Accepted for publication in Studies in History and Philosophy of Science in April 2022

Reframing the Environment in Data-Intensive Health Sciences

Stefano Canali

Department of Electronics, Information and Bioengineering and META - Social Sciences and Humanities for Science and Technology

Politecnico di Milano

stefano.canali@polimi.it

ORCID: 0000-0002-5948-3874

Sabina Leonelli

Exeter Centre for the Study of the Life Sciences (Egenis)

University of Exeter

S.Leonelli@exeter.ac.uk

ORCID: 0000-0002-7815-6609

Abstract

In this paper, we analyse the relation between the use of environmental data in contemporary health sciences and related conceptualisations and operationalisations of the notion of environment. We consider three case studies that exemplify a different selection of environmental data and mode of data integration in data-intensive epidemiology. We argue that the diversification of data sources, their increase in scale and scope, and the application of novel analytic tools have brought about three significant conceptual shifts. First, we discuss the EXPOsOMICS project, an attempt to integrate genomic and environmental data which suggests a reframing of the *boundaries between external and internal environments*. Second, we explore the MEDMI platform, whose efforts to combine health, environmental and climate data instantiate a *reframing and expansion of environmental exposure*. Third, we illustrate how extracting epidemiological insights from extensive social data collected by the CIDACS institute yields innovative *attributions of causal power to environmental factors*. Identifying these shifts highlights the benefits and opportunities of new environmental data, as well as the challenges that such tools bring to understanding and fostering health. It also emphasises the constraints that data selection and accessibility pose to scientific imagination, including how researchers frame key concepts in health-related research.

Keywords: Epidemiology; Big Data; Environment; Exposure.

1. Introduction

It might seem trivial to say that the environment has an impact on population health, and yet traditionally epidemiologists have focused their investigations largely on features of environments that were seen to interact most directly with populations, such as sources of nutrition and housing conditions (Rappaport & Smith, 2010). In recent years, biomedical researchers that study health and disease at the population level have become interested in the crucial role played by broader environmental factors in affecting the development of disease, such as climate, landscape, and socio-economic conditions – an awareness that has yielded conceptual, methodological, and material changes to disciplines such as epidemiology and public health.[1] In epidemiology in particular, this increasing awareness has led to the development of new framings of the concept of health. Notions such as 'global health', 'one health', 'planetary health' have dominated epidemiological discourse in recent years, each advocating for a specific framing of what counts as environment and how it relates to human health. Planetary health has encouraged a more explicit focus on the properties of the whole *physical* environment that populations interact with, including climate and local ecosystems (Horton et al., 2014); one health has emphasised taking account of *multispecies* environments in order to understand co-dependences between human and non-human populations (Gibbs, 2014); and global health has framed health as a result of the needs that different populations experience in *individual and regional* environments, thus stressing the diverse yet interconnected conditions for life around the world (Brown et al., 2006). These expansive conceptualisations of health, which share similar political and economic backgrounds and the backing of national and transnational institutions, suggest a very broad understanding of the scope and scale of environmental risk to humans (Gaudilliere & Gasnier, 2020). At the same time, the emergence of new measurement capabilities such as molecular markers has prompted a renewed and growing emphasis on the effects of environmental exposure at different scales on individual physiology and behaviour (Landecker, 2011; Shostak, 2013). The continuing tensions between population-level and individual-focused approaches signal how the concept of environment, so widely used as an overarching notion and direction of research for contemporary epidemiology and health science, can actually refer to different objects and translate into widely diverse methods and practices of inquiry.[2]

In this article, we are interested in the recent evolution of studies on the relations between environmental exposure and population health and in particular in the conceptual reformulations of the notion of environment in the health sciences. We discuss three case studies to highlight how the use of new and heterogeneous data sources on the environment has led to novel ways of integrating health and environmental data as well as a reframing of the notion of the environment away from reductive approaches privileging molecular data over other sources of evidence. New types of interdisciplinary collaborations are emerging with every-growing abilities to generate, integrate, and analyse data documenting many different aspects of life on the planet, ranging from molecules to climates, to better understand – and intervene in – population health, despite the heterogeneous sources and methods through which such data are generated. These newly expansive and increasingly inclusive data-intensive methods and tools, particularly in projects centred on the integration of data from multiple sources and concerning multiple phenomena, are

[1] Throughout the article we are following (S. Valles, 2020) and using the concept of biomedical science as the "umbrella theoretical framework for most health science and health technology work done in academic and government settings".

[2] For a historical tour de force across some of the multiple and changing meanings associated to the notion of environment within the last two centuries of scientific research, see (Benson, 2020).

in turn affecting researchers' conceptualisation of the environment in the biomedical domain, prompting a dynamic understanding of its significance for health and disease. We identify these conceptual changes and discuss the implications of such changes on the practice, goals, and future scope of epidemiological research.

The last two decades have witnessed a large increase in the volume, diversity, value, veracity, and volatility of data of potential relevance to health research (Hogle, 2016; Leonelli, 2016), due to the rise of data-intensive research as a central model of scientific investigation (Leonelli, 2016) as well as the expansive digitalisation of health-related information powered by internet usage, remote sensing technologies, and personal health tracking (Sharon & Lucivero, 2019; Prainsack, 2020). Epidemiologists have long been concerned with the collection and analysis of large datasets (Morabia, 2004). Yet the availability of new sources of data and analytic tools is often presented as a significant novelty (Holmberg et al., 2013). At the interface of environment and health, this availability affects environmental findings that are brought to bear on the study of population health, for instance by fostering data linkage across vast data collections and reliance on new sources of evidence such as social media, personal health applications, and monitoring devices (Fleming et al., 2017; Hogle, 2016). Extensive literature in data studies has investigated the epistemic, political, and economical role of data in the sciences, discussing the ways in which attempts to link and integrate highly heterogeneous data sources can create new forms of interdisciplinary dialogue, which may breach the fragmented epistemic cultures of the biomedical and environmental sciences and produce innovative results (Leonelli, 2013; Pietsch, 2015; Ratti, 2015). Relying on this work, we ask how the availability of new types of data and related methods has affected views of the environment in population health. Building on insights from data studies, we argue that novel data practices are promoting hitherto unexplored forms of dialogue between biomedical and environmental fields, enabling conceptual and methodological transfers that enrich and expand existing assumptions around the role and characteristics of the environment of relevance to health.

By detailing changes to the notion of environment involved in the use of new sources of data in epidemiology, we also ask which benefits and limitations are connected with such attempts at using data relating to the environmental, social, and molecular spheres. In this way we contribute to the critical literature on discussions on genomics and postgenomics and the limited focus on the environment therein (Shostak, 2013; Richardson & Stevens, 2015; Gibbon et al., 2020). One of the results of the end of the Human Genome Project, in the early 2000s, was the discovery that environmental factors played a more prominent role, in determining human health and disease, than the focus on genetic sequences presupposed (Hilgartner, 2017). This kickstarted discussions on the need to move beyond gene-centrism and enter a 'postgenomic era', where more attention would need to be paid to environmental and external factors of health and disease – to the point that some have called for a conceptualisation of the genome in environmental terms (Rheinberger et al., 2017). Various philosophers, historians, and sociologists have discussed the differences between genomics and postgenomics, the extent to which there is anything new in postgenomics, the continuities between the geneticization of research of the 1990s, and 2000s and the putative shift towards genomic contexts in the 2010s (Barnes & Dupré, 2008; Gibbon et al., 2020; Richardson & Stevens, 2015; Shostak, 2013). In particular, we build on well-documented arguments that depict postgenomic fields – such as, quintessentially, epigenetics – as having defined the environment in relation to sources of environmental exposure that affect organismal development (Landecker & Panofsky, 2013; Shostak & Moinester, 2015). We ask whether the integration of environmental data in epidemiology can reframe the interrogation of what counts as relevant environment beyond reductive approaches and understandings.

To answer this question, in the paper we focus on the ways in which the environment is conceptualised and operationalised in current epidemiology, as an entry point into the analysis of the impact, benefits, and challenges of data-intensive approaches in biomedicine as well as the role of environment as a foundational concept of epidemiological research. We identify three shifts in how the environment is conceptualised by data-intensive approaches to epidemiology, each of which is exemplified through a case study of existing attempts to diversify and increase the use of environmental data as epidemiological evidence. These conceptual shifts are: (1) a novel understanding of the relation between external and internal exposure and thus the location of processes and phenomena of interest in external or internal environments; (2) a reframing of the notion of environmental exposure to expand the focus on internal and individual phenomena and the extent to which these underpin the linkage of data produced by social media, personal health monitoring devices, remote sensing technologies and social/medical services; and (3) an innovative approach to attributing causal power to environmental factors, tied to the degree of resolution and understandings of location facilitated by access to – and computational analysis of – large volumes of digitalised socio-economic data. These shifts involve an expansion of our focus throughout our analysis: first starting from more biological approaches to health, which are benefitting from the increasing volume and tractability of molecular data extracted from individuals and populations; then moving on to studies focusing on the relation between human health and climate, thus building an improved understanding of the role of landscapes in public health; and finally examining work that links climatic and health data with newly emerging socio-economic data, hence providing insight into how specific social as well as environmental conditions may foster or prevent disease. Our conclusion is that despite continuing challenges in the required multidisciplinary dialogue, the use and integration of new environmental data for epidemiology is playing a decisive role in fostering the integration of insights from climate and environmental research beyond existing reductionist leanings.

To ground our discussion in specific research practices, we consider three case studies that in our view exemplify these shifts: (1) EXPOsOMICS, a consortium based at Imperial College London, that run between 2012 and 2017 on the basis of funding from the EU Commission (Vineis et al., 2017); (2) the Medical and Environmental Data Mash-up Infrastructure (MEDMI), which was run between 2013 and 2019 by several leading UK organisations in climate, weather, environment and human health including and funded by the UK Medical Research Council and NERC; (3) the "100 Million Brazilians" cohort, one of the largest cohorts in the world focused on a low-income population, based at the Centre for Data and Knowledge Integration for Health (CIDACS) in Salvador, Bahia-Brazil. For each of these cases, we identify changes to the evidential basis of epidemiological research in connection to an increase of the diversification of relevant sources of data, the scale and scope of the data, as well as the emergence of novel techniques to quantify the impact on environmental factors on health and disease.

