant to the research question. First, the researcher assigns the relevant (for the condition being studied) air quality data to every residential Eircode (e.g. deciles of annual particulate matter of less than 10 micrometres or peak daily nitrogen dioxide) using EPA data. The researcher provides this file to the health data custodian to merge with the health data. Each Eircode (or other appropriate location identifier) occurring in the health data is assigned the corresponding level of pollution. The health data custodian then returns only the aggregated counts of individuals with the condition for each level of pollution.

How can the research findings be used?

The infrastructure and the data linkage expertise required to implement the recommendations provided by the INHALE project are already in place throughout Ireland, specifically within the Central Statistics Office (CSO), as is required by legislation. However, the CSO is not currently tasked with implementing this process, and in order to operationalise and implement the recommendations, additional human resources would need to be allocated to take ownership of the process.

The INHALE team strongly recommends that the EPA is at the table for future discussions of TREs and data linkage because of the organisation's wealth of experience with big data, and because air quality, environment and climate are going to remain leading causes of morbidity and mortality for the foreseeable future.

Project code: 2017-CCRP-MS.47

EPA RESEARCH PROGRAMME 2021–2030

INHALE: Irish Nationwide Health and Air Quality Linkage

(2017-CCRP-MS.47)

EPA Research Report

Prepared for the Environmental Protection Agency

by

University College Cork

Authors:

Éilis J. O'Reilly, Claire Buckley, Christina Dillon and Stig Hellebust

ENVIRONMENTAL PROTECTION AGENCY
An Ghníomhaireacht um Chaomhnú Comhshaoil
PO Box 3000, Johnstown Castle, Co. Wexford, Ireland

Telephone: +353 53 916 0600 Fax: +353 53 916 0699
Email: info@epa.ie Website: www.epa.ie

ACKNOWLEDGEMENTS

This report is published as part of the EPA Research Programme 2021–2030. The EPA Research Programme is a Government of Ireland initiative funded by the Department of the Environment, Climate and Communications. It is administered by the Environmental Protection Agency, which has the statutory function of co-ordinating and promoting environmental research. This project is co-funded by the Environmental Protection Agency and the Health Service Executive.

The authors would like to acknowledge the members of the project steering committee, namely Gerry Brady (Central Statistics Office), Professor Luke Clancy (TobaccoFree Research Institute Ireland), Martin Fitzpatrick (Dublin City Council), Pat Kenny (EPA), Dr Heather Walton (Imperial College London), Micheal Young (Department of the Environment, Climate and Communications) and Dr Matt Robinson (Health Service Executive). In particular, the team wishes to thank the project managers, Lisa Johnson and Siobhán Murphy, for their ongoing help and support.

DISCLAIMER

This report is based on research carried out/data from October 2018 to April 2022. More recent data may have become available since the research was completed.

The EPA Research Programme addresses the need for research in Ireland to inform policymakers and other stakeholders on a range of questions in relation to environmental protection. These reports are intended as contributions to the necessary debate on the protection of the environment.

EPA RESEARCH PROGRAMME 2021–2030
Published by the Environmental Protection Agency, Ireland

ISBN: 978-1-80009-316-4 October 2025

Price: Free Online version

Project Partners

Dr Éilis J O'Reilly
School of Public Health
University College Cork
Cork
Ireland
Tel.: +353 21 420 5529
Email: eilis.oreilly@ucc.ie

Dr Stig Hellebust
School of Chemistry
University College Cork
Cork
Ireland
Tel.: +353 21 490 2680
Email: s.hellebust@ucc.ie

Dr Christina Dillon
School of Public Health (formerly)
University College Cork
Cork
Ireland

Dr Claire Buckley
Health Services in South Lee, Cork South
Health Service Executive
Cork
Ireland
Tel.: +353 21 420 5529
Email: Claire.buckley2@hse.ie

Prof. Ivan Perry
Professor Emeritus, School of Public Health
University College Cork
Cork
Ireland
Tel.: +353 21 420 5529
Email: i.perry@ucc.ie

Greg Kelly
School of Public Health (formerly)
University College Cork
Cork
Ireland

Contents

List of Figures

Executive Summary

Every day, every hour air quality is monitored across Ireland. At the same time, thousands of routine health and social care interactions take place, each generating data that could be used to advance a programme of air quality and health research in Ireland.

Barriers prevent meaningful, best practice pollution and health research. Safeguarding the privacy of the patient or service user is the foremost concern. What this means in practice is that researchers with air quality data, curated and made accessible by the EPA, cannot link the relevant exposure to specific individuals in health data because their names, addresses and other identifiers are (rightly) never revealed.

In this project, best practice for air quality data and health data linkage in other jurisdictions was reviewed and compared with the infrastructure in Ireland. Within Ireland, data collection and storage is advancing rapidly towards fully digital processes, including the continued roll-out of unique health identifiers for health-to-health data linkage. Against this backdrop, the Irish Nationwide Health and Air Quality Linkage (INHALE) project has proposed mechanisms to facilitate linkage of the two data domains. One recommendation is the further prioritisation of digital health and ongoing investment in trusted research environment infrastructure. In the meantime, the team proposes a workable researcher-led solution whereby the researcher provides an aggregation scheme to the data holder, rather than the data holder providing data aggregated on the basis of its customary units of county or service division.

Lastly, the team stresses that the EPA should be at the table for future discussions of investment in, and structuring of, a trusted research environment for data linkage in Ireland because of the organisation's wealth of experience with big data, and because air quality, environmental exposures and climate are going to remain leading causes of morbidity and mortality for the foreseeable future.

1 Introduction

1.1 Funding Call and Project Details

The Irish Nationwide Health and Air Quality Linkage (INHALE) project is a research project funded by the EPA that evaluated methods of linking air quality monitoring data and health data in order to extend the capacity for extensive and ongoing environmental, epidemiological research in the Irish population.

The project identified a low-cost method of linking data and recommends improvements to the existing data infrastructure to enable effective use of health data in research to advance the understanding of the deleterious impact of poor air quality and to facilitate evidence-based policy decisions.

This project was part of the EPA 2017 research call, which highlighted that (i) there are no recent studies published linking Irish health surveillance data to air quality data, (ii) sound Irish data and statistics outlining the health impacts of elevated pollutant levels on the Irish population can assist in policy formation and drive positive change for human health and the environment, and (iii) Ireland needs *a system for collecting and collating* the relevant human health and air quality data in *a manner that allows sensible interrogation so that links between human health and air quality can be identified*.

This report presents the findings of the INHALE research project.

1.2 Background and Objectives

1.2.1 Background and rationale

According to the World Health Organization (WHO), outdoor air pollution is "the single biggest environmental health risk", accounting for almost 4.2 million premature deaths in 2019 (WHO, 2024). Within the EU 27 Member States, it is estimated that almost 400,000 premature deaths were attributable to air pollution in 2019 (EEA, 2021), while lost days at work cost €15 billion, health care costs are estimated at €4 billion and total external costs could be as large as €330 billion (EEA, 2016). Globally, heart disease and stroke account for most premature deaths due to air pollution, followed by lung disease and lung cancer (WHO, 2016). In 2013, the International Agency for Research on Cancer officially classified general air pollution as carcinogenic to humans, with particulate matter (PM) causally associated with the increase in lung and bladder cancer incidence (IARC, 2013). While it is accepted that air pollution contributes significantly to the burden of cardiovascular, cancer and respiratory morbidity worldwide, recent studies have linked air pollution with a range of health outcomes, including hypertension (Fuks *et al.*, 2016; Qin *et al.*, 2021), restricted fetal growth (Pedersen *et al.*, 2013; Fu *et al.*, 2019), pre-term birth (Bekkar *et al.*, 2020), infertility (Mahalingaiah *et al.*, 2016; Vizcaíno *et al.*, 2016) and autism spectrum disorder (Roberts *et al.*, 2013; Xiang *et al*., 2025).

Airborne PM is a mix of pollutants in the form of particles of varying size and chemical composition that contribute to poor air quality. The European Environment Agency estimates that in Ireland in 2019, over 1600 deaths were attributable to all-source $PM_{2.5}$ (EEA, 2021). EU legal limit values set out in the Clean Air for Europe (CAFE) Directive (Directive 2008/50/EC) were not exceeded in Ireland in 2021. However, in the same year approximately 75% of stations measuring $PM_{2.5}$ recorded exceedances of WHO's newly revised daily and annual guidelines (EPA, 2021; WHO, 2021). Across Europe, exposure to particles with an aerodynamic diameter of less than 2.5 micrometres ($PM_{2.5}$) that originates only from wood and coal for residential heating contributed to an estimated 61,000 deaths (Chafe *et al*., 2015). There is no safe threshold for PM exposure; consequently, there is no exposure level at which it will not cause human health impacts. In addition, the varying composition of PM may induce specific health effects. Therefore, it is of policy and scientific interest to assess the health impact of Irish air quality.

Agencies in the environmental and health arenas have long collected data: the outcome of this project adds immediate value to those data collection systems through the provision of a researcher-led linkage method for aggregation of sensitive data. In addition, the ability of the existing network stations

to capture geographical variation in air pollution and their suitability for monitoring population exposure to emissions was evaluated. Longer-term recommendations include solutions similar to those suggested in a 2016 Health Research Board (HRB) report on data sharing (Moran, 2016): specifically, a comprehensive infrastructure for data linkage needs to be established.

1.2.2 Objectives and desired outputs

The project objectives were:

- Identify the relevant available data and statistics for assessing the health impact of air pollutants on the Irish population and highlight data gaps. Perform a capacity analysis of current systems for air quality measurement and routine health information systems in Ireland for assessing the health impact of air quality.
- Review international best practice and assess the feasibility of linking the data domains without compromising patient confidentiality based on best practice.
- Determine *how best to collect and collate the relevant health and air quality data* and *define infrastructural requirements* for developing appropriate systems in Ireland with the capacity to interrogate routine health service utilisation data and *identify linkages* between human health and air quality. In other words, identify a workable system for collecting, collating and interrogating air quality and human health data to support both epidemiological studies and health impact assessment of recent events, and to support integrated workflow for scheduled publications.
- Provide recommendations for health-related and other relevant national agencies on how to develop a comprehensive operational system to link health and air quality data, in line with international best practice.

The INHALE project set out to identify and recommend improvements to the *existing data infrastructure* to enable *effective use of a routine air monitoring and health data in research.* In the following sections, we will address how and to what extent the project has identified workable solutions.

2 Air Quality and Health Data Collection in Ireland

How to measure and use air quality for health studies is discussed. Air and health data sources are reviewed and gaps in data collection are highlighted. The feasibility of linking current data collections is reviewed.

2.1 Air Quality and Health Outcomes

Using air quality data for the prediction of health outcomes can be complex. The earliest studies of air quality and health or mortality were based on aggregate data. These study designs, known as ecological studies in epidemiology, generally followed one of two main approaches. One approach compared health data across geographical regions that had sustained and identifiable differences in measured pollution levels. The other approach involved a comparison of the numbers of health outcomes or deaths in a particular location during periods of acute worsening pollution with the numbers from when the air was cleaner.

An underlying assumption in early ecological studies of air quality and health was that everyone living in the regions studied was exposed to the pollutants in the air to a similar degree. This assumption could be considered to hold reasonably well when regional contrasts in air quality were significant. However, air quality continues to improve, and therefore the potential contrasts are smaller. To reveal subtle aetiological associations, more accurate measurement or estimation of individual exposure is required. (One approach is to adapt the assumption of similar exposure to people living within similar location *types* rather than within the same geographical boundaries.) Nevertheless, these early studies were the impetus for important future work in environmental epidemiology, including a drive to better understand and capture the variable nature of individual exposure.

As the early ecological studies found, in general, individuals living in an area with poor air quality are more likely to be exposed to high levels of air pollutants than individuals who live in an area of good air quality. However, the spatiotemporal variations can be considerable and the resultant measurement error in exposure assessment can be a source of bias that could impact the investigator's ability to find true associations between a pollutant and a health outcome. In short, measurements are made in fixed locations and not per individual. Further complexities arise when the "relevant window of exposure" is unknown. The relevant window of exposure is that time when being exposed to a pollutant confers the most risk of a particular outcome. In parallel to the construct of relevant window of exposure is dose and length of exposure. In other words, one health outcome (including premature mortality) may result from many years of slightly elevated exposure, while another may be the result of days or even hours of considerably elevated exposure. Lastly, in particular for PM, source and composition varies from one location to another and within a location over time. Epidemiological studies that differentiate between sources of pollutants will have immediate policy implications.

To address the causal question "Is exposure to air pollution deleterious to health?", aspects of exposure measurement need to be addressed, including:

- How do relevant pollutant properties vary in space and time between sources and in ambient air?
- What are the implications of these variations for individual exposure?
- What advances have been made in understanding the relationships between level of exposure, both spatially and temporally, and estimates of dose that are linked to health outcomes?

Such questions have been explored by way of direct measurements, source apportionment models and exposure assessment models but require further interrogation in the Irish population.

Personal exposure to air pollutants can be measured directly or estimated using statistical methods based on measured ambient concentrations combined with estimates of time spent commuting, outdoors or engaging in physical activity and air inhalation statistics, etc. However, the relationship between the

ambient concentration of a pollutant and the personal dose is not easily quantified. Studies of personal exposure to $PM_{2.5}$ have been performed in Dublin using portable personal PM monitors (McNabola *et al.*, 2008). In an EPA-funded study (Broderick *et al.*, 2015), the uptake of PM during various activities was estimated using an adaptation of the International Commission on Radiological Protection's human respiratory tract model. The model, its adaptation and application are described in full in McNabola *et al.* (2008), and were used to convert personal exposure concentrations ($\mu g/m^3$) in each micro-environment into uptake (μg) by assigning respiratory rates to different levels of physical exertion along with information on the time spent in particular micro-environments for each sampling period. The model also noted variations in uptake according to the subject's gender, age, height and weight.

As mentioned above, particle compositions and concentrations are extremely variable and several assumptions are needed to model personal exposure; these depend on many factors, such as local emission sources, meteorology or topography, and exhibit diurnal and seasonal changes. This inherent variation may in part explain the heterogeneity observed in the toxicological responses induced by ambient air particulates from different sites and seasons (Becker *et al.*, 2005; Hetland *et al.*, 2005; Kavouras and Chalbot, 2017) and why some studies find associations with a particular outcome while others find weaker or no associations. Source- or site-specific and seasonal variations in ambient PM pollution may confer different health risks, and therefore it is imperative to begin to fully evaluate the health effects of air pollution specific to Ireland.

One of the objectives of the INHALE project was to document the depth and breadth of pollutant data available to predict a wide range of health outcomes.

2.2 Air Quality Monitoring in Ireland

Collection of air quality data is performed by the EPA through the national Ambient Air Quality Monitoring Programme (AAMP), comprising a network of 115 stations across the country (as of September 2024). These data are collected under the CAFE Directive for the purpose of protection of public health and the environment and compliance with the legal thresholds for ambient air pollutants. In this capacity, the air quality monitoring network by design monitors different types of environments, defined as *rural, rural background, urban and suburban,* with further local distinctions including regional background, roadside and urban background. These distinctions are necessary because it can be expected that the air quality in and around Dublin's Pearse Street, for example, will be fundamentally and consistently different from, say, Galway's Mace Head. In the context of assessing population exposure, the legislation places more emphasis on monitoring the air quality in areas of higher population density, resulting in more monitoring stations in Dublin city alone than in some counties combined.

The EPA publishes and makes available to researchers validated air quality data, and provides real-time monitoring station readings at https://airquality.ie/readings. Extensive historical data are available, dating back to 1990. Although the current network is growing, data collection capacity is limited by coverage and the resources needed to operate monitoring stations, and the siting of monitoring stations is based on criteria given in the CAFE Directive, with a view to maximum representation of population exposure (Figure 2.1).

