Data Between Environment and Health:

An Epistemological Study of the Exposome

Von der Philosophischen Fakultät der Gottfried Wilhelm Leibniz Universität Hannover zur Erlangung des Grades Doktor der Philosophie (Dr. phil.) genehmigte Dissertation von M. Sc. Stefano Canali.

Erscheinungsjahr: 2019.

Referent: Prof. Dr. Thomas Reydon, Institut für Philosopie, Leibniz Universität Hannover.

Korreferentinnen: Prof. Dr. Sabina Leonelli, Dept. Of Sociology, Philosophy and Anthropology, University of Exeter; Prof. Dr. Federica Russo, Dept. Of Philosophy University of Amsterdam.

Tag der Promotion: 26.06.2019.

Abstract

This thesis is a philosophical analysis of the epistemic role of scientific data in biomedical research. It is comprised of an introduction (Chapter 1), three articles (Chapter 2, 3, 4) and a conclusion (Chapter 5). I use a case study approach and focus on the epidemiology of the 'exposome', a new line of research based on a reconceptualisation of exposure and the use of new and diverse datasets. I argue that data can sustain the subject matter of exposome research by shaping concepts, strategies, techniques and what counts as evidence. Yet, the epistemic role of data is enacted by the ways in which it is used by epistemic agents and thus constantly connected to and mediated by other artefacts, components and features of scientific inquiry. In Chapter 2, I discuss the innovations and changes of the exposome. I argue that these should be framed as the establishment of a repertoire, as opposed to a paradigm. The exposome repertoire consists in many components transferred from other areas of the life and health sciences: thus, scientific change is the result of the alignment of these components and it is not due to only one of these factors, such as data. In Chapter 3, I discuss data practices in exposome research. I argue that researchers use evidential claims to specify the evidential and representational value of datasets. Three strategies for evidential claims can be distinguished, differing in terms of level of abstraction, lines of work and type of evidential claim and leading to a picture of evidence production as epistemic-intensive labour. In Chapter 4, I discuss how data is classified as evidence in exposome research, in the context of philosophical discussions of the types of evidence used for causal claims in biomedical research. I argue that molecular data collected in exposome research is used to study differences and dependences, as opposed to mechanisms; more generally, the classification of a dataset as a type of evidence is dependent on the ways in which the data is used, rather than its intrinsic properties.

Keywords: Philosophy of epidemiology; Data; Evidence; Exposome; Scientific change.

Acknowledgements

What you can read in the following pages wouldn't have been possible without the support of many people.

First of all, I am deeply grateful to Thomas Reydon and Sabina Leonelli for giving me the opportunity to conduct this project under their supervision. Thanks for your close reading, comments, motivation and support. I feel so privileged to have you as mentors and role models to look up to.

My home during these years as a PhD student has been the Institute of Philosophy of Leibniz Universität Hannover and particularly the DFG Graduiertenkolleg 2073 "Integrating Ethics and Epistemology of Scientific Research" (in Hannover and Bielefeld). A PhD is mostly a solo project and one can easily feel lonely. This is why I feel especially fortunate to have been surrounded by lovely people that have prevented that from happening. Special thanks, first of all, to my fellow PhDs, postdocs and office mates: Markus Ahlers, Rafael Ugarte Chacón, Markus Dressel, Enno Fischer, David Hopf, Daria Jadreškić, Saana Jukola, Anna Leuschner, Anja Pichl, Julia Pfeiff, Leon Schäfer, Tobias Schönwitz, Roel Visser and Martin Wasmer. Thanks for the countless coffees, speculative etymology sessions, beers and table football; you've helped me more than you think. Thanks to all the academic staff of the Institute of Philosophy and the GRK for the chance to do this PhD, feedback and discussions on my work and support through all these years. Special thanks to: Anke Büter, Martin Carrier, Uljana Feest, Mathias Frisch, Dietmar Hübner, Marie Kaiser, Simon Lohse, Cornelis Menke and Torsten Wilholt. Last but not least, thanks to the support staff in Hannover, Güliz Erdem and Jessica Inhestern, especially for the help in organising the Data Workshop in 2017. Thanks to the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation – Project 254954344/GRK2073) for the funding.

Exeter was my second academic home. During longer and shorter stays, I enjoyed the friendly and stimulating atmosphere of Egenis and the Data Studies

group, auditing seminars, attending workshops, eating Data Crunch breakfasts or walking in Devon. Special thanks to: Gemma Anderson, John Dupré, Stephan Güttinger, Gregor Halfmann, Susan Kelly, Staffan Müller-Wille, Gabi Recknagel, Niccolò Tempini and Chee Wong.

I am indebted with Paolo Vineis as PI of EXPOsOMICS and all the members of group for taking the time to discuss their fascinating work with me, surviving philosophical interviews and, most of all, making this case study possible.

Traveling to workshops and conferences has been one of the aspects of my life as a PhD student that I have enjoyed the most. Making new friends is one of the highlights of that, although it is not always easy to keep in touch with friends you might see once a year on the other side of the world (if you're lucky). I am thus particularly thankful to call friends a few fellow and former PhD students. Thanks for the work chats, the non-work chats and for keeping me sane: Thomas Bonnin, Virginia Ghiara, Javier Suárez Díaz, Michele Luchetti and Guglielmo Militello**.** Throughout these years, I also had the privilege to learn from the inspiring work of many advanced researchers, who have read versions of my papers, commented on my talks and helped me with crucial feedback. Thanks to: Rachel Ankeny, Giovanni Boniolo, Raffaella Campaner, Brendan Clarke, Sara Green, Phyllis Illari, Emanuele Ratti and Federica Russo. Special thanks to everyone who bothered to come and ask questions at my presentations.

I wish to thank my parents, Patrizia and Corrado, for inspiring and supporting me, even if this has taken me far away. Thanks for making me feel like I was doing the right thing. Thanks to my sister Franci for the indie music, the GIFs, and making me feel more important than I am. You're one of my "sources of inspiration" (cit.). Thanks to all my family for everything, and all the video calls.

Thanks to all my friends that have made the effort of trying to understand what I have been doing in this thing called PhD. Sorry for failing at that and thanks for pretending to understand. Special thanks to: Elena, Erika, Gab, Stefano and

Yianna. Thanks to Davide for the walks uphill and downhill, taking my mind off things and the occasional mosh.

Above all, thanks to Costanza. I wouldn't be the person who has written a PhD dissertation in philosophy of science if I hadn't met you. Thanks for encouraging me to follow the dream that has led to this dissertation. Thanks for your love because – and not in spite – of that. No one but you will really understand this last sentence. No one really understands me and you and why we have been doing this. And this makes all the difference.

Table of Contents

Chapter 1

Introduction: a Data-Centric Study of the Exposome

This dissertation is about the collection, integration and use of scientific data. It studies scientific data as a focus to understand research in contemporary epidemiology. The dissertation empirically reconstructs the conditions under which data is created, handled and employed, pointing to the material, methodological, theoretical and technological constraints that characterise research about the relation between the environment and health.

Disease and health in human populations are deeply affected by various types of interactions with external entities in their surrounding environments, to the point that the vast majority of disease risks is related to differences in the environment (Rappaport and Smith 2010). Research on these issues is therefore very significant from a societal and political standpoint, is highly funded and is supposed to deliver evidence for policy-making. At the same time, this is highly complex research. The environment influences human health and disease through a variety of different pathways, which may involve significantly different phenomena, contexts and components, take place in diverse locations and temporalities, are often difficult to track and measure, and need to be studied at various scales and levels of abstraction. In turn, research projects in this area are carried out in an interdisciplinary context, which lies in between the life and health sciences and thus comprises diverse epistemic cultures, such as molecular biology, genetics research, animal research, etc. This implies a complex organisation of research, requiring: the collection and dissemination of evidence from and to various sources; the support of institutions, funding bodies and industry; and managing platforms and structures allowing for exchanges and collaborations at an interdisciplinary level.

The primary context for the study of the ways in which disease states are related to and are shaped by environmental factors is the discipline known as epidemiology, which studies the distribution and determinants of health and disease (Morabia 2015; Broadbent 2013). For this dissertation, I have studied a specific approach that has emerged in the last decade of epidemiological research and is presented as a way of overcoming some of the complexities of the issues at stake. The 'exposome' is the totality of all exposures to environmental elements that are experienced by individuals (Wild 2005). Here, totality is firstly meant to include all the exposures experienced throughout a lifetime. Thus, at any point in an individual's lifetime, their exposome will comprise all the exposures experienced from conception onwards: for instance, the exposome of an adult individual includes exposures in utero, which may have an impact on their health only at a later stage (Robinson and Vrijheid 2015). Secondly, totality is meant to comprise all the various elements, substances and processes that an individual is exposed to at a single point in time, both at an external and internal level. In other words, to study an individual exposome requires the study of various levels of investigation, from the macroscopic, external component to microscopic, individual elements (Rappaport 2011). At a given point in time, an individual's exposome comprises exposures to external elements at a general, specific and internal and individual level (Rappaport and Smith 2010). The concept was first introduced by Christopher Wild (Wild 2005), in an attempt to shift attention to the need of better and more precise exposure assessment and integrate various ideas and approaches to the study of the relation between environmental exposure and disease under an umbrella concept (Wild 2012). In the last decade, the concept has increasingly been applied in epidemiology, to the point that large bodies like the European Union have funding streams dedicated to the exposome (Vineis 2018).

The overarching question of this dissertation is the following: in which ways does data inform and sustain the subject matter of epidemiological research on the exposome? In order to answer this question, I study the role of data as an artefact that is produced on the basis of, and in turn used to generate, scientific

knowledge in exposome research.[1] This research question leads me to the analysis of more specific issues. I start by looking at the historical background in which the exposome was developed. I ask the following question: what constitutes the subject matter of exposome research from a conceptual, material, and social point of view and which is the role of data in shaping these components? I discuss this issue in the context of philosophical discussions on what constitutes change in current scientific research (Chapter 2). Further focusing on scientific innovations and their relation with data, I analyse data practices in contemporary research on the exposome. I ask the following question: how are diverse data sources handled, integrated and used, and in which ways conceptual, methodological and material assumptions influence data integration? I position this question as a contribution to current philosophical analyses of data and evidence integration in the sciences (Chapter 3). In the context of discussions on evidence, I investigate approaches to evidence classification in the study of the exposome. I ask the following question: how is evidence distinguished and classified in exposome research and how does data shape knowledge claims about specific aspects of the exposome? I analyse this topic in the context of philosophical and scientific discussions on the classification of medical evidence (Chapter 4).

This 'data-centric' philosophical study is based on the empirical study of a specific research project on the exposome, EXPOsOMICS. The project run between 2012 and 2017 and applied the exposome approach to the assessment of disease risk related to air and water pollution, by studying external and internal components of the exposome (Vineis et al. 2017a). The project was chosen on the basis of an initial study of the philosophical literature on scientific data and online and literary analysis of publications, reports and funding applications of exposome research. The choice was motivated by my philosophical interest

1 I give further details on my research questions in the following subsections (see 1.1.2).

in scientific research that is considered highly interdisciplinary, innovative and based on the collection, integration and use of many different sources of data. I decided that I would focus on the research of a single project, as this allowed me to specify my questions and sample, while at the same time obtaining a good representation of state of the art research on the exposome. EXPOsOMICS was the last in a series of projects that has been carried out by a group of research centres, guided by the Department of Epidemiology and Biostatistics of Imperial College London, that has been very successful at securing funding from the European Union. In addition, research of this group has been the focus of recent work in philosophy of science (Russo and Vineis 2017; Vineis et al. 2017b).

In what follows, I detail the broader context where I conduct my study and I present my research questions (Sect. 1.1). I then specify the methodology used in this dissertation (Sect. 1.2). I conclude by presenting the structure and content of the dissertation (Sect. 1.3).

1.1 Data, Big Data and Scientific (Big) Data

The word 'data' is increasingly present in the public sphere, from discussions in the media to policy reports and all the way to academic debates. In most of these discussions, the notion of data is associated with the adjective 'big', thus forming the phrase 'big data', to refer to the increasing amount of data produced, stored and used for various purposes. Big data is presented as a result of a number of phenomena and processes occurring in modern societies, whereby the increasing use of digital devices, software and online platforms has established activities that have the production of data as one of their main steps, if not their aim and endpoint.

As a notion, yet, big data is vague and has received many critiques by scholars. The notion was introduced between the mid 1990s and the early 2000s in the computing industry (Diebold 2012). At the time, big data was defined on the

basis of three traits: large volume, high velocity and wide variety. This definition, which is also known as the 3Vs, was picked up in various areas of the industry and especially in the business sector, where it is still extensively used (Kitchin and McArdle 2016). Yet, the 3Vs have been extensively criticised by scholars from different fields, to the point that some have argued that it has only increased confusion around big data. According to the Rob Kitchin and Gavin McArdle:

> The 3Vs meme is actually false and misleading and along with the term itself is partially to blame for the confusion over the definitional boundaries of Big Data. (Kitchin and McArdle 2016, p. 9)

In contrast with the 3Vs, many other definitions have been introduced and discussed, adding various dimensions and traits to the notion of big data. It has been argued that definitional aspects of big data include: exhaustivity, as big data can arguably capture and document entire systems; fine-grained resolution, because big datasets can arguably be as detailed as possible in the description of a phenomenon; relationality, since big datasets are arguably easy to integrate.[2] At the same time, scholars have underlined how a dataset that is considered big data in a context might not in another context, as a result of the relational character of the predicate 'big' and the diversity of areas where data is produced and used (Floridi 2012). Therefore, what is considered big data varies significantly in terms of both attribution and ontological characteristics, as different types of data constitute different 'species' of (big) data (Kitchin and McArdle 2016).

On top of discussions about how to properly define big data, why is big data something that we should discuss and care about? An answer to this question is that the increase in quantity of data collected, processed and used is routinely

2 See Kitchin (2013, 2014a) for a detailed discussion of these and other definitions.

associated with revolutionary changes, paradigm shifts and substantial disruptions in various contexts, from the economy to social interactions, from politics to entertainment. This has given rise a certain rhetoric around big data, which extends narratives of the benefits of innovation, revolution and 'disruption' that are typical of the high-tech sector. In academic debates, this rhetoric has led to a broad and interdisciplinary debate where several claims about the benefits, power and features of large datasets have been analysed, assessed and criticised. As part of a debate that has become known as critical data studies, scholars from various disciplines including sociology, philosophy, science and technology studies, information systems and history have criticised various claims of the big data rhetoric (Iliadis and Russo 2016; Kitchin and Lauriault 2018). For instance, scholars have questioned the claims that more data necessarily leads to more objective and accurate claims (boyd and Crawford 2012) and that big data allows for complete descriptions of phenomena and can thus do without sampling and bias issues (Kitchin 2013) or ethical considerations (Simon 2015).

In the context of this debate, scholars have significantly focused on the impact of big data for the scientific context, because various claims of the big data rhetoric have discussed scientific research (Kitchin 2014b). In the sciences, data has always been a cornerstone of the scientific method, at least since the beginning of modern science. At the same time, however, the amount of data scientists can collect, analyse and use has increased significantly in recent years, especially in connection to the development and employment of computing technologies that can collect, store and process large datasets. Between the end of the 2000s and the early 2010s, it was famously claimed that data-intensive approaches were bringing about a fourth scientific revolution (Hey et al. 2009; Bell et al. 2009), as a consequence of which scientific theories were going to be superfluous (Anderson 2008) and correlations were going to triumph over causal knowledge and reasoning (Mayer-Schönberger and Cukier 2013). As a consequence of the epistemological nature of these claims, philosophers of science started to engage in the debate on these issues, pushing back against the view that the use of large datasets in the scientific context was bringing about

revolutionary changes to scientific epistemology (Leonelli 2012a). In this context, philosophers of science have for instance argued against claims according to which data-intensive methods are an instance of theory-free science (Callebaut 2012), can be based on correlations only and do without causal knowledge (Pietsch in preparation), counter issues such as bias and over-sampling (Leonelli 2014) and are a historical novelty in the sciences (Müller-Wille and Charmantier 2012; Strasser 2011).

Most of the 'novelty claims' connected to the collection, processing and use of large datasets have thus been subject to effective criticism in the philosophical literature. I would argue that this has been the first step of the debate on data, a *pars destruens* that has been followed by a *pars contruens* phase. In this second phase, philosophers of science have started to pay attention to data as a philosophically interesting element of scientific methodology, practice and epistemology more generally. Sabina Leonelli (2013) has argued that one of the novelties related to large scientific datasets is the new emphasis placed on the role of data as commodities with crucial scientific, economic, social and political values. Similarly, I would argue that the big data rhetoric has sparked the interest of philosophers of science in data, as epistemic elements of scientific epistemology that should be included in philosophical analyses of the sciences. This is the context where I carry out my research in this dissertation, as I focus on data as the main units of my philosophical analysis. This is also why in the dissertation I analyse the collection, processing and use of data, but I rarely use the phrase 'big data', in an attempt to try to move away from the rhetorical tools used in big data narratives. In the dissertation, I discuss issues that have arisen in the context of what I called the first phase of the debate (such as transformations and changes in science due to data). Still, I intend my research to be located in current philosophical attempts at making sense of data as an element of scientific epistemology that: is something that deserves philosophical attention per se; and has key connections to other elements of scientific epistemology that have more traditionally been discussed in philosophy of science, such as knowledge, models, phenomena and evidence. In the next subsection, I give

further details about the theoretical background of the research conducted in this dissertation.

1.1.1 Data Studies in Philosophy of Science: Data as an Epistemically Salient Artefact

The increasing emphasis placed on data in public debates has thus led to an increase in interest and research on data in philosophy of science. Philosophers have studied scientific data in the context of both innovative data-intensive approaches and more traditional settings. That scientific data is of paramount importance for scientists has always been quite evident, not only to philosophers. Data is one of the primary objects researchers interact with; its production, collection and analysis are daily activities in scientific projects; questions about the quality of data are at the constant centre of scientific research; and data can often be one of the main outputs of scientific projects.

Still, philosophers have historically paid little attention to data in their research. Lack of interest in data can be connected to traditional approaches in philosophy of science and more particularly propositional, theory-centric and syntactic views of science.[3] According to these views, the goal of science is to produce propositions that can be organised in theories. Accordingly, what really matters in science is the justification of these propositions and theories. In this context, non-propositional components of science are important only insofar as they are related to the justification of claims and theories. As a result, non-propositional components of scientific research, such as instruments, methods, models and data are considered of little importance and their role is discussed only insofar as they may be involved in the justification of the truth

3 See Suppes (1976) and Winther (2016) for introductions on these approaches.

of a claim. With the shift from syntactic and semantic accounts of science to pragmatist views (Suppes 2000), philosophers of science started to focus on scientific research not just as a way of justifying claims and theories, but also of understanding, explaining and intervening on the world (Bailer-Jones 2009, pp. 126-158). In this context, philosophers started to pay attention to experimentation, models, understanding and values (Hacking 1983; Morgan and Morrison 1999; Douglas 2009). Still, scientific data was not a key component of these analyses, as data was mostly mentioned in connection with other topics, such as visualisation, analysis and interpretation through models or theories. The topic of data emerged mostly in the context of the focus on experimentation, as philosophers started to discuss the relation between experimental traces, data and evidence (Hacking 1983; Rheinberger 1997; 2010). Similar discussions can be found in science studies more generally, especially science and technology studies (see e.g. Latour and Woolgar 1979; Bowker 2005; Bowker and Star 1999) and the history of science (Chang 2004; Strasser 2012). Studies of historical and contemporary sciences revealed a number of challenges to received philosophical views of the role of data in the sciences. Philosophers highlighted that the same scientific data can be interpreted in significantly different ways by the same or other research groups, which presents a challenge to the idea of data as a mere enabler of the justification of scientific claims and theories (Bogen and Woodward 1988).

In the context of renewed interest in data in the sciences, philosophers have confronted these traditional views and built alternative approaches. Much of this work has been carried out in parallel with novel philosophical interest in scientific practice, where data plays a crucial and constant role.[4] Philosophers have analysed the practices involved in the collection, handling, curation and

4 I give more information on philosophical studies of scientific practice in the next section (especially Sect. 1.2.1), where I discuss the methodology of the dissertation.

storing of data for sharing and re-use within a specific scientific community, with the aim of critically engaging with the big data rhetoric as well as improving philosophical understanding of the epistemic role of data. For instance, Leonelli has extensively documented data practices in the last three decades of the life sciences, focusing on model organism and the setting up of community databases in this context (Leonelli 2016a). Scholars have underlined that a significant portion of time, effort and funding is now dedicated to the storing and dissemination of data in assemblages and infrastructures like databases, in order to allow for the re-use of data by various actors in the research community and beyond (Leonelli 2013a). In addition, the increasing emphasis on data as a very significant output of research in itself has been highlighted, with reference to the establishment of new journals and other publications focused on publishing data only (Leonelli 2013b). The aims of these analysis have had to do with improving philosophical understandings of scientific practice centred on data, documenting practical and methodological issues, and emphasising and distinguishing epistemic strategies designed and applied by researchers (Leonelli 2012b; Leonelli 2013c; O'Malley and Soyer 2012; Green et al. 2018). This line of research has led to more specified and nuanced accounts of the innovative dimension of the use of large datasets in the sciences (Leonelli 2014).

Furthermore, philosophers have studied the relations between data and components of scientific research that have traditionally been considered crucial elements of scientific epistemology. In this context, causality and causal reasoning have been among the most discussed topics. This is connected to one of the main tenets of the big data rhetoric, according to which the vast number of correlations found in large datasets undermines the need for causal analysis and reasoning (Mayer-Schönberger and Cukier 2013). For example, Pietsch and Illari and Russo have explored the nature of causal inference based on large datasets, on the basis of case studies from biomedicine and social sciences, underlining the ever more significant role that causal reasoning plays in these areas of research (Pietsch 2015a; 2014; Illari and Russo 2016a; 2016b).

The role of data has also been discussed in the context of the debate on what is known as the 'theory-ladenness' of scientific experimentation, within which

philosophers discuss how much theory influences the ways in which scientists carry out experiments (Burian 1997; Steinle 1997). As part of this debate, Koray Karaca has for example focused on the large datasets collected and used in high energy physics, distinguishing between data-driven, theory-driven and experimental procedures in data selection (Karaca 2013; 2017).

In addition, data has been discussed with the aim of understanding what constitutes it as a specific component of scientific research. According to an intuitive view, data has a representational and informational content, which is fixed and mind- and context-independent; in this view, data plays the role of providing evidence for phenomena and giving empirical content to claims, models or theories (Bogen and Woodward 1988; Woodward 2000). This representational approach has been challenged by philosophical analyses focused on data practices, and primarily by Leonelli's account of data, which highlights how, in contemporary data practices in the life science, the evidential value and representational content of data depend on the context in which data is used, rather than on its intrinsic and predetermined properties (Leonelli 2009; 2015; 2016a). In this view, data is a crucial element of epistemology, rather than a philosophically uninteresting by-product of scientific research. As a result of this debate, philosophers have also been interested in the relations between data and components of scientific epistemology such as models (Green et al. 2018; MacLeod and Nersessian 2018; Leonelli 2019).