The paper is structured as follows. In Section 2, we briefly review the role that environmental data have played in the history of population health and introduce the significant role of diverse concepts of exposure in shaping biomedical, epidemiological and public health approaches. In Section 3, we discuss newly emerged work on the exposome as an example of the shifting boundaries between internal and external environment in more biomedically-focused research. To exemplify this, we consider the case of the EXPOsOMICS project, which applied the exposome approach to the study of chronic disease in ways that exemplify the use of genomic data as a platform for the linkage of biological and (mostly climatic) environmental data and the study of exposure. In Section 4 we expand our focus beyond molecular approaches, to consider epidemiological research that attempts to combine climate and health data to improve and potentially transform the existing understanding of environmental exposure and its implications for human populations. Our case for this strand of research is MEDMI (https://www.data-mashup.org.uk/), which was devoted to the creation of "data mash-ups" bringing medical records together with climate environmental data and exemplifies the attempt to model spatial and temporal patterns of environmental exposure and its effects on health with high predictive accuracy. In Section 5 we then turn to population studies attempting to consider not only the wealth of biomedical and climate data now available to health researchers, but also the increasing availability of socio-economic data coming from extensive cohort studies and related socio-political interventions, and the impact of such influx of data and related analytic and multidisciplinary methods on existing understandings of environmental factors as causes. Our example for this work is the computational analysis of the 100 Million Brazilians cohort, whose study exemplifies the changing opportunities for causal inference generated by the availability of big data and effective forms of analysis on both a population and its environment. Much of this work is carried out at CIDACS, which was launched in 2016 to ensure the secure and reliable storage, handling, and analysis of sensitive data of potential relevance to public health. In Section 6, we interrogate and discuss the scope of and intersections among the shifts we identified and highlight the extent to which each of the empirical cases we discussed is affected by all three shifts. To close, in the final section of the paper we discuss the benefits and opportunities provided by new sources of environmental data as well as the challenges and issues faced by data-intensive epidemiology.

Our analysis is empirically grounded on reviews of scientific publications, reports and presentations resulting from the projects used as our case studies; research visits to the project sites carried out between 2016 and 2019 (including at the Department of Epidemiology and Biostatistics of Imperial College London in 2017, where EXPOsOMICS was based; the campuses of the University of Exeter where MEDMI was based in 2016-2018; and the CIDACS headquarters in Salvador in 2019), during which we had the opportunity to engage in participant observation, take part in project meetings, and visit the related facilities; and conversations with researchers involved in the project, which ranged from informal collaborative exchanges in the context of preparing joint grant applications to in-depth, semi-structed, qualitative interviews conducted as part of the authors' projects [DETAILS ANONYMISED].

2. How data practices relate to understandings of the environment in the study of human health

While research on population health in epidemiology often involves a study of the environment, we start our analysis with the observation that this study has not typically engaged with direct sampling of environmental features, specific focus on environmental factors, and discussion of their active influence on health. Historically the development and success of epidemiology, as the area of biomedical research that directly focuses on the distribution and determinants of disease and health in populations (Broadbent, 2013), has been deeply connected to the management of public health through environmental interventions, for example with the sanitary revolution and successful management of cholera epidemics through sewage and water supply systems widely recognised as a crucial reference point (Morabia, 2004). Here a way to conceptualise the environment has been as an external and indirect source of exposure and disease risk for individuals and populations. As a consequence, in epidemiology conceptualisations of the environment are deeply connected to what is investigated as exposure to a wide range of different phenomena, which may range from air and water pollution to poverty and levels of education, from occupational settings to dietary conditions, and all the way to the internal chemical features of the body. In other words, the environment is conceptualised relationally, by defining a target object (a landscape, a population, a genome) in relation to the features of its surroundings that are most likely to affect its functions and future behaviour.[3] This corresponds to variations within epidemiology itself, with fields such as environmental, social, and occupational epidemiology focusing on vast environments such as ecosystems, society, and the workplace respectively; while clinical and genetic epidemiology remain mostly concerned with the human body as key environment for investigation, thus supporting a much narrower view of the environment and of what it is an environment of (Rogawski et al., 2016).

What is most notable for our purposes is the tight relation between these disparate conceptions of the environment and the ways in which exposure is measured, documented, and reported – in other words, the data practices and related expertise surrounding the study of environmental exposure, including sampling, selection, storage, and linkage of data of potential relevance (Warde et al., 2018). A crucial aspect of this relation is that the epidemiological study of the environment through the lens of exposure has traditionally led to few interactions with environmental data collected outside the realm of biomedical research, such as data on landscape or climate. Consider, for instance, the wealth of clinical studies which over the last three decades have collected and analysed biological data about molecular mechanisms of disease, genomic markers, and related associations: here the study of population health has focused on individual behaviours and lifestyle choices, rather than the broader socio-ecological settings of different populations (Boniolo & Nathan, 2017). At the other end of the spectrum, many studies focusing on climate and environmental data have prioritised existing techniques to measure pollution, such as the analysis of air quality and its relation to population health – thus again paying little attention to socio-

[3] Benson notes how understanding the notion of environment relationally means accepting that "the history of the concept of environment and the diverse environmentalisms associated with it is also a history of the emergence of surrounded entities, and of how various groups of people have imagined their ideal relationship to their surroundings" (Benson, 2020, p. 13).

ecological dimensions beyond what such data could document (Shostak, 2013). As a result, the methods used to analyse the relation between population health and environmental exposure in epidemiology have traditionally been developed in independent ways from the experimental methods, modelling, and simulations employed in environmental research. Data practices in epidemiology have rather prioritised the collection of observational, biological, and biomedical data on exposure at the individual and population levels, their assemblage in specific configurations, and the study of cohorts over extended periods of time (Bauer, 2008; Morabia, 2004).

Therefore, in epidemiology the study of the environment and population health through the focus on exposure and use of population data has not typically led to the direct study of the environment. For instance, in molecular epidemiology, exposure profiles based on molecular data are used to study the potential exposure of a population to environmental toxicants such as pollutants (Russo & Vineis, 2016). In most cases, no direct sampling of environmental pollutants is conducted and data about the internal biochemical environment are used as a proxy for variables tracking specific features and changes in the environment. A more direct focus on the environment is also made difficult by differences between the types of data collected for environmental and health research, including varied time scales and frequencies (Fleming et al., 2017). While epidemiologists have traditionally been mostly interested in the monitoring and surveillance of population health, climate and environmental scientists have developed tools and methods for the estimate and prediction of climate events, which are not easily redeployed for biomedical research (Parker, 2018). In addition, epidemiological methods usually employ regression approaches that focus on few exposure factors for small groups and single health outcomes, which is problematic when trying to address the overall impact of the environment on health.

This fragmentation and lack of intellectual and social links between research fields involved in environmental research, combined with broader changes affecting the evidential basis of scientific and biomedical research, explain why new attempts to the use and integration of climate and weather data into the epidemiological study of population health are so significant. Epidemiological research and the study of the environment have been distant, and this is connected to important differences at the conceptual, methodological, and disciplinary level. This is why attempts to go beyond these differences by using new and heterogeneous types of environmental data and developing novel ways to integrate them for epidemiology offer a window through which epistemological changes for both biomedical and environmental research can be analysed. We argue that new uses and integration of environmental data in the study of population health have led to re-formulating what count as environment and exposure, conceptually and methodologically, as shown by the three cases we discuss in the remainder of the article.[4]

[4] These and other shifts we identify in epidemiology have run in parallel with similar changes in environmental research, where there has been increasingly more work on the impact of climate changes on population health and interdisciplinary collaborations to integrate environmental and biomedical data (Fleming et al., 2017).

3. Environment and the exposome: Reframing boundaries between external and internal in the EXPOSOMICS project

Considering the centrality of exposure to the study of the environment in epidemiology, we start our discussion of the relation between environmental data and conceptualisations of the environment with the case of the exposome, a new conceptualisation of exposure that was introduced to include and quantify all the different levels of internal and external exposure (Wild, 2005). In the last decade, the exposome has been proposed as a novel approach to conducting epidemiological research, in connection with more focus on environmental determinants of health and disease, the use and integration of molecular and environmental data, the establishment of dedicated funding streams and research institutes (Canali, 2020a), and – as we argue – a reframing of the boundaries between internal and external environments.

The exposome has several connections to discussions on the role of the environment for population health. Among the first proponents of the exposome, one of the rationales for the introduction of the concept was based on the need to focus more on the causal role of the environment in the determination of health and disease. For instance, in a crucial publication for the (re)introduction of the exposome (Siroux et al., 2016), Stephen Rappaport and Martyn Smith explicitly called out that "70 to 90% of disease risks are probably due to differences in environments" and, yet, epidemiologists paid little attention to the environment and mostly relied on gene-based solutions (Rappaport & Smith, 2010, p. 460). The exposome was introduced precisely to shift this focus to the environment and in particular to approach exposure in broader, more comprehensive, and specific ways. In particular, the exposome approach to these issues has been to try and improve the precision of measurements of the environmental impact on population health by including new sources of quantitative data, particularly genomic and climatic data. This has implied moving closer to the methodology and evidential basis of environmental and climate sciences, while also incorporating results from molecular studies. This is why several exposome projects have established collaborations with communities in geography, information science, and genomics to transfer and integrate their data practices. For example, the EXPOsOMICS project established and coordinated a consortium of 13 research centres in Europe and the US including experts in genomics, information systems, personal exposure monitoring, etc. EXPOsOMICS has been among the most prominent exposome projects in Europe, one in a series of projects coordinated by the team at Imperial College London, and was funded by Horizon 2020, the funding programme of the EU Commission. The project focused in particular on the application of the exposome approach on the study of the relations between exposure to air and water pollution and disease risk (see an exemplary study in (Fiorito et al., 2018)).