As well as fixed monitoring stations, several air quality indicators or proxies are available in the form of administrative data that relate to emissions of air pollutants, such as total energy consumption, gas and coal consumption, vehicle use and traffic and transport data, which can be leveraged in public health studies. For example, proximity to a road has been a reliable indicator of individual exposure across several health outcomes (e.g. low birth weight (Dadvand *et al.*, 2014) and childhood respiratory (Freid *et al.*, 2021) and neurological diseases (Yuchi *et al.*, 2020)).

While the air quality network data are readily available and clearly defined, secondary data sources and proxy data streams have not yet been fully exploited for research purposes in Ireland.

2.2.1 Assessing the air quality monitoring network for research, 1988–2017

The current national air quality network had its beginning in 1988 and had grown to 29 active stations by 2017. During over 36 years of active monitoring

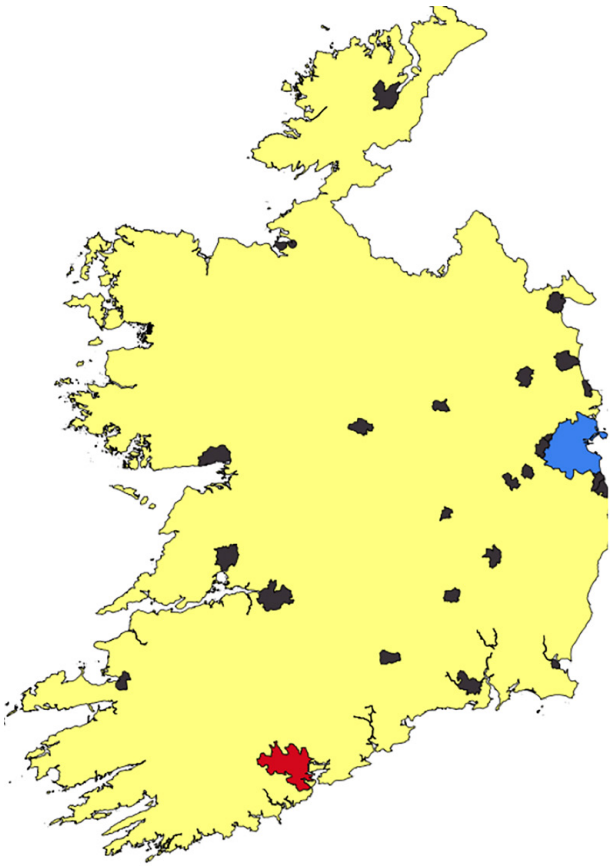

Figure 2.1. Air quality zones in Ireland. Zone A (blue) indicates Dublin, zone B (red) Cork conurbation, zone C (black) other large urban areas (Limerick, Galway, Waterford, Drogheda, Dundalk, Bray, Navan, Ennis, Tralee, Kilkenny, Carlow, Naas, Sligo, Newbridge, Mullingar, Wexford, Letterkenny, Athlone, Celbridge, Clonmel, Balbriggan, Greystones, Leixlip and Portlaoise) and zone D (yellow) rural Ireland (https://gis.epa.ie/EPAMaps/). The location of the monitoring stations in the network is based on the legal requirements set out in the CAFE Directive, which are intended to ensure that the monitoring provides the appropriate coverage to assess population exposure to air pollution.

by the EPA some monitoring locations have been launched and others discontinued. Since 2017, the network has grown rapidly and as of September 2024 includes 115 monitoring locations.

Since the inception of the network, hourly values have been reported for gaseous species. Hourly data have been reported for ozone, nitrogen oxides (NO_x), carbon monoxide (CO) and sulfur dioxide (SO_2) since 2018, whereas particles with an aerodynamic diameter of less than 10 micrometres (PM_{10}) were traditionally reported as 24-hour average values by collection of airborne material on filters followed by gravimetric determination of collected mass. With the addition of automated reference grade instrumentation, automated recording of hourly values of $PM_{2.5}$ and PM_{10} has become available as the technology came on stream. Since the roll-out of AAMP in 2017, more and more locations now include hourly measurements of $PM_{2.5}$, PM_{10}, nitrogen dioxide (NO_2) and ozone, particularly in locations newer to the network (https://airquality.ie/).

Monitoring data are made available through the EPA's Secure Archive for Environmental Research Data (SAFER-Data) (https://eparesearch.epa.ie/safer/) database, where they can be downloaded as separate datasets per individual pollutant, per location, per year.

Air quality can fluctuate substantially within and across days, months and years. Similarly, the aetiological relationship between air quality and health outcomes will vary: acute peak air pollution may be the measure that confers the risk of one outcome, while another health outcome is the result of *in utero* or childhood chronic exposure. For these reasons, assessing and documenting the breadth and depth of coverage matters greatly in determining the capacity of the current routine data collection for epidemiological studies. An initial task of the INHALE project was to collate, harmonise and merge all individual pollutant datasets (6000+ files), which was necessary for assessing coverage, documenting the variation in air quality across the country and across location types and making the historical data easily accessible to researchers.

In this step, air quality data were collated for the years 1988 to 2017 and organised by pollutant and parameter type (24-hour average, hourly, peak, etc.). Pollutant data files can be readily merged by date and location to facilitate two or more pollutant models of health outcomes. An overview of levels and trends is given in Figures 2.2–2.6 for each of the pollutants measured.

A total of 72 monitoring locations have been active for all or part of this period (1988–2017). The most monitored of the regulated pollutants are NO_x, ozone and PM (PM_{10} and $PM_{2.5}$), and, less frequently, SO_2 and CO. Generally, very few locations have

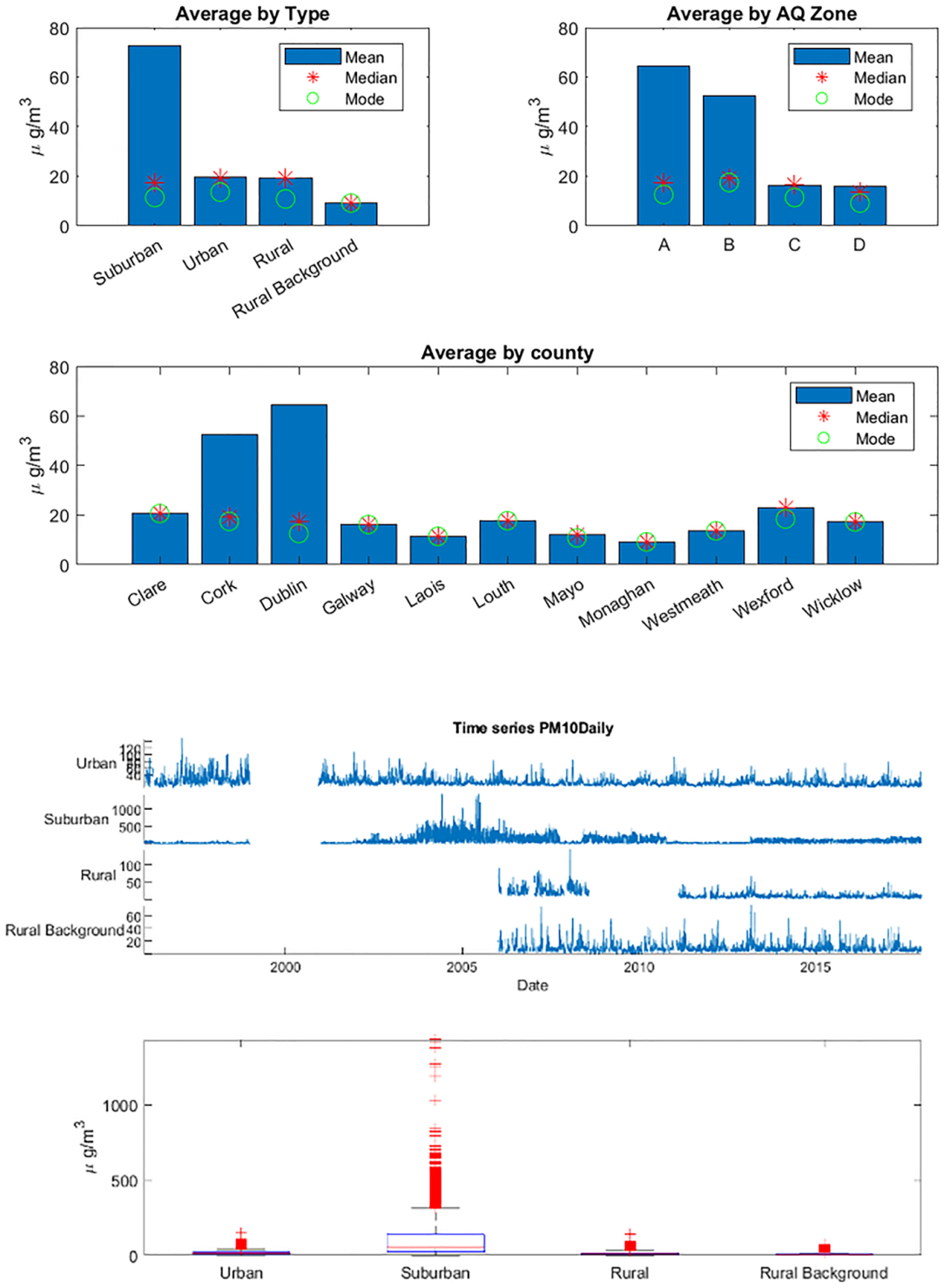

Figure 2.2. Daily PM_{10} levels by location type, 1988–2017. AQ, air quality.

monitored all species. Rathmines in Dublin city is the only location that has been continuously monitoring for the full period with complete readings for NO_x only. Unsurprisingly, the network is strongest in zone A (Dublin).

Figures 2.2–2.6 give a brief overview of pollutant monitoring by the location type.

PM_{10} monitoring was focused on urban and suburban locations until the mid-2000s, when monitoring for PM_{10} also started in rural locations (Figure 2.2). Over the period in question, the daily average of suburban PM_{10} often exceeded urban and rural limits. In suburban locations, levels were higher in summer. Generally, PM_{10} levels are influenced by

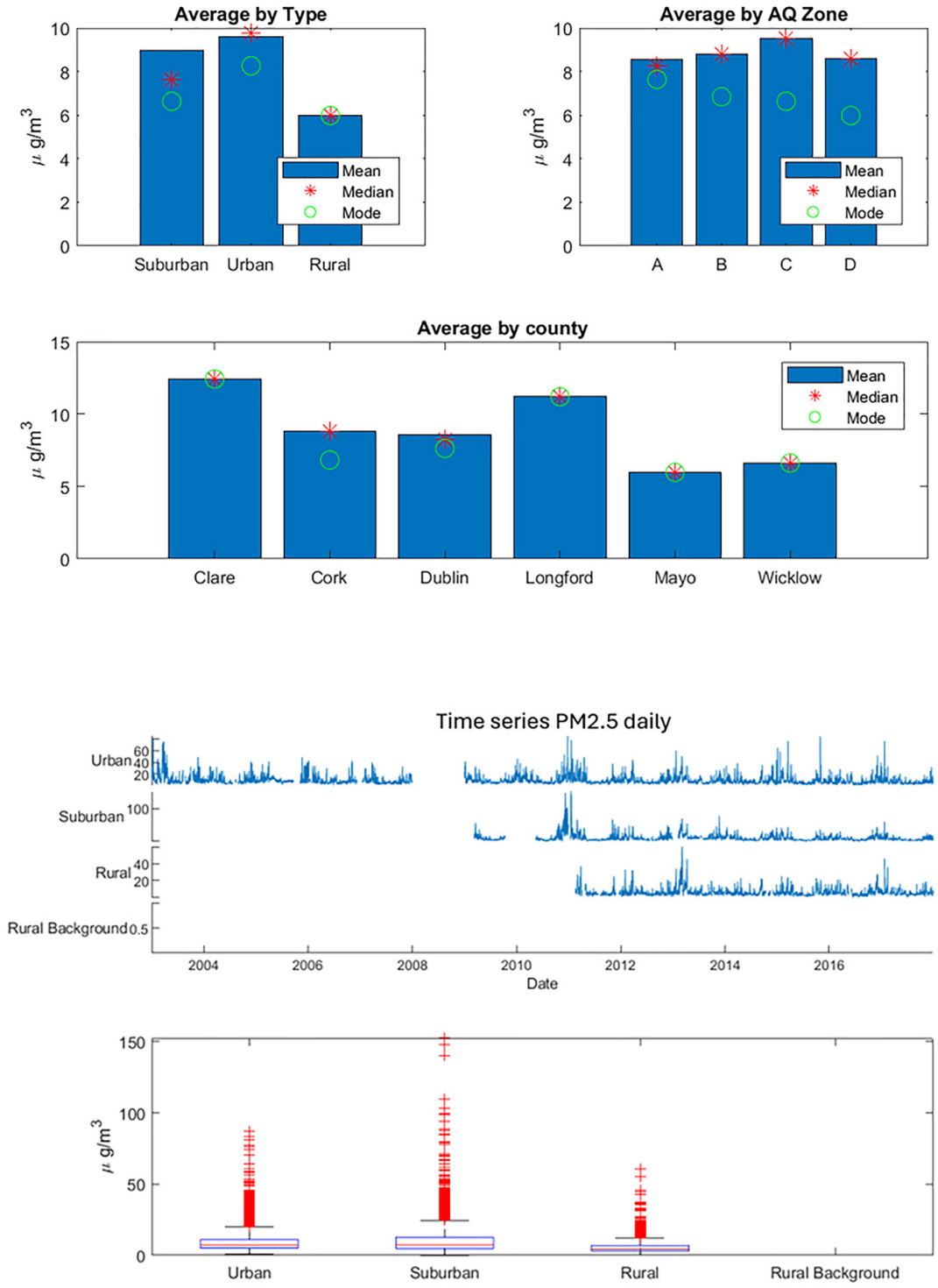

Figure 2.3. Daily $PM_{2.5}$ levels by location type, 1988–2017. AQ, air quality.

road resuspension, which is greater in dry conditions, and it may be the case that stations classified as suburban are capturing roadside levels. The daily variation and the average by location depicted in these figures emphasise the limitations of averaging or approximating individual exposure over time.

Monitoring of $PM_{2.5}$ started in the early 2000s, when it was brought into the regulatory framework, using the same gravimetric method as PM_{10} but with a different-sized cut-off at the sampling inlet for the filter collections (Figure 2.3). The difference across the locations is much less pronounced than for PM_{10}.