A final line of research on data that I want to highlight here has focused on the (largely construed) ethical dimension of data practices in the sciences. These discussions have been elicited by both concerns over the increasing use of data in society more generally, as well as a philosophical debate that has gained in prominence in recent philosophy of science, i.e. the role of values and trust in science (Douglas 2009; Wilholt 2013; Longino 2015). Philosophers have explored issues related to responsibility and accountability (Leonelli 2016b; Rieder and Simon 2016), openness and divides (Levin and Leonelli 2017; Leonelli et al. 2017) and more general ethical frameworks (Simon 2015).

This area of current research in philosophy of science, which I call *data-centric philosophy of science*, is the context where I conduct my own research in the dissertation. Works in data-centric philosophy of science have thus been concerned with questions about: the relation of data and data practices with elements the cognitive and epistemic dimension of scientific research, including explanations, theories and models; the influence that data and data practices can have on and receive from other elements of scientific epistemology; the production of data as scientific artefacts and the relation with the material dimension of the sciences, including technology and methods; and the nature itself of scientific data, in the relation with interactions with the world, phenomena and empirical knowledge. The focus of data-centric philosophy of science is data and data practices, broadly construed and including data collection, storing, analysis, use, dissemination, etc. Depending on the specific context, what counts as data in the first place may vary substantially and 'scientific data' may be physical samples, digital files, analogue documents, etc. As a consequence, the specific units of philosophical analysis of data studies may change accordingly, potentially including the ways in which data is collected and modified, structured and ordered, and more generally used as evidence for various claims. The philosophical aims of data-centric philosophy of science include: the description and documentation of data practices in the sciences; the critical engagement with assumptions, concerns and underlying issues of scientific data; and the philosophical interpretation of the context of data from the perspective of conceptual tools from philosophical work.[5]

At the conceptual centre and basis of this line of research is a specific view of the epistemic role of data in science. In this dissertation, with the term 'data', I refer to both objects produced by the interaction of researchers with an object

5 I will come back to these features of data-centric philosophy of science when discussing methodology and normativity in the next section (see Sect. 1.2.1).

of investigation (such as samples and materials) and their subsequent processing and manipulation (affecting e.g. format, medium and order). In this view, the epistemic value of data as a source of information, evidence and knowledge, is the result of various practical, methodological and conceptual interactions and considerations that take place at various levels of research. In line with Leonelli (2016), the use of a dataset as evidence for knowledge claims is dependent on these interactions and considerations, and not only on the intrinsic properties of the dataset (such as the method through which it was produced, its format, size, etc.). This entails that practices of data collection, processing, integration and use are not just about extracting objective, fixed and context-independent information, but have a significant and specific epistemic character.

Following this approach, I view data as an *epistemically salient artefact* of scientific research, in the sense that it is the result of epistemically salient scientific practices and sits in epistemically salient relations with other products and components of scientific epistemology, such as theories, models and knowledge (Leonelli 2019). As a consequence of these features, using data as units of philosophical analysis provides a window into the interrelations between cognitive and material aspects of scientific research. This does not mean that data plays a more important role than models, theories or knowledge. But it implies that the practices, processes and researchers involved in the collection, integration and use of data have a specific epistemological significance, one that deserves a specific as much as models, theories or knowledge do. This is why, in this dissertation, I will be interested in the conditions and ways in which data collected in biomedical research are processed, integrated and used as evidence by epidemiologists, as epistemically significant aspects of research that shape and influence the study of health, environment and their relation.

1.1.2 A Data-Centric Study of the Exposome

In the expanding context of data-centric philosophy of science, the research of this dissertation is about the collection, integration and use of data in the life

and health sciences. A significant part of the recent philosophical scholarship on data has focused on this area, for a number of reasons. First, in the philosophical literature these disciplines have often been discussed as particularly fragmented, with highly diverse communities, approaches, methodologies, theories, styles of explanation, commitments and goals. These pluralistic features of the life and health sciences have been discussed by historians and philosophers of science in the context of the debate on reduction, unification and integration. For instance, philosophers have argued that different theories in biology cannot be unified nor reduced to more fundamental ones (Mitchell and Dietrich 2006); the use of diverse methods, models and representations is crucial in a number of areas of the life sciences (Mitchell and Gronenborn 2015); many questions and problems require explanations developed in different biological disciplines and with different scientific aims (Brigandt 2013); and the disunity of the life and health science is an inevitable condition (Dupré 1996).

As a result of this fragmentation, scientific data in this context can refer to many different objects, is of significantly different types, is collected by different communities, for different purposes and with diverse commitments. This elicits questions about the ways in which diverse data sources can be interpreted, analysed and used as to constitute a single body of evidence. At the same time, data is an interesting example of integration and diversity in the life and health sciences and provides a significant window into the commitments, assumptions and aims of diverse communities (Leonelli 2013c). In addition, the life and health sciences present an interesting case in which the standardised production of large volumes of data has often been mentioned as a potential game changer (Weinberg 2010; Golub 2010), but practices of data handling, storing and analysis have strong continuity with longstanding approaches (Müller-Wille and Charmantier 2012; Strasser 2012; Leonelli 2016b). In this sense, data also sits at the crossroads of many current trends of the field, especially in the biomedical sciences. For example, personalised medicine (also known as precision medicine), i.e. the attempt to take into account individual variables into the study and prevention of disease, is largely built on the assumption that the

use of large datasets "can account for an increasing number of factors that influence health and disease, and that these data can be used to stratify the population and health problems according to various characteristics" (Green and Vogt 2016, p. 106). Similarly, the molecularisation of medicine, i.e. study and treatment of disease from a molecular point of view (Boniolo and Nathan 2017), and postgenomics, i.e. the attempts of going beyond gene-centric approaches (Stevens and Richardson 2015), are largely based around the use and integration of new datasets, at different levels of abstraction and at increasing volumes. And the classification of evidence into a hierarchy of research designs developed in the context of evidence-based medicine, which has been extensively criticised in philosophy of science for the omission of crucial types of evidence (Worrall 2002; Clarke et al. 2013), can be considered a push for certain data sources over others.

As part of this context, I specify my data-centric study of the life and health sciences context by looking at current research in epidemiology on what is known as the exposome. I choose to focus on epidemiology for a number of reasons. First, epidemiology is an interesting case to look at because it follows the fragmented nature of other disciplines in this context, using approaches and methods from the medical, biological, environmental and statistical sciences (Broadbent 2013). In addition, epidemiology is particularly interesting from a data perspective. While, epidemiologists have traditionally been concerned with the collection and analysis of large datasets (Morabia 2005), the availability of new sources of data and new analytic tools are often presented as a significant novelty, especially in the context of issues at the interface of environment and health (Fleming et al. 2017; Leonelli and Tempini 2018).

In this context, I focus on research on an approach and notion that has been introduced in the last decade: the exposome. The exposome is defined as the totality of individuals' exposures to environmental elements (Wild 2005). The notion is often presented as a new paradigm for the study of the relation between health and the environment (Rappaport and Smith 2010). The exposome is a way to describe and characterise the *totality* of environmental exposures. This 'all-encompassing' approach is considered and presented as innovative

because it distinguishes and includes different levels of exposure, including generic external (e.g. social capital, education, financial status), specific external (e.g. radiation, infectious agents, chemical contaminants and environmental pollutants, diet) and internal exposure (e.g. oxidative stress, metabolism, inflammation, ageing).

As a line of epidemiological research, the exposome is relatively new and young and key questions about how the exposome should be defined and studies are currently discussed in the field. As a consequence, exposome research allows me to analyse contemporary research where new concepts are developed and applied in practice. In addition, this allows me to analyse from a philosophical perspective scientific research that is in a flux and unsettled state, where conceptual and methodological discussions are at the centre of the debate. These features of the exposome also enable me to engage in discussions on innovations and changes in the sciences in general and in connection to the debate on the impact of data in particular. On these issues, I ask this research question: what constitutes scientific change and innovation in the context of the exposome? In order to tackle the question, I then focus on the following questions: What is the conceptual, material and methodological background of the exposome? What is the role of data in shaping theoretical, material, social and infrastructural components of scientific research and bringing about change? In answering these questions, I make a contribution on philosophical debates on change in contemporary biomedical research and provide a thorough analysis of the exposome as a 'paradigm' for epidemiology (see Chapter 2).

The exposome presents an interesting case where data from many and significantly diverse sources of evidence are gathered, including: cohort studies, collecting data on populations of interest over a long period of time with questionnaires, retrieval of physical samples (e.g. blood, cord blood, urine), and follow-up; secondary analysis of primary evidence, including omic analysis to quantify and study the effects of exposure at the level of different molecular processes; experimental studies measuring exposure and responses to expo-

sure at an individual level, through wearable and tracking devices; and environmental studies, producing data on air and water pollution through monitoring stations, geo-spatial models and individual estimates. As a consequence, I consider exposome research as a case study in the integration, handling and use of large datasets. I ask this as my research question on these issues: how are large datasets handled, integrated and used as evidence in exposome research? This leads me to tackle the following questions: Which modes, strategies and approaches have arisen and are employed for the study of the exposome? In which ways do epistemic and non-epistemic commitments and assumptions influence the integration of diverse types of data? I ask this question in the context of philosophical discussions on modes and strategies of data integration in the sciences, especially the life and health sciences, and to paint a picture of data and evidence production as epistemically intensive activities (see Chapter 3).

I study these issues by looking at data in a specific project of exposome research known as EXPOsOMICS and funded by the European Union. The project focused on the assessment of disease risk, and chronic disease more particularly, involved in the exposure to air and water pollution, applying the exposome approach through the study of external and internal components (Vineis et al. 2017a). The study of a specific research project specifies my questions and claims and allowed me to focus on research that was in progress. In addition, this focus gives empirical grounding to my philosophical analysis (as I detail in Sect. 1.2). Funding from the European Union was connected to ongoing revision of quality standards of air and water pollution and points to the need for scientific evidence for policy-making. In relation to these questions, I tackle the following question: what types of evidence are produced in the context of exposome research? This leads me to ask the following, more specific questions: How is data used in claims about aspects of the exposome? Which types of evidence does molecular data provide? I ask these questions from the point of view of philosophical, scientific and policy discussions about the evidence produced in medicine. I contribute to these issues with a new case study of research from an underrepresented discipline like epidemiology and hint at

philosophical discussions on the relation between evidence and data (see Chapter 4).

1.2 Methodology: Qualitative Case Study Research and Empirical Grounding in Philosophy of Science in Practice

My research is empirically-informed philosophy of science, based on a qualitative case study of data in epidemiology. The methodology of this dissertation is based on three main elements: it draws on work in the practice turn in science studies and, more specifically, the approach known as philosophy of science in practice; it is an instance of case study research in philosophy of science; and it uses data collected through qualitative research to empirically ground philosophical claims. In the next subsections, I detail each of these elements in turn.

1.2.1 Following the Practice Turn: Philosophy of Science in Practice

When introducing and discussing data studies in philosophy of science, I have mentioned that philosophical work on data has been carried out in the context of increasing attention to scientific practice. Many philosophers have discussed this increase as a significant change in philosophy of science, and more precisely as a "practice turn" leading to a philosophy of science in practice (Soler

et al. 2014).[6] While there is no general agreement on what the practice turn consists in, there are a few defining aspects that need to be considered and are particularly important for my own work. At a minimum, the philosophy of science in practice consists in philosophical studies of scientific practice. But why is this a turn? The reason lies in the aforementioned propositional views of science, which in Anglophone and analytic philosophy of science have traditionally regarded the sciences as bodies of propositions and have subsequently focused on the truth-value and logical relationships of these propositions, usually as sets of propositions organised in scientific theories (Chang 2015). In this tradition, practice has been largely disregarded because it belongs to the non-verbal and non-propositional dimension of science (Ankeny et al. 2011).

In going beyond this tradition, the analysis of scientific practice as the focus of the philosophy of science in practice leads to significant methodological consequences, which Soler and colleagues discuss in terms of specific "shifts" (Soler et al. 2014, pp. 14–24).[7] Firstly, the focus on practice entails a shift from the study of scientific products to the study scientific processes. As a result, one of the goals of the philosophy of science in practice is the development of descriptively adequate and empirically based accounts of the sciences, as opposed to a priori accounts that are typical of syntactic and semantic approaches. This is closely related to another feature of the practice turn, which consists in a shift from decontextualised, general and macroscopic account to

6 My discussion of the practice turn is limited to philosophy of science, but as argued by Soler and colleagues this turn has concerned science studies more generally (Soler et al. 2014, pp. 1-3).

7 Soler and colleagues list six shifts. For the purpose of the presentation of my own work, I discuss four of them, as the other two concern philosophical analyses of the history of science and experimental activities.

"adopting a *local scale* of analysis and paying close attention to the *specific contexts* in which scientific results are produced, used, and disseminated" (Soler et al. 2014, p. 18; emphasis in original). As I detail in the remainder of this section, these shifts have critical consequences of the methodology of philosophy of science, as they push for the inclusion of empirical methods from social and historical sciences as a way of studying scientific practices. Another aim of the philosophy of science in practice is the documentation of aspects of scientific practice that are contingent of a specific situation and are normally left out of final scientific outputs, which is a shift from traditional philosophical methods such as retrospective, rational reconstructions. In this way, the focus on practice leads to a picture of science as the complex result of multi-faceted, contextual, variable and intertwined processes, with a move away from idealised accounts.[8]

My dissertation follows the practice turn in philosophy of science with the choice of focusing on scientific activities and agents involved in the use of data, which makes my research a philosophical study that focuses on scientific practice, rather than on propositional products of scientific research. Data studies in philosophy of science are closely linked to this shift, since data is a crucial element of scientific practice and this is one of the reasons why it is traditionally considered irrelevant in philosophy of science. Following the practice turn, I ask open-ended questions about the ways in which data may inform and sustain the subject matter of epidemiological research on the exposome. The starting point of my research is thus not an idealised account, and I do not presuppose modes of inquiry, strategies and goals employed by scientists. I engage

8 As highlighted by Bschir and colleagues (Bschir et al. 2018), the philosophy of science in practice is one on the many and new approaches of current philosophy of science, and as such it should not be seen as a rejection of all the other approaches and traditions in the field.

with my case study with the aim of building an account of the practices involved in the collection, analysis and use of data in exposome research. On the basis of this account, I discuss the case at a more theoretical level, by either engaging with conceptual notions discussed in philosophy of science (e.g. evidence, scientific change) or providing conceptual distinctions of data practices (e.g. epistemic strategies, commitments and issues; relations between the cognitive and material dimensions of scientific practice).

Thus, the main goal of the dissertation is largely descriptive, as I aim for an empirically adequate account of the use of data in epidemiological research on the exposome. The normative and descriptive features of philosophical research are among the most debated issues of the philosophy of science in practice, and are indeed discussed by Soler and colleagues as one of the shifts that characterise the practice turn (Soler et al. 2014, pp. 15-16; see also Lynch 2014). Namely, in this context, philosophers have often taken a descriptive approach, aimed at making sense of what scientists do in the first instance. Normativity in philosophy of science in practice has been interpreted in terms of a critique of certain scientific practices or the ways in which they are discussed in the sciences, philosophy and beyond, with the aim of developing and providing normative guidance that may help do better science in specific contexts of investigation (Wimsatt 2007, Chap. 2-3). In this dissertation, I do not present a critique of data practices of exposome research, nor do I offer specific recommendations on ways to improve them. I do engage in discussions at a more normative level, but these are not aimed at the ways in which scientists work, as much as at the ways in which their practices have been discussed in philosophical, sociological and popular accounts of the sciences. In this sense, the normative character of my dissertation can be seen in the fact that I present my own position on these discussions and views, my own perspective on the phenomenon that I study, my own distinctions of what the phenomenon consists

in and my own assessment of how significant this phenomenon is for other components of scientific epistemology.[9]

1.2.2 Case Study Research

At the basis of the philosophy of science in practice is an appreciation of the heterogeneous nature of scientific practice. This, in turn, implies that philosophers employ various approaches to the study and analysis of scientific practice. In this dissertation, I have followed a specific approach based on a qualitative case study of a specific scientific project. This choice is grounded on methodological discussions in the philosophy, history and social studies of science.

According to Wolfgang Pietsch (2016), case studies can be defined as detailed analyses of episodes and their context: they may be analysed with various methods (empirical investigations, statistical analysis, archival work, etc.); and are always case studies for something, in that the episode and its context are analysed from the perspective of specific question, concept or phenomenon. As a consequence, case studies are a crucial aspect of the history and philosophy of science because they "provide the essential link between the history and the philosophy of science" (Pietsch 2016, p. 49). Unsurprisingly, much of the discussion on case studies has taken place in the history of science, where case studies are often used as the main objects of research (Chang 2011). More generally, in philosophy of science case studies can provide grounding to philosophical considerations about various aspects of scientific epistemology; offer challenges and counterexamples to philosophical considerations; and enable philosophers to discover new phenomena and develop ways of accounting for them. Pietsch distinguishes between two modes of reasoning with case studies

9 See similar conclusions drawn by Leonelli (2016, p. 9).

in the sciences and the humanities: a predictive mode, with the goal of predicting other episodes similar to the case study; and a conceptual mode, with the goal of answering questions about the conceptual framework used to analyse and account for the phenomena in the case study. Beyond methodological reasons for why case studies are used in philosophy of science, Adrien Currie (2015) argues that there are historical reasons driving their use in the field. Currie distinguishes between three roles played by case studies in philosophical methodology: as inductive evidence for general claims; for pragmatic and rhetorical purposes, as ways to ground a philosophical debate; and as ways to test the success of philosophical reconstructions of scientific practice and concepts.

In the research of this dissertation, I follow a moderate version of Pietsch's conceptual mode and Currie's view of case studies as tests for philosophical reconstructions of scientific practice and concepts. My primary interest is to map the use of data in exposome research and to understand it by introducing and discussing philosophical accounts of the role of data in scientific research. In this sense, I do not start from preconceived accounts of data practices, but from the empirical reconstruction of my case study, and get to the conceptual level on the basis of this empirical evidence. This use of case studies is grounded in philosophical analyses of the use of case studies in the social sciences. In her analysis of case study research in the social sciences, Mary Morgan has presented a definition of the case study according to which the unit of analysis of case studies is "a bounded whole object of analysis" (Morgan 2012, p. 668). In this view, case studies aim at analysing whole processes, events, or evidence (as opposed to the focus on specific aspects or bits of evidence) and in continuity with their context. Moreover, case studies are open-ended, as there are no preconceived limits on the amount of work, specificity of the topic and separation between subject and context. Case studies also require the study of "a 'real-life' whole", which implies extensive engagement with the subject and collection of evidence on a variety of different aspects (Morgan 2012, p. 668). As a result, the output of case study research is a "complex, often narrated, account that typically contains some of the raw evidence as well as its analysis and that ties together the many different bits of evidence in the study" (Morgan 2012, p.

668). Another element highlighted by Morgan is that case studies may not be used only for theory or hypothesis testing, but rather for the development of "evidence-based concepts" and for "revealing phenomena and developing accounts of them" (Morgan 2012, p. 671).

In my case study research in the dissertation, I draw on these insights and I engage with real-life processes of research in EXPOsOMICS. I aim at analysing data practices as Morgan's whole processes, as I follow the ways in which data is collected, integrated and used as evidence and I situate them in a broader material, social and epistemic background. By closely looking at scientific processes and their context and building what can be considered a microscopic and local account (Soler et al. 2014, p. 18), I also draw inspiration from work in feminist epistemology and philosophy of science. In this line of research, philosophers view knowers as constantly "situated in particular relations to what is known and to other knowers", which implies that "what is known, and the way that it is known, thereby reflects the situation or perspective of the knower" (Anderson 2017). As a metaphilosophical and methodological implication of this view, philosophers have argued for the need of situated accounts of scientific knowledge, aimed at locating epistemic practices, agents and knowledge in the material, institutional, and cultural context of their standpoint (Haraway 1988; Wylie 2012; see also Leonelli 2016a, p. 190). I thus take a bottom-up approach to my case study, in which discussions of issues at a conceptual level are based on empirical accounts of the case I discuss. As a consequence, I discuss and introduce what can be considered Morgan's evidence-based concepts, such as the "exposome repertoire" (see Chapter 2, briefly introduced in the next section) and "evidential claims" (see Chapter 3, briefly introduced in the next section).

As a result, the output of my case study research presents both descriptive and conceptual components. In this sense, I do not use my case study with the aim of general or universal results: rather, I aim at discussing concepts at a 'mesoscopic' level, that is adapted and grounded in a specific context (Pietsch 2016; Burian 2011). I further rely on Pietsch's idea of case studies being always case studies for something, as I take the exposome to be a case study of the use

of data in contemporary epidemiology and the ways in which data practices inform and sustain its subject matter. In this sense, my case study can thus be considered to have a "use-value" (Morgan 2019) that is connection to comparisons with other contexts and cases, to build "bridges" between similarities (Morgan 2014, pp. 1015-1016), and to strengthen the compelling character of a case that has been the focus of other philosophical research.

1.2.3 Empirical Methods for Philosophers: Qualitative Interviews

According to Bschir and colleagues:

> Philosophers of scientific practice typically identify as naturalist philosophers whose methods are continuous with those of the sciences. As a consequence, methods from the arsenal of the social sciences have become more and more integrated into the philosophy of science, blurring the boundaries between philosophy of science and sociology of science. (Bschir et al. 2018)

In my case study research on data in exposome research, I have relied on methodologies from social science research to collect evidence on scientific practices. The use of empirical methods from the historical or social sciences is a recent addition to philosophy of science methodology and is a crucial aspect of the practice turn and the philosophy of science in practice, in whose context it is

discussed as one of the main ways to "acquire insights into and evidence of scientists' research behaviour" (Boumans and Leonelli 2013, p. 260).[10]

For this dissertation, as a result of the choice of case study, I decided to employ empirical methods to study data in the context of EXPOsOMICS. Only very little of data is discussed by scientists in their publications and presentations, which made it necessary to directly engage with EXPOsOMICS researchers. Using an approach that is often applied in the context of philosophy of science in practice (Osbeck and Nersessian 2015) and philosophical studies of data practices (e.g. Leonelli 2010a, pp. 120-122), I carried out qualitative interviews with EXPOsOMICS researchers. Methods such as qualitative interviews are used when social scientisats are interested in gaining insight into the 'worlds' of others, i.e. how participants view, experience or conceptualise an aspect of social life (Flick 2014, Chap. 13). The interviews were semi-structured, with a list of specific topics to cover that was flexible regarding the order and phrasing of the questions (Kelly 2010). They were based on an approach called "informed observation", which Laudel and Gläser define as follows:

> With "informed observation" we refer to social studies of science undertaken by sociologists who acquire a scientific understanding of the field they study by self-education prior to or at the beginning of their empirical study. (Laudel and Gläser 2007, p. 95)

In the literature, this approach is connected to the notion of interviewing as a social interaction and performance that generates new knowledge, as a result of a communication process where the meaning of questions and answers is

---

10 The integration of empirical methods in philosophical analyses has also kickstarted discussions about methodology in philosophy of science, a topic that traditionally has not received detailed consideration in the field. See, for example, the discussion of case studies in the previous subsection and, more generally, Wagenknecht and colleagues (Wagenknecht et al. 2015).

built jointly by interviewers and interviewees (Laudel and Gläser 2007, p. 98). An understanding was necessary for my observations because I wanted to get into the details of the data employed by EXPOsOMICS researchers. Another argument that is often used for informed interviewing is that being informed demonstrates the competence of interviewers and helps with being taken seriously. In my research, I would argue that this has not played a particularly important role, as my preliminary knowledge of exposome research often could not get into the specifics of data practices. As a result, many of the questions I asked were aimed at understanding which data practices are carried out, which types of data and meta-data are involved, which methods are used to collect and stored data, which assumptions are underlying these steps and how these are perceived and discussed by EXPOsOMICS researchers. Questions were thus based on my knowledge of exposome research and the philosophical issues I was interested to tackle.