The exposome approach to the use of environmental data for the study of population health, as exemplified in EXPOsOMICS, has been grounded on the conceptual, methodological, data expansion of the boundaries between internal and external environments and related types of exposures. This expansion has included a new conceptualisation of the internal environment as a source of *internal exposure*, using genomic data as evidence; a new framing of the external environment as a source of both *generic and specific external exposure*, based on the integration of climatic and geographic data; and a new understanding of the strictness and continuity between boundaries and environments, which are interpreted as largely open and flexible categories.

At the internal level, this expansion has focused on extending the developments of sequencing and genomic projects into epidemiology, in particular omics approaches. Omics are techniques used to study and quantify molecules and processes within the cell, such as RNA-transcription, metabolism, proteins, etc. In EXPOsOMICS, omics techniques were used to develop 'exposure profiles' that measure the presence of molecules or processes that can be connected to exposures to environmental elements. For example, on the basis of blood samples collected in longitudinal studies, omics analyses were run to measure adducts that can form between human albumin and pollution toxicants (Canali, 2019). As a result, exposure profiles based on molecular data were used as evidence of the *internal* environment of the human body – and yet, it was clear to EXPOsOMICS researchers that omics data could also be interpreted as evidence of external exposure and environmental processes. For example, exposure profiles were used to study responses to external exposure in the internal processes of the cell, such as inflammation and oxidative stress.

At the other end of the spectrum, the use of omics data for internal exposure in exposome research has elicited the need for the collection and employment of climatic data on external exposure, to match the level of specificity and resolution of omics data. One of the teams in EXPOsOMICS developed Geographical Information Systems (GIS), which used environmental sampling to estimate the specific amount of pollutant that an individual could have been exposed to during a study period (Gulliver et al., 2018). In the project GIS data were analysed as evidence of features and processes in the *external* environment, with a focus on the very specific environment that would have surrounded the specific subjects of molecular analysis, for instance at the level of particular air matter. Thus, data collected from GIS, omics, and other, more traditional epidemiological methods such as longitudinal studies, were analysed in EXPOsOMICS through univariate and multivariate methods in regression models, to study the relation between exposure and disease at different spatial and conceptual levels.[5] For example, this approach was used to study the relation between air pollution and cardiovascular and cerebrovascular disease (Fiorito et al., 2018) by developing regression models that integrate one individual exposure profile and one individual omics feature at a time and look for the strongest statistical associations between omics features and external sources of exposure (Canali, 2020b).

EXPOsOMICS thus exemplifies the use of specific new types of environmental data, which are integrated with new approaches from epidemiology. Conceptually, exposome researchers have approached this way of linking biological and environmental data for the study of population level as a way of bridging gaps between internal and external analyses of the impact of the environment on health. In particular, EXPOsOMICS researchers ordered omics, GIS, and other types of data as ways of quantifying three levels of exposure and interrelated types of environments: generic external exposure (e.g. social capital, education, financial status); specific external exposure (e.g. radiation, infectious agents, chemical pollutants); and internal exposure (e.g. metabolism,

[5] This use of *genomic* data and analytics as a platform to integrate environmental data in EXPOsOMICS is also mirrored in the idea of conducting Exposome-Wide Association Studies, a methodology that quantifies and analyses the totality of exposure and its link to disease risk in similar ways to the Genome-Wide Association Studies (Vineis et al., 2017).

endogenous hormones, physical activity). In this way, the new use and integration of data-intensive sources of data such as omics and GIS has led to new interpretations of divisions between these types of exposure and, more broadly, the distinction between what count as internal and external environments. In particular, the increase in diversity, scale, and scope of exposure data and their linkage in regression models are pushing for new interpretations of exposure as a *continuous* process that is not limited to the external environment and spatially happens at the intersection of different types of environments, to the point that distinctions between types are largely flexible and dependent on the goals and interests of a specific study.[6] This is a significant conceptual formulation in comparison with traditional epidemiology, where as we have seen the environment is considered an indirect source of exposure as an *external* element to which individuals, population, and bodies are exposed to – with the result that boundaries between external and internal are clearly fixed and coincide with boundaries between the body and the environment.

The availability of environmental data at more scales and their use as evidence for exposure have led to re-interpretations of these boundaries: since exposure can be measured at different levels and can happen internally, the environment as a source of exposure can be interpreted as external to very different entities, including populations and bodies but also foetuses and metabolism. For instance, the molecular processes analysed on the basis of omics techniques are clearly internal to the human body, but they are also considered as direct continuations of external processes such as exposure to air and water pollution. In similar terms, phenomena such as pollution are clearly external to individuals and populations, but what is interesting for population health and studied by exposome researchers is their concrete impact on the specific environment surrounding individuals and as such these processes can be interpreted as pertaining to both the internal and external environments. While regression models of the type used in EXPOsOMICS assign either internal or external levels to exposure, through the ordering and modelling in terms of omics and exposure profiles, the processes under study are conceptualised as concerning both levels. The study of omics is in this sense intended to capture the continuity of these processes and in particular the internal component of processes elicited at external levels.

Challenging simplistic distinctions between internal and external in these ways is a significant step beyond the traditional understanding of fixed boundaries between internal and external environments and has enabled exposome researchers to operationalise environmental data as evidence of different types of exposure and environments.[7] The approach to data integration of exposome research and EXPOsOMICS is thus an attempt to reframing the environment in terms that are not only genetic or molecular and thus reductive (Shostak & Moinester, 2015). As a result, in exposome research the notion of environment is used to discuss entities at very different levels of analysis and has been considered to include anything that is non-genetic (Wild, 2008), what is relevant to a specific analysis (Rappaport, 2011), as well as the maternal body (Robinson & Vrijheid,

[6] Many thanks to an anonymous referee for pushing us to revise our use of the notion of continuity, which here is intended to capture the continuity between processes of exposure at the internal and external level, and thus not in terms of the contrast between discrete and continuous or interpretations being continuous.

[7] As one anonymous reviewer helpfully suggested, this can be considered a “transgressive interpretation”, changing the definition and strictness of the boundaries between internal and external environments.

2015). This juxtaposition between the external and internal dimensions of the analysis runs parallel with interpretations in continuous terms of the processes that population health studies intend to capture. In this sense, one of the main goals of the exposome approach in EXPOsOMICS was indeed to trace the development of disease as moving through stages and levels that are difficult to ascribe clearly to either an external or internal level.[8] Our analysis of the EXPOsOMICS project shows an approach to integrating environmental and health data that is based on the use of climatic and molecular data and the reframing of the conceptual boundaries between internal and external environments.

4. Environment and the semantics of exposure: Data mash-ups in MEDMI

While the biomedical turn towards data-intensive gene-environment analysis has clearly affected how researchers understand the role of the environment vis-à-vis the human body, this is by no means the only direction from which change has come to the health sciences. Another important movement has been towards integrating climate and health data on an unprecedented scale, through the semi-automated linkage and analysis of very large volumes of heterogenous data. In this section, we argue that this attempt has shifted the very semantics of what researchers mean by exposure, expanding the boundaries of what are considered as relevant aspects of the environment to include information about climate (e.g. data about altitude, temperature, humidity, rainfall and related weather conditions) and territory (such as population density, residence information and characteristics of urban, rural and coastal landscapes).

This expansion is exemplified by the idea of "data mash-ups", which has recently acquired popularity as a technical approach to managing big data integration from multiple and heterogeneous sources, with particular relevance to the domains of biomedicine, climate and environmental science and with the aim of informing research on social and environmental challenges that require an interdisciplinary knowledge base. A data mash-up consists of the methods used to process, mix, and analyse different types of data to produce a unified and unique output which can be potentially more useful than and accessed independently of the original individual datasets (Fleming et al., 2017). This data integration is feasible not only thanks to increasing data volumes, but also and most importantly thanks to novel – and often open source – software packages, cloud computing specifically, and interoperability standards designed to manage geospatial and temporal data used as reference points for calibration and triangulation (Leonelli & Tempini, 2021). In turn, the development of such novel technologies necessitates input from experts on the specifics of each data source and potential application, which involves the establishment of new forms of collaboration across relevant domains of expertise, including data, health, climate, and environmental science.

The MEDMI project was one of the first to attempt to merge climate and territory data made available by the MET Office (the main meteorological research institute in the UK). The main goal

[8] It should be noted that this remains one of the tensions of the exposome approach: the focus on internal exposure and conceptualisations of exposure as a process crossing boundaries exists vis-à-vis the reliance on more discrete and possibly reductive tools such as omics. See our discussion in Sect. 6 for more on reductionism.

was to integrate data garnered from patient groups by hospitals and general practitioners in specific regions of the UK with information about the biology of pathogens known to be harmful to humans (their life cycle, nutrition requirements and physiology, as relevant to estimate the conditions under which the pathogen is most likely to be damaging to human hosts, see (Fleming et al., 2014)). Just as in the case of EXPOSOMICS, the early applications of MEDMI were also focused on the effects of air pollution on the spread and impact of pathogens on human populations, such as in studies of the seasonality of respiratory diseases like asthma and hay fever. Rather than focusing on molecular data, however, MEDMI exemplifies yet another way of using and integrating new sources of environmental data for epidemiology. MEDMI paid the most attention to the intersection between medical data extracted from hospitalisations and samples collected from patients, and broader features of the climate and territory, such as shifts in humidity levels, temperature, and wind speed/direction, as well as the degree of urbanisation of the landscape and the types of vegetation in the affected areas. To this aim, MEDMI put together a data infrastructure that facilitates access to – and joint analysis of – climate, territory, and health data, which in turn required years of consultations among different types of experts concerning the choice and/or development of the right formats, metadata, software and analytic tools to be associated with this effort (Fleming et al., 2014). This resulted in a fine-grained understanding of the environmental factors that could be most reliably correlated with the local population being most strongly affected. In turn, this translated into predictive models for the climatic conditions under which medical services could expect an uptick in the incidence of respiratory disease (Djennad et al., 2018, 2019).