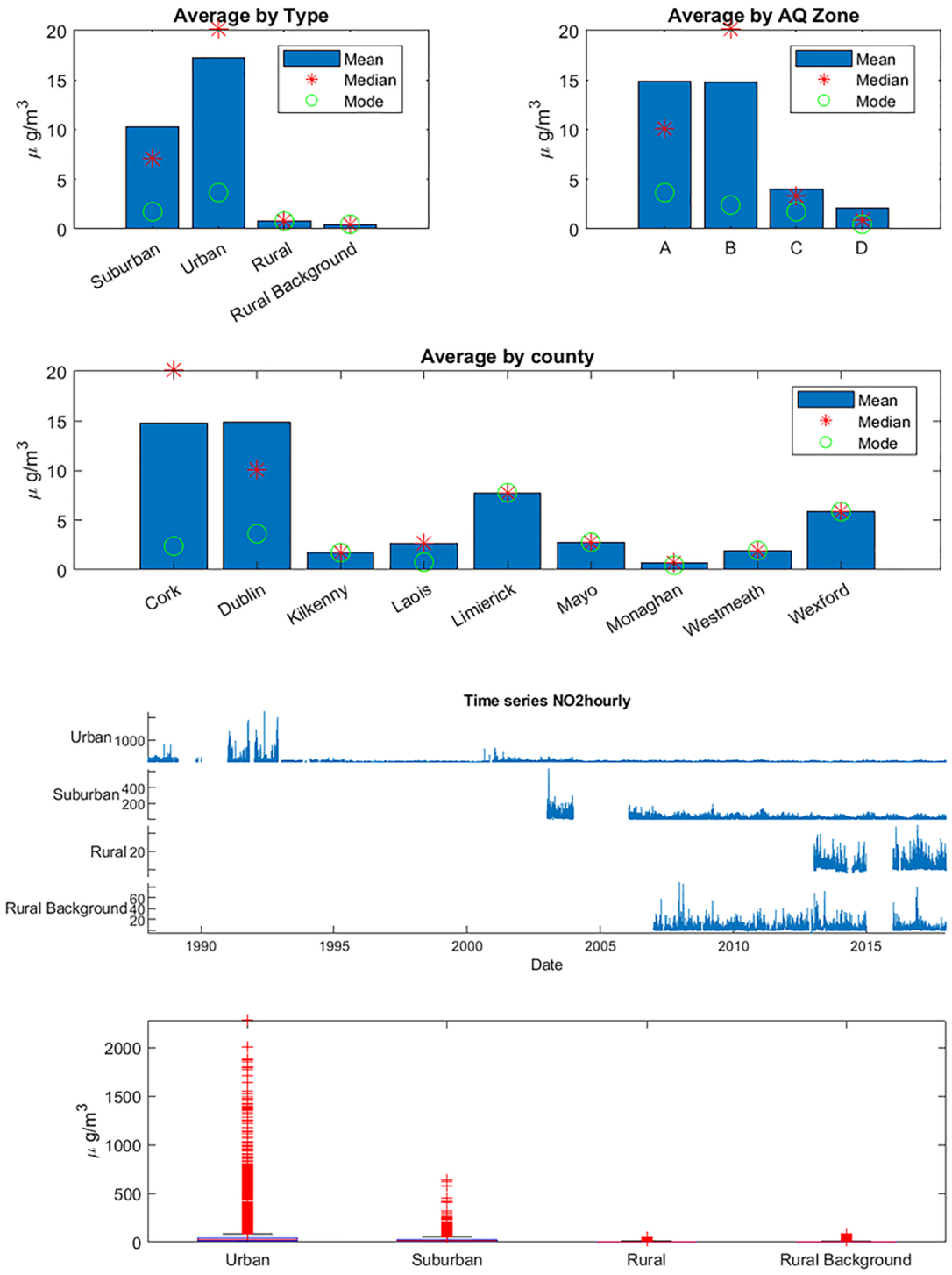

Figure 2.4. Hourly NO_2 levels by location type, 1988–2017. AQ, air quality.

While urban and suburban locations have higher recorded levels than rural locations, the average levels across the four air quality zones are comparable.

Hourly measurements of NO_x levels have been available since the first monitoring station of the network was commissioned, in Rathmines in Dublin (Figure 2.4). This pollutant is higher in urban and suburban locations, particularly in the two main cities, Dublin and Cork. The seasonal variation in NO_2 concentrations within location types is similar, with levels lowest in summer months. Diurnal variation in hourly NO_2 concentrations is bimodal in urban and suburban locations and appears to lag behind traffic burden. In all locations there is daily peak before

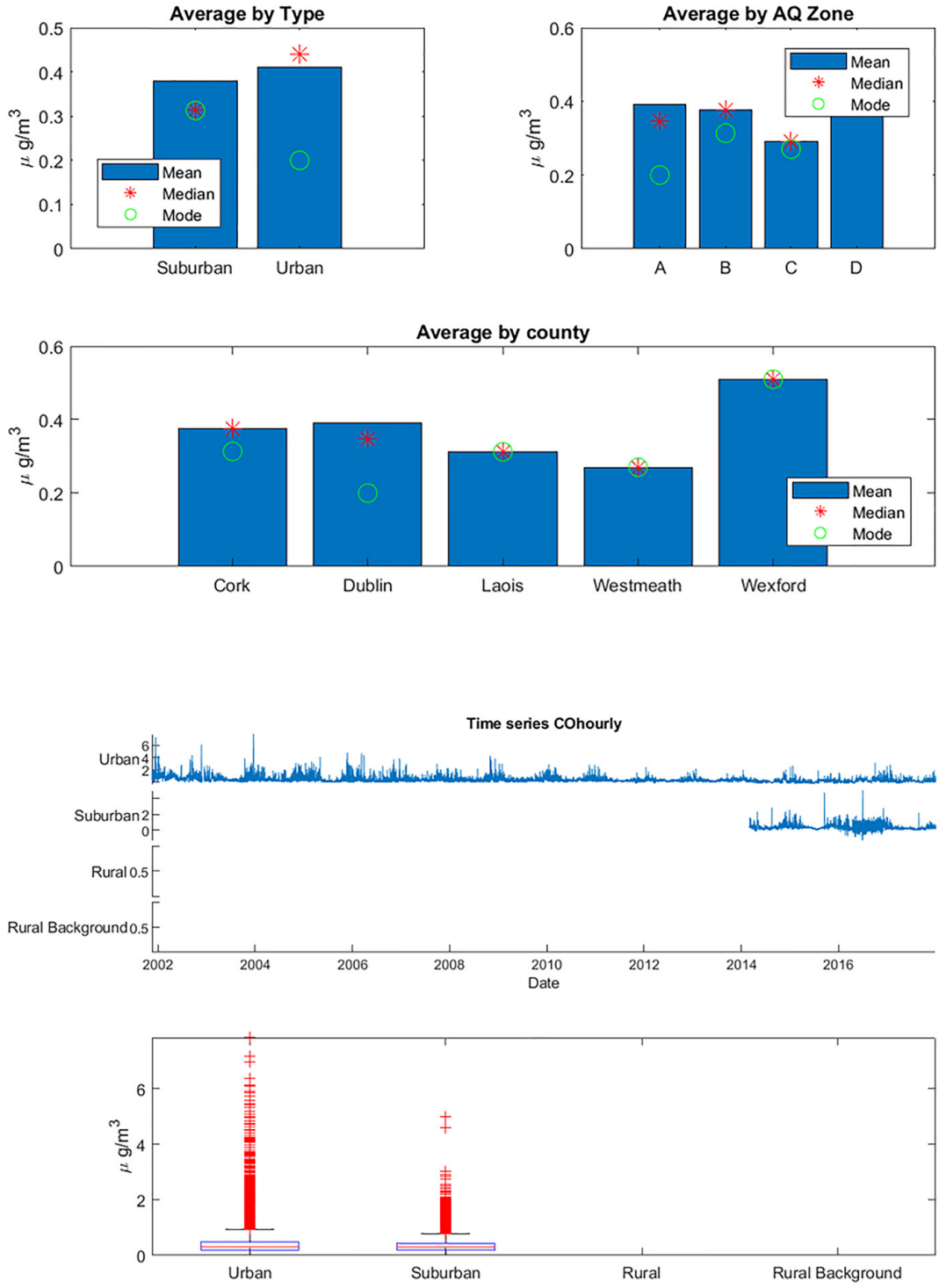

Figure 2.5. Hourly CO levels by location type, 1988–2017. AQ, air quality.

midday. Hourly variations emphasise the levels of granularity needed in epidemiological studies of pollutants, particularly for acute outcomes.

CO has been measured well historically in urban settings since 2002 but not in other location types (Figure 2.5). The seasonal pattern is similar in urban and suburban locations and is lowest in summer months. Diurnal variation in hourly CO concentrations is bimodal, with slightly more elevated peaks in urban areas than in suburban locations.

Ozone, a secondary pollutant that is influenced by sunlight, volatile organic carbon compounds, NO_x

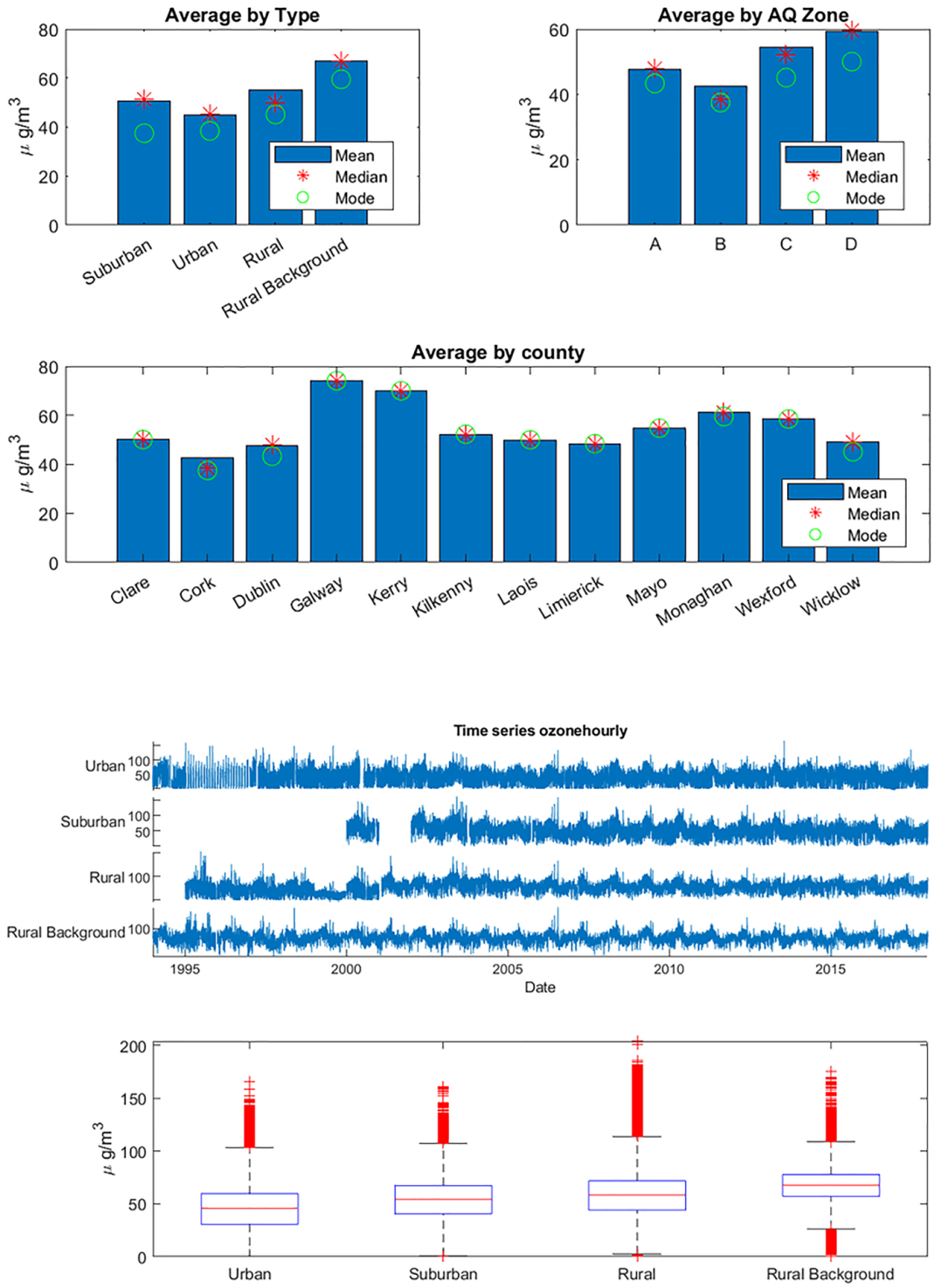

Figure 2.6. Hourly ozone levels by location type, 1988–2017. AQ, air quality.

levels and other pollutants, has been measured since the inception of the monitoring programme. Hourly average ozone levels by location are presented in Figure 2.6. Levels tend to be slightly elevated in rural locations with similar daily patterns across location types.

2.2.2 Data from the extended network (ongoing since 2017)

As the field of environmental epidemiology grows it has become apparent that further refinement of pollution measurements is needed to fully elucidate aetiological relationships, whether this is distinguishing

$PM_{2.5}$ exposure from PM_{10} or apportioning the source. Since 2017, the EPA has been rapidly expanding the monitoring network through the AAMP, and new stations commissioned since the expansion frequently include contemporaneous hourly measurements of the levels of different PM size fractions, NO_x and ozone. Figures 2.7–2.9 show data for $PM_{2.5}$ and NO_2 from the expanded network locations since 2017.

The seasonal pattern is similar in urban and suburban locations and lowest in late summer/early autumn months. (April 2019 was a previously documented anomaly with very high $PM_{2.5}$ levels across the country, seen as a peak in month 4.) Diurnal variation in hourly $PM_{2.5}$ concentrations is similar in urban and suburban settings, peaking in late evening.

The seasonal pattern of NO_2 concentrations is similar in urban and suburban locations (Figure 2.9), with higher levels in winter. Hourly NO_2 concentration peaks in the morning and evening, following traffic patterns.

The measurements across the air quality monitoring network, shown in the figures above, indicate that there are clear and consistent differences, of varying degrees, in pollution levels between the locations classified as one of urban, suburban, rural or rural background, and between locations in each of the four Irish air quality zones, A, B, C and D. This is as should be expected because it reflects the rationale behind the design of the monitoring network: to capture measurements relevant to protecting the human population in each of the geographical categories while optimising the monitoring resources for compliance monitoring. Also evident in the figures are the gaps in measurements over time, which will limit historical epidemiological study of the recent past. The extent to which the monitoring network has grown and the differentiation of pollutant species is also evident. These data will enrich future epidemiological enquiry.

2.3 Health Data for Epidemiological Research

Routine interactions with public health services generate millions of rich records of data annually. In addition to routine data sources, periodic surveys of the population are undertaken, including a whole population census, and several research cohorts are ongoing. These data sources and their utility in health and air quality research are discussed below.

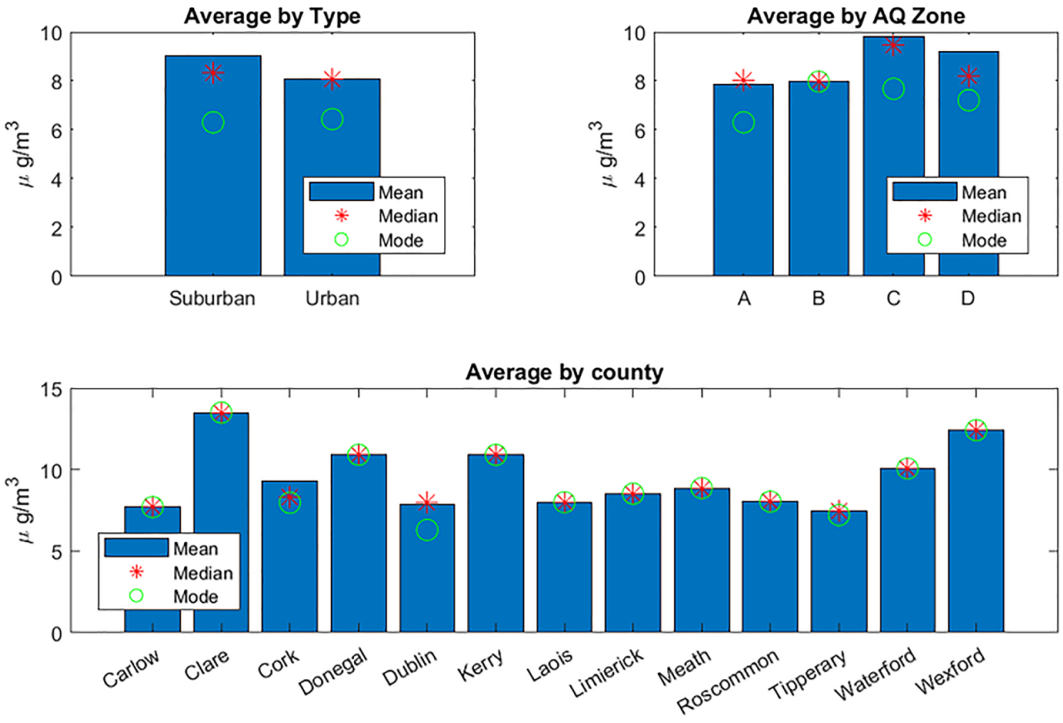

Figure 2.7. Hourly $PM_{2.5}$ levels by location type, 2018–2020. AQ, air quality.