I conducted five interviews. Participants were recruited on the basis of the suggestions of the principal investigator of EXPOsOMICS and following a snowball method. Interviews lasted around an hour and were audio-recorded, following consent by participants. All the interviews took place at the Department of Epidemiology and Biostatistics of Imperial College London, UK, where I was hosted for a week in January 2017. During the week, I did some participant observation through informal discussions with researchers and by attending a weekly meeting of the group. The background, skillset and role in the project of the interviewees varied, including molecular biology, statistics, epidemiology and medical science. I argue that this provided me a good representation of the work of EXPOsOMICS.

In the months following the fieldwork, I transcribed and analysed the interviews and I took the data collected through the interview as providing the empirical grounding of my philosophical analysis. I made a list of excerpts and themes that emerged from the interview transcripts and developed in the course of my analysis. This included: types of data used in EXPOsOMICS; the management of large datasets and modes of data integration; the current landscape of innovations in epidemiology; the concept of the exposome and

changes to other approaches and foundational notions of epidemiology; interdisciplinarity and collaboration in contemporary science. In carrying out the interviews, my interest was to gain an understanding of the ways in which data is used in the project. At the same time, I did not take what interviewees said only at face value, as I connected their claims with publications and reports of exposome research. Throughout the dissertation, when I discuss insights or information from the interviews, I provide references to published articles, as a way of giving further grounding to my claims. In addition, I gave my own philosophical interpretation to the material I collected from the interviews. Namely, the main claims of the subsequent chapters were based on a mix of input from the interviews data, the philosophical literature and my own analysis. The claims that I present in the dissertation are my own, as for example I did not discuss the notion of repertoire (Chapter 2), evidential claims (Chapter 3) and evidential pluralism (Chapter 4) directly with the interviewees.

1.3 Overview of the Chapters

My dissertation consists of this introductory chapter, three articles that I reproduce as the main chapters of the dissertation[11], and a concluding chapter. The articles all focus on data in the context of epidemiological research on the exposome, that I have accounted for on the basis of empirical research on the

11 I reproduce the articles in their current publication or submission form. Turning articles into chapters, I have only edited page, section, footnote, figure and table numbering and referencing (i.e. a reference to Section 1 in article form is now a reference to Section 1.1, the phrase "In this paper I will" has been changed to "In this chapter I will", etc.). Following university regulations, I have not made any other change to the content of the articles. This creates some repetition of references and premises across the central chapters, as these are primarily intended to function as stand-alone articles.

EXPOsOMICS project. I present an account of how data practices are aligned to and inform the cognitive, material and social dimensions of scientific research on these issues.

I start by providing an account of the conditions under which the exposome was conceived, developed and established in epidemiology, in Chapter 2 "The Exposome as a Postgenomic Repertoire: Exploring Scientific Change in Contemporary Epidemiology".[12]

I start by discussing the ways in which the exposome has been introduced (Wild 2005). The exposome is a way to describe and characterise the totality of environmental exposures. This 'all-encompassing' approach is considered and presented as highly innovative because it distinguishes different levels of exposure, including generic external (e.g. social capital), specific external (e.g. environmental pollutants) and internal exposure (e.g. oxidative stress). As such, it has been presented as a new "paradigm" in the scientific literature (Rappaport and Smith 2010). However, I show that the paradigm framework does not work here, because: the exposome has strong continuity with longstanding approaches in epidemiology and environmental science; and the plurality of exposome approaches, solutions and procedures lacks the theoretical coherence of paradigms. Instead of a new paradigm, I argue that the innovations of the exposome are better captured by an account of the exposome as a *repertoire*. The repertoire framework has recently been introduced in the philosophical literature as a way to understand scientific change beyond traditional approaches like Kuhnian paradigms (Ankeny and Leonelli 2016). A repertoire is the result of the alignment of conceptual, institutional, material, technological, organisational and economic elements of scientific research.

---

12 This chapter will be submitted for publication in an academic journal at a later stage.

In the repertoires framework, scientific change may happen when a new repertoire is developed, components of a repertoire change or when a repertoire is re-instantiated in another context and varied. I use this feature of the framework to show that the exposome is the result of the merging of elements from three other lines of research: the sequencing repertoire, which emerged in the genomic context and shares with the exposome repertoire technical, conceptual and funding elements; exposure science, which studies human contact with external agents and shares with the exposome repertoire interdisciplinary approaches and teams, the conceptualisation of the internal components of exposure and rhetorical tools from risk assessment; biomarkers research, which shares with the exposome repertoire laboratory techniques and conceptualisations of disease. I show that the exposome approach is the alignment of these components, which are repurposed for new audiences, as well as newly introduced components and elements.

As a second step, I use the repertoires framework to draw an account of scientific innovations in the exposome and argue that the exposome is a postgenomic repertoire. I use the term here with a historical meaning, i.e. as a way to describe research that employs genomic-based technologies, is increasingly aware of the complexity in interpreting genomic results and has a critical engagement with gene-centric approaches (Stevens and Richardson 2015). The historical use of the term postgenomic leads me to a nuanced account of scientific change. First, in the context of conceptual consequences of the exposome, I focus on the notion of environment. I argue that the pluralistic approach to conceptualising and operationalising the environment in the exposome is in continuity with traditional epidemiology and can hardly be seen as a paradigmatic change. Second, I focus on technical innovation of the exposome and I discuss the role of data. Discussions on data are another context where scientific change has been discussed recently, as many have argued that paradigmatic changes are due to increasing quantities of data used in the sciences. My account of the exposome as a repertoire points to the alignment of data with other ingredients of the repertoire. While the availability of new and different

types may lead to the revision of notions like environment and exposure, these changes are due to the alignment of various components, rather than only data.

On the basis of this initial analysis of the background conditions of the exposome, I present an empirical reconstruction of data practices employed and aligned with the other ingredients of the repertoire. In Chapter 3 ("Making Evidential Claims in Epidemiology: Three Strategies for the Study of the Exposome"[13]), I closely analyse the data practices involved in exposome research and show how different epistemic strategies have arisen in this context.

I start by introducing the main theoretical tool I use in the article: evidential claims. I view evidential claims as claims that identify the datasets that are to be used as evidence for the investigation of phenomena. I draw inspiration from philosophical inquiries on archaeology, where evidential claims are discussed as the main results of archaeological research (Chapman and Wylie 2016; Wylie 2017). Using the notion, I highlight that different types of evidential claims are generated at different stages of exposome research and through different approaches, methods and lines of work. I claim that various aspects of exposome research can be interpreted as relying on or building towards evidential claims, beyond the explicit claims made by EXPOsOMICS researchers. In particular, I distinguish three epistemic strategies. The macro strategy identifies the initial evidence platform, generating scoping claims that restrict the sample and provide an initial understanding of the phenomena to focus on. As part of the micro strategy, significantly different kinds of data are collected, with the common aim of identifying structure at microscopic level and elaborating evidential claims about microscopic structures. On the basis of the evi-

13 This chapter is currently under revision and being considered for publication by *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*.

dential claims generated by the first two strategies, the association level provides evidential claims at the statistical level, concerning associations between exposure to the environment and outcomes of interest.

I argue that these strategies are a crucial step of research, where datasets are given value and representation content through evidential claims. They differ in terms of level of abstraction, line of work and type of evidential claim, and are ways of dealing with the increasing complexity and diversity of datasets. In addition, I argue that viewing data practices from the perspective of evidential claims and distinguishing strategies for evidential claims yield significant insights. It enables to unpack the epistemic issues and challenges that concern each strategy and, in turn, influence research done at a different stage. It gives a characterisation of the context of data practices in terms of evidential claims, which shows how much epidemiological research is not necessarily about causal claims, but neither is to be overlooked as producing 'raw data'.

Furthermore, I argue that this focus on evidential claims provides a new perspective on epidemiological research. Firstly, in the context of discussions on epistemic changes related to data, while some of these strategies present a novelty that is closely related to the availability of large datasets, I paint a picture of evidence production as epistemic-intensive labour, in contrast to views of big data shaping scientific research towards automatic and theory-free approaches. Secondly, I contrast my approach with existing philosophical accounts of epidemiology, that have vastly focused on causality and causal inference (e.g. Broadbent 2013). I argue that it would not be charitable to interpret all epidemiological research as primarily concerned with causality or to focus only on causal claims. An analysis based only on causality focuses for the most part on the end results of epidemiological research and may thus end up overlooking the crucial epistemic role of other elements and claims that proceed final results but make them possible and significantly influence them. Investigating the role of data and evidence in epidemiology as one of the central outputs and results of research enables to shed lights on these issues, which are particularly important in the context of life and health sciences contributing to evidence used in policy.

In the context of discussions on evidence, in Chapter 4 ("Evaluating Evidential Pluralism in Epidemiology: Mechanistic Evidence in Exposome Research"[14]), I connect my analysis of data practices in exposome research with philosophical accounts of evidence and evidence classification. In current discussions on evidence in the medical sciences, epidemiology has been used to exemplify a specific version of evidential pluralism. According to this view, known as the Russo-Williamson Thesis, evidence of both difference-making and mechanisms is produced to make causal claims (Russo and Williamson 2007). This approach has been employed to describe the types of evidence that are produced in epidemiological and, more specifically, exposome research (Russo and Williamson 2012; Russo and Vineis 2016).

In the chapter, I analyse approaches to evidence classification in exposome research and I cast doubt on the extent to which evidential pluralism holds in this case. I start by focusing on the claim that molecular data allows for the production of mechanistic evidence. According to Russo and Vineis (2016), exposome research can produce – and it does not only use – mechanistic evidence, and that this production is to be seen in the molecular features of exposome research. Within the Russo-Williamson Thesis framework, mechanistic evidence is evidence of the existence of mechanisms, evidence that is about mechanism and has mechanisms as its' object. Following Illari and Williamson (2012), mechanisms evidence thus refers to evidence of the entities, activities or the way these are organised to produce the phenomenon for which the mechanism is responsible. I thus closely analyse data practices in the context of molecular data in exposome research, by looking at omics data and exposure profiles. I argue that molecular evidence has been used for claims about the

14 This chapter is published in *History and Philosophy of the Life Sciences*, 41(4).

difference that exposures make, which is in contrast with the notion of mechanistic evidence proposed by the Russo-Williamson Thesis. I therefore caution against interpretations in terms of mechanistic evidence.

Then, I focus on another aspect of the Russo-Williamson Thesis, i.e., the way evidence is classified and distinguished between different kinds. I expand my critical remarks on the thesis by empirically reconstructing data practices in EXPOsOMICS and addressing the conditions under which data is categorised as evidence in exposome research. I argue that these show that the classification of a dataset as a type of evidence is dependent on the ways in which the data is used and therefore researchers may have substantially different views and approaches to what counts as a type of evidence. This is in contrast with the approach used by evidential pluralism, where evidence is classified in different types on the basis of intrinsic properties of the type of evidence and a specific type of evidence is usually linked to a specific method used for the generation of evidence. In these views, data counts as evidence on the basis of properties that are fixed, inherently local and stand in a representational relation with aspects of reality, independently of the context where the data is used. I compare and contrast this approach to relational accounts of data, where data is taken to be a relational notion with a non-representational character (Leonelli 2016a, pp. 69-92).

Finally, I come back to what I consider the core of the thesis and suggest that exposome research, and epidemiology more generally, indicate different interpretations of evidential pluralism and its applicability in the health sciences. I propose an interdisciplinary and use-based interpretation of the thesis, which takes into account the critical points I have raised in the chapter. Additionally, it tries to push the thesis forward, towards a direction where it is capable of accounting for the use of data as evidence in current research and can therefore make fitting suggestions on the consideration and appreciation of a plurality of medical evidence.

Finally, Chapter 5 presents the conclusions of the dissertation. I introduce and discuss the main themes discussed in the dissertation, including the epistemic

role of data, the epistemic-intensive character of evidence production and the status of empirical knowledge and data in contemporary research. I conclude by presenting an outlook of this research and pointing to topics that have been explored in the dissertation but will require more research. I discuss the notion of evidence and compare my characterisation of data practices in exposome research with conceptualisations of the notion in the philosophical literature. Similarly, I reflect on what counts as disease and health in epidemiological research and contrast these approaches to more traditional views and accounts in the literature on the philosophy of the life and health sciences.

The dissertation concludes with Appendices related to my empirical research, including a copy of the consent form.

Chapter 2

The Exposome as a Postgenomic Repertoire: Exploring Scientific Change in Contemporary Epidemiology

Abstract: In the last decade, a new concept has emerged in epidemiology: the 'exposome', defined as the totality of exposures experienced by individuals. The concept is often presented as a new paradigm for the study of the relation between health and the environment. In the chapter, I analyse the conditions under which the exposome was conceived, developed and established. I argue that these point to the establishment of an exposome repertoire, not a paradigm. I use this framework to show the alignment of the epistemic elements of the exposome with material, financial, institutional and technological factors. I argue that some of these factors were transferred from other areas of the life and health sciences, including sequencing, exposure science and biomarkers research. I then analyse the conceptual and material innovations of the exposome through the lens of broader discussions in the context of the life and health sciences, by arguing that the exposome can be considered a postgenomic repertoire.

2.1 Introduction

In the last decade, a new concept has emerged in the context of epidemiological research: the 'exposome'. The exposome is defined as the totality of individuals' exposures to environmental elements. This includes the various internal and external elements and substances an individual is exposed to at any given

time, as well as the different exposures that individuals experience throughout their lifetime (Wild 2005). For instance, studying the effects of air pollution through the exposome involves studying health outcomes at the population level, air pollution at the environmental level, as well as internal exposure at the molecular level. The concept is usually presented as a new "paradigm" for the study of the relation between health and the environment (Rappaport and Smith 2010).

In this Chapter, I analyse the conditions in which the exposome was conceived, developed and established in the wider context of the life and health sciences. I argue that, rather than a new paradigm, these conditions point to the establishment of a 'repertoire', where epistemic and conceptual elements are aligned with material, financial, institutional and technological factors. I draw on Rachel Ankeny and Sabina Leonelli's repertoires framework, which is aimed at broadening traditional views of scientific change and capturing the role of elements of scientific research that have usually been overlooked in the philosophical literature. I use the repertoires framework to specify the role of material, technological and institutional components in the development of the exposome. This specification allows me to draw connections between many of these components and other disciplines or repertoires in life and health science research. Furthermore, I discuss the changes brought about by the introduction of the exposome repertoire by arguing that it can be considered a postgenomic repertoire, whose innovative status is to be seen in continuity with other approaches in the field.

My analysis of the exposome repertoire is grounded in the study of publications, reports and presentations in the scientific literature on the exposome; as well as empirical research that I carried out in the context of a particular expo-

some project, EXPOsOMICS, which included a series of interviews with exposome researchers.[15] I proceed as follows. I start by introducing the concept of the exposome and the ways in which it has been presented in the literature. I argue that the exposome has been established as a repertoire and I specify its material, social and epistemic components (Sect. 2.2). I then argue that three traditions played a crucial role in the development of the exposome repertoire though the transfer and modification of its components: the sequencing repertoire, biomarkers research and exposure science (Sect. 2.3). Thus, I analyse the innovations of the exposome through the lens of broader discussions in the context of the life and health sciences. I argue that the exposome can be considered a postgenomic repertoire, which leads me to discuss conceptual and material forms of innovation (Sect. 2.4).

2.2 Using Repertoires to Understand the Exposome

The term 'exposome' was first introduced by Christopher Paul Wild in 2005, with an article that appeared on *Cancer Epidemiology Biomarkers and Prevention* with the title "Complementing the genome with an "exposome": The outstanding challenge of environmental exposure measurement in molecular epidemiology" (Wild 2005). Wild characterised the exposome as a way of describing the totality of environmental exposures that individuals are exposed to in their life-course. The exposome was to "encompasses life-course environmental exposures (including lifestyle factors), from the prenatal period onwards", as a

15 EXPOsOMICS ran between 2012 and 2017 and applied the exposome approach to the assessment of disease risk related to air and water pollution, by studying external and internal components of the exposome (Vineis et al. 2017).

"highly variable and dynamic entity that evolves throughout the lifetime of the individual" (Wild 2005: 1848).

The exposome is a way to describe and characterise the totality of environmental exposures, including exposures horizontally and vertically: all the environmental exposures experienced at any given point in life, at both the internal and external level (horizontal perspective); and all exposures experienced by individuals, from their conception onward (vertical perspective). In the field, this 'all-encompassing' approach is considered and presented as highly innovative. Innovation here is connected to going beyond the traditional study of exposure in epidemiology, that focused mostly on the external level of exposure by measuring the presence and interaction with environmental elements and chemicals. The exposome, instead, distinguishes between different levels of exposure: generic external exposure (social capital, education, financial status, psychological and mental stress, urban–rural environment, climate, etc.); specific external exposure (radiation, infectious agents, chemical contaminants and environmental pollutants, diet, lifestyle factors, occupation, medical interventions, etc.); and internal exposure (metabolism, endogenous hormones, body morphology, physical activity, gut microflora, inflammation, lipid peroxidation, oxidative stress, ageing, etc.). Exposure is measured through different types of sampling, including: indirect sampling through proxies that can be correlated to otherwise unmeasurable variables of generic external exposure (e.g. eligibility for free school meals as a proxy for socio-economic status); sampling of sources of specific external exposure, through specific-measurement campaigns (e.g. air and water pollution); and sampling of molecular markers of presence or effects of internal exposure (e.g. molecular analysis of blood samples).

On this basis, the goal of the approach is to identify associations between these different components, as a way of studying disease as it develops through the different levels and moves through potential pathways. For example, The EXPOsOMICS project studied the relation between exposure to ultrafine particles and asthma, on the basis of pollution estimates of ultrafine particles de-

rived from air quality measurements and metabolomic analyses of internal exposures from blood samples (Jeong et al. 2018). The investigation showed a strong association between some ultrafine particles and adult-onset asthma. Moreover, through the focus on internal exposure, it identified metabolic pathways which are associated with both air pollutants and health outcomes and could be considered to mediate the effects of air pollution on disease. The inclusion of this internal perspective, that is considered a component of the exposome and thus a type of exposure as important as the external ones, is seen as one of the innovations of the exposome (Rappaport and Smith 2010). Beyond the inclusion of the internal component of exposure, the simultaneous study of different levels of exposure and chemicals at the same time is considered innovative in contrast to the traditional focus on single chemicals or single types of exposure. Finally, the all-encompassing approach is also a way of implementing a life-course and dynamic approach, thus moving away from measurements at single points and towards an understanding of issues of exposure and disease as developing dynamically throughout lifetime.

In presentations, introductions and discussions of the exposome, the term 'paradigm' is often used. The exposome has been presented as: "a new and exciting paradigm for improvement and integration of currently scattered and uncertain data on the environmental component in disease aetiology" (Vrijheid 2014, p. 876); an "operational paradigm" for exposure science (Rappaport 2011, p. 5); a "paradigm shift" for public health (Juarez et al. 2014, p. 12868); a new "research paradigm" for environmental epidemiology (Stingone et al. 2017, p. 316), planetary health (Logan et al. 2018) and biomonitoring (Dennis et al. 2017). In philosophy of science, the term paradigm was introduced and popularised by Thomas Kuhn (1962). Kuhnian aradigms are the "shared commitments of a scientific group" (Kuhn 1977, p. 294), i.e. the concepts, theories, assumptions, generalisations, values, exemplary problems and solutions, etc. that provide a specific identity and constitution to research communities in the

sciences.[16] The term is very popular in the sciences, where it often used with a loose and generic meaning to refer to theoretical shifts and changes, whose dramatic character can sometimes be overstated (Hoyningen-Huene 1993, p. 131). This loose use of the term seems close to the way in which the exposome is discussed as a paradigm. Namely, as we will see in the chapter, the exposome is in continuity with longstanding topics and approaches in epidemiology and environmental science and has a plurality of conceptual approaches, which seem in contrast with the idea of a radical paradigmatic shift and the theoretical coherence of Kuhnian paradigms. More importantly, while the paradigms framework could arguably be shaped to account for elements of the exposome, I argue that many innovative aspects of the exposome are related to the technological, material and institutional dimensions of scientific research, that are not at the centre of the paradigms framework.

In order to clarify what scientific innovation in the case of the exposome consists in, I present the claim that the exposome has been established as a new repertoire, and not a paradigm. The repertoires framework has been introduced by Rachel Ankeny and Sabina Leonelli as a way to understand scientific collaboration, community-building and change (Leonelli and Ankeny 2015). Ankeny and Leonelli argue that the success of a research community in establishing durable, stable and coherent change is often dependent on the development of repertoires, which are defined as:

> The well-aligned assemblages of skills, behaviors, and material, social, and epistemic components that groups may use to practice certain kinds of sci-

16 For a detailed analysis of the notion of paradigm, including its evolution in Kuhn's work after 1962, see Kuhn (1977, Chap. 12) and Hoyningen-Huene (1993, p. 131-162).

> ence, and whose enactment affects the methods and results of research, including how groups practice and manage research and train newcomers. (Ankeny and Leonelli 2016, p. 20)

The constituents of a repertoire are thus of three main types: material, including specimen, resources, data, skills and training; conceptual, including theoretical commitments, norms, values and goals; and social, including funding, institutions and infrastructures.[17] In the framework, repertoires are successfully established when these elements are aligned, both internally and externally. Namely, this occurs when members of the community know their roles, possess the relevant skills, are able to perform them, operate in the context of beneficial funding and institutional support, and can communicate and promote their approach and outputs inside and outside their community. For example, Ankeny and Leonelli argue that the successful discovery of the Higgs boson by CERN scientists was not only due to epistemic components, but also the alignment of their research practices with funding and institutional support (Ankeny and Leonelli 2016, p. 26). Repertoires do not necessarily arise as new approaches within a community, but are often transferred from other lines of research and disciplines; like franchising businesses, transfer implies the modification and adaptation of the repertoire, which retains its identifiability while evolving and developing from its original instantiation. For instance, Ankeny and Leonelli contrast the CERN success with attempts at implementing the same repertoire in the United States, which did not succeed because of the failure in the adaptation of crucial components like funding and governmental policies (Ankeny and Leonelli 2016, p. 26).

The successful transfer, adaptation and alignment of material, conceptual and social components into a repertoire is thus one of the ways in which scientific

---

17 This is not a strict classification, as elements a repertoire can often be connected to more than one type: for instance, professional training of researchers can be considered a material as much as a social component.

change can manifest itself. As a consequence, in the context of repertoires, change does not necessarily happen in dramatic, paradigmatic ways (Ankeny and Leonelli, forthcoming). Moreover, several repertories may exist simultaneously within a single discipline or community. Ankeny and Leonelli use the framework to put the emphasis on the role of social, institutional, material, technological, organisational and economic elements in the successful establishment of communities, moving away from a privileged focus on theoretical and propositional knowledge. In this chapter, I use the repertoires framework to account for the various components and innovative elements of the exposome. I argue that the exposome does not hinge on conceptual components only – instead, the exposome should be analysed as an assemblage of aligned material, social and epistemic components, which constitute the *exposome repertoire*.