To understand just how transformative such research has been, consider that in biomedicine exposure has long been characterised as the proximity and/or contact with something that might transmit disease or other outcomes of interest and has been measured by quantifying the extent to which an individual or a population are exposed to a specific factor. One of the main conceptual and methodological challenges for the study of exposure in epidemiology is that individuals are exposed to endless sources of exposure throughout a lifetime, yet it is hard to trace and measure several types of exposure at the same time, and hence to investigate the cumulative impact of these factors on health. For example, it seems trivial to say that features of the climate and territory are sources of exposure for a population. Still, when these are considered together, the intertwined effects of these types of exposure are difficult to quantify and thus in epidemiology they been studied mostly from an internalist perspective, as we have seen with exposure profiles in the exposome. The increasing availability of precise and local environmental data on these features of environments have the potential to revolutionise this area of study. Yet this requires methodological and technological innovation as well as the capacity to cross disciplinary, conceptual, and technical boundaries in order to facilitate communication across the different epistemic cultures that generate the data and relevant analytic tools. The promise of projects such as MEDMI is precisely to enable such cross-disciplinary alignment and link local, environmental data in the study of population health, thus making it possible to consult and use such data as a single body of evidence. One important strategy to achieve this in MEDMI has been the choice of treating some data types – such as geolocation data – as invariant parameters through which highly heterogeneous data could be compared and integrated. This work exemplified the extensive labour required to make diverse datasets compatible with each other, typically by exploring their respective histories and devising ways of linking them without losing sight of their specific characteristics and provenance (Leonelli & Tempini, 2021). The progressive refinement of such techniques is transforming the scope of data mash-ups, making it possible to devise studies that integrate across medical and climate data to enhance understandings of population health at specific locations and times of the year (Fleming et al., 2017).

As a consequence of this use and integration of environmental data for population health in data mash-ups, exemplified by MEDMI, the epidemiological notion of exposure has become a broader framing for the joint consideration of new internal and external processes. In this context, exposure can of course still be conceptualised as the contact between specific individuals and their surroundings, at single points in time. Yet data mash-ups and the use of environmental data to study both health and environment at once are instigating a view of exposure as a process that evolves throughout the life course and can therefore be treated as a proxy to study elements of interest depending on the specific study. For example, one of the traditional principles of exposure sciences has been the need to identify a specific pathogen to model as cause of disease. New conceptualisations of exposure based on the use of environmental data are pushing for more dynamic views, where exposure is continuously defined through the specific and unfolding relations between hosts, pathogens, and relevant elements of their surroundings. Pathogens are considered in relation to their complex microbial and organismal ecologies, which are in turn highly susceptible to broader climatic and environmental conditions. Which aspect of the environment is most relevant to understanding host-pathogen interactions, and even the very characteristics and causal role of the pathogen, can and does vary over time; and approaches such as data mash-ups provide novel avenues for researchers to align information documenting such changes in great detail and at different levels of resolution. New forms of data linkage are thus moving away from the idea of obligate pathogen, which are modelled as causes of disease no matter the circumstances, and towards a more dynamic understanding of health and disease as functions of the relation between organisms, their microbiome, and their surroundings.[9] Our analysis of the MEDMI project shows a specific and new way of using and integrating environmental data for epidemiology, which reframes the role of exposure for notions of the environment.

5. Environment, sociality, and causality: Linking health and social data at CIDACS

When considering such a broad nexus of environmental factors that could be regarded as sources of exposure, an obvious epistemic question that emerges is how to determine the causal significance of such diverse factors, as well as their respective causal role in relation to the effects being studied. And indeed, causality continues to be at the centre of theoretical discussions in epidemiology and population health: particularly in the age of big data, where so much emphasis has been placed on correlations as potential sources of causal understanding (Broadbent, 2015;

[9] The connections between more traditional approaches of disease aetiology, such as germ theory, and these forms of data linkage require more work, which however we do not have space to do in this article. See work from Lauren Ross and James Woodward on disease aetiology in the context of discussions of causal specificity (Ross & Woodward, 2016; Ross, 2018).

Pietsch, 2015; P. Illari & Russo, 2016), and in epidemiology, where explicit causal claims and terms such as cause and effect are often avoided (Russo, 2009). In this section, we consider how conceptions of causality – and particularly of what may constitute a causal factor – have shifted in relation to data-intensive practices for the use and integration of environmental data, as well as related changes in conceptualising exposure that we discussed in the previous sections.

The availability of new and different sources of data on the environment, together with new methods of data analysis and linkage, has led to a shift in population thinking, whereby environmental factors are attributed a specific and distinct causal role. This has been further strengthened through the introduction of novel, extensive sources of longitudinal socio-economic data on specific populations, which can then be combined with environmental and biomedical data to produce ever sharper analyses and innovative approaches to causal inference. As an example, we consider the epidemiological studies carried out on the incredibly rich dataset derived from the "100 Million Brazilians" cohort study, which includes over 114 million low-income Brazilian citizens enrolled in the Unified Registry for Social Programmes (CadUnico) since 2003. This Brazilian government registry, which serves as a primary data source for CIDACS, routinely collects data on citizens wishing to obtain access to social protection programmes such as Bolsa Familia. In order to access such programmes, individuals are required to provide up-to-date data on a regular basis, resulting in a large repository of high-quality, reliable longitudinal information on several socio-economic and demographic parameters including family units, education levels, employment history, housing conditions, and access to social and medical services. Data derived from the 100 Million Brazilians cohort has been linked to other governmental databases containing information on deaths and births, the incidence of infectious diseases, nutritional status, etc. These data are of the highest possible sensitivity given their confidential nature and the potential they have to damage data subjects and their communities in case of breaches. The data are therefore anonymised and securely stored by the CIDACS centre, which is also the key Brazilian institution with responsibility for regulating access to the data and conducting a wealth of studies on their basis (Almeida et al., 2018).[10]

Given the breadth and scope of these data, the 100 Million Brazilians cohort arguably constitutes one of the most extensive longitudinal data repositories ever assembled, and a precious resource to investigate the long-term implications of specific policies and socio-environmental conditions for the health of the poorest members of the population in a setting plagued by large social inequalities. Unsurprisingly, given the new opportunities offered by data integration strategies and the availability of extensive environmental datasets, as already discussed in previous sections, CIDACS has become increasingly interested in developing methods to link such administrative data to

[10] The security conditions at CIDACS are worthy of a paper in and of themselves, given the care and attention devoted by the CIDACS team to ensuring that no breach of privacy and confidentiality occurs – especially given the ongoing political tensions in the country as well as the ease with which these data, given their comprehensiveness, could lead to subject re-identification and discrimination against communities and minorities. In the words of CIDACS staff, "the minimal risk is already too high" (pers. comm, 2019). The scrutiny associated to data access and re-use is therefore extensive and detailed, with procedures aimed at monitoring the goals, methods, and results of those authorised to work with the data (Almeida et al., 2018).

environmental as well as biomedical (including genomic) data, which in turn involve creating strategies to foster multidisciplinary work. Hence CIDACS has formed extensive collaborations with other specialist centres around the world to devise in-house algorithms that can increase the accuracy and scalability of data linkage and modelling over the cohort data, as well as pools of experts from diverse disciplines to help evaluate the reliability and plausibility of results given existing knowledge of the social contexts in question. Both algorithms and multidisciplinary collaborations are seen as crucial ways to improve the reliability of related causal inferences, particularly for the purpose of informing policy (Harron et al., 2017; Pita et al., 2018).[11] Practitioners have argued that these data linkage studies and interdisciplinary collaborations are indispensable to obtaining the kind of granular understanding of population segments required by precision medicine, and that this is in turn increasing the robustness and accuracy of causal inferences from big data (Barreto & Rodrigues, 2018). This is not solely a matter of data volume, though the richness and comprehensiveness of the 100 Million Brazilian cohort certainly plays an important epistemic role. The emphasis is on how data are managed once they are stored on the CIDACS servers, including which forms of data governance are used to regulate the conditions for data use, including access, visualisation, analysis and even the goals of research.

There is a strong interest, for instance, in identifying conditions which may play a causal role in creating or reproducing social inequalities, with a view to support specific interventions. For example, CIDACS researchers have demonstrated how juxtaposition of data documenting mortality rates by ethnic group with data about the territorial distribution of primary healthcare yields strong evidence for the effectiveness of primary care in offsetting existing racial inequalities in Brazil (Hone et al., 2017). There is also an interest in increasing the evidence base – and particularly the scope and size – of such epidemiological investigations by developing new methods to harmonise and link data from different sources across national borders. An example is a recent study of mortality registries from seven different countries, to which CIDACS participated, which mined data pertaining to 1.7 million individuals to obtain an increased understanding of risk factors (Stringhini et al., 2017). Last but not least, CIDACS is invested in increasing confidence in the causal power of data analysis through ongoing innovation in mathematical modelling tools, whose import and reliability can be expanded to take advantage of the dramatic growth in scale and quality of data available as empirical input. Recent work on the impact of COVID-19 on the population of the Bahia region is a particularly poignant example of the power and speed of deployment of such tools, once they are applied to a well-maintained, high quality dataset (Oliveira et al., 2021).