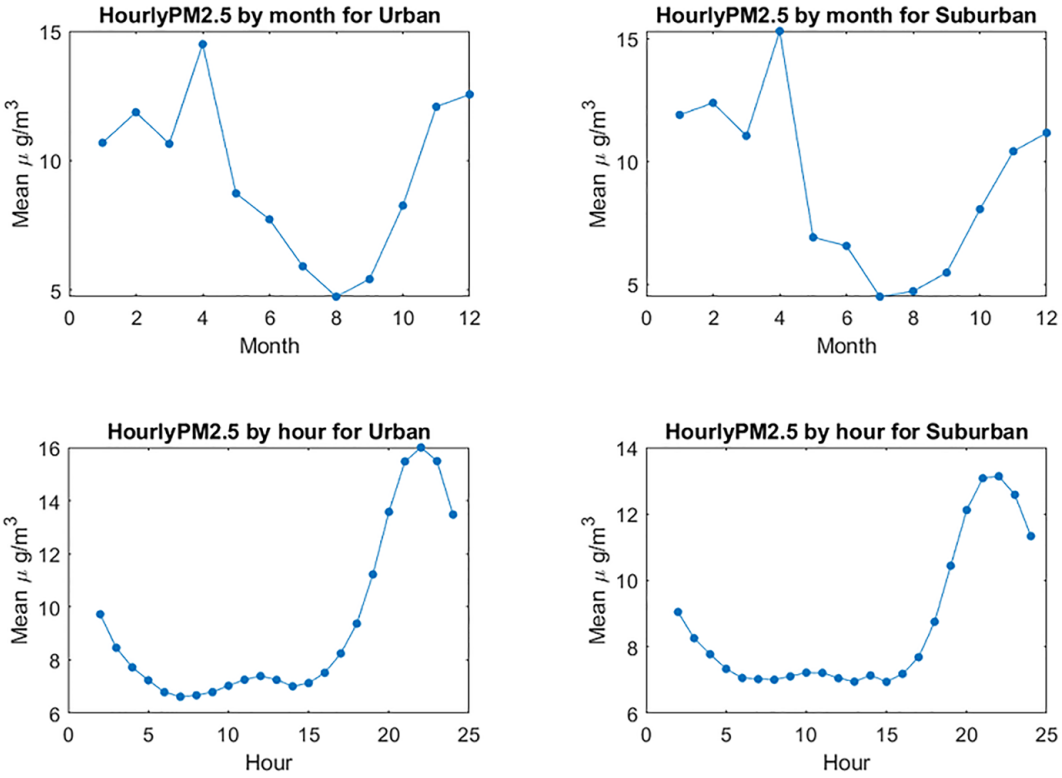

Figure 2.8. Seasonal and diurnal variation in hourly $PM_{2.5}$ levels by location type, 2018–2020.

2.3.1 Sources of data

Routine health and social interaction data and registries

Every day in Ireland thousands of interactions occur between members of the public and health and social care services. For each of these interactions data are generated. Many of these data reside in repositories owned by the Health Service Executive (HSE) but are often unconnected to each other and have been historically unavailable for external research. Data collected and stored by HSE ranges from large repositories like the Hospital In-patient Enquiry (HIPE) database (1.7 million annual records of discharge or death records with associated International Classification of Diseases code, length of stay, patient demographics and other clinical indicators) to the Primary Care Reimbursement Service (PCRS) (general practice service, drug payment scheme, dental services for public users entitled to free or reduced cost service, childhood immunisation scheme) to country-wide screening programmes (e.g. BreastCheck, newborns' bloodspot screening). Other sources of data include specific disease registries like the Irish Motor Neurone Disease Register and Cystic Fibrosis Registry Ireland, both of which are registered charities operating in agreement with the HSE to access and amalgamate paper and electronic patient data on an opt-in basis. Patient data are also collected and collated within private health care providers and hospitals. Furthermore, several distinct disease biobanks have opened in the last decade, usually spearheaded by academic and medical/hospital research groups.

As indicated above, routine health data are maintained by separate extra- and intramural agencies (e.g. Primary Care Eligibility & Reimbursement Service, Healthcare Pricing Office) and remain fragmented. Unlike in other jurisdictions, Ireland lacks fully electronic health data collection at point of contact and has not fully rolled out unique national identifiers or health numbers. In 2014, the Health Identifiers Act 2014 was introduced to "provide for the assignment of a unique number to an individual to whom a health service is being, has been or may be provided" and, shortly thereafter, the national Individual Health Identifier (IHI) National Register was created. Initial systems available to seed with the new IHIs were selected based on evaluation of candidate systems and include general practice management systems, CervicalCheck, the schools immunisation programme and the COVID-19 vaccination programme (Walsh

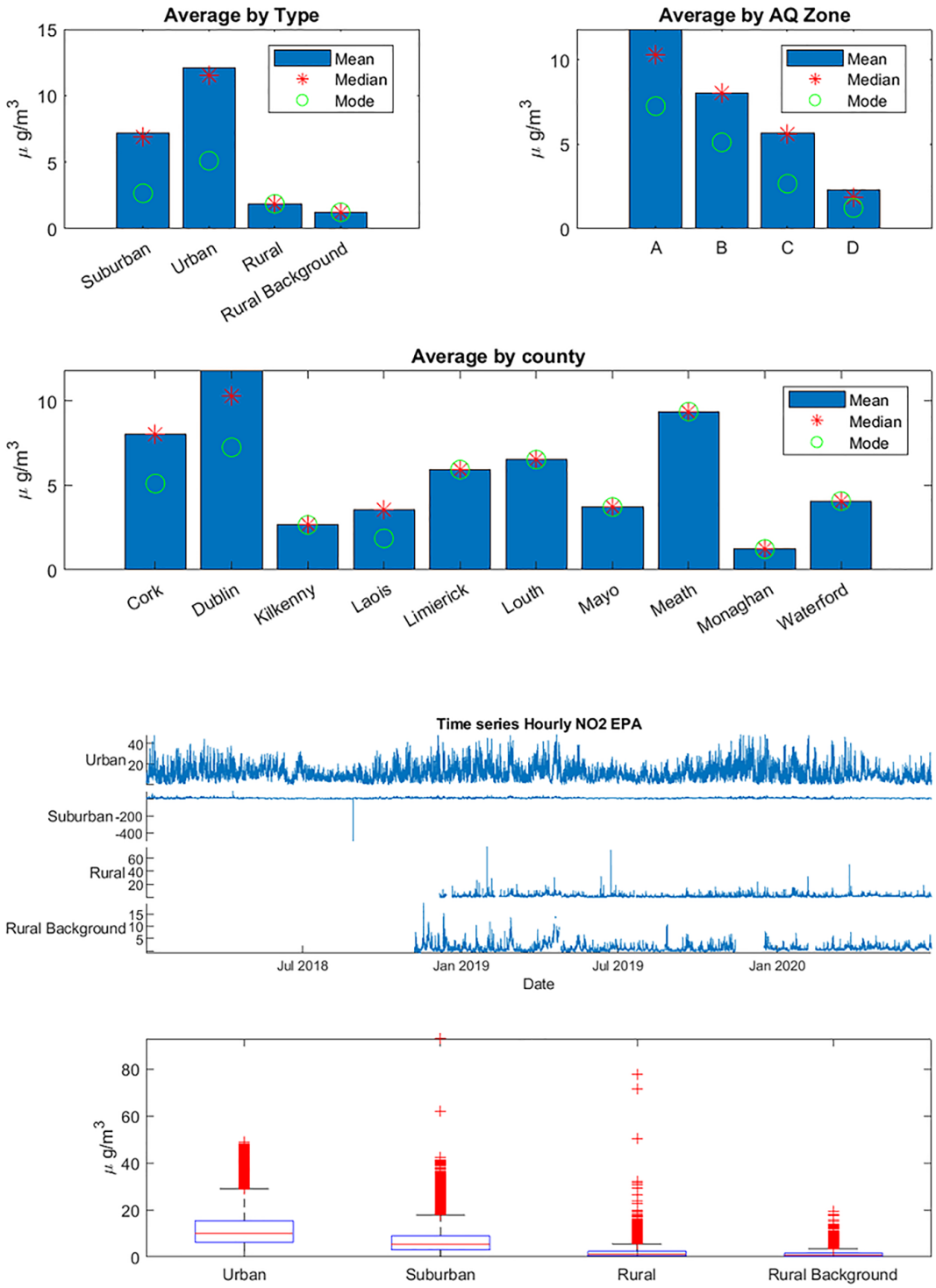

Figure 2.9. Hourly NO_2 levels by location type, 2018–2020. AQ, air quality.

et al., 2021). The National Strategic Plan 2022 lists the systems chosen for the next phase. Major advances in eHealth have been made in the past decade in Ireland, and COVID-19 precipitated the acceptance of and rapid change towards digital health (e.g. eReferrals system). Further roll-out of IHIs and integration of data collection systems will greatly enhance future health research.

Surveys and cohorts

In addition to point-of-contact data, several surveys and statistical compilations can be used to characterise population health and well-being. The broadest survey in Ireland is the Central Statistics Office (CSO) National Census, which occurs every 5 years, capturing data from all residents and including

questions on disability, long-term illness, caring and perception of general health. Smaller representative surveys include the Health Behaviour in School-aged Children survey (from the University of Galway and designed to estimate health and well-being in children aged 9 to 17) and the Healthy Ireland Survey (surveying 7500 residents aged 15 and over and including questions on health, health service use, caring and behaviour). These surveys give useful snapshots of health, but each wave recruits new respondents, meaning that longitudinal changes in health are missed and data are irrecoverably deidentified; therefore, linkage with other data is not possible. To fill this gap, several research units have established cohort studies that enrol participants who are administered repeat surveys and health assessments at regular intervals. Examples include the Irish Longitudinal Study on Ageing, which involves a cohort of almost 9000 people aged over 50 enrolled in 2009 who have now completed five biennial questionnaires, and Growing Up in Ireland, involving a cohort of over 8000 9-year-olds and 10,000 9-month-olds enrolled in 2008 who have taken part in five waves of the survey to date.

Lastly, there are sources of health and well-being statistics collected in Ireland by agencies with specific remits. Examples include the Road Safety Authority, which collects and collates data on road usage, collisions and fatalities in conjunction with the Garda Síochána, and the Health and Safety Authority, which similarly collects and maintains a national database on workplace accidents.

2.3.2 Access to health data

In 2017, the Health Information and Quality Authority (HIQA) published its third catalogue of national data collections; it maintains a corresponding up-to-date searchable website (https://www.hiqa.ie/areas-we-work/health-information/data-collections). Many data files listed by HIQA (especially those resulting from research activity, rather than interaction with government services) are held by the Irish Social Science Data Archive (ISSDA). Facilitated by University College Dublin Library, ISSDA is a reputable organisation focused on data acquisition and sharing, and depositing data with ISSDA ensures its preservation, particularly for historical data collections. Data on ISSDA are stored as anonymised microdata, and the process for applying to access data in easily used formats has been streamlined.

eHealth Ireland is an organisation within the HSE and Department of Health tasked with delivering digital health for Ireland. As part of its remit, eHealth Ireland provides data hosted by the Open Data Policy Unit at the Department of Public Expenditure, NDP Delivery and Reform to facilitate access to data from the health services (https://data.gov.ie/dataset/). Over 400 data files are easily searched and in usable formats but are often at a high level of aggregation, which limits their use in primary research.

Researchers can negotiate with data owners for access to microdata or disaggregated data. Understandably, many of the structures for routine data collection and collation were not designed with linkage and data sharing in mind. Organisations can lack the resources to provide data in a format that works outside the system it is stored on, and data owners are often unclear about what they can share and how. These issues can result in protracted and convoluted negotiations. Even when data are available, to ensure protection of patient privacy, data are provided in aggregations that prevent relevant linkage and analyses. This is especially true in the context of air quality research, which requires disaggregation across specific locations and times due to the changing levels of pollutants, since the data holder might use an aggregation unit that includes people with heterogeneous air pollution exposures, such as county or hospital attended and/or longer time periods.

2.3.3 Limitations of health data

The extensive body of health data in current formats and aggregation has limited value for air quality and health research in Ireland. For example, data that are more readily accessible, like those on ISSDA, are so by virtue of having had identifiers removed and so cannot be linked to location-specific air quality data.

Several routine health data collections can be accessed by applying to the data custodian, for example HIPE or PCRS. But there are historical sharing protocols that limit these data too. Two aggregation units are commonly used when data are sought: (i) patients' county of residence or (ii) patients' service division. There are several layers

of service division that, until 2022, were not nested: hospital, hospital-group ($n=7$), community health care organisation ($n=9$) and HSE region ($n=4$). Since 2024, six regional health areas have been introduced across Ireland to replace hospital groups and community health care organisations. Data in these formats have been shown to work well for time-series analyses of air quality and health in more densely populated areas or after an intervention that substantially changes exposure levels across the population. For example, using HIPE data, decreases in black smoke pollution levels following coal bans in Dublin in 1990 and subsequently in other urban areas were shown to be associated with lower age-standardised death rates (Clancy, 2002; Dockery, 2013) and hospital admissions (Dockery, 2013). However, methodological issues arise as the measured air quality moves further from the actual exposure experienced, particularly if the range of exposure levels is small (noise overwhelms signal). Using large aggregation units such as hospital or county to determine exposure increases measurement error.

In addition, there is an established practice for sharing sparse data that hamper analyses of rare outcomes (or rare within a sparsely populated area, in a given time frame). When events are rare, the aggregation unit is often collapsed over time, so instead of daily counts (which might be relevant for an acute health event), the researcher might receive monthly counts for each geolocation. Furthermore, all instances of count data between 1 and 5 will have been removed from the data acquired by researchers. It is not always clear how providing the exact count for an event in an aggregation unit can reveal information about patients; it depends whether there is a unique combination of other information available.

Complex studies of air quality and health often require health data from two or more domains (e.g. data on acute and long-term illness could be linked and further linked to prescription data). Health-to-health data linkage is constrained in the absence of full roll-out of a unique health identifier and digital medical record. It is important to note, however, that researchers have acquired multiple-source linkage with assistance from the CSO, which acted as a trusted third party. However, a resourced, streamlined workable process is not in place. Even if some researchers can access microdata in some restricted cases for a specific project after extensive negotiations, this does not constitute a workable operational system for monitoring air pollution's impact on health.

An HRB-funded report concerned with creating an environment for health research (Moran, 2016), included interviews with key users about attitudes and practices for project-specific health data access and linkage. Interviewees identified fear of infringement of privacy legislation; lack of unique identifiers; limited consent from patients or study participants for new use of data; lack of clarity on data ownership (including inability to identify points of contact); and a hard-to-navigate paperwork process involved in acquiring data as barriers to timely research.

In summary, a trusted third party can facilitate safe data sharing, providing a framework that protects patients as well as data custodians, who must comply with the recently enhanced General Data Protection Regulation.