By arguing that the exposome is a repertoire, I am arguing that the exposome has emerged in the context of contemporary epidemiology (also) thanks to the crucial role played by non-conceptual components at the material and social level. Namely, the exposome is built on a conceptual commitment to the understanding of exposure as a dynamic, multifaceted and global issue and the study of the totality of exposures. Focusing on the conceptual implications of the exposome is important: for instance, as I analyse in Sect. 2.4, the exposome implies an expansion of the concept of exposure and a broad characterisation of environment. Yet, an analysis of the conceptual level should not overlook the role of performative components, which influence and enable conceptual commitments, and can be innovative elements per se, independently or in relation with the conceptual level. For the exposome to be a successful line of research, the social, material and conceptual components of the repertoire have to be intertwined and aligned.

At the material level, the repertoire is based on the employment of omic technologies and the biomarkers approach. Omic techniques are used to quantify and study molecules at the level of: intermediary functioning of metabolism (metabolomics), protein or DNA adducts (adductomics), epigenetic changes

(epigenomics), mRNA (transcriptomics) and proteins (proteomics). Biomarkers are elements that can be precisely measured and used as indicators or traces of various processes within an organism. The use of these techniques has significant influences on material specimen and data. Data sources used in exposome research include: cohort studies, collecting data on populations of interest over a long period of time with questionnaires, retrieval of physical samples (e.g. blood, cord blood, urine), and follow-up; secondary analysis of primary evidence, primarily omics techniques; experimental studies measuring exposure and responses to exposure at an individual level, through wearable and tracking devices; and environmental studies, producing data on air and water pollution through monitoring stations, geo-spatial models and individual estimates. The diversity of sources of data, which span from population level cohort data to micro-level data about the internal component of the exposome and specific features of an individual's environment, has significant influences at the methodological level, as researchers use various epistemic strategies to handle complexity (see Chapter 3). In turn, these features of material specimen and data influences the skillset of researchers, which is increasingly interdisciplinary and includes: wet lab molecular biology, for the use and analysis of omics techniques and data; statistics, for the development of regression models, analysis of associations and data handling; project and consortium management for internal communication, funding application, scheduling, etc.; information science for data curation and integration. The resulting assemblage of skills in a way further blurs the disciplinary boundaries of epidemiology, which in itself has arguably always escaped strict disciplinary definitions.[18] Moreover, it reinforces interpretations of the current landscape of life sciences as profoundly interdisciplinary, especially in connection with an increasingly data-rich context (Richardson and Stevens 2015, pp. 236-237). At the same time, the material, methodological and disciplinary composition of the

18 See Alfredo Morabia's presentation of epidemiology as a "theory" that is based on "principles" (Morabia 2005, p. 5).

exposome repertoire has a consequence on key conceptual components of the repertoire. It pushes for molecular and biological conceptualisations of what counts as a disease state, which is significant in the context of epidemiological research that has traditionally looked at the issue of disease as a statistical and macroscopic matter (Broadbent 2013).

At a social level, the exposome repertoire is organised in short term research projects and consortia with framing and funding for public health, disease risk and environmental health (see the conceptual, material and social elements of the repertoire in Table 1). Institutionally, the organisations that have been more responsive to the exposome repertoire are large funding agencies and schemes focused on public health and the environment. Funding for the exposome has been directed to different – but often interrelated – research projects, including cohort studies collecting data that can then be used to study specific aspects and components of the exposome and projects aiming at analysing and integrating data with other datasets. While in the US there has been a significant trend in establishing research centres about the exposome, which focus extensively on training and education, in Europe the exposome repertoire has been particularly successful at securing short-term funding from the European Commission. The exposome has a dedicated track in Horizon 2020, the next funding programme of the EU Commission, where the approach is defined as a "toolbox for assessing and addressing the impact of environment on health" (see European Commission 2019). The funding context plays an important role in the repertoire because it influences the way in which the exposome is conceptualised and framed as a way of addressing the challenge to improve health and reducing the burden of disease. At the same time, the social components of the repertoire have an influence on the ways in which the exposome is implemented. For example, the short lifespan of exposome projects is a challenge to how much can be achieved in the repertoire. In a project like EXPOsOMICS, that run for five years, one of the challenges had to do with data analysis: for all the data that was collected in the project, five years were not enough for analysis and so much data had to be left for future research.

Type of Ingredient	*Elements of the repertoire*
Conceptual	Adoption and use of the exposome notion Expansion of notion of 'exposure' Broad characterisation of 'environment'
Methodological	Omic technologies Biomarkers and intermediate biomarkers Interdisciplinary teams (molecular biology, data curation, epidemiology, medicine, statistics, management)
Financial and institutional	Short term research projects Long term cohort studies Consortia and research centres Funding priorities and policy mandates: public health; disease risk; environmental health

Table 1. Elements of the exposome repertoire

2.3 The Lineage of the Exposome as an Emerging Repertoire

In the previous section, I have argued that the exposome is the result of the alignment of conceptual, material and social components. I now use the repertoires framework to analyse the conditions in which the exposome has been conceived, developed and established and connect the components of the exposome as a repertoire to other repertoires and traditions in epidemiology and beyond (see Figure 1).

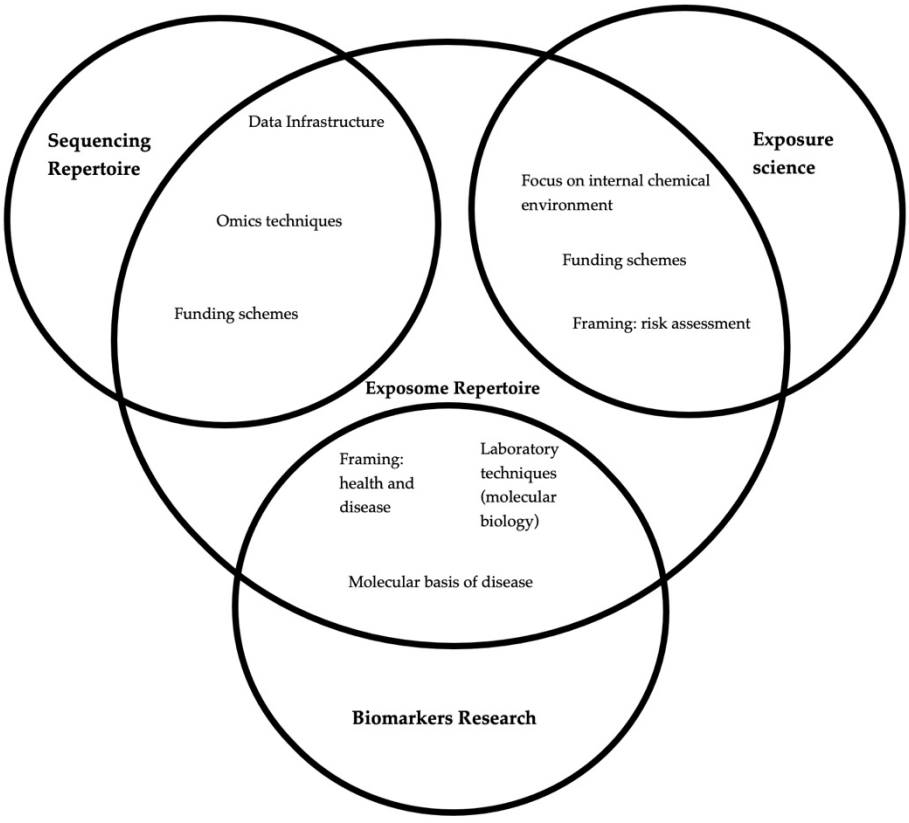

Figure 1. The sequencing repertoire, exposure science and biomarkers research merged in the exposome repertoire.

When Wild first introduced the exposome, he situated the concept in the context of several debates in the life and health sciences (Wild 2005). With the introduction of the exposome, Wild hoped that the concept could help shift attention to the "need for methodologic developments in exposure assessment" and for "methods with the same precision for an individual's environmental exposure as we have for the individual's genome" (Wild 2005, p. 1848). Wild praised the development made in genomics, but also noted that the low penetrance of genetic variants – as opposed to their high prevalence – implies that their contribution to disease burden is crucially linked to the presence of some environmental exposure and therefore argued for a broader consideration of and a focus on the environmental side. At the same time, Wild also noted that many disease-exposure interactions were ill-defined and interactions with the genome had not yet been discussed in depth. In Wild's intentions, the exposome was to bring propositions together for epidemiology in a similar way to what the genome had done for genetic research. Namely, 2005 was just a few years after the end of the Human Genome Project. The Human Genome Project, which officially began in 1990, with the goal of sequencing the complete human genome, is a fascinating case of contemporary science, where economic, political, epistemic, societal elements are intertwined and which was a major breakthrough for the life sciences, with significant impact on funding, conceptualisation of genes and genetics, science management, concept of disease, technology, data infrastructures, etc. (Guttinger and Dupré 2016; Gannett 2016; Hilgartner 2017). I argue that the exposome draws significantly on the *sequencing repertoire* that emerged in the context of genomic projects and has since then increasingly spread in the life and health sciences.

From a technical and material perspective, the sequencing repertoire is based on sequencing and mapping techniques, which are used to individuate the sequences that make up the genetic material and thus understand the molecular composition, functions and inter-relations of genomes (Hilgartner 2017, Chap. 1). In the context of the Human Genome Project, these techniques were developed with the aim of increasing the resolution of genetic maps and their speed of production. The use of sequencing techniques and their high-throughput

character increased the importance and role of sequence databases (Leonelli 2010b). Together with large data infrastructures, the sequencing repertoire includes specific norms and values around data sharing (Maxson Jones et al. 2018). Beyond the Human Genome Project, the sequencing repertoire has been successful at securing large funding in connection with the need for sustained development of infrastructures and technologies, as well as the rhetorical framing of the benefits for biomedical research. Sequencing repertoire projects benefit from large, blue-skies, medium or long-term governmental funding, especially in the US and Europe. From a conceptual perspective, the sequencing repertoire is centred around the role of the genome as the "epistemic thing" connecting various approaches (Rheinberger and Müller-Wille 2017, pp. 5-6). As a result of the complex combination of these ingredients, the size and structure of many sequencing projects, the repertoire gathers interdisciplinary teams of researchers with a background in the life sciences as well as information science, statistics, physics, etc.

The exposome shares and relies on various elements of the sequencing repertoire. Conceptually, the exposome lies in a two-fold relation with the genome. On the one hand, the various epistemic breakthroughs of genome-sequencing are constantly mentioned in discussions on the exposome and the wording itself indicates a close relation to the genome.[19] On the other hand, the idea of the exposome as the necessary complement of the genome puts it into a critical position to the genome and points to the need for new and different approaches, beyond genomics. This is evident in Wild's considerations about the low penetrance of genetic variants and their contribution to disease risks, which calls for a more significant consideration of the role of the environment (Wild 2005). In an article that played a crucial role for the exposome, Stephen

19 The term can be seen in conjunction with the use of -omes and -omics as suffixes in the life sciences, which has significantly increased in the last decade (Guttinger and Dupré 2016).

Rappaport and Martyn T. Smith connected the idea of an exposome as a new "paradigm" for research on disease risks to the need to put emphasis on the role of environmental exposure, in contrast to gene-centrism (Rappaport and Smith 2010). Rappaport and Smith noted how, while 70-80% of disease risks are due to changes in the environment, instead of genetic variations, the latter have nonetheless received more attention than the former. Whilst also retaining a critical perspective on genomic approaches, the hope is that the exposome can do for environmental epidemiology what the genome did for genomics, i.e. to collect and organise under an umbrella concept various ideas and approaches to the study of the relation between environmental exposure and disease.[20]

The close relation between the exposome and the sequencing repertoire is also shown by the transfer and use of other ingredients of this repertoire. From a material and technical point of view, the exposome draws on the advancement of omics techniques that can be traced back to sequencing and mapping technologies. Techniques that are referred to as omics expand on traditional the sequencing toolkit by applying a high-throughput approach to the analysis, quantification and characterisation of biological molecules in the cell and its environment. The exposome repertoire is to a large extent about bringing the developments in omics techniques from genomics to epidemiological research, as a way of trying to get a similar level of precision in measuring the presence, variation and impact of exposures (Rappaport 2011). Through omics, the exposome repertoire also draws on many infrastructures for the handling and or-

20 The article by Rapport and Smith (2010) was published in the "Insights" section of *Science* and arguably served as a re-introduction of the exposome to a larger audience (see also Siroux and colleagues on the increase in number of citations of exposome in PubMed after 2010, Siroux et al. 2016).

ganisation of data from the sequencing repertoire. For instance, exposome projects such as EXPOsOMICS use bioinformatics providers often used in the genomic setting, like Genedata (https://www.genedata.com).

Another feature shared between the sequencing repertoire and the exposome is the increasingly interdisciplinary nature of research, in particular in relation to the use of large and diverse datasets. In this interdisciplinary context, a second area of research that I argue played a significant role in the development and establishment of the exposome repertoire is *exposure science*. Exposure science is an interdisciplinary area of research that "addresses the intensity and duration of contact of humans or other organisms with those agents (defined as chemical, physical, or biologic stressors) and their fate in living systems" (US National Research Council 2012, p. 3). While its roots are connected to early work in industrial and occupational hygiene, exposure science can be considered a relatively new field of research, whose definition and specification took place in the US context around the time of Wild's initial proposal of the exposome (Lioy 2010; van Tongeren and Cherrie 2012). I have mentioned earlier the role played by an article published in *Science* in spreading the exposome to a larger audience (Rappaport and Smith 2010). Both authors come from an exposure science background. Rappaport is currently one of the major proponents of the exposome in the US and directs the Berkeley Center for Exposure Biology.

Until the 1920s exposure scientists and environmental epidemiologists collaborated on the study of workplace exposures in the context of occupational disease. This changed with the establishment of two different agencies in the US in the 1970s (on exposures in the workplace and in the ambient environment), which created two distinct paths of research in the field (Lioy and Rappaport 2011). At the same time, the focus shifted from human health towards risk assessment and compliance with standards, from measurement of personal and individual exposure to predictions of exposure levels and presence of chemicals and toxicants from deterministic models (Rappaport 2001). Exposure science is now centred on the focus of the "fundamental issues of whether and how human contact with toxicants occurs after release into the environment or

workplace" (Lioy 2010, p. 1081). Around the same period, environmental epidemiologists started to focus on the links between genetic and environmental factors and the genetic determinants of disease.

Exposure science was very important for the exposome repertoire from a methodological and material point of view, especially for the focus on internal chemical environment as one of the most "relevant" foci of analysis and the use of chemical tools such as mass spectrometry (Rappaport 2001). The exposome repertoire also transferred from exposure science some funding schemes, especially in the US, where exposome projects are funded by the same governmental and national-level agencies that fund exposure science projects. Additionally, the inclusion of some of the rhetorical tools of exposure science also allows to pitch the exposome in terms of occupational and public health issues (Juarez et al. 2014). In turn, the exposome has also been important for exposure science, where it is considered a major shift from the focus on genetic to environmental factors (Lioy and Rappaport 2011).

The benefits of the exposome approach are increasingly discussed for health and biomedical issues, for instance in the context of nutrition science and personalised medicine (Prescott and Logan 2017; Verma et al. 2018), air pollution research (Vineis 2018), the nature-nurture debate (Miller and Jones 2014) and pregnancy research (Robinson and Vrijheid 2015). Wild himself has introduced the concept to different lines of biomedical research, including cancer research (Wild 2009, 2011, 2012).[21] In this context, I argue that *biomarkers research* played a crucial role in the shaping of the exposome repertoire.

Biomarkers are used as indicators of normal and pathogenic processes and responses to stimuli within an organism (Strimbu and Tavel 2010; see also Russo

21 Since 2009, Wild has been director of the IARC, the International Agency for Research on Cancer, an intergovernmental agency that works under the World Health Organization of the United Nations.

2017, pp. 153-154). In a broad sense, any medical sign is a biomarker, including measurements such as pulse, blood pressure and more or less sophisticated blood tests; thus, biomarkers can be any indication of any medical state. The use of biomarkers has had a steady increase since the 1980s, especially in the context of clinical trials, as surrogate outcomes of diseases including cancer and heart disease, basic and clinical drug research (Strimbu and Travel 2010, p. 466). In this sense, the development of molecular techniques and their application and use in the context of biomedical research has played a crucial role, specifying the notion of biomarkers in a research context and pushing the level of abstraction to a molecular level (Boniolo and Nathan 2017). Unsurprisingly, the use of molecular biomarkers for epidemiological purposes has been established in the context of molecular epidemiology, with the development and incorporation of laboratory techniques to study the molecular basis of disease (Bonassi and Au 2002).[22]

Biomarkers research has provided many ingredients of the exposome repertoire. From a methodological perspective, biomarkers are explicitly mentioned as crucial elements in introductions of the exposome as a new paradigm for epidemiology. Together with omics, the use and development of biomarkers research is supposed to give more precision to the techniques of the exposome repertoire. The connection between biomarkers and the exposome is often discussed as a way of improving precisions in the context of environmental exposure, in order to try and adjust the differences in precision between genetic and environmental measurements (Wild 2005, p. 1848; Wild 2009). The extent to which biomarkers are influential on the methodological ingredients of the

22 Wild himself was chair of the Molecular Epidemiology Unit of the University of Leeds when he published the first article on the exposome. He has a background in pharmacology, oncology and environmental and molecular epidemiology, where he has mostly done research on the interplay between environmental and genetic risk factors in cancer causation.

exposome repertoire is such that, practically, in exposome projects most of the work consists in the search for biomarkers of external and internal exposure and biomarkers of disease and in the study and analysis of their associations. This has several interesting consequences for the exposome repertoire. First, the study of biomarkers is connected to the use of molecular tools for the study of both the external and internal component of the exposome, with the aforementioned aim of bridging the gap between the accuracy and precision of measurements of external and internal exposures (Turner et al. 2017). In addition, the molecular components of biomarkers research mark another difference with more traditionally genomic and sequencing methods, such as GWAS, as these are considered to enable more direct analyses of what lies in between associations and shed light on molecular pathways and mechanisms (Russo and Vineis 2016). Thus, the connections between biomarkers research and the exposome repertoire extend beyond methodological aspects, and, at the conceptual level, inform etiological approaches used in the exposome repertoire and push for molecular conceptualisations of health and disease (Chadeau-Hyam et al. 2011).

2.4 Tracing Scientific Innovation in Repertoires: The Exposome as a Postgenomic Repertoire

I have used the repertoires framework for my account of the material, social and epistemic components of the exposome and the connection of these components to the conditions under which the exposome was conceived, developed and established. What I haven't discussed in depth yet is scientific innovation and change in the context of the exposome. Understanding scientific change is indeed one of the primary uses of the repertoires framework.

Connecting discussions on change to the context of contemporary epidemiology and health and life science research, I argue that the exposome is a *postgenomic repertoire*. The term postgenomic has been used to describe a 'new' era in the life and health sciences after the completion of the Human Genome Project. This has led to an interdisciplinary debate where philosophers, historians, sociologists and scientists have discussed the extent to which postgenomic research can be considered innovative (Richardson and Stevens 2015). Science scholars in this debate have mostly cautioned against claims depicting postgenomics as a new paradigm and a revolution in the sciences, pointing to the continuity between the genomic and postgenomic era and criticising enthusiastic tales of scientific innovation (Gibbon et al. 2018). As a result, 'postgenomic' should be used with caution, as it is a loaded term that could be interpreted to describe define something as highly innovative and 'revolutionary' (Richardson 2011). By claiming that the exposome is a postgenomic repertoire, I use a historical meaning of the term, whereby postgenomic refers to research that: employs genomic and genomic-based technologies; is increasingly aware of the complexity in interpreting genomic results and data; and has a critical engagement with gene-centric approaches (see e.g. Leonelli 2018). As we have seen, critiques of the extensive use of genomic data and gene-centrism to explain human disease are constantly present in introductions and presentations of the exposome. The exposome repertoire extensively employs omic techniques, which are based on technological development of the genomics era. Rhetorically, the exposome is often pitched as a 'post-genome notion', in the sense that it pushes for research that should provide solutions that the genome did not deliver. Yet, more specifically, I argue that postgenomic features of the exposome repertoire can be identified at the conceptual and methodological levels, because, similarly to what has been argued by science scholars in the postgenomic context, these are innovative but have strong continuity with longstanding lines of research and approaches.

Consider, for example, the role played by the concept of environment in the exposome. The operationalisation of the environment and its role in human health and disease is often considered among the triggering factors for moving

into a postgenomic era. One of the 'paradoxical' results of the Human Genome Project was the discovery that the environment played a role, in determining human health and disease, that was more significant than what gene-centric approaches had assumed. Yet, as highlighted by Shostak and Moinester, in postgenomic settings the environment is conceptualised in several different ways, as it might refer to the cell, hormonal profiles, ambient environments, the body, social networks, etc. (Shostak and Moinester 2015). This is leading to blurring distinctions between internal and external environments or between social and biological environments (Landecker and Panofsky 2013; Lloyd and Raikhel 2018) and calls for reconceptualisations of the genome in environmental terms (Keller 2015).

As we have seen, the push for more consideration of the role of environment in disease is indeed one of the reasons for the introduction of the exposome. In addition, in line with the broader postgenomic context, exposome researchers understand the environment in a variety of different ways. For example, Wild talks of the environment as anything that is "non-genetic" (Wild 2009); Rappaport identifies different ways of interpreting the role of the environment and therefore discusses the concept of "relevant environment" (Rappaport 2011); Robinson and Vrijheid discuss the body as "an environment" (Robinson and Vrijheid 2015). Are these considerations of the environment and plurality of conceptualisations postgenomic innovations of the exposome? I would argue that, in the epidemiological context, the pluralistic approach of the exposome is not a radical innovation, but rather a continuation of the various ways in which the environment has been conceptualised in epidemiology. Namely, in epidemiology, environmental features are traditionally measured indirectly through the collection of exposure data, as 'environmental exposures'. This makes it fairly easy to operationalise much, if not anything, as environmental, from air and water pollution and the socio-economic status, to occupational settings, dietary conditions, and all the way to the internal chemical features of the body. The exposome repertoire merges the pluralism of postgenomic approaches with the tradition of epidemiology, thus developing a framing of the

environment that, rhetorically, speaks to both epidemiological and post-genomic audiences.[23]

This leads me to an account of the conceptual innovations of the exposome that emphasises its pluralistic approach to the environment and continuity with longstanding approaches. I thus come to conclusions on the conceptual innovation of postgenomics that are similar to the arguments of many science scholars. My analysis is an expansion of these arguments, as it focuses on the influence of postgenomics beyond fields like biology and genomics, that have usually been the focus in this context (Green 2016). The use of the repertoires framework can also be considered an expansion of the debate, as it emphasises that non-conceptual components are crucial both as enablers of scientific change and as ways in which scientific innovation and change can manifest themselves. One of the non-conceptual elements that is often discussed in postgenomics is the availability of new and large datasets, or 'big data' (Ankeny and Leonelli 2015, pp. 128-129; Gibbon et al. 2018).