The ways in which more causal consideration has been given to environmental factors and data in CIDACS are thus based on the acquisition of observational data on various socio-environmental conditions, a sustained focus on the quality and multidisciplinary contexts of data practices, and triangulation with biomedical datasets. This aligns CIDACS with other approaches in epidemiology

[11] This approach is common to other epidemiologists and public health researchers involved in the analysis of cohort studies, including for instance the UK Biobank – a data collection exercise encompassing the longitudinal acquisition of a vast set of biometric measurements and samples (including genetic data) on half a million UK-based participants since 2006, which has been used by over twenty thousand researchers to date to perform a variety of studies, many of which aiming to link the biobank data with other data sources.

and biomedicine, where causal attribution to social, economic, and environmental factors is considered more robust when observational evidence is supported by laboratory studies, clinical trials, meta-analyses, etc. (Clarke et al., 2013)). CIDACS thus exemplifies an additional use of new types of environmental data, which are integrated with new approaches to re-formulate notions of the environment. This is particularly evident in the type of causal inference applied in CIDACS, which resembles 'multifactorial' models of causality that frame disease as resulting from the aggregation of several different causes instead of focusing on specific causes (Vineis, 2003). These models have received significant criticism in the philosophical literature: for instance, according to Alex Broadbent the fact that disease ethology is traceable to several causal factors (e.g many socio-environmental factors) does not mean that epidemiologists should embrace multifactorialism and abandon the strategy of identifying some causes that have a special status (Broadbent, 2013, 2015). In contrast with this position, Sean Valles has argued that multifactorial thinking can be successfully refined and applied in particular to socio-environmental causes of the incidence of disease in a given population (Valles, 2021). This is the direction that CIDACS efforts to triangulate and link environmental data seem to go towards. Rather than aiming to identify a specific cause or focusing on a factor that needs special focus, for instance environmental pollution, data integration in CIDACS has aimed to identify a variety of factors that can, by operating together, have significant causal effects and may thus serve as special loci of policy interventions. This parallels Valle's argument for multifactorialism and a focus on the intersection of several different causes, particularly socio-environmental causes. Even in cases where disease has a central biological cause (e.g. viral infection), other factors and particularly socio-environmental factors crucially influence the development of disease (e.g. developing symptomatic disease after infection, (Valles, 2021). In addition, socio-environmental factors can often lead to multiple disease outcomes, by influencing multiple risk factors and multiple mechanism (Valles, 2019 Chapter 5). Integrating diverse sources of data on socio-environmental factors in this direction, as exemplified by CIDADS, is not necessarily new – multifactorial models have been very influential in epidemiology over the last century (Vineis, 2003) – but constitutes an innovative way of implementing this specific approach to causal inference.[12]

While causal inference is a clear focus of the project, we must note the reluctance of CIDADS staff to even use the language of "causality" to discuss their research. A notable feature of their published work and group discussions is a preference for terms such as "determinants" over the term "cause", which is meant to highlight the uncertain and possibly contingent nature of the correlations being uncovered, and thus the fallible nature of any causal generalization derived from data triangulation (pers. comm. 2019). For example, a recent overview of CIDACS research contributions presents their work as investigating "the social determinants of health and the effects of social and environmental

[12] Notably, CIDACS staff integrates such studies with a methodological reinvention of natural experiments in epidemiology, a fascinating approach which we cannot however cover within the scope of this paper (Pescarini et al., 2020).

policies on different health outcomes" (Barreto et al., 2019).[13] In our analysis, we take on board this carefulness in attributing causal powers to social factors, which follows on a well-established tradition of epidemiological thinking. We however see the reference to "determinants of health" as a clear reference to causal power,[14] especially given that researchers working with CIDACS data typically present their hard-earned insights on such determinants as ground for policy interventions. What is notable in CIDACS work is, on the one hand, the willingness to expand the range of environmental factors considered as potential candidates for causal attributions (hence our previous point on multifactorial models); and on the other hand, to strengthen the confidence with which causal attributions are made in relation to specific determinants and populations. As CIDACS staff put it, their methodological sophistication, newly developed software, data richness and emphasis on working across disciplines, publics, and countries "enable the addition of new exposures or outcomes, the study of outcomes at different times of exposure, including over the long term, and the evaluation of various social protection policies on health outcomes" especially in relation to the poorest populations in Brazil (Barreto et al., 2019).

6. How strands intersect: A closer look at links between our case studies, and their conceptual implications

An obvious question emerging from our analysis is the extent to which the shifts we identified are compartmentalised and contained within the specific cases and areas of health research that we have discussed. Our answer is that, while these shifts are readily apparent in relation to the examples and domains we identified, they are much wider-ranging in scope and may be argued to appear within each of our empirical cases.

As we briefly noted already, the work carried out within MEDMI bears some parallels to the case of exposome research with respect to the reframing of boundaries between external and internal environment. Just as in the EXPOSOMICS project, the possibility of linking medical, environmental, and climate data in MEDMI has enabled a conceptualisation of disease as involving both internal mechanisms in the interactions between pathogens and hosts and the external environments of the pathogen and the host. Even more evidently than in exposome research, what counts as external and internal within data mash-ups ends up depending on the interests, goals, and level of abstraction of specific analyses: depending on how much detail is necessary, processes can be

[13] A well-known subfield devoted to the analysis of administrative data such as those stored by CIDACS is "health technology assessment", which specialises in data-intensive research towards evaluating the effectiveness and implications of health-related innovation (Ali et al., 2019). CIDACS staff and users of their data resources, however, go well beyond this approach, with ongoing projects spanning several areas of epidemiology and biomedicine, and a strong emphasis on supporting multidisciplinary and interdisciplinary research teams (Barreto et al., 2019).

[14] We read this as an instance of the understanding of causality as probabilistic (often interpreted in terms of "difference-making" in philosophical literature (Hitchcock, 2021)), and particularly the ways in which "difference-making helps us discriminate the different effects of multiple or complex causal paths", often in the absence of a mechanistic understanding for the phenomenon under study (P. M. Illari & Russo, 2014, p. 55).

interpreted as internal or external to the interactions under study.[15] A more general consequence of this shift is a reframing of the boundaries between healthy and diseased individuals. In MEDMI, the linkage and integration of data has shaped a dynamic understanding of health and disease, according to which states of health or disease come in degree and are a function of the changing relations between organisms and their environments. A similar dynamic understanding of health and disease is at the basis of the exposome approach, which has been presented as a "highly variable and dynamic entity that evolves throughout the lifetime of the individual" since its initial introduction and as a way of characterising the totality of exposure that individuals experience in their life-course (Wild, 2005, p. 1848). The conceptual basis for these approaches is clearly connected to increases in the diversity, scales, and scopes of data sources, which can be used to move away from single measurements at individual points in a lifetime and towards bio-monitoring that covers a complete lifetime (potentially).

In addition, this diversification of data sources has led to moves beyond the dichotomy between externalist or internalist approaches to the environment in population health. In exposome research, the focus on the external dimension of exposure has been expanded to include different types of external environments as well as an internal component. In EXPOsOMICS, this was achieved with the inclusion of omics techniques, data, and analytics to develop exposure profiles at the individual level. A similar expansion is at the basis of data mash-ups too, although in a opposite direction to EXPOsOMICS: in MEDMI, the external levels of exposure have been studied by combining diverse data relating (also) to territories, climate, and microbiomes. As we have shown, these reformulations of the environment are crucially tied to new uses and the integration of environmental data. In turn, this brings the validity of these shifts to bear on the robustness of the linkage and integration of diverse datasets (Leonelli, 2013). While the inclusion of new perspectives and dynamic understandings beyond dichotomy are promising, we also find significant limitations tied to data integration. For instance, omics data are still difficult to integrate with more traditional exposure data and their use often implies assumptions and values that are not transparent to epidemiologists. This has been an important issue for MEDMI, where the combination and linkage of large datasets was brought to bear on a critical approach towards the limitations, scopes, and assumptions incorporated in the datasets that are being mashed together (Leonelli & Tempini, 2021). The documentation and discussion of contextual features of data has allowed MEDMI researchers to assess correlations and causal relations between biological, climatic, biomedical, and environmental factors and integrate markedly diverse sources of data. The case of data mash-ups is thus a way of dealing with expanding distinctions between boundaries and environments without blurring significant differences in the qualities, limitations, and evidential values of individual datasets.

Another element that runs across our case studies is causal inference, which is a crucial concern well beyond the work of CIDACS. For example, in EXPOsOMICS researchers employed the statistical approach that is known as the Meet-In-The-Middle approach (MITM) to "investigate the temporal

[15] In MEDMI, this interpretation of internal and external dimensions was closely aligned with an idiographic conceptualisation of locality (Leonelli & Tempini, 2021).

sequence of exposure, biological pathway perturbation and disease onset" (Vineis et al., 2017, p. 143). The MITM is based on the use and identification of biomarkers, i.e. elements of an organism or its surrounding environment that can be measured and at the same time used to measure other entities and processes, which in the case of exposome research are used to track the development of disease at different stages (Strimbu & Tavel, 2010). When biomarkers of disease and biomarkers of exposure are associated, the idea of approaches such as the MITM is to look for an intermediate biomarker that can connect exposure and disease (Chadeau-Hyam et al., 2013). In the philosophical literature, Phyllis Illari and Federica Russo have argued that intermediate biomarkers lie in the middle of a causal continuum that link exposure to disease (Illari & Russo, 2016). According to their informational account of causation, the MITM should be interpreted as a way of moving beyond correlations and associations and towards identifying whether exposure caused the disease and through which pathways the effects were produced (P. M. Illari & Russo, 2014). This is one way of attributing causal power to environmental factors in data-intensive epidemiology, which goes hand in hand with a more explicit focus on *how* the environment brings about effects. These causal attributions can constitute a significant shift from the traditional epidemiology, where the focus has rather been on associations and determinants and epidemiologists have been reluctant to use causal inference vocabulary (Russo, 2009).[16] A similar combination and shift in the approach to causality and the environment is evident in the case of MEDMI. As we have seen, one of the main benefits of data mash-ups is the possibility of integrating data that were not brought together before, with the goal of exploring correlations and causal links between environmental, biological, and medical factors. In this direction, the employment of data-mash-ups is an attempt to develop a comprehensive analysis, which can capture the development of disease through possible causal links between human health and the environment. Similarly, the approach to integration of CIDACS exemplifies efforts to go beyond the study of variations and correlations and improve our understanding of the socio-economic factors that determine health and disease in populations.