2.4 Utility of Current Data for Epidemiological Research

A major limitation of sharing routine health data in Ireland is the need to aggregate data on large geographical or service units. As described previously, while it is known that measured individual air pollution exposure is not readily available, it is also accepted that exposure can vary greatly within small geographical areas and is further influenced by individual behaviour (time outdoors, level of exercise, time away from home/at work, mode of commuting). The reason this needs to be stated explicitly is that it makes air quality different from other factors that are routinely linked to health data in epidemiological studies. Diet (current, long-term, childhood, etc.) is another complex exposure that affects health, but, unlike air quality, many instruments exist for individuals to report their current or past diet. On the other hand, data on socio-economic indicators of poverty, deprivation, access to facilities, green spaces, etc., are collected and published in aggregate form with a geographical distribution that *can* be linked to published geographical distributions of health indicators.

Aggregate or population-level data are not only subject to measurement error when used to estimate risk of disease, they are often prone to confounding bias as well, because individual habits may differ in each of

the units being compared. Cohort studies that follow individuals' health outcomes over time can minimise some of this bias. A time-series approach can also be used, where the same geographical location is compared with itself at different time points, thereby circumventing some confounding bias. Time-series analyses work well for acute changes in air quality, but care should be taken in long time-series studies to account for secular changes other than in air pollution. For example, a comparison of air pollution and hospitalisations in Dublin in 2006 and 2016 should consider the temporal change in income as a result of the abrupt economic downturn in 2008, which could result in a possible decrease in health care sought by individuals who were under financial constraint (and a potential increase in disease incidence in the longer term). At the same time, the EPA documented a reduction in road transport emissions during the recession. In this example, there are simultaneous changes in air quality and in health data that are, in part at least, not directly related. Statistical methods exist to account for changing trends over time, such as spline functions, whether the cause and timing of that change (slopes) are known or not (data driven).

An optimal data linkage solution is one that allows data from many domains to come together: health and demographics (age, illnesses, smoking habits, body mass index, education, etc.), air quality at an individual's address or workplace (directly measured or modelled), social and income data and aggregate data associated with the address (deprivation index, green space).

In summary, the current situation in relation to linking air quality and health data is that accessible health data are not relevant, while relevant health data are not accessible.

3 Review of International Best Practice for Data Linkage

The aim of the review of international best practice was to inform the development of a feasible method for linking routinely collected and research-led health data with routinely monitored air quality data. In many jurisdictions data access and linkage is operationalised by an independent body, sometimes referred to as an honest broker, a data safe haven or trusted research environment (TRE), that works within the regulatory and legal frameworks for use of personal data.

TREs provide secure data access and reliable data record matching with a commitment to protecting the identification of individuals and their information at the centre of their design and operations. Creating data repositories for secure access and linkage across multiple administrative domains offers more than the sum of the individual datasets and creates opportunities for new insights and policy planning in the public interest.

3.1 Overview of the Functions of Trusted Research Environments

A TRE or honest broker is an information technology (IT) organisation that creates a wall between identifiable data and a data user. As a first step, access to desired identifiable data is given to the TRE by the data owner. The TRE will house many datasets from a wide array of administrative bodies and academic research groups and will have the IT infrastructure and staff to safely maintain, clean, link and share (ideally remotely) these data with suitable users. A TRE can securely provide de-identified or limited datasets to a researcher, merge existing data from *several domains* via an identifier, merge newly collected research data with existing administrative data, use an identifier to provide small-area-level aggregated data and act as third party to process data for a databank. In Ireland, the CSO performs these functions for specific data sources.

Where data owners might only provide data to researchers that have been de-identified, or provide them at an aggregate level, preventing linkage to other data, a TRE can provide individual-level data that have been merged to other relevant data and then stripped of identifiers. Personal data may be indirectly identifiable from combinations of other variables including age, sex, place of work or school, name of GP or vehicle registration. A TRE can determine when a person could be indirectly identified and will work with a researcher to avoid this scenario.

When the nature of data means that they cannot be wholly de-identified, a TRE will decide an appropriate aggregation unit (when is a small area too small, for example?). When working well, the researcher can change the aggregate variable and the TRE can readily re-create a data file. While data are de-identified for the researcher, the TRE can provide coded individual-level data if codes or re-identification are not available to the researcher (one-way encryption) to allow future linkage.

3.2 Review of Trusted Research Environments

The following organisations were selected for review based on an initial review of the literature: Secure Anonymised Information Linkage (SAIL) Databank, Wales, and Administrative Data Research Northern Ireland (ADR NI). The review involved desktop research to understand how sensitive data linkage is organised in other jurisdictions.

3.2.1 Secure Anonymised Information Linkage Databank, Wales

SAIL Databank is a Wales-wide research resource focused on improving health, well-being and services for the Welsh population since 2007 (Figure 3.1). Its databank of billions of anonymised person-based records is an ISO 27001-certified and UK Statistics Authority-accredited environment (https://saildatabank.com/).

A salient feature of SAIL Databank is that identifiable data cannot be sent to or reside in the databank. A trusted third party, Digital Care and Health Wales (and formerly the NHS Wales Informatics Service), provides

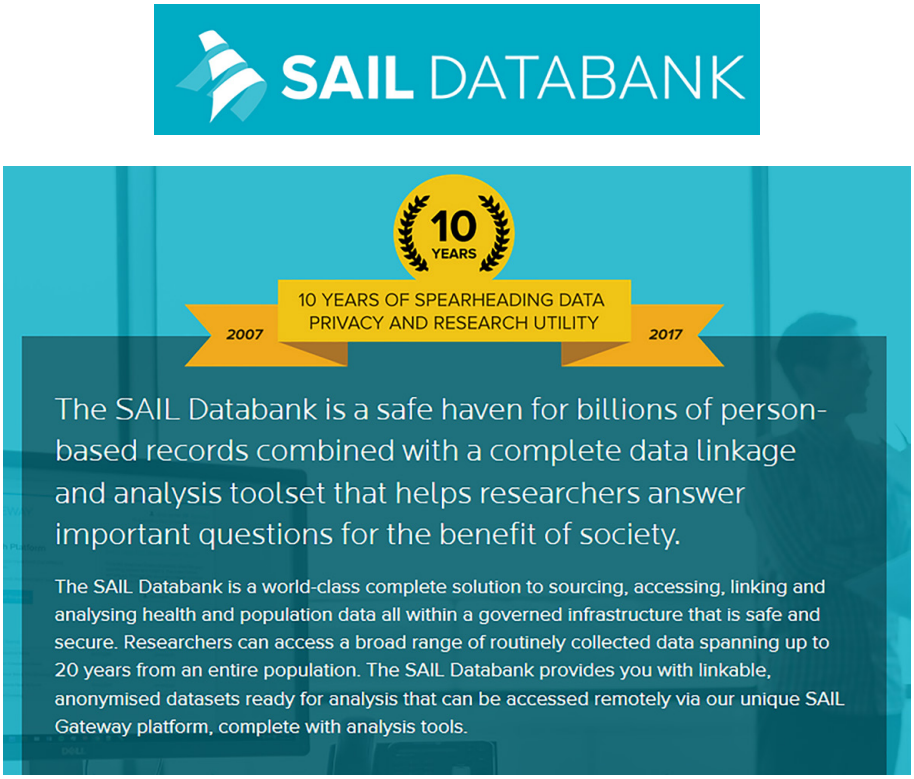

Figure 3.1. Overview of SAIL Databank, Wales.

anonymisation and encryption for SAIL Databank. Before a dataset is housed, the organisation that owns the data splits the file into two datasets with a key attached so the files can be re-joined. One file contains the substantive data variables (e.g. clinical/educational/census variables) and is sent directly to SAIL Databank. The other file contains the demographic variables (direct and indirect identifiers) and is sent to Digital Care and Health Wales, where a unique encrypted code, called an anonymous linking field (ALF), is added in place of the identifiable data; the minimum data needed for research are retained, such as week and year of birth, sex and area of residence (lower-layer super output areas are small areas of approximately 1500 residents or 650 households). This newly de-identified file is then sent to SAIL Databank where it is merged back onto the content file. A further encryption is applied to the ALF (ALF-e) and then the complete dataset is ready for use. Although anonymised, by using ALF-e, data can be anonymously linked together without compromising individual privacy. Data currently held at SAIL Databank include birth and death registry data, census data, primary care and hospital record data, screening programme data, justice department record data, education attendance and attainment data and social care data.

Researchers can interact with SAIL Databank in three ways. They can apply to use only data already residing in SAIL Databank; they can request that their own research data is added to SAIL Databank; and they can ask that data residing in SAIL Databank is merged with their own research data. To ensure proper use of data, research projects are reviewed by the independent Information Governance Review Panel and researchers undergo safe researcher training. Once a project is approved, researchers access data remotely via the SAIL Gateway. More details on the technical safeguards at SAIL Databank are available at https://saildatabank.com/governance/privacy-by-design/.

3.2.2 Administrative Data Research Northern Ireland

At the beginning of the INHALE project a federated system of data linkage was in operation in Northern Ireland. A federated system means data owners retain data and share it on a project-by-project basis with a trusted third party for linkage and de-identification.

In the recent past, data access and linkage was managed by ADR NI, an organisation comprising the Northern Ireland Statistics and Research Agency (NISRA) (a TRE for Northern Ireland), Queen's University Belfast and the University of Ulster. Historically, ADR NI did not operate like SAIL Databank, where data are held centrally when provided by the data owner. Within NISRA there were two distinct networks: one was used for trusted third-party functions and the other was used by a research support unit (O'Reilly *et al.*, 2020). Projects were approved in three stages. As a first step, the feasibility of the project was determined in consultation with ADR NI support staff, where the data requirements were reviewed. If the project was deemed feasible, the ADR NI support staff guided researchers through the application for approval. An approval panel adjudicated on the project and, if approved, a series of data sharing agreements were put in place with data owners. Then, data were linked and de-identified by the trusted third-party staff for use *only* by the research team that devised the project.

Health data are curated and maintained by the Business Services Organisation (BSO) within the Health and Social Care Board (HSC) in Northern Ireland. The BSO has provided the TRE, the Honest Broker Service (HBS), for health-to-health data across the health trusts, primary care and registers since 2014 (https://bso.hscni.net/directorates/digital/honest-broker-service/).

The HBS manages a data warehouse that receives data from a wide array of services within the HSC and link in some data sources held outside the HSC, such as the Northern Ireland General Register Office (Figure 3.2).

Up until the COVID-19 lockdown in March 2020, data held by the HBS were only accessible to researchers at the safe haven suite at the BSO offices in Belfast. A pilot programme was launched to trial remote access. The UK Secure eResearch Platform, also used to build the SAIL Gateway system, was tested and subsequently implemented for new projects.

ADR NI joined Administrative Data Research United Kingdom (ADR UK) in a pilot programme in September 2018. The mission of ADR UK is linking "the abundance of administrative data already being created by government and public bodies across the UK and making it available to approved researchers

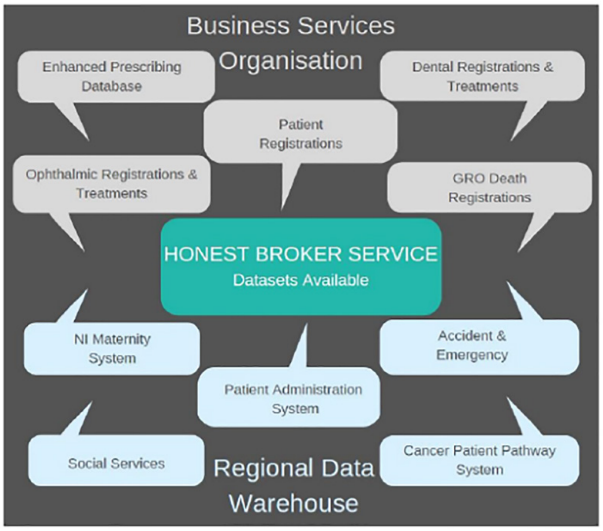

Figure 3.2. Overview of the HBS, Northern Ireland. GRO, General Register Office; NI, Northern Ireland.

in a safe and secure way". In May 2022, ADR NI published its strategy for 2022–2026 (ADR NI, 2022). A relevant strategic goal states that "ADR NI will engage with data owners to *acquire* administrative data related to NI and make these into sustainable research datasets that can address areas of policy need". In other words, there is a move away from the federalist, create-and-use-once approach to data linkage and sharing. Further details on ADR UK are available at https://www.adruk.org.

3.2.3 OpenSAFELY

OpenSAFELY is a relatively new initiative comprising a secure platform that allows safe and secure data analyses of over 17 million electronic health records. The platform was developed to facilitate rapid data analyses during COVID-19, on behalf of NHS England, by the Bennett Institute for Applied Data Science at the University of Oxford, the Electronic Health Records Research Group at the London School of Hygiene and Tropical Medicine and entities managing UK NHS data. Briefly, the platform contains tools and code that data holders can use to convert raw data into "research-ready" data. If code is amended by the user, it is added to a repository for future users and for transparency. A second layer of security is in place such that data users write the analysis code on a "dummy dataset" that replicates the real dataset in structure. Code is then run against the real patient data and results returned. Thus, the user never has

access to patient data and only summary data and analysis results, tables or graphs are released (https://www.opensafely.org/about/).

3.2.4 Examples of trusted research environments facilitating air quality and health research

Impacts of air pollution on educational attainment in children and young people

Children with respiratory tract illnesses may be more vulnerable to educational loss when exposed to poor air quality. Mizen *et al.* (2018) describe how they combined data from five separate systems via the SAIL Databank as part of the Cognitive Development, Respiratory Tract Illness and Effects of Exposure (CORTEX) study. Three of the data resources required were already available in the SAIL Databank, where they had been de-identified but tagged with ALF-e, which allowed retrospective linkage to new data sources. The available datasets were (i) the Welsh Longitudinal General Practice Dataset (containing data on asthma and seasonal allergic rhinitis treatment), (ii) the Welsh Demographic Service Dataset (containing week of birth, gender, address and multiple move dates) and (iii) the Welsh Government's National Pupil Database (containing students' capped points score (derived from their best eight General Certificate of Secondary Education marks standardised for mean of their exam year), gender, free school meal eligibility, special educational needs, attendance data and school-level data). Hourly measures of air pollutants were used to model exposure at each home and school location in Cardiff. Within the secure environment of SAIL Database, home and school air quality and pollen data were linked to individual demographic, health and education files. Data on a pupil's home and school location were combined to create individual-level air pollution and pollen exposure data. The richness of the data allowed researchers to vary a pupil's location in the 3 pm to 5 pm window on school days because it was not known if a pupil was commuting home or participating in after-school activities.

In Mizen *et al*. (2020), the authors present an analysis using the data prepared for CORTEX. They reported that among 18,241 students aged 15–16 an increase in short-term exposure to NO_2 modestly reduced the standardised capped point score after adjusting for individual and household factors.

Infant health and mothers' exposure to air pollution

This project was led by Queen's University Belfast. The approved project requests linkage of data from the Northern Ireland Maternity Services and modelled air pollution data. Mother and infant demographics and clinical history were linked to the annual average level of ambient air pollution in a 1 km × 1 km grid associated with the mother's address while pregnant to examine the role of air quality in infant health (e.g. birth weight) (Jahanshahi *et al*., 2024).

4 Recommendations for Data Linkage

Air quality data are highly variable, even within geographically close areas, such that within a city or a small town the exposure levels of the individuals living there can vary significantly. Sophisticated models can estimate exposure at the home, work, school and commute levels. To link the correct air quality exposure to the individuals represented in health data would require that health data include multiple identifiable variables, but legitimate privacy concerns mean that microdata cannot be given over to researchers for the purpose of data linkage.