In the exposome, the use of omics techniques and approaches from exposure science expands the types and size of exposure data that are collected and used, which range from individual exposure estimates to data about internal presence and responses to exposure. Traditionally, in epidemiology exposure data has been collected by sampling of environmental conditions or, indirectly, through interviews, surveys, questionnaires asking study participants about features of their surrounding environment. The shift to high-throughput omics is innovative because it expands exposure with the inclusion of new dimensions, most importantly the internal dimension: to be exposed does not only mean to be exposed to external elements, but also to the internal chemical environment. In abstract terms, we can think of exposure as the property of being

23 This framing also shows in the integration of various sources of data on the environment in terms of exposure data (Chapter 3; Leonelli and Tempini 2018).

exposed to something. In the context of the health sciences and epidemiology, exposure has more specific connotations: it refers to a characterisation of proximity and/or contact with something that might transmit disease or other outcomes of interest, to be measured through an amount of a factor that an individual or a population are exposed to (Lee and Pickard 2013). The availability of technologies providing data on the internal chemical environment leads to the expansion of this property, which in the exposome has both external and internal components, classified as generic external exposure, specific external exposure and internal exposure. It is the shift from the traditional externalist focus to the inclusion of the internal level of the exposome which is considered an innovation by epidemiologists and leads them to discuss the exposome as the totality of exposures, a global approach to exposure and a holistic way of studying the individual (Vrijheid 2014; Robinson and Vrijheid 2015; Stingone et al. 2017).

Shifts at the material and technological level are thus significant scientific innovations of the exposome, as well as postgenomics more generally.[24] At the same time, arguing that the exposome is a postgenomic *repertoire* points to the alignment of material, technological and conceptual components. This is particularly significant when discussing the epistemic implications of large datasets. Science scholars have criticised many claims about revolutionary changes due to the use of large, new types of datasets in the sciences (Leonelli 2014; Leonelli 2016a). Similarly, as part of this debate, my use of the repertoires framework points to the alignment of data with other ingredients of the repertoire: the availability of new types of data can lead to the revision and expansion of conceptual aspects of the exposome repertoire, but scientific innovation is the result of the alignment of various components. Data can thus hardly be

24 See also Gross and colleagues' analysis of the emergence, in contemporary biology, of new approaches and fields in relation to the development of high-throughput data technologies (Gross et al. 2019).

a difference maker on its own, as its role of data is connected to other elements of the repertoires, such as conceptual commitments, infrastructures, funding and skills.[25]

2.5 Conclusions

This chapter set out to analyse the exposome, as a new concept that has recently been introduced in epidemiology, and to understand its innovative character. I have used the conceptual framework of repertoires to distinguish the material, social and conceptual components that constitute the exposome as a specific line of research, approach and community of researchers. This distinction has enabled me to connect some of these components to other traditions and disciplinary contexts of research in the life and health sciences. In this way, I have developed an account of the innovation of the exposome that is closely connected to the conditions in which it was developed, introduced and established, and thus problematises views of the exposome as a new paradigm for research in epidemiology. As a result of my analysis, I hope to have expanded and contributed to the repertoires framework, by showing a case in which a repertoire is the complex assemblage of various other repertoires and traditions, in addition to cases of transfers of one repertoire in another context. In

25 My point here is close to critiques of technological determinism, i.e. the view that the emergence of new technologies can be the sole determinant of social changes (broadly construed). I would argue that technological determinism is often an implicit component of claims about the revolutionary changes that data-intensive methods can have on scientific epistemology (see e.g. Hey et al. 2009). Technological determinism has been extensively discussed in the science and technology studies literature (see e.g. Sismondo 2010, pp. 96-105).

addition, the exposome helps to illustrate another key characteristic of repertoires: no one component of the repertoire is central, primary or fundamental. I would emphasise that this is one of the most interesting and possibly innovative aspects of the framework, which can be considered a synthesis of various traditional views on the primacy of theoretical components in philosophy of science and social components in science studies. Keeping all these components together in analyses of scientific research yields situated and local accounts of scientific practice (Leonelli 2016a, p. 190), which aim at looking at the very local details of research as part of the critical assessment of what constitutes change, innovation and success in contemporary science.

Chapter 3

Making Evidential Claims in Epidemiology: Three Strategies for the Study of the Exposome

Abstract: In this chapter, I present an account of contemporary epidemiology based on the notion of evidential claims and show how this helps specify and differentiate data practices, methods, approaches and results. Analysing current research on the exposome, I identify three different strategies to generate evidential claims. The macro strategy, which identifies the initial evidence platform and generates scoping claims that restrict samples and provide an initial understanding of phenomena. The micro strategy, which collects data of significantly different types to generate evidential claims on structures at the microscopic level. The association strategy, which uses evidence from the two other strategies to generate evidential claims at the statistical level of associations between exposure to the environment and outcomes of interest. Differentiating between these strategies enables to identify and unpack the epistemic commitments involved at various steps of research; gives a characterisation of the context of data collection, interpretation and use; and provides a different philosophical perspective on epidemiology and epidemiological evidence.

3.1 Introduction

The health sciences have a long tradition of studying the ways in which the environment has an influence on human health. Epidemiology is the area of clinical medicine that comprises the various approaches to the study of these

phenomena by focusing on the relation between outcomes of interest and the exposure to the environment, broadly construed to include diet, external chemicals, lifestyle, etc.[26] More precisely, epidemiology can be defined as the study of the distribution and variation of exposure and disease in populations (Russo and Vineis 2017, p. 252).

Relatively little philosophical attention has been dedicated to epidemiology and most of the focus has been in the context of discussions on causality and causal inference. Some philosophers have studied ways of thinking about the aetiology of disease in epidemiology by analysing the historical development of causal models of disease (Broadbent 2009, 2013, 2014; Fuller 2018). Others have looked at the use of causal and non-causal terms by epidemiologists (DeVreese 2009; Russo 2009), which has led to discussions on causal explanations in epidemiology (Campaner 2011) and the interpretation of epidemiological results in causal terms (Broadbent 2013, pp. 26–55). Epidemiology has also been used as a case study for philosophical accounts of causal inference, such as the Russo-Williamson Thesis (see Chapter 4).

In this chapter, building on recent philosophical scholarship on data practices in the life sciences (Leonelli 2016a) and on evidential claims and evidential reasoning in the historical sciences (Chapman and Wylie 2016), I look more closely at data practices employed by epidemiologists and I present a typology of their methods, approaches and results based on the notion of evidential claims. I start by explaining my use of the notion of evidential claim in Section 3.2. Then, in Section 3.3, I analyse data practices in contemporary epidemiological research from the standpoint of evidential claims. This allows me to show the multiple layers that characterise research, where significantly different approaches, lines of work and types of claims can be distinguished. I argue that these differences can be seen in terms of distinct *strategies* for evidential claims.

---

26 Here I am following Thompson and Upshur's distinction between clinical medicine and bench medicine (Thompson and Upshur 2018, pp. 2-4; p. 24).

I identify three strategies. The *macro strategy*, which generates scoping claims that restrict the sample and provides an initial understanding of the phenomena under study. The *micro strategy*, which is applied to collect different types of data and generate evidential claims on microscopic structures. The *association strategy*, that uses evidence from the two former strategies to generate evidential claims at the statistical level of associations between exposure to the environment and outcomes of interest. As I hope to show, this typology yields a new perspective on the epistemology of epidemiological research and current philosophical discussions on the relation between data, evidence and knowledge (Sect. 3.4).

As a case study in current environmental epidemiology, I focus on the line of research known as the exposome. I base my philosophical analysis on data and experiences I gathered through qualitative interviews, participatory observation and discussions with epidemiologists in EXPOsOMICS, one of the leading projects on the exposome. Focusing on this line of research is important for a number of reasons. First, it enables to look at an area of epidemiological inquiry that is young and in a particularly interesting state, as central notions are still under discussion (Stingone et al. 2017), technology is new and sometimes used for the first time in this context (Turner et al. 2017) and there is talk of the exposome as a new "paradigm" (Vrijheid 2014). The exposome is defined as the totality of exposures individuals face during their lifetime, which includes external and internal aspects of exposure.[27] The notion can be considered a synthesis of the traditional focus on the external element of exposure and the innovative use of molecular, 'omic' technologies for the study of internal aspects. This also puts the exposome approach at centre of medical innovations discussed by philosophers of science, such as the molecularisation of medicine,

27 As we have seen in Chapter 2, the notion was first conceived and introduced by Wild (2005), and then further developed by Rappaport and Smith (2010) and Wild (2012).

data-intensive methods and genomics and post-genomics (Boniolo and Nathan 2017; Gibbon et al. 2018). In addition, this is a line of research which is significantly funded, especially in Europe and the United States, in connection to the potential benefit on public health issues (Juarez et al. 2014).[28] Researchers look for associations between environmental pollutants and disease with the aim of identifying intermediate elements or features that are associated both retrospectively and prospectively, that is both with the exposure and with the disease. These are referred to as intermediate biomarkers, which are considered crucial to the understanding of the pathways through which the environment influences disease, thus potentially allowing for the identification of disease-related reactions at an early stage. Assessing exposures helps determining levels of risks associated with the presence of specific elements in the environment and, consequently, results are used to inform policy-making regarding the accepted standards. Focusing on exposome research therefore makes it possible to study an area of epidemiology that works at the interface of environment and health issues, with integrated approaches and interdisciplinary teams.[29]

3.2 Focusing on Evidential Claims

Before going into the details of exposome research, let me briefly specify the central notions that I use in the chapter. In my analysis, I rely on the notion of

---

28 At the end of 2012, EXPOsOMICS was funded by the European Commission, as part of the seventh framework for the European Union's Research and Innovation programme.

29 EXPOsOMICS was structured as a consortium of 13 research centres in Europe and the US, coordinated by the Department of Epidemiology and Biostatistics of Imperial College London and working in collaboration with other research teams in Europe and the US.

evidential claims to interpret data practices in exposome research. I view evidential claims as claims that identify the datasets that are to be used as evidence for the investigation of phenomena. I highlight that different types of evidential claims are generated at different stages of exposome research and through different approaches, methods and lines of work. I claim that various aspects of exposome research can be interpreted as relying on or building towards evidential claims, beyond explicit claims made by researchers.

For the notion of evidential claims, I draw inspiration from philosophical inquiries on archaeology. In the work of Alison Wylie, evidential claims are discussed as the main results of archaeological results. In Wylie's account, evidential claims are relations relying on "mediating assumptions that establish a link between the material traces that survive archaeologically and the past events or conditions of life that are presumed responsible (in part) for producing these traces" (Wylie 2000, pp. 231-232). The relation is established through a chain of inferences, warrants, postulates and interpretations that provide the necessary "scaffolding" to determine how and if material traces can be used as evidence (Wylie 2017).[30] A simple example of the evidential claims discussed in this context may be of the kind "archaeological observation of ceramics is evidence that

30 The notion of scaffolding has been developed in the context of discussions on cultural evolution by William Wimsatt and James Griesemer (2007). Scaffolding is the result of "structure-like dynamical interactions with performing individuals that are means through which other structures or competencies are constructed or acquired by individuals or organizations" (Wimsatt 2013 p. 568). In recent years, the notion has been used by many philosophers of science to refer to elements of scientific practice that enable and canalise research.

the site under investigation was occupied for 10-15 years".[31] Differently from Wylie, I take the "factual ground" of the evidential claims I discuss in the chapter to be datasets and not necessarily only material traces. In other words, the evidential claims I analyse here are of the kind "data collected through this procedure is evidence for the relation between air pollution and cardiovascular disease". In this way, the epistemic role of an evidential claim consists in the identification of a dataset ("data collected through this procedure") and the specification of its evidential value ("is evidence for") in the context of the investigation of a phenomenon ("relation between air pollution and cardiovascular disease"). This is important to stress, because Chapman and Wylie sometimes use the word 'data' to describe the factual ground of evidential claims, but data is sometimes also used as a synonym to 'evidence', which – as also suggested by Leonelli (2017) – may lead to some confusion. For the notion of data, I follow Sabina Leonelli's account, according to which data is any material artefact that can be used as evidence for claims on phenomena and can be circulated among individuals (Leonelli 2016a, pp. 69-92). When I discuss data as the factual ground of evidential claims, the term is thus to be understood as encompassing data generated digitally, material traces and physical samples. The use of data as a unit of philosophical analysis has been developed in recent years in the context of a multidisciplinary and critical debate, in part originated as a response to the rhetoric surrounding 'big data', to which philosophy of

31 This is an adapted and simplified version of one of the evidential claims that Chapman and Wylie analyse (Chapman and Wylie 2016, p. 153). Here, I am mostly interested in highlighting the general features that constitute evidential claims in their account, which is why I am not delving into the specifics of their reconstruction.

science is increasingly contributing.[32] Epidemiology is particularly interesting from a data perspective because it has traditionally been concerned with the collection, storing and analysis of (big) data, but the availability of new sources of data and new analytic tools are often presented as a significant novelty, especially in the context of issues at the interface of environment and health (Leonelli and Tempini 2018).

3.3 Three Strategies for Evidential Claims

I now turn to present a typology, based on the notion of evidential claims, of the ways in which epidemiologists acquire knowledge on the exposome. I interpret data practices in exposome research as building various evidential claims that identify datasets as evidence for the study of phenomena or the development of subsequent, higher-level claims. The evidential claims I identify are not necessarily explicit, but still play a crucial role in the study of the exposome. The framework of evidential claims allows me to identify the use of three strategies for evidential claims (the macro, the micro and the association strategies), which differ in terms of distinct approaches to the phenomena under study, distinct kinds of work that researchers carry out and distinct types of evidential claims (see Table 2). My empirical reconstruction is based on discussions with researchers in EXPOsOMICS on what they called the "data workflow" of the project, which is used as an organising principle for the different sources and journeys of data and related activities. I expand on this primary classification of the data practices of EXPOsOMICS, with the aim of showing that various aspects of the project can be interpreted as aimed at the

32 See Chapter 1 and, for instance, the special issues of *Big Data & Society* edited by Iliadis and Russo (2016), *Studies in History and Philosophy of Biological and Biomedical Sciences* edited by Leonelli (2012) and *Science, Technology, & Human Values* edited by Leonelli, Rappert and Davies (Leonelli et al. 2017).

generation of evidential claims. I view these strategies as ways of mobilising datasets on the basis of which evidential claims are generated. Thus, the strategies are about data collection as much as about retrieving data from other repositories or building diverse and integrated datasets.[33]

	Approach	Lines of work	Type of evidential claim
Macro strategy	Population level	- Selection of cohort studies - Retrieval from cohort studies	Evidential claims about scopes of investigation
Micro strategy	Microscopic level	- Omics - GIS - Experimental studies	Evidential claims on microscopic structures
Association strategy	Association level	- Regression models	Evidential claims about statistical associations

Table 2. The macro, micro and association strategies along three dimensions

In the next subsections, I delve into the details of each of the three strategies and I illustrate the strategies using examples from research, in EXPOsOMICS, on the relation between health and exposure to environmental pollutants. In order to show how these strategies are interrelated in the context of research

33 The integration of a variety of sources of data in EXPOsOMICS was in itself one of the goals of the project, and is usually connected to its innovative character (Vineis et al. 2017a, p. 2; Fleming et al. 2017, p. 12).

published as a result of EXPOsOMICS, I refer to the work of Fiorito and colleagues on the relation between air pollution and cardio- and cerebrovascular disease as an example (Fiorito et al. 2018). I call 'macro strategy' the strategy that is implemented at the starting point of research, as part of the selection and retrieval of samples from cohort studies, and individuates the dataset that serves as the primary evidential platform for individual studies. This is where we can see some of the traditional tools of epidemiology at work, most prominently cohort studies. What I call 'micro strategy' is on the other hand the strategy that comprises the novel approaches employed in exposome research, such as omic techniques and geographical information systems. This strategy works on data specified by the macro strategy and is used to perform a microscopic analysis of the structures of data, which generates high resolution data on the internal component of exposure and the surrounding environment. I call 'association strategy' the strategy that is at work to statistically analyse data generated through the micro strategy, in order to identify associations between elements and features of the environment and health outcomes of interest.

3.3.1 The Macro Strategy: Primary Data Retrieval from Cohort Studies

The first strategy for evidential claims works at the macroscopic level of investigation. It is based on data collected in observational, cohort studies, that follow a population of interest for an extended period of time. Cohort studies track variables including clinical states, outcomes of interest, features of participants' lifestyle and their surrounding environment (city's district, house, work, family, etc.). Data collection is based both on the extraction of physical samples (e.g. blood, cord blood, maternal milk, hair, etc.) and the use of questionnaires and one-to-one interviews. Most of the data of cohort studies is collected in hospitals, by physicians, nurses or trained interviewers. The goal is to bio-monitor or follow participants throughout their life, so as to track the potential development of outcomes of interests. This is usually achieved through

what is known as follow-up, whereby participants are tracked at different points of their life, as they may for instance be asked to answer a new questionnaire years after the initial recruitment or data may be retrieved through record linkage with hospital, mortality and local health authority registries.

Several cohort studies are currently in operation and widely used as sources of evidence in epidemiology. Generally, they tend to differ in terms of: areas covered, as some are national whilst others are transnational and/or continental; focus of research, as some are interested in tracking a variety of disease states and influential factors, whilst others are more specifically focused on only a handful of phenomena; and time, as some are set up to run constantly and without a predetermined end point, while others come to an end at a specific time but still provide data that can be further analysed and re-used. For instance, one of the larger studies used in EXPOsOMICS was EPIC (the European Prospective Investigation into Cancer and Nutrition), which followed 521,000 participants across 10 European countries for almost 15 year starting in the 1990s, with a broad focus looking at the relationships between diet, nutritional status, lifestyle and environmental factors, and the incidence of cancer and other chronic diseases. An example of a smaller cohort is INMA (INfancia y Medio Ambiente), which is a currently operating study that focuses on the national level and looks at the exposome during pregnancy, as to investigate the relations between environmental exposures and child development.

A significant portion of epidemiological inquiry is primarily concerned with the establishment and management of cohort studies. However, these data collection activities usually precede projects like EXPOsOMICS, where researchers selected and withdrew datasets from specific cohorts, as opposed to setting up their own cohorts. They will want to focus on a specific outcome of interest (such as cardio-vascular disease, child development, increase in oxidative stress) and, consequently, they will ask for a subset of the data that is collected in the chosen cohort study. Researchers in charge will thus create a dataset meeting the requests and will upload it to the data infrastructure which is used in EXPOsOMICS.

The strategy that I want to describe here is thus not about setting up cohort studies, but rather about generating the initial evidential space. Through this strategy, researchers develop evidential claims that are essentially scoping claims at the macro level and concern trends in the populations under study and the surrounding environment. In other words, they have the epistemic goal of identifying and describing the initial dataset and do not provide a precise analysis of the links between trends in populations and trends in environmental pollutants. Rather, they are established to identify which kind of data is to be used for further research. The macro strategy thus employs a population-level approach to the phenomena under study, involves the work of selecting and receiving data and is used to make evidential claims that specify the scope and space in which research is to be developed and shaped. These results require the work of different kinds of individuals and groups, including: health professionals in hospitals, epidemiologists setting up and coordinating the cohort, and, in EXPOsOMICS, epidemiologists who selected and received the data and perform an initial analysis, statisticians, and project managers.[34]

The main results of this strategy are usually framed in exposome research as either "health data" or "covariates". The first category includes the physical samples as well as other pieces of information, collected via questionnaire or interviews, that concerns health states of participants. The latter category covers what is considered additional elements, surrounding participants' health: this can be both of qualitative and quantitative kind, as it includes information about participants' lifestyle, home environment, occupation, diet, as well as the

34 The category of 'epidemiologist' is in itself very interesting, as it is difficult to associate with unique career paths and backgrounds. The epidemiologists working in EXPOsOMICS I interacted with had slightly different backgrounds, including medical school, molecular biology and medicine.

postcodes of participants, smoking status, etc.[35] The macro strategy is thus aimed at establishing the dataset comprising health data or covariates as evidence for the study of a specific aspect of the relation between health states and the environment. For example, in a recent EXPOsOMICS study published by Fiorito and colleagues (Fiorito et al. 2018), the study was performed on data collected in the Italian sub-cohort of the EPIC study. The initial step consisted in the setup of a case-control study, that is the identification of incident cases of cardio- and cerebrovascular disease that arose during the follow-up period and the identification of control cases (i.e. free of the outcome of interest), with 18,982 individuals in total (Fiorito et al. 2018, p. 3). The macro strategy consisted in this identification, supported by background knowledge and other work in the literature on the association of cardio- and cerebrovascular disease with exposure to air pollution, which yielded evidential claims specifying the dataset as evidence for the study of the relation between air pollution and cardiovascular disease.

The evidential claims developed through the macro strategy are thus not statistical claims, nor causal claims, but are rather implicit claims about the initial evidential platform for research. This may be one of the reasons why this level of investigation is rarely discussed by philosophers of science. However, the macro strategy is particularly interesting from a philosophical standpoint because the epistemic commitments and assumptions made at this stage have a significant influence on subsequent research. In EXPOsOMICS, this strategy was indeed connected to challenges regarding which kind of data was best suited for the study, including considerations about available evidence, choice of which cohort studies to focus on and which data to select. Other conceptual

35 Notice that this distinction does not necessarily overlap with a distinction between health and environment data. As we will see in the following subsections, exposome researchers have a broader characterisation of what counts as environment data, which requires a different strategy.

challenges concern the internal comparability of the data collected in cohort studies, as often data from cohort studies is collected at quite different points in time and may regard participants with quite different backgrounds, home environment, etc. One of the ways in which the evidential claims developed through this strategy shape the kind of research that is carried out at later stages in the project is thus connected to generalisability. Namely, data collected in cohort studies is local by default, having been collected in person and/or physically extracted from participants, and this locality is also important for exposome research, because researchers aim to connect features of specific kinds of environment with specific health states of the participants. This dialogue between locality and generalisability is stressed multiple times in the project, as researchers aim at finding results that are general enough, but, at the same time, also sensible to local features.

3.3.2 The Micro Strategy: Omics, Gis and Experimental Studies

One of the ways in which the exposome approach is considered innovative is the use of molecular data and geographical information systems. These can be interpreted as the result of what I call the micro strategy, that is based on an approach focused at a lower level of investigation and is aimed at producing evidential claims at the individual level of the participants to the studies, their surrounding environment and internal molecular environment. The micro strategy is thus applied at different points of exposome research and operates to generate evidential claims at both ends of the spectrum, i.e. evidence on environmental pollutants on the one hand and, on the other hand, evidence on outcomes of interest (health and disease states). This strategy is particularly complex, because it requires significantly different lines of work and generates significantly different types of data.

A first line of research where this approach is applied is what EXPOsOMICS researchers called the 'OMICS labs'. Once the physical samples which form a

portion of the health data collected in cohort studies are acquired, they are analysed through omics technologies. These technologies are used to study the internal component of the exposome, that is to look for traces of external exposures and potential initial reactions to exposure at the molecular level. Omics measurements are performed using methods such as mass spectrometry, whose results are visualised in terms of plots with lines and peaks; these peaks – also known as features – are indicative of the chemical composition of the sample. The result obtained through omic analysis of the samples is, thus, a list of molecules per each sample. In exposome research, this is used to get a picture of the molecular structure of the samples and to thus try to understand the potential influence of pollutants on the internal, molecular component of the exposome: by looking at the internal level of exposure, researchers aim at spotting the presence of toxicants or reactions to toxicants which is due to exposure to pollutants in air or water. The resulting tables of data are usually called exposure or omics profiles; their type will depend on the omics technique used for the analysis, which in turn depends on the kind of molecular features researchers intend to focus on for their particular project, on the nature of the samples or the process they want to study, as well as on considerations about the cost and availability of the chosen omics technologies.[36] Some of these technologies are still at an experimental stage, to the point that a project like EXPOsOMICS was in some cases the first project using one of the techniques on a large scale. Choices and assumptions will be connected to the specific omic technique to be used, as researchers might have to assume the effectiveness

36 EXPOsOMICS researchers spoke of the data produced through omic analyses as one of the areas of research where 'big data' has an important influence. Here, the notion of big data is connected to the possibility of getting a very large amount of information for a small number of individuals. This is usually compared and constrasted to more traditional scenarios – such as, for example, cohort studies – where many individuals are studied through a smaller number of variables.

and validity of the results obtained and often they also face the challenges involved in the availability, cost and effectiveness of the technique.