In all our three cases, we thus see a shift towards a more direct focus on causal inference and the causal connections between the environment and health, as a result of new sources and integrations of data and formulations of notions of the environment. Still, the extent to which environmental factors are interpreted in causal terms remains a challenge for all of our case studies. For instance, EXPOsOMICS researchers insisted that the MITM approach is a statistical method that can be causally interpreted but is primarily aimed at validating statistical associations, rather than detecting causality – as a result, omics techniques were only rarely used to detect mechanisms and pathways of disease development (Canali, 2019). Similarly and even more strikingly, given their investment on providing evidence for policy, we discussed how some of the more methodological publications by CIDACS researchers avoid the language of causality altogether and highlight how

[16] As suggested by an anonymous reviewer, traditional notions of cause involve an ordered series of relationships between boundaries, such as between inside and outside. The reframing of boundaries and particularly boundaries between inside and outside in exposome research, combined with approaches such as the MITM, can go beyond these notions of cause. This is the direction presented by (Illari & Russo, 2016), with their informational view of causality as flowing throughout the series of relationships between boundaries. This is a promising and innovative direction for causal inference in the health sciences, but we cast doubt on the extent to which it has been adapted in research practice and is actually supported by new data sources.

problems with the very structure and management of data linkage can affect the robustness of the analysis, and yet remain hard to avoid (in their words: "the dynamic, error-prone and incomplete nature of administrative data makes a certain level of linkage error inevitable, and this is compounded when data are required to be anonymised before linkage" (Harron et al., 2017, p. 5)).

Another worry related to causality is the extensive use of genomic platforms in the study of population health. The new opportunities to study and document physiological shifts at the molecular level are arguably fostering a "molecularization" of epidemiology, where the focus is increasingly on microscopic pollutants and toxicants produced by environmental exposure and detected in the internal, chemical environment. This can have the effect of prioritising the study of only those environmental stressors that affect molecular processes, which is particularly problematic given the significance of social determinants that might be difficult to study at the molecular level (Shostak & Moinester, 2015; S. A. Valles, 2019; Ghiara & Russo, 2019). We see the attribution of causal power to a broader range of environmental factors in our cases as a step in the direction of more dynamic views of the environment and focus on processes and interventions at the broader social level. This is possible when molecular studies are conducted in combination with environmental and climate research, such as informing data mash-ups like MEDMI, as well as public health research, such as expansive administrative data linkage like CIDACS. The importance of expanding and diversifying both data sources and relevant methods of analysis is ever more evident given the continuing risk that the wide availability and high evidential status attributed to genomic data and related analytics fosters a renewed, data-related version of biological reductionism.

7. Conclusions

How is the use of data-intensive methods and environmental data shaping the health sciences and the study of environmental health impacts therein? We have identified three major shifts and argued that current strategies for the integration of environmental data in epidemiology are yielding new approaches to the conceptual and material boundaries between environments, methods for the study of environmental exposure, and attributions and modes of causal inference. The use of genomic and climatic data in the study of the exposome is connected to a revised view of the relation between external and internal exposure, as exemplified by EXPOsOMICS, and of the boundaries between external and internal environments. Shifts in the conceptualisation of the environment do not concern biological approaches to health only, but also studies of the relation between human health and climate – such as in MEDMI, where the integration and linkage between new sources of data has yielded to a framing of environmental exposure to include internal and individual processes as well. The collection, computational analysis, and integration of digitalised socio-economic data with climatic and health data in projects such as CIDACS are expanding the range of social and environmental factors considered as potential causes, and thereby shifting epidemiological attention towards innovative analyses of complex phenomena such as poverty and disease transmission.

We have framed this article as a contribution to the study of the epistemic role of data in scientific research (Leonelli, 2016). Focusing on the interconnections between changes at the level of sources of data and shifts at the conceptual and methodological level, our analysis yields a picture where conceptual framings often need to be updated vis-à-vis the availability of new data – yet it is typically the alignment of data with other material and conceptual components of existing research repertoires that yields change (Ankeny & Leonelli, 2016). The conceptual expansions of exposure that we have identified in data mash-up studies and exposome research are connected to the expansion in data sources and techniques to link and analyse them. Yet these re-framing of environmental exposure are not solely a result of evidential strategies employed to integrate traditional and new sources of environmental data. Just as crucial has been the work employed by research consortia such as MEDMI to foster and facilitate interdisciplinary dialogue and the forging of new conceptual frameworks within which researchers in public health, biomedicine, climate science, data science and environmental science could exchange insights and expertise. For example, in the exposome context the notion of 'internal exposure' has originally been transferred from biomarkers research, where it is used with reference to the concentration of external chemicals in tissues. Moving this notion into epidemiology and using it to characterise omics data has required considerable conceptual and organisational labour, including an expansion of the original interpretation of the notion in biomarkers research (Canali, 2020a). One of the upshots of our analysis is the role of data as an 'asset' that can create interdisciplinary and dialogue, which in the cases we have discussed has often led to a contextual and situated approach to data linkage. These are not automatic results of the integration of diverse datasets: as new notions of health and trends in the context bring in different approaches and try to integrate different disciplinary perspectives, we need to pay attention to the types of data that are prioritised by different notions and how we can foster more transdisciplinary dialogue therein.

We have also framed our focus on the notion of the environment in epidemiology as a contribution to the literature on health in genomics and postgenomics: our analysis shows that the notion of environment in epidemiology is changing, and this matters for public health and public policy. Hence this paper points towards new questions around the role that boundaries play in approaches to the environment from a research and policy standpoint. The classic idea of the *milieu intérieur*, developed by French physiologist Claude Bernard, was based on the view that various mechanisms maintain boundaries between the body and the environment and regulate the state of body in an equilibrium with external changes. This notion of boundary is still influential in the health sciences and epidemiology (Vineis, 2003), as exemplified by current attempts to label, annotate, and specify environmental data in the Environment Ontology, where the environment is defined as a "certain sort of system which has the disposition to environ" (Buttigieg et al., 2013, p. 2). Similarly, environmental and health policy are still largely separate areas of intervention and policy-making often presupposes the existence of boundaries between individuals and the environment (Cousins et al., 2021). The rise of new framings and flexible uses of boundaries between environments do not mean that tensions between the focus on populations and individuals are settled, nor that reductionistic approaches cannot be reintroduced with the use of new types of data. Studying data practices can offer insight on these issues, especially as the biomedical context is filled with promises of new 'revolutions' thanks to the collection and integration of new types of data (Prainsack, 2020; Leonelli, 2021).

Our findings point to a dramatic shift in the scope of public health recommendations in the future and a much tighter link between health-related policies and policies on climate change. With the increasing relevance of wide-ranging environmental data to population health and shifts in what are considered to be exposure and relevant causes, the role given to the environment in a data-intensive context has become more prominent. While the study of population health is environmental in the sense that epidemiologists study the influence and effects of actions and processes of the environment on populations, this has rarely rendered the role of the environment in dynamic terms. While epidemiologists have started to show the relevance of their work with actions aimed at changing the environment, public health priorities focused on changes at the individual and behavioural level, rather than actions at the level of the context and environment that surrounds populations (Reis et al., 2015). The distance we have identified as a starting point of our analysis, between epidemiology and environmental health, only grows larger when socio-economic factors are taken into account: both epidemiology and environmental science have traditionally neglected the causal influence of socio-economic factors on health, and these translate to conceptual, epistemological, and methodological difficulties at integrating and coordinating research on these issues (Valles, 2019; Lohse & Canali, 2021). New approaches to boundaries between internal and external environments, health and disease, social and biological, such as the ones we discussed, can help identify and work through this distance and understand the role played by data to this end.

Acknowledgements

Many thanks to Jan Baedke and Tatjana Buklijas as special issue editors and two anonymous reviewers for their very helpful feedback on our work in the paper. SC wishes to thank Federica Russo, Saana Jukola, Virginia Ghiara and Thomas Reydon for several related discussions on the exposome, epidemiology and philosophy of medicine. SC also thanks the generous hospitality of Paolo Vineis and the EXPOsOMICS project at Imperial College London in 2017, where most of the empirical materials were collected. SC's research was funded by the Deutsche Forschungsgemeinschaft– Project 254954344/GRK2073, as part of the graduate research group "Integrating Ethics and Epistemology of Scientific Research" (https://grk2073.org) that was of crucial support during the PhD research at the basis of this work. SL thanks Lora Fleming, Bethânia Almeida and Maurício Barreto for enriching and productive collaborations, as well as their scientific leadership and their openness to working with philosophers. SL also thanks Egenis colleagues, particularly Niccolò Tempini who collected most of the empirical materials relating to the MEDMI case study, and colleagues working on global health – especially Kaushik Sunder Rajan, Thomas Cousins and Michelle Pentecost – for illuminating discussions around the different guises of environmental and public health discourse. SL's research was funded by the European Research Council under award number 335925, the Alan Turing Institute under the EPSRC grant EP/N510129/1 and the generous hospitality of CIDACS in 2019.

References

Almeida, B. A., Santos, P. A., & Barreto, M. L. (2018). Dados governamentais na perspectiva da Ciência Aberta: Potencialidades e desafios para saúde pública a partir de um estudo de caso. *Cadernos De Biblioteconomia, Arquivistica E Documentacao*, *1*, 172–179.

Ankeny, R. A., & Leonelli, S. (2016). Repertoires: A post-Kuhnian perspective on scientific change and collaborative research. *Studies in History and Philosophy of Science Part A*, *60*, 18–28. https://doi.org/10.1016/j.shpsa.2016.08.003

Barnes, B., & Dupré, J. (2008). *Genomes and what to make of them*. University of Chicago press.