As described in Chapter 3, TREs or similar infrastructure are required to do this linkage in a safe way. The INHALE project makes two main recommendations. The first is for a medium- to long-term solution based on establishing (or expanding existing, for example, CSO) TRE infrastructure where de-identified individual-level linked data are made available to researchers and government departments. A second, researcher-led, solution is one in which researchers do not receive individual-level data but instead are provided with aggregated data on a unit that is relevant to their research question.

4.1 Recommendation 1

4.1.1 Investment in a centralised trusted research environment for data linkage

Linked data can be used to improve the lives of the residents of Ireland. Implementing or expanding a safe environment to facilitate linkage for research and for government department use should remain a national priority.

The data access, storage, sharing and linking (DASSL) approach to linking health data was recommended in an HRB-funded report in 2016 and was closely modelled on Northern Ireland's HBS (Moran, 2016). The INHALE study team recommends building on the DASSL approach and moving towards a model, similar to the SAIL Databank and the new ADR NI/ ADR UK approaches of centralised linkage for optimal security and efficiency, which prevents duplication of collection, processing, linking and research efforts and will have greater buy-in from a wide array of health and administrative data custodians. Moreover, data pertaining to potential confounders or modifiers of associations between air quality and health could also be provided to the researcher in a centralised system (e.g. type of indoor heating, nearest green space or education level of head of household), ensuring excellence in research methods without the need to approach additional data custodians. In Ireland, the CSO has the legal authority, infrastructure, expertise and experience, as well as the reputational standing, to link and house sensitive data in a manner that will always allow retrospective linkage to new data streams; it would therefore be reasonable to suggest that, rather than establishing a wholly new TRE, the CSO could be funded to expand its current work in this area.

The Irish Centre for High-End Computing, National University of Ireland Galway, was awarded an HRB pilot project to "design and test the major infrastructural elements for safe use and linkage of different [administrative] data sets using synthetic data" based on the DASSL model. The research group published a technical prototype of infrastructure required for safe data sharing and linkage using a trusted broker paradigm that adheres to regulatory safeguards (Fennelly *et al*., 2022; Keogh *et al*., 2024).

4.1.2 Continued prioritisation of the roll-out of individual health identifiers to facilitate linkage of health-to-health data

The inability to link the many domains of health data is a barrier to air quality and health research. Unlike in the UK, where every participant in the NHS is assigned a unique number that has been used across all services since the mid-1990s, in Ireland the IHI is used in a limited (but growing) number of services. Detailed reports support and recommend models for the ongoing efforts to transition to digital health (Moran, 2016; Walsh *et al*., 2021; Health Service Executive, 2022). In addition, we recommend that Eircodes are collected and regularly updated in all administrative systems to ensure that health data, like air quality

data, are spatiotemporal. Indeed, the HSE National Strategic Plan 2022 indicates a ring-fenced budget to meet the 2022 key deliverable of IHI and Eircode integration in several large health IT systems.

4.2 Recommendation 2

In the absence of a working TRE for health data linkage, like those suggested in recommendation 1, data requested are released by data custodians to researchers in aggregated form to protect the privacy of patients and service users. The geographical unit for aggregation is usually county or hospital or community health care organisation or hospital group, as described in Chapter 2. Studies of air quality and health are thus limited to time-series and acute outcomes or hospitalisation events in the main cities. It is not feasible to reliably study medium- to long-term effects of air pollution on chronic disease or effects of exposure during critical windows like pregnancy (see Chapter 2).

The second recommendation is a dynamic researcher-led work-around solution to linking air quality to relevant health data that can be easily implemented, does not compromise privacy and can evolve with the needs of the research and policy community. Furthermore, this approach could eventually be implemented in a routine workflow, within the CSO or another TRE, to produce an annual report on air quality and health-linked statistics to assist in the monitoring of policy impacts.

The recommendation is illustrated with a reductive example. A very simplified approach to linking air quality to health data without compromising privacy is to amend the health data collection process to include specific geocodes recorded at point of contact. For example, by implementing the mandatory recording of Eircodes one could immediately pull out proxy parameters like “nearest air quality monitoring station to patient’s home address”, or a geographical identifier based on environment descriptors in the CAFE Directive, like “urban background”, “city centre” or “rural”. Availability of this additional information in the raw data would enable health data controllers to publish health statistics aggregated on units (like air quality zone, electoral district or nearest air quality station) that were considerably more relevant to air quality than the current units. Importantly, there would be no new issues in relation to privacy.

While this is a simplified example, it illustrates that the missing link between existing health data and existing air quality data is simply a translation of addresses in individual records into geographical identifiers or units that are linkable to measured air quality data, allowing the aggregation of health data in a meaningful way.

An alternative approach to embedding translated geographical identifiers into health data recording systems, as described above, is to merge geographical identifiers with the microdata records after collection, as an intermediate step before making aggregate data available to researchers. This could require an intermediary agent that sits between the data controller and the researcher such as a TRE. Alternatively, the researcher could provide the data controller with a data file of all Eircodes or addresses, each tagged with the level of the aggregate unit.

In summary, making health data aggregates relevant to air quality means that the original data must be either amended at source in the health IT system, in a permanent fixed fashion, or amended before aggregation with auxiliary data in a dynamic and flexible fashion. It is the view of the INHALE project team that the latter option is more feasible, flexible and future-proofed. With this approach, microdata are needed only to generate relevant aggregates, i.e. obtain the appropriate level of aggregation relevant to air quality. The solution proposed by the INHALE team is to publish relevant aggregate data on higher-level variables (e.g. air quality zones, Clean Air Act locations) and to implement a system in which health data controllers provide data on *researcher-defined* aggregated units. This approach alleviates the need for researchers to access personal, sensitive, identifiable data.

The advantages of this approach are that (i) it has a low technical barrier, because no change in existing infrastructure and IT requirements are required, (ii) it is flexible and adaptable to changing research questions, (iii) there is no legal barrier, as there is no need for the researchers to access original microdata with personal information, (iv) it can be relatively easily operationalised, (v) it has a relatively low cost because no new infrastructure or legal instruments are required, only human resources, (vi) additional area-level data could also be linked and, lastly, (vii) it is dynamic and can evolve with the needs of the research, health and policy communities. With this approach,

individual-level variables cannot be included in data analyses, but research and data linking that requires access to individual-level variables for purposes other than generating relevant aggregates will require further negotiations with the data controller, as is currently the case, and that process cannot be operationalised.

The following gives a description of how the project team proposes that this could be achieved in practical terms.

4.3 Operationalising Linkage of Air Quality and Health Data without a Trusted Research Environment

4.3.1 Publishing mortality and health data on researcher-defined aggregated air quality metrics

In Figure 4.1, the original microdata held by the data controller are presented in the top left corner. Each record will have some form of geographical identifier, such as an address, an Eircode or something less specific, like a town name. Currently, the data controller produces aggregate health data, which are made available on request. However, as emphasised above, none of the current health data aggregation geolocation parameters is helpful in the study of air quality and health and, in addition, when data are sparse, the time metric for aggregation can be too broad (e.g. cases aggregated by location or by month).

But in the proposed scheme the addition to the current situation that is needed to make the original microdata useful without compromising privacy is the introduction of an intermediary step, which can translate an individual's geographical identifiers in the original microdata into one or more geographical identifiers that are relevant to the research questions under investigation. The means by which this is done is a *linking file,* or *a key*, which translates one geographical identifier into another.

A trusted third party could be engaged to perform the data transformation and would need a legal basis on which to access, process and link the data, as well as technological competence, infrastructure and procedures for safely handling sensitive data, and also the agreement of the data controller to receive the data. Currently, in Ireland the CSO is an institution with the expertise, legal basis and public trust to perform this function. Furthermore, the CSO has in place procedures and infrastructure for routinely publishing aggregate statistics tables safely via the CSO's Statbank/PxStat using existing aggregation schemes. In this proposal, the only departure from the existing situation is that the scheme proposed here would involve an aggregate unit relevant to air quality exposure. The appropriate level of aggregation may

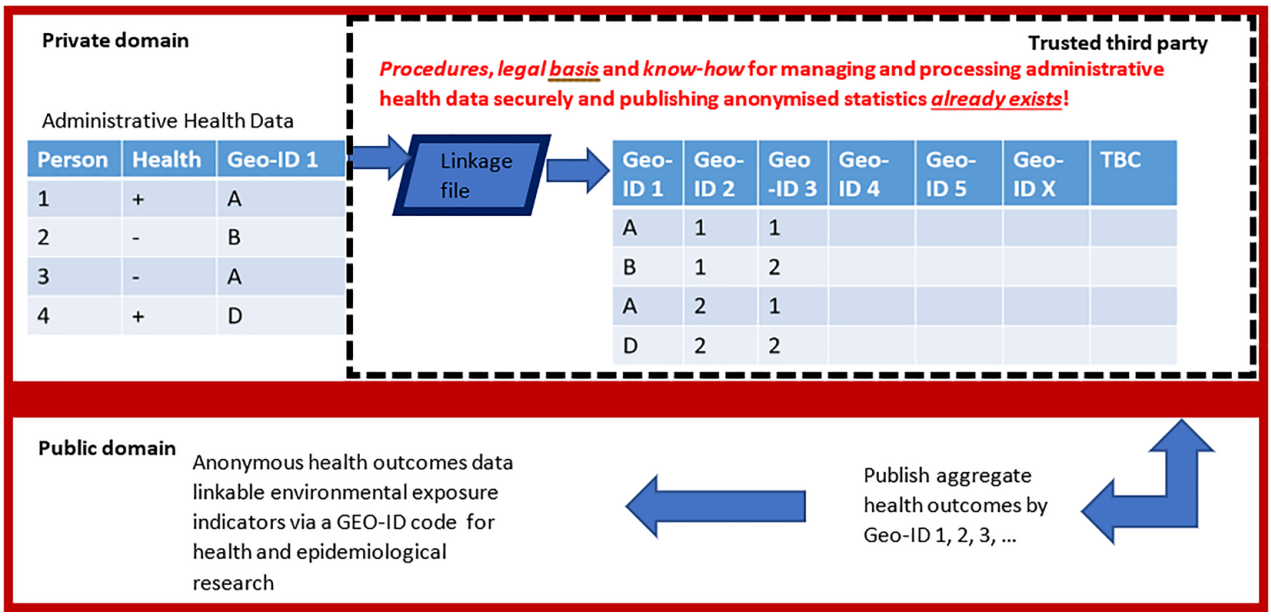

Figure 4.1. Proposed scheme for publishing health data in a format that is meaningfully linkable to air quality data without compromising privacy.

be high, such as air quality zones or geographical restrictions on sales of smoky coal, air quality index regions, proximity to roads or even proximity to radon levels across the country.

The advantages of this approach include minimum investment, since no additional infrastructure would be required. Furthermore, no new legislation would be needed. The only investment needed would be human resources and setting up a new production dataflow within the CSO or eHealth Ireland.

While publication of health-by-air quality aggregate data in a resource like CSO's PxStat, the EPA's SAFER-Data or the eHealth Ireland Open Portal is consistent with these organisations' mandates to disseminate to the public information relating to the state and citizenry, this approach presents some limitations for epidemiological research. Although the variability of exposure across individuals *within* these aggregation units is considerably less than the current aggregation units provided by health organisations (like county or hospital attended), there remains some heterogeneity. On the other hand, data cannot be disseminated at a level of spatiotemporal granularity that reinstates the risk of re-identifying individuals. In addition, there is a large and growing number of diseases and other social outcomes of interest to researchers that are potentially linked to air quality.

4.3.2 Researcher-led aggregation

In section 4.2, a proposal for publishing aggregate health data by a meaningful geocode was discussed. The lynchpin for transforming sensitive, identifiable health data to data useable for environmental epidemiological research is a linking file. A linking file can be dynamic and have a level of granularity suitable for research and can also be public.

The INHALE project linking file is based on the Eircode directory (2021), and the prototype version presented in this report is augmented in a process that involves geographic information system (GIS)-based information, air quality network details and information from the census. Every Eircode, electoral district or small area contains readings for the range of species monitored and has been categorised by the same criteria as the air quality monitoring locations, using the definitions that are given in the CAFE Directive. As well as the Eircode and classification of the location, the file includes information about the small area and the electoral district, such as population, households, cars, roads, heating types and the names and coordinates of the three nearest air quality monitoring stations, for a given year. Because the air quality network and the Eircode database are continuously expanding, the file should be updated to the most recent year of health data when used. Likewise, census data can be updated as they become available.

Examples of how the linkage file can be used for research-led aggregation

Consider the research question "Is there a higher risk of low birth weight if mothers are exposed to, say, PM_{10} in the first trimester?". Let's assume that birth weight data are available for every birth in Ireland over the past 10 years. The researcher first must decide how to attribute PM_{10} exposure to every household in the linking file for each week of the past 10 years. For the sake of simplicity in the example, we will not add work location to the estimation of PM_{10} exposure.

The researcher may decide to use location type as the aggregation unit. Each Eircode sits in one of four location macroscales: rural, rural background, suburban or urban (in future versions of the file, this can be refined further to specify roadside, inner city, agricultural, etc.). These geoidentifiers are almost static for each address across the study duration. The researcher creates a file that contains the Eircode and location type for the data controller. The data controller merges this linking file with the health data by the Eircode (or other geoidentifying) variable. The data controller may then send the researcher a file containing only the week and year of birth, birth weight and location type for each birth or, at the most discrete, a 52 (week) by 10 (year) by 8 record file of low birth weight (yes/no) by four location types, with the frequency of births. The researcher can now analyse the frequency of low birth weight by location type or can merge back to the de-identified file average (or peak) PM_{10} for each location type for each of the weeks over the past 10 years. The researcher now has a near-continuous variable of weekly PM_{10} to predict low birth weight and fewer instances of cases being removed due to *minimum frequency* threshold, such as five events per aggregation unit (Linehan and Dineen, 2024).

Researchers are continually improving estimated individual exposure using models for air quality. *How* to model or estimate weekly $PM_{2.5}$ or PM_{10} exposure at all Eircodes, small areas or electoral divisions is not part of the current project. However, we can explicitly state that, regardless of how a researcher decides to estimate exposure, once the linking file is prepared, the health data can be aggregated on a variable relevant to the researcher.

For example, a researcher might use a model to estimate average $PM_{2.5}$ and ozone for every Eircode (or for every small area and therefore for every Eircode in that area) every day for the past 3 years. Next, within each day the researcher calculates quintiles or cut-points for $PM_{2.5}$ and ozone. There are now 25 possible combinations of exposure (e.g. on a given day some percentage of households are in the highest group for $PM_{2.5}$ and for ozone). The researcher also believes that deprivation level is an important variable in their analysis so also links one of five deprivation levels to each household, giving 125 aggregation units. The researcher can now send a file of all Eircodes, each with 1 of 125 geocodes ($PM_{2.5}$ × ozone concentration by deprivation level) to the health data owner. The health data are merged by Eircode with the air quality file. Then the health data controller can aggregate the health outcome on the geocode and return this file to the researcher. Also known to the researcher is the total number of people residing in each aggregation unit on a given day (via census data), and therefore the incidence of cases is now calculable (number of cases divided by number of people) for each category of $PM_{2.5}$ concentration, ozone concentration, deprivation level and any combination of the three.

An advantage of this approach is that the researcher can tailor the file to meet their research question, whether the exposure is long or medium term or acute. Additionally, the dynamic nature of the approach means that the researcher can derive solutions in consultation with the data provider in instances where, depending on numbers, there could still be a potential identification issue if there are many aggregate units resulting in unique combinations of the group of variables. (Because maps of air quality concentrations are often publicly available it could be argued that it is theoretically possible to infer location from a combination of pollution variables and other data.) The potential barrier is the willingness of and investment from health data controllers to create aggregate data on these dynamic aggregates.