The approach I want to highlight here is the focus on a lower level of investigation, as a way of getting to know about the biological structure of samples. This is in contrast to the population-level approach used in the macro strategy. It produces evidential claims that identify omics data as evidence for the study of phenomena connecting environmental exposures and health at the microscopic level, on top of the population level data identified as evidence through the first strategy.

The same strategy is used for evidential claims about environmental pollutants, at the other side of the spectrum of phenomena analysed in exposome research. In this case, on the basis of the covariates collected in cohort studies, and more precisely the postcodes of the areas where participants lived during the study period, exposome researchers generate a new source of evidence to describe the structure of the environment that individual participants experienced during the study.

This line of work within EXPOsOMICS was carried out by the 'GIS team', whose task is to develop Geographical Information Systems (GIS) to visualise, model and analyse geographical data. In the context of EXPOsOMICS, these systems were used to generate individual estimates of the chemicals and pollutants that could have been present in the environment and to which each of the participants to the cohort could have been exposed (see e.g. Gulliver et al. 2018). Per se, the postcodes provide an initial picture of where participants lived during the study, which is in itself already more specific than most other data collected in cohort studies. Cohorts can be very broad and track large areas of a country, if not more than one country, which may have very different kinds of environment and levels of pollutants. On the basis of this more specific picture provided by postcodes, the GIS team retrieved data from maps of characteristics of populations that are routinely collected by monitoring stations, which are usually located at various city locations in Europe and provide information on the conditions and features of the area. All these variables are

used to inform and tweak a geo-spatial model (developed in the form of a regression model) which, taking into account these variables and differences in populations, will assign an estimate of the exposures which every participant could have experienced during the study.

Why is the micro strategy used for environment data? Because, whilst many variables about environmental pollutants and exposure are indeed tracked as part of cohort studies – also at the individual level –, researchers get a better and more precise understanding of exposure through evidence from GIS data. Namely, whilst environment data is collected in cohort studies at the individual level, in questionnaires and one-to-one interviews, its level of resolution and precision is not comparable to that of GIS data. At the same time, other data available about the environment and presence of chemicals and pollutants is rather large, but not very precise and tied to the individual participant. In current epidemiology, geo-space modelling is therefore considered the best solution to track pollutants such as particulate air matter and have detailed information about exposure to it, as no blood biomarkers of toxicants including particular matter have been discovered so far. In addition, the epistemic goal of the collection of GIS data is to get a level of detail and resolution which is on par with omics data. The job of the GIS team was not only to give more detail to the environment and exposure data available, but also to level and balance the types of data on environment and exposure. The resulting data identified by this strategy will indeed be tables of data with estimates for different chemical compounds assigned to the members of the cohort, to an extent mirroring the exposure profiles produced by the omics labs. Yet, it is important to note that GIS data remains substantially different from omics data, in terms of the methods used to generate the data, the richness of the results and the level of interdisciplinarity required. Omics profiles can be used light on a variety of features of the structure of the molecules under study (see Chapter 4), while GIS data provides estimates of the concentrations of single pollutants. On the one hand, omics data is the result of analysis of physical samples; GIS data, on the other hand, is estimates. Additionally, omics tables are very wide (because of all the omics features analysed by these technologies), whilst exposure data

on environmental factors will be just one column wide. In EXPOsOMICS, omics data was usually collected by epidemiologists working at the core of the research team of the project, whilst the work on GIS data is responsibility of a specific team, whose members have significantly different skills and backgrounds from epidemiology. Still, I argue that the underlying strategy and the rationale for its use are the same: understand the microscopic structure of the phenomena under study. This complexity is indeed one of the features that sets this strategy apart from the other two elements of the typology that I am presenting in the chapter. Complexity also entails that groups of researchers with different backgrounds and epistemic skills are involved in the application of the strategy, as GIS is a sub-discipline of information systems, the use of omics required skills in molecular biology, and statistics is continuously part of the picture.

The same strategy is also applied in the context of the experimental studies conducted in exposome research. In EXPOsOMICS, one of the lines of research used personal monitoring and tracking devices in relatively controlled environments to perform real-time measurements of exposure levels and physiological and immunological variables for a brief period of time (Vineis et al. 2017a, pp. 148-149). These studies included tracking participants while they were, for example, walking in areas with clearly contrasting pollution levels, such as Oxford Street and Hyde Park in London, or in a swimming pool, whilst exposed to a variety of chemicals (see Espín-Pérez et al. 2018). In these studies, the same participants experienced several different conditions, which allowed EXPOsOMICS researchers to look at high effects of exposure, as opposed to long-term effects that are tracked through cohort studies.

The micro strategy thus enables exposome researchers to develop evidential claims about microscopic structures at the individual level of participants and their surrounding environment. The strategy is designed to produce claims about the presence or absence of pollutants, to be integrated with the evidential claims at the macro level and for tuning the regression model that I introduce in the next subsection. In the case of Fiorito and colleagues example (Fiorito et al. 2018), the internal and the external components of the exposome – external

exposure to air pollution and internal presence of pollutants or reactions – were measured using omic analyses and GIS. For omics data, proteomics and whole DNA methylation were performed to identify inflammation-related proteins to be used as biomarkers for inflammation and other reactions to external exposure. For environment data, the GIS team generated estimates of air pollution concentrations at the postcode areas of the participants to the study, more specifically for the exposures to NO_2, NO_x and $PM_{2.5}$. This allowed researchers to study the features of the exposome and environment at the individual level of the participants to the study, elaborating evidential claims of the kind "these exposure profiles are evidence for the pathways of air toxicants and cardiovascular disease". In other words, these evidential claims identify a dataset ("exposure profiles") as evidence of the existence or the study of phenomena at the microscopic level of investigation ("pathways of air toxicants and cardiovascular disease"). The example shows one of features of the micro strategy that may be connected with causality. One way of interpreting the evidential claims of this strategy is namely in causal terms, in the sense that, to an extent, these evidential claims can be used to later identify causal relations in the complicated web of relations between presence of pollutants in the environment, exposure, reactions to exposure and (potentially) development of disease. In exposome research, the theoretical background of the use of molecular data is usually discussed in terms of a methodological approach known as the ''meet in the middle approach'' (Chadeau-Hyam et al. 2011), whose goal is to investigate what lies 'in the middle' of the associations between exposure and outcomes of interest and therefore complement and test the causal nature of statistical associations. A way of interpreting the strategy is thus in terms of the

study of causal relations, together with the more general goal of getting to know about the structural elements of the samples to be used as evidence.[37]

3.3.3 The Association Strategy: Regression Analysis and Data Integration

The epistemic goals of epidemiological research on the exposome are connected to public health and, more specifically, exposure assessment. This is why EXPOsOMICS researchers usually discussed 'results' for their work only for the evidential claims produced by what I call the association strategy. This strategy is based on an approach focused at the level of associations between the environment and outcomes of interest and can be seen as an integration of the claims that are made as part of the macro and micro strategies. The data specified by the evidential claims generated through the other strategies is used in a regression model, looking for associations between environmental features and toxicants on the one hand and health outcomes on the other. In EXPOsOMICS the regression model typically looked at one exposure and one omics feature at a time. This means that, for example, each of the 4000 features that can be measured in metabolomics was modelled in association with environment data. Then, the data analysts of EXPOsOMICS looked at those models where the association was statistically significant, after taking into account that, whilst doing thousands of tests, some would look statistically significant by chance. From that, EXPOsOMICS researchers could for instance find that about 170 out of the 4000 features were associated with a specific outcome or set of phenomena.

---

37 Vineis (2015) says that the meet in the middle approach is a "very rudimentary approach to causality", but links it to the work of Wesley Salmon on the propagations of marks and causal processes (Vineis 2015, p. 720).

These models are typically developed through a mix of univariate and multivariate methods.[38] They are based on the assumption that the prediction of the outcome of interest is a matter of multiple factors: for example, when using omics data, models will be based on the assumption that a pool of different omic features will predict the outcome of interest better than each of them separately or their sum. These models are thus not primarily causal: their goal is prediction, and more specifically exposome researchers are after the model that best predicts the outcomes of interest, taking into consideration the omics variables jointly, as opposed to using one by one or summing the predictions separately. The models will be adjusted on the basis of the data generated at the individual level. For example, the data produced through the second strategy and the experimental studies is used to try and see, when the exposure is modelled, how well the actual exposure a participant was exposed to can be predicted, as a way to try to make sure that what is modelled is in the range of what was measured.

The models generate the space of correlations and associations pertaining to the context of the specific study that exposome researchers are conducting. The resulting evidential claims thus have a statistical character because they specify statistical data. The goals and importance of these evidential claims are to be seen in in the context of the ongoing revision of exposure, such as air or water quality standards. They are evidential claims about the associated trends between environment and health/disease, in the sense that they look for the constituents and internal features of the datasets built through the other two strategies. More specifically, they are evidential claims which identify a specific set of statistical data as evidence for a phenomenon, such as the relation between exposure to air pollution and cardio- and cerebrovascular disease risk, as mediated by oxidative stress and inflammation, in Fiorito and colleagues' study.

---

38 For an overview of the statistical methods used for the study of the exposome, see Chadeau-Hyam et al. (2013).

These claims have a primarily statistical nature but may potentially be interpreted causally thanks to a causal interpretation of the results obtained by the micro strategy.

Although the results of this strategy are statistical claims, they are not the result of the work of statisticians only. Statisticians do play a key role at this stage, especially in the building of the tools and the models, but it is data analysts who run the models and perform the data analysis that will deliver the evidential claims. Like the other two strategies, the association strategy involves the working of different groups of individuals (as can be inferred by the long list of authors of published results). Researchers face various conceptual challenges when using this strategy. One involves the level of comparability between the estimates of exposure developed through the model and the omics data based on the samples collected in cohort studies, as we have seen in the context of the micro strategy. Other questions concern the role of integration and especially data integration. In EXPOsOMICS, researchers used a number of statistical tools to first of all integrate and analyse data collected in the cohort studies and, in a second step, integrate health and exposure data in order to try and find associations. The kind of data integration used at this stage regarded data concerning different phenomena – outcomes of interest, exposure, the environment, disease – and was aimed at making comparisons, identifying common elements or associations and studying relations stretching across this set of phenomena.

The ways in which exposome researchers tend to frame and present their results might suggest that the only evidential claims that actually matter are those presented as the results of the study, which in Fiorito and colleagues' case would be the evidence for the association between exposure to air pollution and cardio- and cerebrovascular disease risk. This would in a way come close to EXPOsOMICS researchers' view of their research, but I argue that it would provide a too simple and, in a way, misleading picture of their work. Namely, whilst what are presented as results of a study are indeed evidential claims, these are just one type of evidential claims, which is in turn based on

other types of claims and relying on different strategies. Although the association-level strategy might end up being the one that produces claims that have the most impact on further research, it should not be overlooked that this is crucially linked to the claims generated by the macro and micro strategies.

3.4 Epidemiology, Evidence and Causality

I have here reconstructed data practices in contemporary epidemiological research on the exposome in terms of evidential claims, presenting a typology of different approaches, methods, lines of work.

The way in which I have structured the subsections and I referred to examples from research in EXPOsOMICS might suggest that the strategies are necessarily subsequent and that they are exhaustive of all the aspects of an exposome study. However, as summarised by Figure 2, this is not necessarily the case. Firstly, the strategies are interrelated in both subsequent and non-linear ways, through at least two different types of relations. The first is rather straightforward: the evidential claims generated through the macro and micro strategy provide data that is used as evidence in, respectively, the micro and association strategy (see the curved arrows). Yet, the macro and micro strategies are also individually connected to the association strategy through a relation of influence and specification, as the epistemic commitments and assumptions made in the macro and micro strategies limit the kind of the evidential claims that can be generated through the association strategy (see the dashed arrows). Namely, the evidential claims of the macro and micro strategies provide a constrain to the generalisability of the evidential claims of the association strategies, for instance in terms of the size and type of case-control study specified by the macro strategy and the assumptions of data production of the micro strategy. As a result, the strategies are not mutually exclusive and are interrelate in non-linear ways. Figure 2 summarises these two relations and, in addition, visualises within each individual strategy the relation between data and

evidential claims, in terms of a two-fold relation of evidential support and data specification.

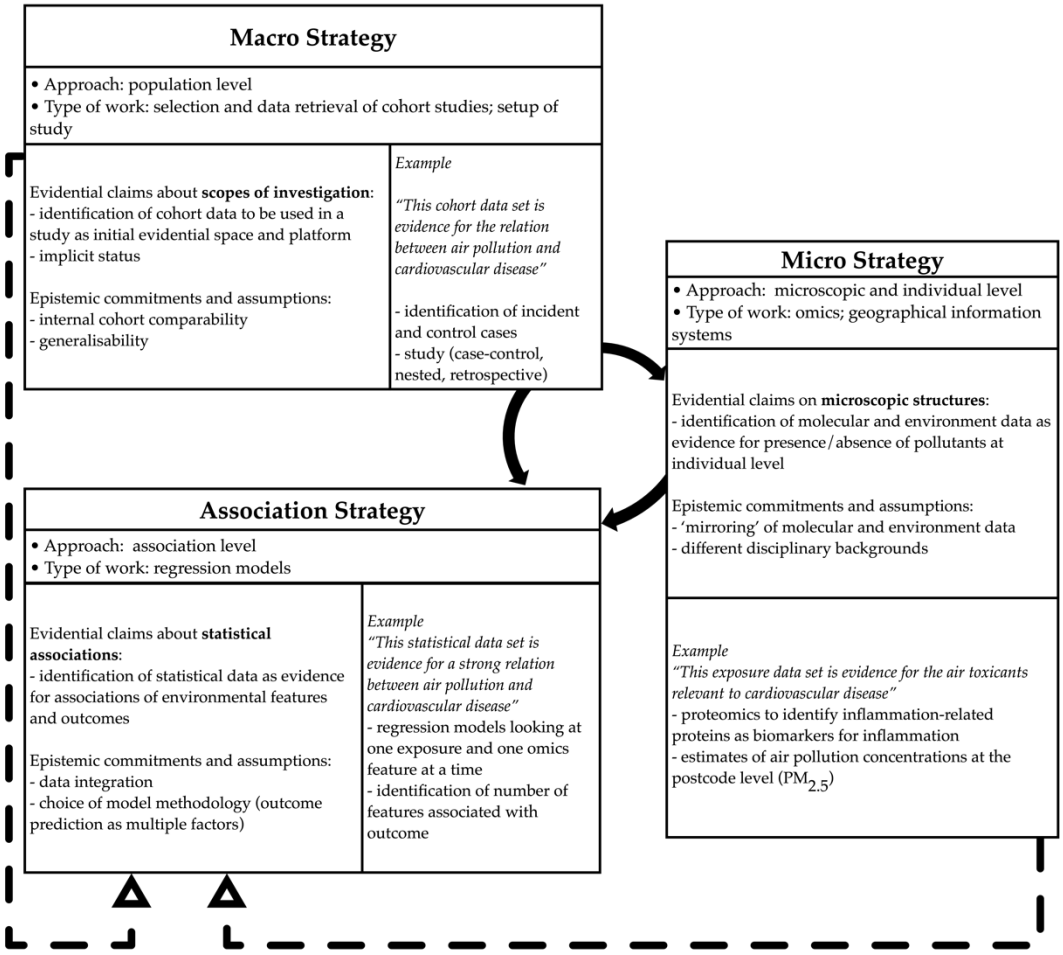

Figure 2. Reconstruction of the relations between the macro, micro and association strategies.

As for exhaustivity, it is important to note that this typology is a philosophical analysis of the epistemic characteristics and goals of exposome research, that I reconstruct in terms of three epistemic strategies used to generate evidential claims. As a result of this focus, my typology is not exhaustive of all the ways in which exposome research can be performed. Moreover, some of the strategies could stand on their own and be used to generate evidential claims whose value does not have to be linked to another strategy (such as the macro strategy). Sometimes, researchers may as well be unable to use all the three strategies, for instance as a consequence of the type of data they work with, and may have to employ other approaches.

Nonetheless, I claim that my approach is a fruitful perspective which distinguishes between different strategies for evidential claims and emphasises the complexity of research carried out in contemporary epidemiology. In doing so, it improves philosophical understandings of epidemiology. My emphasis on evidence production as a crucial goal of epidemiological research is different from many philosophical analyses, which have so far mostly interpreted epidemiological research in causal terms. This focus on causation and causal inference has yielded important work and has had the merit of putting epidemiology on the map of disciplines of interest to philosophy of science. Additionally, it has led philosophers to study epidemiological practice and methodology, to collaborate with epidemiologists and directly engage in current discussions in the scientific literature, which is in and of itself a significant result.[39] Plus, this focus is coherent with the 'end goal' of epidemiology, that is intervening on populations to improve public health: one wants to know about causes of disease affecting public health to intervene on them. Whilst I do not want to contrast this, I also think that it would not be charitable to interpret *all* epidemiological research as primarily concerned with causality or to focus only on causal claims. An analysis based only on causality focuses for the most part on the end results of epidemiological research and may thus end up overlooking the crucial epistemic role of other elements and claims that proceed final results but make them possible and significantly influence them. Investigating the role of data and evidence in epidemiology as one of the central outputs and results of research enables to shed lights on these issues, which are

39 See, for instance, the work of Federica Russo and Phyllis Illari on their information transmission account of causation, in collaboration with Paolo Vineis (Illari and Russo 2016; Russo and Vineis 2016; Vineis et al. 2017b), and the work of Alex Broadbent with various colleagues on the potential outcomes approach (Vandenbroucke et al. 2016; Fuller et al. 2016).

most important in the context of life and health sciences contributing to evidence used in policy.

In addition, by bringing forward the notion of evidential claims, I have tried to sketch a particular philosophical perspective on evidence. In Wylie's view, evidential claims relate material traces to phenomena of the past, identifying traces as evidence for the existence of the phenomena. In the chapter, I have used this notion to interpret data practices in contemporary epidemiology as relying on claims that identify a dataset as evidence for the study of phenomena. The use of this notion of evidential claims in the context of exposome research points to a view evidence as: the product of various epistemic practices, claims and considerations (epistemic strategies, levels of abstraction, assumptions, types of data, commitments, etc.); and as a category that identifies specific data to be used for specific purposes. In this sense, evidence is not to be seen as a necessarily fixed and stable entity, and neither as synonym with data. Rather, claims that identify a dataset as evidence of something are a way of expressing knowledge on the world. In the context of discussions about the relation between data and knowledge, Leonelli distinguishes between embodied knowledge, i.e. the knowing how that is required to interpret data and use it as evidence, and propositional knowledge, that is knowledge about reality (Leonelli 2013, p. 505). The strategies for evidential claims that I have presented here are indeed aimed at expressing knowledge, but at the same time retain a very close link to the data and have a rather loose and implicit propositional character. The evidential claims I have analysed are centred around the specification of datasets that can be used as evidence in the context of a research hypothesis. They crucially require the deployment of various aspects of embodied knowledge, both in terms of skills to be deployed during research and the use of existing knowledge and literature. They involve activities including the production, collection and retrieval of data. To use Leonelli's framework, the evidential claims generated in EXPOsOMICS would therefore be an instance of propositional knowledge, but understanding these claims as evidential claims provides a specification of the rather broad notion of propositional knowledge: evidential claims are very close to the data used as evidence and

are not full-fledged knowledge claims. At the same time, the interpretation of one of the goals of data practices in terms of evidential claims enables to specify some of the processes that lie within Leonelli's account of data, as material artefacts that can be used as evidence for claims about phenomena and can be circulated among individuals (Leonelli 2016a). I have suggested that this use can be interpreted as the generation of evidential claims, which gives a more specific perspective of what it means for data to get used as evidence in contemporary epidemiology.

3.5 Conclusions

In this chapter, I have argued that several aspects of epidemiological research on the exposome can be interpreted in terms of the development of evidential claims and can be reconstructed in terms of strategies for generating evidential claims. I have distinguished between three strategies. The macro strategy identifies the initial evidence platform, generating scoping claims that restrict the sample and provide an initial understanding of the phenomena to focus on. As part of the micro strategy, significantly different kinds of data are collected, with the common aim of identifying structure at microscopic level and elaborating evidential claims about the microscopic structure of things. On the basis of the evidential claims generated by the first two strategies, the association level provides evidential claims at the statistical level, concerning associations between exposure to the environment and outcomes of interest.

I have argued that this focus on evidential claims provides a different perspective if compared with existing philosophical work on epidemiology, that has vastly focused on causality and causal inference. Choosing strategies for evidential claims as units of a philosophical analysis makes it possible to identify the epistemic role of lines of work that are usually neglected in philosophical analyses (such as the macro strategy), thus allowing for a broader appreciation of research carried out in projects like EXPOsOMICS, their relevance to achieving epistemic goals and building a larger body of knowledge. And it enables

to identify and unpack conceptual choices and issues involved at almost every step of research. An additional insight of my analysis is connected to the epistemic dimension of the evidential claims generated on the exposome, and where they fit in considerations about data and scientific knowledge. I have argued that my approach sketches a view of evidence as a category which identifies the use of data for specific purposes and thus is positioned in between data and empirical knowledge. I have argued that this gives a specific perspective of how data is used as evidence in contemporary epidemiology, with a special focus on the data-centric research on the exposome. By focusing on this case, I also hope to contribute to the improvement of philosophical understandings of the epistemic role of data in the sciences, particularly in data-intensive settings, which run the risk of interpreting an increase in the volume of data as an automatic improvement in the quality and quantity of evidence.

Chapter 4

Evaluating Evidential Pluralism in Epidemiology: Mechanistic Evidence in Exposome Research

Abstract: In current philosophical discussions on evidence in the medical sciences, epidemiology has been used to exemplify a specific version of evidential pluralism. According to this view, known as the Russo–Williamson Thesis, evidence of both difference-making and mechanisms is produced to make causal claims in the health sciences. In this paper, I present an analysis of data and evidence in epidemiological practice, with a special focus on research on the exposome, and I cast doubt on the extent to which evidential pluralism holds in this case. I start by focusing on the claim that molecular data allows for the production of mechanistic evidence. On the basis of a close look at the ways in which molecular data is used in exposome research, I caution against interpretations in terms of mechanistic evidence. Secondly, I expand my critical remarks on the thesis by addressing the conditions under which data is categorised as evidence in exposome research. I argue that these show that the classification of a dataset as a type of evidence is dependent on the ways in which the data is used. This is in contrast with the approach of evidential pluralism, where evidence is classified in different types on the basis of its intrinsic properties. Finally, I come back to what I consider the core of the thesis and suggest that the epidemiological research analysed in the paper indicates different interpretations of evidential pluralism and its applicability in the health sciences.