Barreto, M. L., Ichihara, M. Y., Almeida, B. de A., Barreto, M. E., Cabral, L., Fiaccone, R., Carreiro, R. P., Teles, C., Pita, R., Penna, G., Barral-Netto, M., Ali, M. S., Barbosa, G., Denaxas, S., Rodrigues, L., & Smeeth, L. (2019). The Center for Data and Knowledge Integration for Health (CIDACS): An Experience of Linking Health and Social Data in Brazil. *International Journal of Population Data Science*, *4*(2). https://doi.org/10.23889/ijpds.v4i2.1140

Barreto, M. L., & Rodrigues, L. C. (2018). Linkage of Administrative Datasets: Enhancing Longitudinal Epidemiological Studies in the Era of "Big Data". *Current Epidemiology Reports*, *5*(4), 317–320. https://doi.org/10.1007/s40471-018-0177-5

Bauer, S. (2008). Mining data, gathering variables and recombining information: The flexible architecture of epidemiological studies. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *39*(4), 415–428. https://doi.org/10.1016/j.shpsc.2008.09.008

Benson, E. (2020). *Surroundings: A history of environments and environmentalisms*. University of Chicago Press.

Boniolo, G., & Nathan, M. J. (Eds.). (2017). *Philosophy of molecular medicine: Foundational issues in research and practice*. Routledge, Taylor & Francis Group.

Broadbent, A. (2013). *Philosophy of Epidemiology*. Palgrave Macmillan UK. https://doi.org/10.1057/9781137315601

Broadbent, A. (2015). Causation and prediction in epidemiology: A guide to the "Methodological Revolution". *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *54*, 72–80. https://doi.org/10.1016/j.shpsc.2015.06.004

Brown, T. M., Cueto, M., & Fee, E. (2006). The World Health Organization and the Transition From "International" to "Global" Public Health. *American Journal of Public Health*, *96*(1), 62–72. https://doi.org/10.2105/AJPH.2004.050831

Buttigieg, P., Morrison, N., Smith, B., Mungall, C. J., Lewis, S. E., & the ENVO Consortium. (2013). The environment ontology: Contextualising biological and biomedical entities. *Journal of Biomedical Semantics*, *4*(1), 43. https://doi.org/10.1186/2041-1480-4-43

Canali, S. (2019). Evaluating evidential pluralism in epidemiology: Mechanistic evidence in exposome research. *History and Philosophy of the Life Sciences*, *41*(1), 4. https://doi.org/10.1007/s40656-019-0241-6

Canali, S. (2020a). What Is New about the Exposome? Exploring Scientific Change in Contemporary Epidemiology. *International Journal of Environmental Research and Public Health*, *17*(8), 2879. https://doi.org/10.3390/ijerph17082879

Canali, S. (2020b). Making evidential claims in epidemiology: Three strategies for the study of the exposome. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *82*, 101248. https://doi.org/10.1016/j.shpsc.2019.101248

Chadeau-Hyam, M., Campanella, G., Jombart, T., Bottolo, L., Portengen, L., Vineis, P., Liquet, B., & Vermeulen, R. C. H. (2013). Deciphering the complex: Methodological overview of statistical models to derive OMICS-based biomarkers: Statistical Approaches for OMICS-Based Biomarkers. *Environmental and Molecular Mutagenesis*, *54*(7), 542–557. https://doi.org/10.1002/em.21797

Clarke, B., Gillies, D., Illari, P., Russo, F., & Williamson, J. (2013). The evidence that evidence-based medicine omits. *Preventive Medicine*, *57*(6), 745–747. https://doi.org/10.1016/j.ypmed.2012.10.020

Cousins, T., Pentecost, M., Alvergne, A., Chandler, C., Chigudu, S., Herrick, C., Kelly, A., Leonelli, S., Lezaun, J., Lorimer, J., Reubi, D., & Sekalala, S. (2021). The changing climates of global health. *BMJ Global Health*, *6*(3), e005442. https://doi.org/10.1136/bmjgh-2021-005442

Djennad, A., Lo Iacono, G., Sarran, C., Fleming, L. E., Kessel, A., Haines, A., & Nichols, G. L. (2018). A comparison of weather variables linked to infectious disease patterns using laboratory addresses and patient residence addresses. *BMC Infectious Diseases*, *18*(1), 198. https://doi.org/10.1186/s12879-018-3106-9

Djennad, A., Lo Iacono, G., Sarran, C., Lane, C., Elson, R., Höser, C., Lake, I. R., Colón-González, F. J., Kovats, S., Semenza, J. C., Bailey, T. C., Kessel, A., Fleming, L. E., & Nichols, G. L. (2019). Seasonality and the effects of weather on Campylobacter infections. *BMC Infectious Diseases*, *19*(1), 255. https://doi.org/10.1186/s12879-019-3840-7

Fiorito, G., Vlaanderen, J., Polidoro, S., Gulliver, J., Galassi, C., Ranzi, A., Krogh, V., Grioni, S., Agnoli, C., Sacerdote, C., Panico, S., Tsai, M.-Y., Probst-Hensch, N., Hoek, G., Herceg, Z., Vermeulen, R., Ghantous, A., Vineis, P., Naccarati, A., & for the EXPOsOMICS consortium‡. (2018). Oxidative stress and inflammation mediate the effect of air pollution on cardio- and cerebrovascular disease: A prospective study in nonsmokers: Effect of Air Pollution on Cardio- and Cerebrovascular Disease. *Environmental and Molecular Mutagenesis*, *59*(3), 234–246. https://doi.org/10.1002/em.22153

Fleming, L., Haines, A., Golding, B., Kessel, A., Cichowska, A., Sabel, C., Depledge, M., Sarran, C., Osborne, N., Whitmore, C., Cocksedge, N., & Bloomfield, D. (2014). Data Mashups: Potential Contribution to Decision Support on Climate Change and Health. *International Journal of Environmental Research and Public Health*, *11*(2), 1725–1746. https://doi.org/10.3390/ijerph110201725

Fleming, L., Tempini, N., Gordon-Brown, H., Nichols, G. L., Sarran, C., Vineis, P., Leonardi, G., Golding, B., Haines, A., Kessel, A., Murray, V., Depledge, M., & Leonelli, S. (2017). Big Data in Environment and Human Health. In L. Fleming, N. Tempini, H. Gordon-Brown, G. L. Nichols, C. Sarran, P. Vineis, G. Leonardi, B. Golding, A. Haines, A. Kessel, V. Murray, M. Depledge, & S. Leonelli, *Oxford Research Encyclopedia of Environmental Science*. Oxford University Press. https://doi.org/10.1093/acrefore/9780199389414.013.541

Gaudilliere, J.-P., & Gasnier, C. (2020). From Washington DC to Washington State: The Global Burden of Diseases Data Basis and the Political Economy of Global Health. In S. Leonelli & N. Tempini (Eds.), *Data Journeys in the Sciences* (pp. 351–369). Springer International Publishing. https://doi.org/10.1007/978-3-030-37177-7_18

Ghiara, V., & Russo, F. (2019). Reconstructing the mixed mechanisms of health: The role of bio- and sociomarkers. *Longitudinal and Life Course Studies*, *10*(1), 7–25. https://doi.org/10.1332/175795919X15468755933353

Gibbon, S., Prainsack, B., Hilgartner, S., & Lamoreaux, J. (2020). *Routledge handbook of genomics, health and society*.

Gibbs, E. P. J. (2014). The evolution of One Health: A decade of progress and challenges for the future. *Veterinary Record*, *174*(4), 85–91. https://doi.org/10.1136/vr.g143

Gulliver, J., Morley, D., Dunster, C., McCrea, A., van Nunen, E., Tsai, M.-Y., Probst-Hensch, N., Eeftens, M., Imboden, M., Ducret-Stich, R., Naccarati, A., Galassi, C., Ranzi, A., Nieuwenhuijsen, M., Curto, A., Donaire-Gonzalez, D., Cirach, M., Vermeulen, R., Vineis, P., … Kelly, F. J. (2018). Land use regression models for the oxidative potential of fine particles (PM 2.5) in five European areas. *Environmental Research*, *160*, 247–255. https://doi.org/10.1016/j.envres.2017.10.002

Harron, K., Dibben, C., Boyd, J., Hjern, A., Azimaee, M., Barreto, M. L., & Goldstein, H. (2017). Challenges in administrative data linkage for research. *Big Data & Society*, *4*(2), 205395171774567. https://doi.org/10.1177/2053951717745678

Hilgartner, S. (2017). *Reordering life: Knowledge and control in the genomics revolution*. The MIT Press.

Hitchcock, C. (2021). Probabilistic causation. In *Stanford Encyclopedia of Philosophy: Vol. Spring 2021 Edition*. Metaphysics Research Lab, Stanford University.

Hogle, L. F. (2016). Data-intensive resourcing in healthcare. *BioSocieties*, *11*(3), 372–393. https://doi.org/10.1057/s41292-016-0004-5

Holmberg, C., Bischof, C., & Bauer, S. (2013). Making Predictions: Computing Populations. *Science, Technology, & Human Values*, *38*(3), 398–420. https://doi.org/10.1177/0162243912439610

Hone, T., Rasella, D., Barreto, M. L., Majeed, A., & Millett, C. (2017). Association between expansion of primary healthcare and racial inequalities in mortality amenable to primary care in Brazil: A national longitudinal analysis. *PLOS Medicine*, *14*(5), e1002306. https://doi.org/10.1371/journal.pmed.1002306

Horton, R., Beaglehole, R., Bonita, R., Raeburn, J., McKee, M., & Wall, S. (2014). From public to planetary health: A manifesto. *The Lancet*, *383*(9920), 847. https://doi.org/10.1016/S0140-6736(14)60409-8

Illari, P. M., & Russo, F. (2014). *Causality: Philosophical theory meets scientific practice* (First edition). Oxford University Press.