5 Conclusions

In this project, data repositories in Ireland relevant to air quality and health research, and methods for safe, secure, researcher data linkage were reviewed. The INHALE project team concluded that there are many data sources that could be exploited to understand the effect of air quality on the Irish population and to perform innovative, world-class environmental epidemiology.

There is an abundance of excellently curated air quality data, both historical and contemporary, collected by the EPA. Data are well organised and accessible to researchers. Health data are rich but difficult to navigate and access in forms that allow interrogation of pertinent research questions.

The team also notes that there exist infrastructural limitations, such as fragmentation of health data and restricted use of the unique health identifier, and some established practices of health data aggregation that actively inhibit research. However, the infrastructure and expertise, in particular within the CSO, already exist to perform individual-level data linkage. The team recommends that this expertise is built upon to create a new centralised research data TRE with streamlined access protocols or to resource the CSO to expand its services.

In the absence of these resources, the team has developed an innovative researcher-led aggregation approach. A linkage file with every Eircode tagged with aggregation variables that are based on the research question is created by the researcher. Then, the data controller returns aggregate data on this unit, rather than on patients' county or service division, such as their hospital.

Lastly, the INHALE team strongly recommends that the EPA is at the table for future discussions of TREs and data linkage because of the organisation's wealth of experience with big data and because air quality, environment and climate are going to remain leading causes of morbidity and mortality for the foreseeable future.

References

ADR NI (Administrative Data Research Northern Ireland), 2022. *ADR Northern Ireland Strategy 2022–2026*. ADR Northern Ireland, Belfast.

Becker S, *et al.*, 2005. Seasonal variations in air pollution particle-induced inflammatory mediator release and oxidative stress. *Environmental Health Perspectives* 113:1032–1038.

Bekkar B, Pacheco S, Basu R and DeNicola N, 2020. Association of air pollution and heat exposure with preterm birth, low birth weight, and stillbirth in the US: a systematic review. *JAMA Network Open* 3, e208243. https://doi.org/10.1001/jamanetworkopen.2020.8243

Broderick B, *et al.*, 2015. *PALM: A Personal Activity–Location Model of Exposure to Air Pollution*. Environmental Protection Agency, Johnstown Castle, Ireland.

Chafe Z, *et al.*, 2015. *Residential Heating with Wood or Coal: Health Impacts and Policy Options in Europe and North America*. World Health Organization Regional Office for Europe, Copenhagen.

Clancy L, Goodman P, Sinclair H and Dockery DW, 2002. Effect of air-pollution control on death rates in Dublin, Ireland: an intervention study. *Lancet* 360:1210–1214.

Dadvand P, *et al.*, 2014. Residential proximity to major roads and term low birth weight: the roles of air pollution, heat, noise, and road-adjacent trees. *Epidemiology* 25:518–525.

Dockery DW, *et al.*, 2013. *Effect of Air Pollution Control on Mortality and Hospital Admissions in Ireland*. Health Effects Institute Report No. 176. Heath Effects Institute, Boston, MA.

EEA (European Environment Agency), 2016. *Air Quality in Europe – 2016 Report*. *EAA Technical Report No*. 28/2016. EEA, Copenhagen.

EEA (European Environment Agency), 2021. *Health Impacts of Air Pollution in Europe*. Available online: https://www.eea.europa.eu/publications/air-quality-in-europe-2021/health-impacts-of-air-pollution

EPA (Environmental Protection Agency), 2021. *Air Quality in Ireland 2021*. EPA, Johnstown Castle, Ireland.

Fennelly O, *et al.*, 2022. DASSL “Data Access Sharing Storage & Linkage” proof-of-concept: health and related data linkage in Ireland. *International Journal of Population Science* 7:1908. https://doi.org/10.23889/ijpds.v7i3.1908

Freid RD, *et al*., 2021. Proximity to major roads and risks of childhood recurrent wheeze and asthma in a severe bronchiolitis cohort. *International Journal of Environmental Research and Public Health* 18:4197. https://doi.org/10.3390/ijerph18084197

Fu L, *et al.*, 2019. The associations of air pollution exposure during pregnancy with fetal growth and anthropometric measurements at birth: a systematic review and meta-analysis. *Environmental Science and Pollution Research International* 26:20137–20147.

Fuks KB, *et al.*, 2016. Long-term exposure to ambient air pollution and traffic noise and incident hypertension in seven cohorts of the European study of cohorts for air pollution effects (ESCAPE). *European Heart Journal* 38:983–990.

Health Service Executive, 2022. *National Service Plan 2022*. Health Service Executive, Dublin. Available online: https://www.hse.ie/eng/services/publications/serviceplans/hse-national-service-plan-2022.pdf

Hetland RB, *et al.*, 2005. Cytokine release from alveolar macrophages exposed to ambient particulate matter: heterogeneity in relation to size, city and season. *Particle and Fibre Toxicology* 2:4. https://doi.org/10.1186/1743-8977-2-4

IARC (International Agency for Research on Cancer), 2013. *Air pollution and Cancer*. IARC Scientific Publication No. 161. IARC, Lyon, France.

Jahanshahi B, *et al*., 2024. Prenatal exposure to particulate matter and infant birth outcomes: evidence from a population-wide database. *Health Economics* 33(9):2182–2200. https://doi.org/10.1002/hec.4862

Kavouras IG and Chalbot MG, 2016. Influence of ambient temperature on the heterogeneity of ambient fine particle chemical composition and disease prevalence. *International Journal of Environmental Health Research* 27:27–39.

Keogh A, *et al*., 2024. Breaking down the digital fortress: the unseen challenges in healthcare technology – lessons learned from 10 years of research. *Sensors* 24(12):3780. https://doi.org/10.3390/s24123780

Linehan T and Dineen K, 2024. *CSO Best Practice for Statistical Disclosure Control of Tabular Data*. Central Statistics Office, Cork, Ireland. Available online: https://www.cso.ie/en/aboutus/lgdp/csodatapolicies/dataforresearchers/resourcesforresearchers/

Mahalingaiah S, *et al.*, 2016. Adult air pollution exposure and risk of infertility in the Nurses' Health Study II. *Human Reproduction* 31:638–647.

McNabola A, Broderick BM and Gill LW, 2008. Relative exposure to fine particulate matter and VOCs between transport microenvironments in Dublin: personal exposure and uptake. *Atmospheric Environment* 42:6496–6512.

Mizen A, *et al.*, 2018. Creating individual level air pollution exposures in an anonymised data safe haven: a platform for evaluating impact on educational attainment. *International Journal of Population Data Science* 3:412. https://doi.org/10.23889/ijpds.v3i1.412

Mizen A, *et al.*, 2020. Impact of air pollution on educational attainment for respiratory health treated students: a cross sectional data linkage study. *Health & Place* 63:102355. https://doi.org/10.1016/j.healthplace.2020.102355

Moran R, 2016. *Proposals for an Enabling Data Environment for Health and Related Research in Ireland*. Health Research Board, Dublin.

O'Reilly D, Bateson O, McGreevy G, Snoddy C and Power T, 2020. Administrative Data Research Northern Ireland (ADR NI). *International Journal of Population Data Science* 4(2):1148. https://doi.org/10.23889/ijpds.v4i2.1148

Pedersen M, *et al.*, 2013. Ambient air pollution and low birthweight: a European cohort study (ESCAPE). *Lancet Respiratory Medicine* 1:695–704.

Qin P, *et al.*, 2021. Long-term association of ambient air pollution and hypertension in adults and in children: a systematic review and meta-analysis. *Science of The Total Environment* 796:148620. https://doi.org/10.1016/j.scitotenv.2021.148620

Roberts AL, *et al.*, 2013. Perinatal air pollutant exposures and autism spectrum disorder in the children of Nurses' Health Study II participants. *Environmental Health Perspectives* 121:978–984.

Vizcaíno MAC, González-Comadran M and Jacquemin B, 2016. Outdoor air pollution and human infertility: a systematic review. *Fertility and Sterility* 106:897–904.

Walsh B, Mac Domhnaill C and Mohan G, 2021. *Developments in Healthcare Information Systems in Ireland and Internationally*. ESRI Survey and Statistical Report Series No. 105. Economic and Social Research Institute, Dublin.

WHO (World Health Organization), 2016. *Ambient Air Pollution: A Global Assessment of Exposure and Burden of Disease*. Available online: https://apps.who.int/iris/handle/10665/250141

WHO (World Health Organization), 2021. *WHO Global Air Quality Guidelines: Particulate Matter ($PM_{2.5}$ and PM_{10}), Ozone, Nitrogen Dioxide, Sulfur Dioxide and Carbon Monoxide*. Available online: https://apps.who.int/iris/handle/10665/345329

WHO (World Health Organization), 2024. Fact sheet: ambient (outdoor) air quality and health. Available online: https://www.who.int/en/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health

Xiang AH, *et al*., 2025. Discordant sibling analysis of autism risk associated with prenatal exposure to tailpipe and non-tailpipe particulate matter pollution. Environmental Research 275:121449. https://doi.org/10.1016/j.envres.2025.121449

Yuchi W, Sbihi H, Davies H, Tamburic L and Brauer M, 2020. Road proximity, air pollution, noise, green space and neurologic disease incidence: a population-based cohort study. *Environmental Health* 19:8. https://doi.org/10.1186/s12940-020-0565-4

Appendix 1

The INHALE synthetic linking file is based on the Eircode directory (2021), and the prototype version presented in this report is augmented in a process that involves GIS-based information, air quality network details, CORINE (Coordination of Information on the Environment) land cover data for Ireland (2018) and information from the Irish Census of Population (2016). In this file, every Eircode, electoral district or small area, and the names and coordinates of the three nearest air quality monitoring stations for a given year are included, such that data on air quality across time can also in used as an aggregate.

In the example below, we found that some Eircodes in one administrative file (Property Price Register database) were not in the 2021 Eircode databases. Because the air quality network, the Eircode database and the population are continuously expanding, the file should be updated to the most recent year of health data when used. Naturally, location type classifications that are modelled in the linkage file using geographical, economic and demographic information to generate a judgement as to whether any given Eircode is in a rural, urban or suburban location and assign it to an air quality zone is work that would be in need of regular refinement and updating due to demographic/economic developments and changes in land use.

A.1 Prototype Linking File Coverage

For each Eircode in the linking file, the distance to each of the air quality stations in the monitoring network can be calculated from coordinates, and the nearest station can be assigned to individuals based on their address of residence.

Aggregation based on the location type or modelled exposure, rather than the distance, is preferable because the distance between an Eircode and the nearest monitoring location can traverse different location types with very different air quality profiles.

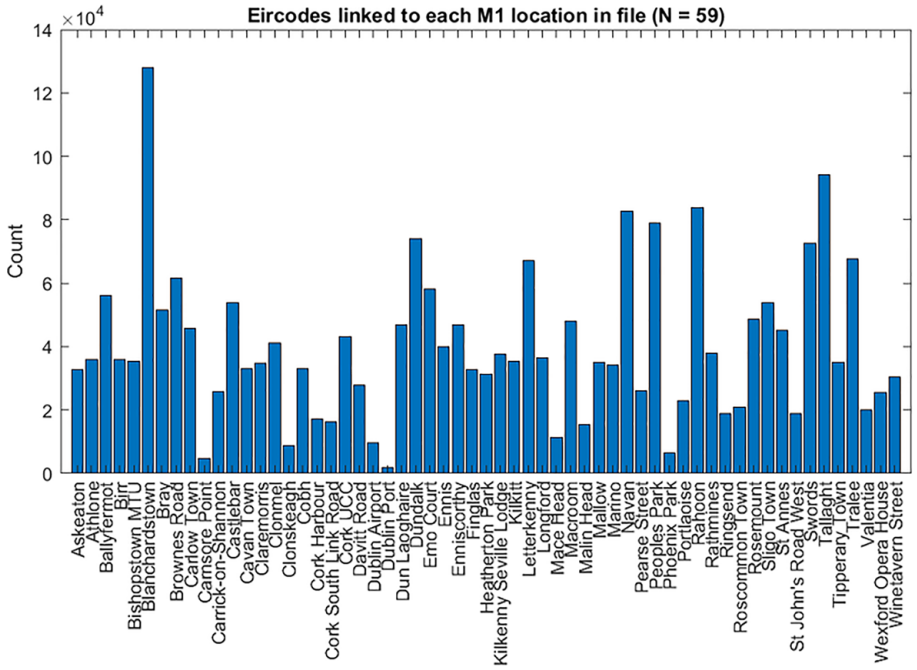

Figure A1.1. Number of Eircodes per nearest monitoring station (*n* = 59). MTU, Munster Technological University; UCC, University College Cork.

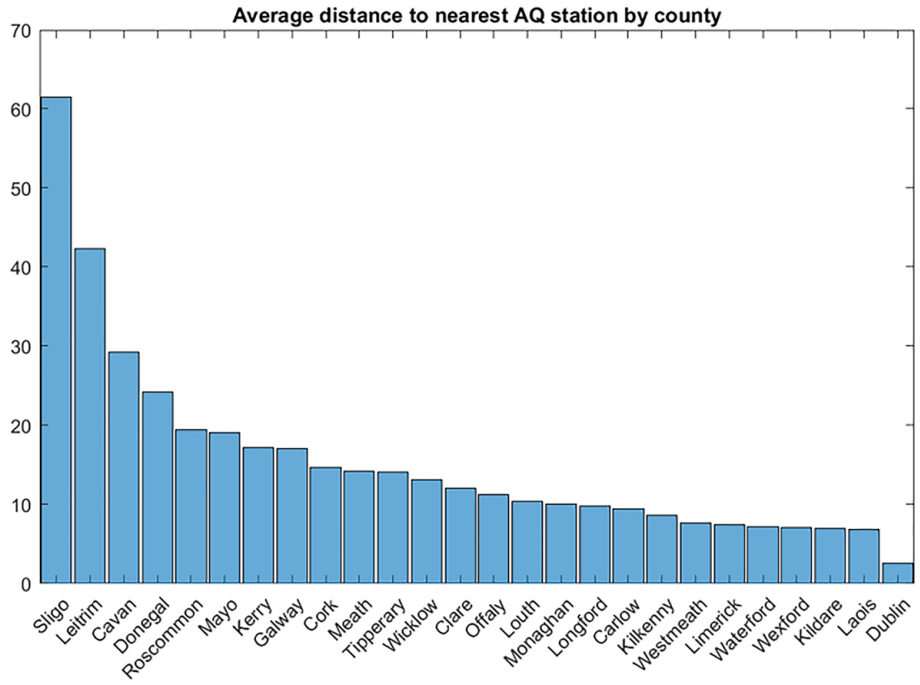

Figure A1.2. Average distance from each Eircode to the nearest monitoring station within the county. AQ, air quality.

A.2 Trial of Linkage Workflow Using Location Type as the Aggregation Variable

As outlined in the main report, individual identifying data cannot be easily accessed or used by researchers. The INHALE team recommends a research-led aggregation approach using a linking file.

Step 1: Health data owner prepares data file of case status and patient Eircode (this might also include the date of the health event). To trial the prototype file, a synthetic health data file was created. The 2017 Property Price Register (a file that contains more than 50,000 sales with associated Eircodes) was used to generate a health file that spanned Ireland and included each of the location types or air quality zones. One researcher generated a randomly assigned case status (yes/no) to each Eircode ($n = 1500$ cases). The date of sale was used as the date of the health event. The linkage trial, therefore, will not produce a meaningful result and is merely used to illustrate the process of linking the cases/health outcomes with air quality information on an aggregate variable.

Step 2: The researcher prepares a linkage file that contains the aggregation variable and the associated Eircodes. Using the linking file, we created a sub-file that contained all Eircodes and a modelled location type.