4.1 Introduction

Evidence is a notion that we associate with multiple activities of our societies, including journalism, law, criminal investigations, etc. Evidence is an important issue when we want to use or produce knowledge. Science is therefore one of the areas where we would intuitively say that evidence plays a crucial role, and one may be even tempted to say that the use of evidence is among the features that characterises practices as scientific in the first place.

The topic of evidence has received philosophical attention primarily in epistemology and philosophy of science (Kelly 2016). In philosophy of science, many discussions have focused on the health sciences. This is partly a consequence of the rise of the methodological approach known as evidence-based medicine, which has brought about new ways of assessing the quality of evidence and provoked many critical reflections in the literature (Worrall 2002; Clarke et al. 2013; Stegenga 2014). As a result of this debate, a number of authors started to discuss how to classify evidence produced in medical research, connecting the issue to more general issues in philosophy of science, such as causality and causal inference (Campaner and Galavotti 2012).

Within this debate on causal inference and evidence classification in medicine, Federica Russo and Jon Williamson introduced a version of evidential pluralism according to which:

> To establish causal claims, scientists need the mutual support of mechanisms and dependencies. … The idea is that probabilistic evidence needs to be accounted for by an underlying mechanism before the causal claim can be established. (Russo and Williamson 2007, p. 159)

The 'Russo-Williamson Thesis' (RWT) identifies evidence of two kinds – evidence of causes making a difference to effects and evidence of underlying mechanisms – which has to be provided to make causal claims in the health

sciences.[40] A number of examples have been used to illustrate the thesis, including research from epidemiology (Russo and Williamson 2012). Epidemiology can be defined as the area of the health sciences concerned with the study of the distribution and determinants of health states – particularly disease – in human populations (Broadbent 2013). Philosophers of science have recently turned their attention to epidemiology and related problems, including: the notion of cause in epidemiology (Russo 2009; De Vreese 2009); the causal interpretation of epidemiological results (Broadbent 2013); causal explanations (Campaner 2011); and the consequences of methodological novelties on causal inference (Broadbent 2015; Vandenbroucke et al. 2016). In this context, the RWT has played an important role, connecting philosophical work on epidemiology with the broader debate on medical evidence, where epidemiology had largely been overlooked, despite being among the most important sources of medical evidence.

In this paper, I present a critical analysis of the RWT from the perspective of contemporary epidemiological research on the exposome. The exposome approach combines data on the molecular and external environment to study the internal and external components of exposure and improve the assessment of disease risks associated with environmental elements. It has recently attracted significant attention and has been used as a source of examples in the literature on the RWT (Russo and Williamson 2012; Russo and Vineis 2016). My analysis

40 As noted by one reviewer, a few different versions of the thesis have been proposed since the original formulation by Russo and Williamson, and this can generate some confusion. Thus, let me state that this general interpretation of the thesis is the one I shall adopt throughout the chapter.

of this line of research is based on the study of a specific project: EXPOsOMICS.[41] The project received funding by the European Commission in 2012 and, between 2012 and 2017, it was carried out by as a consortium of research centres in Europe and the US, coordinated by Imperial College London. EXPOsOMICS aimed at improving the assessment of exposure to air and water pollution and their impact on chronic disease, by studying and integrating internal and external components of exposure (Vineis et al. 2017a). The project can be considered a paradigmatic example of exposome research, as it was built on data collected by several other projects studying the exposome, with which it shares approaches and methods.[42] At the same time, the aim of integrating data from these diverse sources to produce evidence on environmental exposure makes the project a particularly significant case in the context of discussions on data and evidence (Fleming et al. 2017).

---

41 My study is empirically grounded on the study of EXPOsOMICS publications, reports, presentations. In addition, I conducted a series of qualitative interviews, which took place at the MRC-PHE Centre for Environment and Health of Imperial College London and aimed to document practices of data collection, production and integration and notions and classifications of evidence. Quotations from these interviews are taken from transcripts; full transcripts of the interviews are available upon request. In this chapter, I will use the empirical material to provide a reconstruction of uses of molecular data and approaches to evidence classification in EXPOsOMICS. A more detailed empirical analysis of the ways in which data is collected, interpreted and used in the project can be found in Chapter 3.

42 See for instance the Helix project, a cohort study looking at the 'early-life exposome', which some EXPOsOMICS researchers helped set up (Maitre et al. 2018).

My analysis will proceed as follows. I start by introducing the exposome approach and the way in which the RWT has been presented in this context (Sect. 4.2). I then turn to my critical remarks on the extent to which the RWT holds up in this case. I present challenges to two main aspects of the interpretation of exposome research based on the RWT, one pertaining to the interpretation of molecular data as mechanistic evidence and another to the way evidence is classified in the RWT framework. Firstly, in Sect. 4.3, I consider the claim that new sources of molecular data used in exposome research count as the production of mechanistic evidence and that, therefore, exposome research shows the interplay between mechanistic and difference-making evidence suggested by the RWT. On the basis of a detailed look at molecular data in EXPOsOMICS, I show that molecular data is used to study differences and dependencies of exposure and cast doubt on interpretations in terms of mechanistic evidence. Secondly, in Sect. 4.4, I consider the way in which the RWT distinguishes between different types of evidence. I focus on the ways in which data is used and classified as evidence in EXPOsOMICS and I argue that this presents more general challenges to the approach to evidence classification of the RTW framework. On the basis of these criticisms, in Sect. 4.5, I come back to the 'core' of the RWT and ask if it can be reconciled with the challenges highlighted in the paper. I argue that this may be possible and the case of exposome research indicates different ways of interpreting evidential pluralism and its applicability.

4.2 Exposome Research and Evidential Pluralism

First introduced by Christopher Wild (2005), the exposome refers to the totality of exposures experienced by individuals throughout their lifetime. Totality, here, means that the exposome encompasses exposures experienced at different levels, which include exposures to elements and features of the external environment (such as pollution, diet, stress, etc.) as well as the internal chemical environment (such as inflammation, oxidative stress, gut flora, etc.). These

different levels exist because individuals' exposure to external elements can produce toxicants and/or reactions at the internal level, which can in turn lead to the development of disease. In contrast with the more traditional focus on the external components of exposure, the 'all-encompassing' approach to the exposome is often presented as a new paradigm in the context of environmental epidemiology (Rappaport and Smith 2010).

Practically, much of the work on the exposome consists in the search for biomarkers, i.e., elements of the environment or the organism that can be measured and in turn measure the presence of some toxicant. Biomarkers are considered capable of signalling the development of disease, thus guiding researchers in the understanding of the impact of the environment on disease.[43] In this context, molecular data is among the primary tools used to identify biomarkers of the internal component of exposure. More precisely, exposome researchers use omics technologies to measure the different levels at which exposure from environmental elements leaves traces in the body, including: processes that involve small molecules playing intermediary roles in the functioning of the metabolism (metabolomics); parts of proteins or DNA that are bound to specific chemicals (adductomics); epigenetic changes in the cells' genetic material (epigenomics); mRNA expressions (transcriptomics) and proteins (proteomics).

Recently, Federica Russo and epidemiologist Paolo Vineis have connected the use of molecular data in exposome research to the RWT. They say that "molecular epidemiology improves on traditional epidemiology also in another important respect: it goes beyond associations and is in principle able to shed light on the *mechanisms* of disease causation" (Russo and Vineis 2016, p. 254; emphasis in original). Similar considerations have been developed elsewhere in the literature on the RWT. In one of the first publications including a philosophical

43 For a discussion of working definitions and the conceptual status of biomarkers, see Strimbu and Travel (2010).

analysis of exposome research, Russo and Williamson (2012) claim that this line of research shows an interplay between difference-making and mechanistic evidence. Together with Phyllis Illari, Russo and Vineis use the exposome approach to cancer research to argue for an information transmission account of causality and talk of the need for evidence of both the difference that exposure makes to the development of cancer and of how exposure leads to the development of cancer (Vineis et al. 2017b).[44]

In this context, the significance of Russo and Vineis' claim is that exposome research can *produce* – and it does not only *use* – mechanistic evidence, and that this production is to be seen in the molecular features of exposome research.[45] In this sense, exposome research is innovative compared to traditional epidemiology (also) because, thanks to molecular data, it does not only produce new difference-making evidence but also mechanistic evidence. In other words, the claim is not that exposome research *uses* mechanistic evidence that has been gathered in other areas of biomedical research and that this, together with the difference making evidence it produces, shows the interplay maintained by evidential pluralism. Rather, the argument is that the exposome approach improves on traditional epidemiology because it can study the mechanisms of disease causation and, thus, produce mechanistic evidence. It is this claim about the production of a specific kind of evidence that will be the target of my critique in the next section.

Before presenting my remarks, let me briefly summarise the distinction between difference-making and mechanistic evidence developed by the RWT.

---

44 For a presentation of Illari and Russo's information transmission account, see Illari (2011b) and Illari and Russo (2016).

45 I will come back this distinction between use and production of evidence and why it is important in Sect. 4.5.

Evidence of difference-making refers to evidence that is used to show that effects depend on causes and that causes make a difference to effect (Russo and Williamson 2007, p. 158). Randomised controlled trials are among the most cited examples of a methodology producing difference-making evidence. Other examples, from Illari (2011a), include evidence gathered in observational studies, probability distributions in Bayesian networks, as well as counterfactual and invariant relationships. As such, Russo and Williamson say that difference-making often comes in a probabilistic form and is usually employed to make predictions.[46]

As for mechanistic evidence, Russo and Williamson suggest that we should think of it as evidence of the mechanism(s) underlying the relation between causes and effects. Mechanistic evidence is what provides the productive component of causal claims, which gives information on *how* causes bring about effects. They also argue that this is the kind of evidence which plays the crucial role of helping with issues possibly affecting difference-making evidence, such as confounding (Russo and Williamson 2012, pp. 249–250). This notion of mechanistic evidence has often been criticised because of its vagueness and ambiguity, which has in turn led to disambiguating work in the literature on the RWT. For example, Illari argues that we should not think of the distinction between difference-making and mechanistic evidence as a distinction between the methods we use to gather evidence, but rather as a difference between objects of evidence (Illari 2011a). In her view, mechanistic evidence is not necessarily the evidence we get through, say, interventions, as opposed to observational studies. Rather, mechanistic evidence is evidence of the existence of

46 Russo and Williamson sometimes use the term probabilistic evidence instead of difference-making evidence. For matters of consistency, I will follow their original claim on the connection of probabilistic evidence with causes making a difference to effects and, in line with Illari (2011a), I will only use the term *difference-making* evidence.

mechanisms, evidence that is about mechanism and has mechanisms as its' object. Following her and Williamson's characterisation, evidence of mechanisms thus refers to evidence of the entities, activities or the way these are organised to produce the phenomenon for which the mechanism is responsible (Illari and Williamson 2012). In this view, potential mechanistic evidence for, say, protein synthesis would be evidence of the organised entities (DNA, RNA, ribosomes, etc.) and activities (e.g. transcription, regulation, etc.) involved in the mechanism that brings about the synthesis of proteins.

Despite these disambiguating efforts, the evidential pluralism proposed by the RWT has received significant criticism in the literature, especially with regard to the strong stance on mechanistic evidence. For example, Howick (2011) has casted doubt on the claim that mechanistic evidence has always to be provided in order to establish causal claims. In the philosophy of epidemiology, Broadbent (2011) has argued against mechanistic interpretations of the causal inference strategies used by epidemiologists. Gillies (2011) has shown that the epidemiological discovery of the causal relation between smoking and lung-cancer was first established on the basis of difference-making evidence only, much earlier than any mechanism was found.

4.3 Mechanistic Evidence in Exposome Research

The use of molecular data in exposome research has been interpreted as the production of mechanistic evidence, which is considered necessary by the RWT to support a causal claim. I now give a close look at how molecular data is collected and used in exposome research, focusing on examples of the use of exposure profiles from molecular data in EXPOsOMICS. This analysis leads me to argue that, in the context of EXPOsOMICS, evidence from molecular data is not really used as mechanistic evidence, at least if we are to follow the notion of mechanistic evidence developed in the RWT literature.

In EXPOsOMICS, molecular data was collected through the analysis of the microscopic structure of samples from cohort studies. For example, blood plasma samples retrieved from cohort studies were analysed using the adductomics technique, i.e., by collecting data on protein adducts, which are formed by the binding of proteins to chemicals, in order to study the effects of exposure to air pollution on proteins. As one EXPOsOMICS researcher explained:

> Adductomics is an omic technique and the specificity of the method is that it measures adducts of human serum albumin and toxicants. The toxicants present in your serum tend to have a very fast turnaround, so they disappear and they become metabolite and they're very difficult to measure. So, for example, if you've been exposed to air pollution it's going to be difficult to find the related toxicant in your serum. But the same molecules can bind to albumin and then they are conserved in your serum for a longer period of time. So, by measuring these adducts, you can track exposure for longer time. (Researcher A1)

Results consist in spectrometry data, i.e., plots with lines and peaks, where different peaks correspond to different molecules and break down the sample into its molecular elements and compounds. These are usually called 'exposure profiles' or 'omic profiles'. On the basis of exposure profiles, researchers try to identify toxicants whose presence can be considered a consequence of exposures to external pollutants and can in turn be used as an intermediate biomarker to track the effects of exposure and possibly the development of disease. Exposure profiles are usually presented as tables where lines are dedicated to chemicals and the content of the columns provides information on the measured mass of the chemical within the sample.

As we have seen, according to the disambiguated notion provided by Illari, mechanistic evidence is the kind of evidence that has mechanisms as its object (Illari 2011a); and, following Illari's work with Williamson, a mechanism for a phenomenon consists in the entities and activities organised in order to be responsible for the phenomenon (Illari and Williamson 2012). The first problem

I see in interpreting evidence from exposure profiles as mechanistic evidence is precisely that the 'object' is not a mechanism, to use Illari's language. The rationale behind the use of molecular data in exposome research is to look for chemicals whose presence may be due to an exposure to the external environment, not to look for the ways in which these chemicals are organized to form mechanisms. From the perspective of Illari and Williamson' notion of mechanism, one may say that researchers do look for entities – yet, they do not really look at the activities these carry out, as molecular profiles provide a static picture, and not even at the organisation of these entities as part of a mechanism. In EXPOsOMICS, exposure profiles were rather used to provide evidence for claims about the *dependency* of the molecular presence of toxicants on external exposure and to find out about the *difference* that the environmental exposures experienced by individuals make to the presence of these toxicants. These features place molecular evidence very close to difference-making evidence, of which dependency and difference are core elements, and quite apart from evidence of mechanisms linking exposure and disease.

More generally, the context where molecular evidence was used in EXPOsOMICS is the identification of associations between pollutants and aspects of health and disease, rather than the discovery of mechanisms. Exposure profiles were used in regression models as evidence about toxicants at the molecular level, in order to find associations with the presence of toxicants at the external level. One EXPOsOMICS researcher explained this whilst talking about modelling molecular data from metabolomics to study the effects of environmental exposures on birth weight:

> Each model models the relation of birth weight to that metabolic feature, so what's the *change* of birth weight with that metabolite, once you take into account covariates like the mum's weight, the mum's education and all these things (Researcher B1; my emphasis).

The focus here is on changes and differences, rather than on how these differences are brought about or through which mechanisms.

In addition, one of the characteristic features of mechanistic evidence, as defined in the RWT framework, does not seem to be supported by molecular data. As argued by Illari (2011a), mechanistic evidence is crucial for causal inference because it helps with issues affecting difference-making evidence, including confounding. However, this feature is not supported by exposure profiles, which on the contrary are affected by issues that typically affect difference-making evidence. I take this point from the scientific literature on molecular evidence, where for instance Vineis and Perera say that one of the problems related to the use of omics as intermediate biomarkers is confounding:

> One of the main challenges with intermediate biomarkers is to understand whether they belong to the causal pathway between exposure and disease, whether they are simply a side effect of exposure or disease, or whether their measurement is confounded by some other exposure. (Vineis and Perera 2007, p. 1961)

The emphasis on difference making and problems of confounding in exposome research are the reasons why I caution against interpreting molecular data as mechanistic evidence. One may still say that there is a simpler and weaker claim that can be defended, namely that molecular data in exposome research *hints* at where to look for mechanisms and this is enough to say that it counts as mechanistic evidence. In this view, evidence from exposure profiles would count as mechanistic because it hints at the presence of entities involved in underlying mechanisms. The claim, however, is problematic even in this weaker form, as it is not clear how molecular data hinting at a mechanism would count as the kind of mechanistic evidence defined by the RWT. To call mechanistic evidence something that hints at mechanisms would imply to stretch the notion of mechanistic evidence too much, up to a point where the notion itself loses most of its power. It would imply going beyond the notion of mechanistic evidence put forward by Illari (2011a), which is already quite broad. One may say that my contention about mechanistic evidence may be too stringent and suggest that, in constrained areas of research like exposome research, even just providing hints at mechanisms could count as production of mechanistic evidence or mechanistic explanations. Still, as I explain here, I struggle to see how

evidence hinting at mechanisms could count as the mechanistic evidence defined by the RWT. If one wants to look at the case from the perspective of mechanistic explanations, I would rather say that the molecular data used in EXPOsOMICS plays the role of a heuristic for finding mechanisms, as no mechanism seems to be discovered directly, or that is part of how-possibly mechanistic explanations (Forber 2010; Reydon 2012). Moreover, mechanisms can often be discovered on the basis of types of evidence such as difference-making evidence and manipulative evidence, rather than mechanistic evidence (Campaner and Galavotti 2012).

4.4 Classifying Evidence in Exposome Research

Whilst critically analysing interpretations of molecular data as mechanistic evidence, I have put significant emphasis on the context of the *use* of molecular data as evidence in EXPOsOMICS. Building on this approach, in this section I expand my analysis and make a more general point on evidence classification. I intend to show that, in the context of data practices in EXPOsOMICS, what makes a difference for the classification of data as evidence of a specific kind is the way in which the data is used. I argue that this is in contrast with how evidence is differentiated in the RWT framework. I therefore expand my critical remarks on the RWT by pointing to more general issues in the way medical evidence is framed in the thesis.

In the literature on the RWT, a specific type of evidence is usually linked to a specific method for the generation of evidence. This suggests that the notion of evidence employed by the RWT is close to traditional philosophical accounts of evidence and data. In these views, data counts as evidence on the basis of intrinsic properties, that are fixed and inherently local and stand in a representational relation with aspects of reality, independently of the context in which

the data is used.[47] These views have recently been challenged by Sabina Leonelli, who has argued that data should rather be considered a relational notion with a non-representational character (Leonelli 2015, 2016, pp. 69–92). When asked about notions and classifications of evidence, EXPOsOMICS researchers discussed these issues in the context of considerations about the data used in the project, and more specifically the ways in which they use, analyse, aggregate and disseminate their data. Thus, I now introduce the various types and sources of data researchers work with in exposome projects like EXPOsOMICS.

In the previous section, whilst analysing the use of molecular data, I have mentioned that the starting point for those analyses is data retrieved from cohort studies. In EXPOsOMICS (as in most epidemiology), cohort studies were one of the crucial and initial sources of data. Cohort studies follow a population of interest – the cohort of participants – for a long period of time, during which extraction of physical samples, questionnaires and one-to-one interviews are performed to collect data on a number of variables, including clinical states, features of participants' lifestyle, house environment, etc. The types of data used in EXPOsOMICS and retrieved from cohort studies varied significantly, including: quantitative data, such as measures of chemicals and compounds, as well as qualitative data, such as information provided in interviews; data about specific individuals as well as data about the surrounding environment; physical samples stored in low temperature as well as spreadsheets saved in databases.

As we have seen, some of the data retrieved in cohort studies is analysed to study the molecular composition of the internal chemical environment and produce exposure profiles. Together with what is directly acquired through cohort studies, molecular profiles constitute what EXPOsOMICS researchers

47 A paper by Bogen and Woodward is often cited as a prime example of this view (Bogen and Woodward 1988). For more recent and nuanced takes on this approach, see e.g. Rheinberger (2011).

defined as evidence on population health, or 'health data', which is thus composed of various kinds of data, collected at different levels of abstraction, at different times, and with different kinds of expertise involved. EXPOsOMICS researchers distinguished health data from what they called 'environment data', i.e., data on the conditions participants were exposed to during the study period. As with health data, the starting point is cohort studies, where participants are asked about the environment where they live and/or work, their lifestyle, diet, etc. One of the basic questions that participants to cohort studies are asked to provide is the postcode of the area where they live. In EXPOsOMICS, this was used to design geographical information systems to produce new data about the quantity of toxicants in the cohort's environment (Vineis et al. 2017a, p. 143; Fiorito et al. 2017, pp. 237–238; see also Chapter 3, Sect. 3.3.2). Starting from postcodes, and on the basis of maps of characteristics of the population (such as density of the population, traffic intensity, whether the area is an industrial area, etc.), a geo-spatial regression model is designed to take into account these variables and differences in populations and assign exposure estimates to every participant in the study (Gulliver et al. 2018).

I argue that the context of data practices in EXPOsOMICS shows that the classification of a dataset as a specific kind of evidence crucially depends on the ways in which the dataset is *used* as evidence. For example, data from cohort studies can be evidence for a variety of phenomena, analyses and claims, depending on the way it is used. In EXPOsOMICS, it was indeed used as evidence for the study of different phenomena, including: as molecular evidence for exposure; as evidence for health states of a population; and, as part of geo-spatial models, to produce further evidence for the environment the cohort could have been exposed to. This is at odds with the idea that evidence should be classified on the basis of the method with which it is produced. Collection methods do play a crucial role for the evidential value of a dataset, but they do not determine it on their own. Another important feature of data practices in EXPOsOMICS is that the evidential content of a specific type of data is determined by the mobilisation and use of other types and sources of data. This is particularly noticeable in the case of environment data, where the evidential

value of the initial dataset (postcodes) is determined through the mobilisation of various other sources of data.

I argue that these features point to a general approach to evidence, where considerations about evidence are indeed determined by the data collected and retrieved in specific projects like EXPOsOMICS. Yet, what plays a crucial role is not just the intrinsic properties of a dataset (such as its origin, type, content, format, etc.), but rather the ways in which the dataset is used, fits into research questions and is mobilised and put in relation with other datasets.

This approach to evidence is also reflected in researchers' reactions when asked about the classification of data as mechanistic evidence. One researcher said that he would find it necessary to perform various further experiments and have other types of data, for molecular data to be used as mechanistic evidence. His way of framing mechanistic evidence closely linked classification to the context of use of data. For example, he pointed to research in the metabolomics literature, published in *Nature Medicine* (Wang et al. 2011). In this work, Wang and colleagues first identified molecules that were correlated with cardiovascular disease through omics analyses of metabolites. However, the data produced in this way was not used as evidence of mechanisms. In order to test the presence of mechanisms underlying associations and thus whether the metabolites were rightfully predictive of the disease, researchers carried out further tests, including animal studies. It was only as a result of the mobilisation of these other sources of data that researchers claimed they had managed to link molecules both to an external exposure (a dietary source) and disease, thus granting the use of their molecular data as mechanistic evidence. In the words of one EXPOsOMICS researcher:

> Through all these different experiments – and you can tell, a Nature paper is only two sides, but it brings together lots and lots of work –, using these, they were able to show the dietary source, phosphatidylcholine (which is found in meat, eggs, etc.) and they were able to link it all the way though to these metabolites and also show these

> metabolites directly caused atherosclerotic, as the blood vessels are directly linked to cardiovascular disease. (Researcher B1)

What I am arguing here is not that a definition of mechanistic evidence closer to Wang and colleagues' work is superior to the RWT. Rather, I am pointing out that views of what counts as a specific kind of evidence can vary significantly and depend on the context of use of data as evidence. Therefore, what counts as a type of evidence may shift considerably and clear-cut distinctions between pieces of evidence, as if these were inherently different entities, seem too strict. The kind of evidence a dataset may provide is not determined ex ante, and it rather depends on the heterogeneous ways in which data is used and disseminated across different sites.