Illari, P., & Russo, F. (2016). Information Channels and Biomarkers of Disease. *Topoi*, *35*(1), 175–190. https://doi.org/10.1007/s11245-013-9228-1

Landecker, H. (2011). Food as exposure: Nutritional epigenetics and the new metabolism. *BioSocieties*, *6*(2), 167–194. https://doi.org/10.1057/biosoc.2011.1

Landecker, H., & Panofsky, A. (2013). From Social Structure to Gene Regulation, and Back: A Critical Introduction to Environmental Epigenetics for Sociology. *Annual Review of Sociology*, *39*(1), 333–357. https://doi.org/10.1146/annurev-soc-071312-145707

Leonelli, S. (2013). Integrating data to acquire new knowledge: Three modes of integration in plant science. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *44*(4), 503–514. https://doi.org/10.1016/j.shpsc.2013.03.020

Leonelli, S. (2016). *Data-centric biology: A philosophical study*. The University of Chicago Press.

Leonelli, S. (2021). Data Science in Times of Pan(dem)ic. *Harvard Data Science Review*. https://doi.org/10.1162/99608f92.fbb1bdd6

Leonelli, S., & Tempini, N. (2021). Where health and environment meet: The use of invariant parameters in big data analysis. *Synthese*, *198*, S2485–S2504. https://doi.org/10.1007/s11229-018-1844-2

Lohse, S., & Canali, S. (2021). Follow *the* science? On the marginal role of the social sciences in the COVID-19 pandemic. *European Journal for Philosophy of Science*, *11*(4), 99. https://doi.org/10.1007/s13194-021-00416-y

Morabia, A. (Ed.). (2004). *A History of Epidemiologic Methods and Concepts*. Birkhäuser Basel. https://doi.org/10.1007/978-3-0348-7603-2

Oliveira, J. F., Jorge, D. C. P., Veiga, R. V., Rodrigues, M. S., Torquato, M. F., da Silva, N. B., Fiaccone, R. L., Cardim, L. L., Pereira, F. A. C., de Castro, C. P., Paiva, A. S. S., Amad, A. A. S., Lima, E. A.

B. F., Souza, D. S., Pinho, S. T. R., Ramos, P. I. P., & Andrade, R. F. S. (2021). Mathematical modeling of COVID-19 in 14.8 million individuals in Bahia, Brazil. *Nature Communications*, *12*(1), 333. https://doi.org/10.1038/s41467-020-19798-3

Parker, W. (2018). Climate Science. In *Stanford Encyclopedia of Philosophy: Vol. Summer 2018 Edition*. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2018/entries/climate-science/

Pescarini, J. M., Williamson, E., Nery, J. S., Ramond, A., Ichihara, M. Y., Fiaccone, R. L., Penna, M. L. F., Smeeth, L., Rodrigues, L. C., Penna, G. O., Brickley, E. B., & Barreto, M. L. (2020). Effect of a conditional cash transfer programme on leprosy treatment adherence and cure in patients from the nationwide 100 Million Brazilian Cohort: A quasi-experimental study. *The Lancet Infectious Diseases*, *20*(5), 618–627. https://doi.org/10.1016/S1473-3099(19)30624-3

Pietsch, W. (2015). Aspects of Theory-Ladenness in Data-Intensive Science. *Philosophy of Science*, *82*(5), 905–916. https://doi.org/10.1086/683328

Pita, R., Pinto, C., Sena, S., Fiaccone, R., Amorim, L., Reis, S., Barreto, M. L., Denaxas, S., & Barreto, M. E. (2018). On the Accuracy and Scalability of Probabilistic Data Linkage Over the Brazilian 114 Million Cohort. *IEEE Journal of Biomedical and Health Informatics*, *22*(2), 346–353. https://doi.org/10.1109/JBHI.2018.2796941

Prainsack, B. (2020). The political economy of digital data: Introduction to the special issue. *Policy Studies*, *41*(5), 439–446. https://doi.org/10.1080/01442872.2020.1723519

Rappaport, S. M. (2011). Implications of the exposome for exposure science. *Journal of Exposure Science & Environmental Epidemiology*, *21*(1), 5–9. https://doi.org/10.1038/jes.2010.50

Rappaport, S. M., & Smith, M. T. (2010). Environment and Disease Risks. *Science*, *330*(6003), 460–461. https://doi.org/10.1126/science.1192603

Ratti, E. (2015). Big Data Biology: Between Eliminative Inferences and Exploratory Experiments. *Philosophy of Science*, *82*(2), 198–218. https://doi.org/10.1086/680332

Reis, S., Morris, G., Fleming, L. E., Beck, S., Taylor, T., White, M., Depledge, M. H., Steinle, S., Sabel, C. E., Cowie, H., Hurley, F., Dick, J. McP., Smith, R. I., & Austen, M. (2015). Integrating health and environmental impact analysis. *Public Health*, *129*(10), 1383–1389. https://doi.org/10.1016/j.puhe.2013.07.006

Rheinberger, H.-J., Müller-Wille, S., & Bostanci, A. (2017). *The gene: From genetics to postgenomics* (German edition). The University of Chicago Press.

Richardson, S. S., & Stevens, H. (Eds.). (2015). *Postgenomics: Perspectives on biology after the genome*. Duke University Press.

Robinson, O., & Vrijheid, M. (2015). The Pregnancy Exposome. *Current Environmental Health Reports*, *2*(2), 204–213. https://doi.org/10.1007/s40572-015-0043-2

Rogawski, E. T., Gray, C. L., & Poole, C. (2016). An argument for renewed focus on epidemiology for public health. *Annals of Epidemiology*, *26*(10), 729–733. https://doi.org/10.1016/j.annepidem.2016.08.008

Ross, L. N. (2018). The doctrine of specific etiology. *Biology & Philosophy*, *33*(5–6), 37. https://doi.org/10.1007/s10539-018-9647-x

Ross, L. N., & Woodward, J. F. (2016). Koch’s postulates: An interventionist perspective. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *59*, 35–46. https://doi.org/10.1016/j.shpsc.2016.06.001

Russo, F. (2009). Variational Causal Claims in Epidemiology. *Perspectives in Biology and Medicine*, *52*(4), 540–554. https://doi.org/10.1353/pbm.0.0118

Russo, F., & Vineis, P. (2016). Opportunities and challenges of molecular epidemiology. In G. Boniolo & M. J. Nathan (Eds.), *Philosophy of Molecular Medicine*. Taylor & Francis.

Sharon, T., & Lucivero, F. (2019). Introduction to the Special Theme: The expansion of the health data ecosystem – Rethinking data ethics and governance. *Big Data & Society*, *6*(2), 205395171985296. https://doi.org/10.1177/2053951719852969

Shostak, S. (2013). *Exposed science: Genes, the environment, and the politics of population health*. University of California Press.

Shostak, S., & Moinester, M. (2015). The Missing Piece of the Puzzle? Measuring the Environment in the Postgenomic Moment. In S. S. Richardson & H. Stevens (Eds.), *Postgenomics: Perspectives on Biology after the Genome*. Duke University Press.

Siroux, V., Agier, L., & Slama, R. (2016). The exposome concept: A challenge and a potential driver for environmental health research. *European Respiratory Review*, *25*(140), 124–129. https://doi.org/10.1183/16000617.0034-2016

Strimbu, K., & Tavel, J. A. (2010). What are biomarkers? *Current Opinion in HIV and AIDS*, *5*(6), 463–466. https://doi.org/10.1097/COH.0b013e32833ed177

Stringhini, S., Carmeli, C., Jokela, M., Avendaño, M., Muennig, P., Guida, F., Ricceri, F., d'Errico, A., Barros, H., Bochud, M., Chadeau-Hyam, M., Clavel-Chapelon, F., Costa, G., Delpierre, C., Fraga, S., Goldberg, M., Giles, G. G., Krogh, V., Kelly-Irving, M., … Zins, M. (2017). Socioeconomic status and the 25 × 25 risk factors as determinants of premature mortality: A multicohort study and meta-analysis of 1·7 million men and women. *The Lancet*, *389*(10075), 1229–1237. https://doi.org/10.1016/S0140-6736(16)32380-7

Valles, S. A. (2020). Philosophy of Biomedicine. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2020 Edition). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2020/entries/biomedicine/

Valles, S. A. (2019). *Philosophy of population health science: Philosophy for a new public health era*. Routledge.

Valles, S. A. (2021). A pluralistic and socially responsible philosophy of epidemiology field should actively engage with social determinants of health and health disparities. *Synthese*, *198*, 2589–2611. https://doi.org/10.1007/s11229-019-02161-5

Vineis, P. (2003). Causality in epidemiology. *Sozial- Und Präventivmedizin*, *48*(2), 80–87. https://doi.org/10.1007/s00038-003-1029-7

Vineis, P., Chadeau-Hyam, M., Gmuender, H., Gulliver, J., Herceg, Z., Kleinjans, J., Kogevinas, M., Kyrtopoulos, S., Nieuwenhuijsen, M., Phillips, D. H., Probst-Hensch, N., Scalbert, A., Vermeulen, R., & Wild, C. P. (2017). The exposome in practice: Design of the EXPOsOMICS project. *International Journal of Hygiene and Environmental Health*, *220*(2), 142–151. https://doi.org/10.1016/j.ijheh.2016.08.001

Warde, P., Robin, L., & Sörlin, S. (2018). *The environment: A history of the idea*.

Wild, C. P. (2005). Complementing the Genome with an ‘Exposome’: The Outstanding Challenge of Environmental Exposure Measurement in Molecular Epidemiology. *Cancer Epidemiology Biomarkers & Prevention*, *14*(8), 1847–1850. https://doi.org/10.1158/1055-9965.EPI-05-0456

Wild, C. P. (2008). Environmental exposure measurement in cancer epidemiology. *Mutagenesis*, *24*(2), 117–125. https://doi.org/10.1093/mutage/gen061