Step 3: The health data owner merges the linkage sub-file to health data. In this step, in the trial, the sub-file that contained *all* Eircodes and location types is merged with the synthetic health data file, by Eircode. Now the synthetic health data file has a location subtype for each "case".

Step 4: The health data owner aggregates the health data. The synthetic health data are now collapsed on the aggregate variable and a count of cases per location type is returned to the researcher. The synthetic health data could also be returned as a file of dates of events per location type which could be used to examine the health impact of short-term deterioration in air quality. (Aggregation on location type is illustrative and more detailed contrast in exposure would be required using the same principles and steps.)

Step 5: The aggregated data are returned to the researcher. In the above example, there will be four aggregate types of environment and the number of cases in each will now be returned to the researcher, as illustrated in Figure A1.3, without any identifiable information, but crucially without loss or identification of cases.

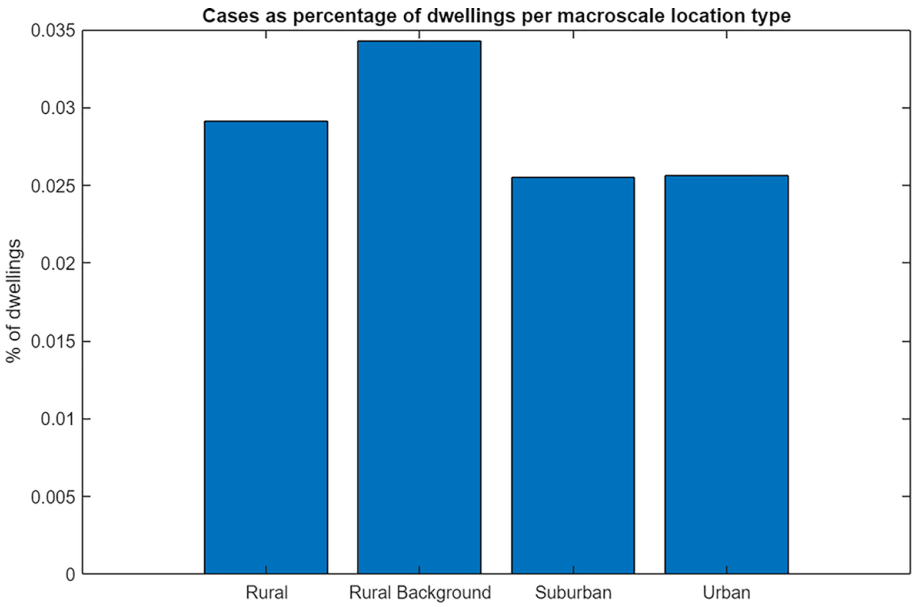

Figure A1.3. Cases in each aggregate location type using the synthetic health dataset.

In Figure A1.3, the total number of cases has been compressed to four location types (instead of county or hospital, for example). However, the proposed scheme allows for an infinite number of permutations and stratifications to inform the research questions, while the data owner remains in full control of the identifiable information in the health data.

The researcher can now explore the anonymised data in the context of their research question and generate visualisations and examine associations as required.

For example, Figure A1.4 shows the linkage of health outcomes and air quality by location type and time of event generated using health data received on a researcher-defined aggregation (location type and month). Note that the health data are synthetic and, as expected, there is no meaningful pattern to observe.

In this process, both the researcher and the data owner achieve their respective objectives. The data owner protects the privacy of the subjects in their data as no identifiable information is provided to the researcher. The researcher is able to harness the full power of the available data because they are able to extract the information using 100% of the cases and not just a subset, but without receiving any identifiable information and therefore without compromising privacy rules, while aggregating patients on a relevant exposure variable.

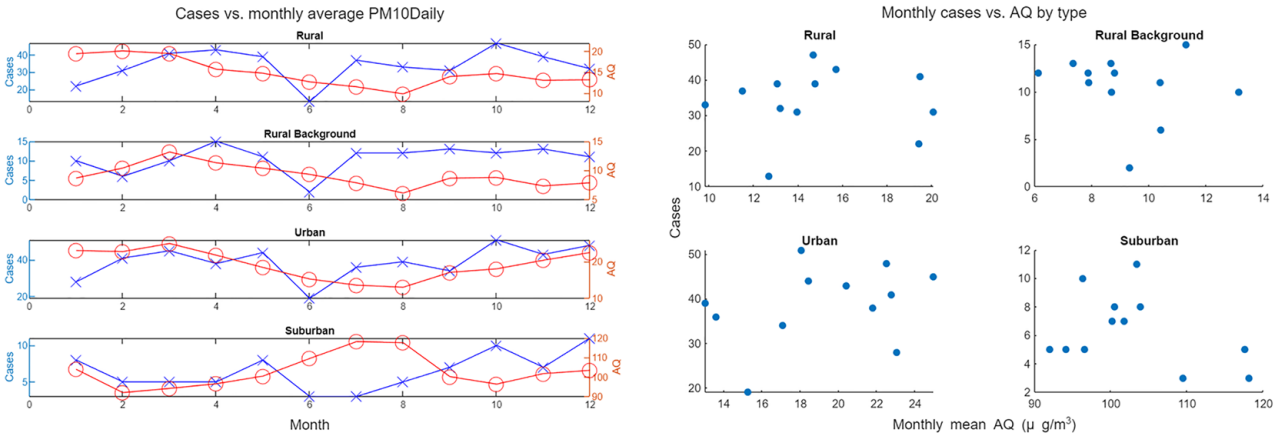

Figure A1.4. Visualisation of the number of cases each month relative to the air quality level using the synthetic health dataset. AQ, air quality.

Importantly, the researcher remains in control of the research question because the researcher defines the aggregate type. The data owner remains in control of the information published because they perform the aggregation and will therefore be in a position to ensure that no potentially identifiable information is returned.

In the context of the INHALE project, this workflow achieves one of the main objectives of the project, which was to find a way to operationalise air quality and health linkage in a workflow that enables regular publication of statistical estimates of the impact of air quality on population health. By regularly updating the linkage file, which in itself contains no private information, and merging it with administrative health data in a standardised workflow, annual estimates can be easily produced and used as basis for policy development and policy evaluation.

Abbreviations

AAMP	Ambient Air Quality Monitoring Programme
ADR NI	Administrative Data Research Northern Ireland
ADR UK	Administrative Data Research United Kingdom
ALF	Anonymous linking field
BSO	Business Services Organisation
CAFE	Clean Air for Europe
CORTEX	Cognitive Development, Respiratory Tract Illness and Effects of Exposure
CSO	Central Statistics Office
DASSL	Data access, storage, sharing and linking
EPA	Environmental Protection Agency
GIS	Geographic information system
HBS	Honest Broker Service
HIPE	Hospital In-patient Enquiry
HRB	Health Research Board
HSC	Health and Social Care Board
HSE	Health Service Executive
IHI	Individual health identifier
INHALE	Irish Nationwide Health and Air Quality Linkage
ISSDA	Irish Social Science Data Archive
IT	Information technology
PCRS	Primary Care Reimbursement Service
PM	Particulate matter
SAFER-Data	Secure Archive for Environmental Research Data
SAIL	Secure Anonymised Information Linkage
TRE	Trusted research environment
WHO	World Health Organization

An Ghníomhaireacht Um Chaomhnú Comhshaoil

Tá an GCC freagrach as an gcomhshaol a chosaint agus a fheabhsú, mar shócmhainn luachmhar do mhuintir na hÉireann. Táimid tiomanta do dhaoine agus don chomhshaol a chosaint ar thionchar díobhálach na radaíochta agus an truaillithe.

Is féidir obair na Gníomhaireachta a roinnt ina trí phríomhréimse:

Rialáil: Rialáil agus córais chomhlíonta comhshaoil éifeachtacha a chur i bhfeidhm, chun dea-thorthaí comhshaoil a bhaint amach agus díriú orthu siúd nach mbíonn ag cloí leo.

Eolas: Sonraí, eolas agus measúnú ardchaighdeáin, spriocdhírithe agus tráthúil a chur ar fáil i leith an chomhshaoil chun bonn eolais a chur faoin gcinnteoireacht.

Abhcóideacht: Ag obair le daoine eile ar son timpeallachta glaine, táirgiúla agus dea-chosanta agus ar son cleachtas inbhuanaithe i dtaobh an chomhshaoil.

I measc ár gcuid freagrachtaí tá:

Ceadúnú

- Gníomhaíochtaí tionscail, dramhaíola agus stórála peitril ar scála mór;
- Sceitheadh fuíolluisce uirbigh;
- Úsáid shrianta agus scaoileadh rialaithe Orgánach Géinmhodhnaithe;
- Foinsí radaíochta ianúcháin;
- Astaíochtaí gás ceaptha teasa ó thionscal agus ón eitlíocht trí Scéim an AE um Thrádáil Astaíochtaí.

Forfheidhmiú Náisiúnta i leith Cúrsaí Comhshaoil

- Iniúchadh agus cigireacht ar shaoráidí a bhfuil ceadúnas acu ón GCC;
- Cur i bhfeidhm an dea-chleachtais a stiúradh i ngníomhaíochtaí agus i saoráidí rialáilte;
- Maoirseacht a dhéanamh ar fhreagrachtaí an údaráis áitiúil as cosaint an chomhshaoil;
- Caighdeán an uisce óil phoiblí a rialáil agus údaruithe um sceitheadh fuíolluisce uirbigh a fhorfheidhmiú
- Caighdeán an uisce óil phoiblí agus phríobháidigh a mheasúnú agus tuairisciú air;
- Comhordú a dhéanamh ar líonra d'eagraíochtaí seirbhíse poiblí chun tacú le gníomhú i gcoinne coireachta comhshaoil;
- An dlí a chur orthu siúd a bhriseann dlí an chomhshaoil agus a dhéanann dochar don chomhshaol.

Bainistíocht Dramhaíola agus Ceimiceáin sa Chomhshaol

- Rialacháin dramhaíola a chur i bhfeidhm agus a fhorfheidhmiú lena n-áirítear saincheisteanna forfheidhmithe náisiúnta;
- Staitisticí dramhaíola náisiúnta a ullmhú agus a fhoilsiú chomh maith leis an bPlean Náisiúnta um Bainistíocht Dramhaíola Guaisí;
- An Clár Náisiúnta um Chosc Dramhaíola a fhorbairt agus a chur i bhfeidhm;
- Reachtaíocht ar rialú ceimiceán sa timpeallacht a chur i bhfeidhm agus tuairisciú ar an reachtaíocht sin.

Bainistíocht Uisce

- Plé le struchtúir náisiúnta agus réigiúnacha rialachais agus oibriúcháin chun an Chreat-treoir Uisce a chur i bhfeidhm;
- Monatóireacht, measúnú agus tuairisciú a dhéanamh ar chaighdeán aibhneacha, lochanna, uiscí idirchreasa agus cósta, uiscí snámha agus screamhuisce chomh maith le tomhas ar leibhéil uisce agus sreabhadh abhann.

Eolaíocht Aeráide & Athrú Aeráide

- Fardail agus réamh-mheastacháin a fhoilsiú um astaíochtaí gás ceaptha teasa na hÉireann;
- Rúnaíocht a chur ar fáil don Chomhairle Chomhairleach ar Athrú Aeráide agus tacaíocht a thabhairt don Idirphlé Náisiúnta ar Ghníomhú ar son na hAeráide;
- Tacú le gníomhaíochtaí forbartha Náisiúnta, AE agus NA um Eolaíocht agus Beartas Aeráide.

Monatóireacht & Measúnú ar an gComhshaol

- Córais náisiúnta um monatóireacht an chomhshaoil a cheapadh agus a chur i bhfeidhm: teicneolaíocht, bainistíocht sonraí, anailís agus réamhaisnéisiú;
- Tuairiscí ar Staid Thimpeallacht na hÉireann agus ar Tháscairí a chur ar fáil;
- Monatóireacht a dhéanamh ar chaighdeán an aeir agus Treoir an AE i leith Aeir Ghlain don Eoraip a chur i bhfeidhm chomh maith leis an gCoinbhinsiún ar Aerthruailliú Fadraoin Trasteorann, agus an Treoir i leith na Teorann Náisiúnta Astaíochtaí;
- Maoirseacht a dhéanamh ar chur i bhfeidhm na Treorach i leith Torainn Timpeallachta;
- Measúnú a dhéanamh ar thionchar pleananna agus clár beartaithe ar chomhshaol na hÉireann.

Taighde agus Forbairt Comhshaoil

- Comhordú a dhéanamh ar ghníomhaíochtaí taighde comhshaoil agus iad a mhaoiniú chun brú a aithint, bonn eolais a chur faoin mbeartas agus réitigh a chur ar fáil;
- Comhoibriú le gníomhaíocht náisiúnta agus AE um thaighde comhshaoil.

Cosaint Raideolaíoch

- Monatóireacht a dhéanamh ar leibhéil radaíochta agus nochtadh an phobail do radaíocht ianúcháin agus do réimsí leictreamaighnéadacha a mheas;
- Cabhrú le pleananna náisiúnta a fhorbairt le haghaidh éigeandálaí ag eascairt as taismí núicléacha;
- Monatóireacht a dhéanamh ar fhorbairtí thar lear a bhaineann le saoráidí núicléacha agus leis an tsábháilteacht raideolaíochta;
- Sainseirbhísí um chosaint ar an radaíocht a sholáthar, nó maoirsiú a dhéanamh ar sholáthar na seirbhísí sin.

Treoir, Ardú Feasachta agus Faisnéis Inrochtana

- Tuairisciú, comhairle agus treoir neamhspleách, fianaise-bhunaithe a chur ar fáil don Rialtas, don tionscal agus don phobal ar ábhair maidir le cosaint comhshaoil agus raideolaíoch;
- An nasc idir sláinte agus folláine, an geilleagar agus timpeallacht ghlan a chur chun cinn;
- Feasacht comhshaoil a chur chun cinn lena n-áirítear tacú le hiompraíocht um éifeachtúlacht acmhainní agus aistriú aeráide;
- Tástáil radóin a chur chun cinn i dtithe agus in ionaid oibre agus feabhsúchán a mholadh áit is gá.

Comhpháirtíocht agus Líonrú

- Oibriú le gníomhaireachtaí idirnáisiúnta agus náisiúnta, údaráis réigiúnacha agus áitiúla, eagraíochtaí neamhrialtais, comhlachtaí ionadaíocha agus ranna rialtais chun cosaint chomhshaoil agus raideolaíoch a chur ar fáil, chomh maith le taighde, comhordú agus cinnteoireacht bunaithe ar an eolaíocht.

Bainistíocht agus struchtúr na Gníomhaireachta um Chaomhnú Comhshaoil

Tá an GCC á bainistiú ag Bord lánaimseartha, ar a bhfuil Ard-Stiúrthóir agus cúigear Stiúrthóir. Déantar an obair ar fud cúig cinn d'Oifigí:

1. An Oifig um Inbhunaitheacht i leith Cúrsaí Comhshaoil
2. An Oifig Forfheidhmithe i leith Cúrsaí Comhshaoil
3. An Oifig um Fhianaise agus Measúnú
4. An Oifig um Chosaint ar Radaíocht agus Monatóireacht Comhshaoil
5. An Oifig Cumarsáide agus Seirbhísí Corparáideacha

Tugann coistí comhairleacha cabhair don Ghníomhaireacht agus tagann siad le chéile go rialta le plé a dhéanamh ar ábhair imní agus le comhairle a chur ar an mBord.

EPA Research

Webpages: www.epa.ie/our-services/research/
LinkedIn: www.linkedin.com/showcase/eparesearch/
Twitter: @EPAResearchNews
Email: research@epa.ie