The problem of the RWT approach is that it does not seem flexible enough to accommodate this attitude to evidence classification, as the way in which evidence is classified according to the RWT is based on a view of evidence as something that is produced in a form which makes it of a specific and stable kind, more precisely either difference-making or mechanistic evidence. This approach seems to overlook the context of data practices, where considerations about what counts as evidence actually take place in practice; and more generally, questions on the relation between data and evidence do not seem to be considered in the RWT framework. In this context, the work of Illari (2011a) on specifying the notion of mechanistic evidence is particularly significant, as it moves away from a classification based on evidence-gathering methods towards a distinction based on what we use evidence for. This is close to my argument, but Illari also argues that evidence should be defined on the basis of its 'object'. For instance, a piece of evidence is mechanistic evidence insofar as it has the existence of a mechanism as its object. As such, Illari's view seems similarly based on a view of evidence as something with a fixed and representational content, which – I have argued – does not hold in the context of the data practices of exposome research.

4.5 Exposome Research and the Core of the Russo-Williamson Thesis

The case of exposome research presents a number of challenges to the conceptual framework of the RWT. On the basis of the critical remarks of the paper, what should we make of the evidential pluralism proposed by the RWT? As we have seen, discussions on causal inference in the health sciences are the main context of the RWT. The RWT is a claim about what we need to know to make causal claims. The arguably most important and controversial element of the thesis – its core idea – is that production is a necessary aspect for inferring causality and that we need to know about production when making causal inferences. This productive component is framed in terms of mechanistic evidence.

The core of the RWT seems to grasp an idea that is in general terms shared in the health sciences, including exposome research. EXPOsOMICS researchers often commented on their work saying that it is mostly focused on getting information about dependences between variables, but knowing what lies *in between* dependence relations makes a crucial difference and is the ultimate way of testing causal claims:

> And the ultimate way of validating your result and making sure that it's something that is general is to understand the biology behind it. We're not at this stage, but the ultimate way would be to do some mechanistic wet lab experiment, to understand the mechanisms. And, once you understand the mechanisms, then there's no reason why there should be differences between one individual and another. So that's your ultimate way of generalising. (Researcher D1)

How can we reconcile this intuition, at the core of the RWT, with the critical remarks made in the paper? The problem is how to interpret the claim on the productive aspect of causality. The RWT frames it in terms of a need for mech-

anistic evidence. But does that mean that mechanistic evidence should be *produced* when causal claims are made, or is the *use* of mechanistic evidence enough? The RWT is not clear on this issue.

If we take it to mean that mechanistic evidence has to be produced, the thesis seems too strong and hard to maintain in many lines of research in the life sciences. For instance, in exposome research and other areas of epidemiology, it seems often difficult to identify mechanisms and mechanisms are not always the focus of research. At the same time, we will probably want to say that epidemiologists can make causal claims, even if no mechanistic evidence is produced. Many public health policies are based on epidemiological studies where no mechanistic evidence seems to be produced. For example, environmental epidemiologists are often concerned with determining air quality standards, that is measurements of how much air pollution we can and should tolerate.[48] Public health policies on these standards are developed on the basis of epidemiological studies about, say, the relation between ambient air pollution and the incidence of cardiovascular disease, often with no primary focus on mechanisms. We will probably want to say that these polices are based on causal claims of some sort, as knowledge of causal relations is usually considered to provide among the best justifications for effective policy interventions (see e.g. Cartwright 2011). Hence, requiring mechanistic evidence to be produced, each time causal claims are made in the health sciences, seems problematic. This suggests that the core of the RWT should be rather interpreted as the need for use, rather than production, of mechanistic evidence.

At the same time, it is important to stress that, in exposome research and epidemiology more generally, new studies are designed, data is analysed and re-

48 Projects like EXPOsOMICS, which focuses on air pollution and water contamination, are precisely funded with the aim of gathering evidence for updating and revising these standards.

search is shaped (also) on the basis of mechanistic knowledge. When environmental epidemiologists assess, for instance, associations between air pollution and cardiovascular disease, background knowledge on potential mechanistic pathways linking the two phenomena plays an important role. However, in cases such as exposome research, evidence of mechanisms often comes from research in other areas of the health sciences. This suggests an interdisciplinary interpretation of the thesis. So far, most of the examples used in the literature on the RWT come from single disciplines and/or single projects in the life sciences. Yet the thesis can be interpreted to apply on a higher level: some disciplines may provide evidence for the probabilistic angle of causal inference, whilst others may provide the productive one. In this view, one could say that a project like EXPOsOMICS provides some of the evidence that is necessary to put together to solve the puzzle of disease causation. Other kinds of evidence, possibly including evidence of productive causation, may need to be provided by other disciplines and lines of research. This interpretation would take into account my characterisation of evidence classification in the previous section, and especially the need for mobilising various lines of evidence for data to be used as evidence. As such, the thesis would allow to interpret epidemiology as contributing in part to explanations of disease that, in the long run, may be mechanistic, without running the risk of 'over-interpreting' all disciplines or lines of research in productive or mechanistic terms. In other words, in this view one may be able to say that exposome research provides hints at mechanisms without having to say it produces mechanistic evidence.[49]

---

49 Personally, I find it more charitable to say that epidemiological research on the exposome is more than an instantiation of research intended to describe a mechanism, because it can potentially inform mechanistic research but produces evidence that is also important on its own.

This interpretation of the RWT on the basis of my critical remarks might be a promising project. Still, a few problems remain. One problem concerns the normativity of the thesis. Whilst this interpretation of the thesis would pose weaker requirements on single disciplines or lines of research, the claim on mechanistic evidence may still be too strong. Namely, in some areas of the health sciences, researchers might be tempted to make a causal claim even if they do not have mechanistic evidence and before carrying out research aimed at discovering underlying mechanisms. For example, in the aforementioned case of the epidemiological discovery of the relation between smoking and lung cancer, the mechanism was discovered long after the initial causal claim was made and the initial smoking bans were enforced (Gillies 2011). A second and related problem is the notion of mechanistic evidence, which may not be the best way of framing the need for the productive component of causality. In disciplines like epidemiology, mechanisms are indeed used to shape research. However, more than evidence of mechanisms per se, it is rather knowledge of mechanisms that is used to guide research.

4.6 Conclusions

The central object of research of this paper has been data practices and ways of classifying and framing evidence in epidemiology, with particular reference to research on the exposome. The principal issue I have dealt with is the extent to which the RWT can be applied to this line of research.

I have cautioned against the claim that the use of molecular evidence can be considered as providing mechanistic evidence as defined by the RWT. I have shown that, in the EXPOsOMICS project, molecular evidence has been used for claims about the difference that exposures make, in contrast with the notion of mechanistic evidence proposed by the RWT. Then, I have focused on another aspect of the RWT, i.e., the way evidence is classified and distinguished between different kinds. On the basis of a reconstruction of data practices in

EXPOsOMICS, I have shown that, in contrast with the RWT, there are substantially different views and approaches to what counts as a type of evidence. Finally, I have looked back at the core of the RWT from the perspective of my case study and critical remarks, and I have proposed an interdisciplinary and use-based interpretation of the thesis.

This interpretation takes into account the critical points I have raised in the paper. Additionally, it tries to push the thesis forward, towards a direction where it is capable of accounting for the use of data as evidence in current research and can therefore make fitting suggestions on the consideration and appreciation of a plurality of medical evidence.[50] I have pointed to some aspects of the core of the thesis, and the interpretation that I have put forward, that need further review and consideration. It also remains to be seen to which extent these considerations are applicable in other areas of epidemiological and, speaking more broadly, health science research. Another related issue concerns how the integration of evidence from different sources, produced through different methods and mobilised through different pathways, is and should be carried out – a topic that has recently been discussed in the literature, but has yet to be explored in the context of epidemiology.[51]

---

50 In this sense, see also the work on evaluating evidence of mechanisms by the EBM+ consortium (Parkkinen et al. 2018).

51 See the special issues of *Synthese* on "Evidence Amalgamation in the Sciences" (e.g. Bertolaso and Sterpetti 2017) and *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* on "Integration in Biology: Philosophical Perspectives on the Dynamics of Interdisciplinarity" (Brigandt 2013).

Chapter 5

Conclusion and Outlook

This dissertation set out to discuss the collection, integration and use of data in the context of epidemiological research on the relation between exposure and disease. My main research question concerned the ways in which data practices inform and sustain the subject matter of the epidemiology of the exposome. I have investigated the conditions under which data is retrieved, collected, integrated and used as evidence for scientific claims, connecting the context of data practices with material, technological, institutional and rhetorical components. In this way, the three central chapters of this dissertation converge into an account of scientific practices about data, that is positioned in the local context of epidemiological research as well as broader philosophical discussions on the life and health sciences.

I pursued my research questions by examining data in contemporary research on the exposome, an all-encompassing approach to the understanding of the ways in which exposure shapes health of human populations. I have taken EXPOsOMICS, an EU-funded project that studied the effects of air and water pollution on chronic disease, as the main empirical focus of my dissertation. I have employed a qualitative case study approach, based on the analysis of documents, reports, presentations and publications and a series of qualitative interviews and participant observation that I have carried out with researchers in the EXPOsOMICS project. I have used exposome research as a case study of research that:

- is centred around the integration of data of significantly different types, about health and environment, collected in experimental and observational studies, at the macroscopic and microscopic level;
- employs interdisciplinary teams and approaches, which include molecular biology, medicine, statistics, geography, information science, etc.;

- is considered highly innovative as it is built around the development of new concepts, methods and technology;
- and deals with issues at the intersection of health and environment, thus with significant political, social and economic consequences.

As a result of my analysis of exposome research, I argue that one of the epistemically salient features of data is that it can sustain the subject matter of epidemiological research by shaping notions and concepts, approaches and strategies, techniques and technologies and what counts as evidence. In the case I have studied, this epistemic role played by data is enacted by the ways in which it is used by epistemic agents and communities. As a consequence, the epistemic role of data in exposome research is connected to and mediated by other artefacts, components and features of the scientific inquiry.

In particular, in the central chapters of the dissertation I have claimed that:

- The scientific innovations of the exposome should be framed as the establishment of a repertoire, as opposed to a paradigm, in which various conceptual, social and material components are aligned. The exposome repertoire consists in many components transferred from other areas of the life and health sciences. Scientific change is a result of this alignment, and it is not due to only one of these factors, such as data (Chapter 2).
- In the context of data practices in exposome research, researchers use evidential claims to specify the evidential and representational value of their datasets. Three epistemic strategies for evidential claims can be distinguished, differing in terms of level of abstraction, lines of work and type of evidential claim and leading to a picture of evidence production as epistemic-intensive labour (Chapter 3).
- Philosophical views on how to distinguish and classify evidence in health research, such as the Russo-Williamson Thesis, fall short in the way they have been applied in exposome research. New types of data collected and analysed in exposome research are not used as mechanistic evidence of the

kind maintained by the thesis. Exposome research suggests a different approach to classifying evidence and different ways of interpreting the thesis (Chapter 4).

To conclude, I would like to point to some issues that I have discussed in various parts of the dissertation and collect in one place the claims I have made in the chapters of the dissertation. In the remainder of this chapter, I will discuss: the philosophical significance of epidemiology; the relations between data and other components of scientific epistemology; scientific change and innovation; the role of technology; notions of evidence; and conceptualisations of health and disease. I will also present an outlook of my research, pointing to evidence and health and disease as topics that could be considered for future research.

As I have often mentioned in the dissertation, *epidemiology* has started to receive attention by philosophers of science only very recently. In this dissertation, I hope to have shown that epidemiology is an interesting area of research from a philosophical viewpoint. I think that epidemiology is especially interesting for philosophical questions about the role of integration in the life and health sciences, i.e. integration of methods, technologies, data and theoretical commitments of diverse communities, disciplines and lines of research, and about the nature of contemporary scientific research that intertwines political, epistemic and societal dimensions. I hope to have given a perspective on some of these issues in this dissertation, where I have engaged in a data-centric philosophical study to understand, document and account for the conditions in which data is used in the context of contemporary research. I have discussed these and other general features of my dissertation in Chapter 1, where I have positioned my work in the context of academic debates in philosophy of science and science studies. I have discussed the broad, interdisciplinary debate on the role and effects of data in various processes and activities of modern societies, including the sciences; and, more specifically, what I have called data-centric studies in philosophy of science, looking at the epistemic role of data and data practices in the sciences. I have specified my research by introducing the philosophical scholarship on data in the life and health sciences, in

order to explain and justify the reasons why I have chosen to pursue a study of data in the context of the exposome. I have finally connected my project to a recent turn in philosophy of science, as a consequence of which philosophers have started to give more attention to the study of scientific practice and have argued for the need of empirical grounding through methods from the social sciences. I have used this discussion to introduce the methodology I have used during the research for this dissertation, including case study and qualitative interviews.

In Chapter 1, I have argued that data is to be considered as one of the (many) material and epistemic artefacts that are produced by scientists throughout their research. I have argued that an epistemology of science focused on scientific practices should discuss the epistemic role of these artefacts and the relation between them, including data, methods, models, evidence, claims, theories, knowledge, etc. In this dissertation, I have focused on scientific data and have discussed relations between the context of data practices and some of these artefacts, including theoretical commitments, knowledge claims and evidence. I have started my analysis of these *relations* by discussing the background and wider context in which the exposome has emerged as a new concept in the last decade of epidemiological research. In Chapter2, I have investigated the conditions under which the exposome was conceived, developed and established. I have used Ankeny and Leonelli's (2016) conceptual framework of 'repertoires' to highlight the role of non-conceptual components and make sense of the ways in which material, financial, institutional and technological elements are aligned in exposome research. I have shown that many of these factors were transferred from other areas of the life and health sciences, including: the sequencing repertoire, which emerged in the genomic context and has since then increasingly spread in the life and health sciences; exposure science, which studies human contact with various external, chemical agents; and biomarkers research, i.e. an approach that studies elements or characteristics that can be precisely measured and used as indicators of various processes. This analysis has led me to a nuanced account of *change and innovation* in the context of the exposome, as I have argued that forms of innovation include

non-theoretical components of the repertoire and are to be seen in continuity with traditional and longstanding approaches in the field. Questions on change and innovation have been explored in other chapters of the dissertation. In Chapter 3, I have given a close look at data practices in EXPOsOMICS and argued that different epistemic strategies have arisen to collect, analyse and use new sources of data as evidence. In Chapter 4, I have focused more specifically on the use of molecular data on the internal component of the exposome, and I have explored claims according to which the use of this type of data provides a new kind of evidence for epidemiology. As I explained in Chapter 1, data studies have often pushed back against views ascribing revolutionary epistemic powers to the use of large datasets, by looking at specific contexts of data use. My analysis goes in the same direction: the account of scientific change that can be drawn from the dissertation suggests that innovation is the complex result of the alignment of various components of a line of research, thus pushing back against the view that only one component – data – could be responsible for paradigmatic changes.

Discussions on scientific innovation are often connected to analyses of the role of *technology* in recent scholarship on scientific research. In the dissertation, I have discussed technology as a form of innovation when analysing material and technical components of the exposome as a repertoire, in Chapter 2, where I have connected the origin of technologies used in exposome research, such as omics, to other areas of life science research, such as genomics and sequencing. In Chapter 3 I have presented an account of data practices in the EXPOsOMICS project, highlighting the specific approaches and techniques used to collect, retrieve and analyse data for the study of the exposome and how assumptions and commitments connected to different techniques are integrated. I have argued that evidential claims are employed at various points of data practices in exposome research in order to specify and identify the datasets used in research. I have used evidential claims as a conceptual tool to provide a distinction between different types and levels of data practices in EXPOsOMICS. I have shown how different epistemic strategies have arisen in this case: the macro strategy, which identifies the initial evidence platform and generates

scoping claims that restrict samples and provide an initial understanding of phenomena; the micro strategy, which collects data of significantly different types to generate evidential claims on structures at the microscopic level; and the association strategy, which uses evidence from the two other strategies to generate evidential claims at the statistical level of associations between exposure to the environment and outcomes of interest. In Chapter 4, I have focused more explicitly on omics technologies used in EXPOsOMICS, as way of analysing views of molecular data as providing mechanistic evidence. On the basis of a close look at the ways in which molecular data is used in exposome research, I have cautioned against interpretations in terms of mechanistic evidence. I have shown that new types of molecular data collected and analysed in EXPOsOMICS do not necessarily yield mechanistic evidence of the kind maintained in current philosophical classifications of evidence.

Evidence is one of the topics that I have discussed the most throughout the dissertation. Chapter 2 is based on the notion of evidential claims, for which I have taken inspiration from recent work in the philosophy of archaeology (Wylie 2017; Chapman and Wylie 2016) and that I have taken to mean implicit or explicit claims that identify a dataset as evidence for something, this being a phenomenon or a claim. I have used this conceptual tool to distinguish between different strategies, that I have emphasised as a crucial step of research, where datasets are given evidential value and representation content, which differ in terms of level of abstraction, type of work and kind of evidential claim and are ways of dealing with the diversity of techniques used for data generation. In Chapter 4, I have analysed approaches to evidence classification in exposome research and compared and contrasted them with philosophical views on evidence classification such as evidential pluralism. I have argued that current philosophical views on evidence classification in the health sciences fall short in the ways in which they have been applied in exposome research. In current philosophical discussions on evidence in the medical sciences, epidemiology has been used to exemplify the view referred to as the Russo-Williamson Thesis (Russo and Williamson 2007, 2012), according to which evidence of both difference-making and mechanisms is produced to make causal claims. I have

addressed the conditions under which data is categorised as evidence in exposome research and argued that the variety of data used as evidence in exposome research suggests a different approach to classifying evidence, indicating that what counts as a type of evidence depends on the ways in which a dataset is used, in contrast with the approach of evidential pluralism, according to which evidence is classified in different types on the basis of its intrinsic properties. In the chapter I have sketched an alternative approach to this view, arguing that evidence is rarely classified as a specific kind that shows distinct features and what counts as evidence, and more precisely as a specific type of evidence, depends on the purpose for which the data is used.

I think that the dissertation could serve as a platform for more in-depth exploration of an account of evidence in biomedical research. In the epidemiological context that I have looked at in the dissertation, evidence appears to be a pragmatic and relational notion, a category that is assigned to specific sets of data depending on the way they are used and for what they are used. This would entail that classifications of evidence depend not only on the data itself but also on the ways in which the data is used and for which claims it is used. The approach would be in contrast with more traditional views of evidence, which tend to be representational (Woodward 2000) and closer to philosophical work on scientific data, by suggesting a relational interpretation of evidence (Leonelli 2015).

Exploring this account would be especially interesting in the health sciences. In this context, data and evidence have both been linked to the future of research, often in a way that suggests that the two terms are equivalent. An approach such as evidence-based medicine is usually presented as a way of optimising and advancing decision-making in medical practice. Much of the recent philosophical work on evidence developed in the context of the health sciences is connected to the rise of evidence-based medicine, which brought about new ways of assessing sources of evidence and tackling evidence amalgamation and provoked many critical reflections in the literature. Within this debate, a number of authors have started to think in more general terms about the kinds of evidence produced in medical research, connecting these discussions to

other issues of philosophy of science, such as explanation, causation and causal inference (Campaner and Galavotti 2012). This line of research has successfully emphasised the need for a wider consideration of medical evidence, by showing the importance of types of evidence that are poorly considered within the evidence-based medicine framework, for instance in the context of causal inference (Clarke et al. 2013). Yet, in this context philosophers have not really engaged with data and its relation with evidence. The fact that the concept of data does not play much of a role in discussions on medical evidence seems odd, as similar promises about revolutionising biomedical research have been connected to the use of data, especially in the context of large quantities and new sources of data, such as omics data or personal health records, and the increasing implementation of data-intensive approaches (e.g., systems medicine, personalised and precision medicine). At the same time, from the point of view of philosophical accounts of data, evidence has been discussed as the result of practices of collecting, ordering and use of data for claims about phenomena (Leonelli 2009; Leonelli 2016a). Here, the emphasis placed on data entails that anything that counts as data and is operationalised in knowledge claims is evidence. A potential follow-up project on these issues could thus look at: the relations between data and evidence and the conditions under which data is given evidential value and content, possibly in connection to evidential claims, as discussed in Chapter 3; and the relations with other elements of the scientific inquiry, including empirical claims, models and knowledge (see e.g. Leonelli 2019.

Another set of issues to which results of my work can be applied to is the conceptualisation of *health and disease* in contemporary life and health sciences. In Chapter 2, I have discussed the innovations of the exposome in the context of other debates on change in these disciplines, including postgenomics. I have used the term postgenomics here with a historical meaning, i.e. as a way to describe research that employs genomic-based technologies, is increasingly aware of the complexity in interpreting genomic results and has a critical engagement with gene-centric approaches. Investigating the relations between data and conceptual changes in the exposome repertoire, I have argued that

current changes in epidemiology should be framed as alignments of various conceptual and material components: scientific change is the result of this alignment, and it is not only due to one of these factors, such as data. I have argued that the exposome should not be analysed only as a notion, at the conceptual level, nor as a methodological approach, nor for its technological consequences – it should rather be considered as the alignment of these and other components, which as a whole constitutes a repertoire. Using this approach, I have considered some conceptual implications of the exposome repertoire, primarily in the context of notions of exposure and environment. Something that could be discussed in more depth are the ways in which health and disease are conceptualised in the exposome repertoire and what consequences this can have on how health and disease should be accounted for and explained. For example, one of the aspects of the exposome that I have often mentioned in the dissertation is the all-encompassing approach of the exposome, which is aimed at capturing all the exposures experienced by individuals throughout their lifetime. In the exposome literature this is connected to an approach referred to as 'course of life approach', whereby disease or health appear as outcomes of interest for analysis in a continuum and sets of states in which the exposome can be in (Robinson and Vrijheid 2015). This seems close to recent discussions in philosophy of science on process-oriented views of organisms, as opposed entity-oriented views (Nicholson and Dupré 2018). In this sense, an avenue for further research could be an exploration of postgenomic approaches to disease and health, on the basis of processual instead of entity-based terms, possibly also in relation to discussions on pathways and mechanisms (Boniolo and Campaner 2018) and disease states as points to be tracked in a continuum (Russo 2017).

The project would imply engaging with theories and accounts of disease, that are a longstanding and traditional debate in philosophy of science and philosophy of medicine. Concepts of disease and health have attracted the attention of philosophers not only for the fundamental and basic role they play in medical research and care, but also because of specific conceptual features. The distinction between what counts as health and disease state is famously vague, as

a consequence of the diversity of human populations and therefore experiences of disease, the (debated) social construction of many conceptualisations of disease, the uncoupling of care and disease and the fact that interpretations of health and disease usually involve both normative and descriptive elements (Reiss and Ankeny 2016). Thus, perhaps unsurprisingly, the philosophical literature on these issues is very extensive (Cooper 2016). It includes accounts based on the distinction between pathological and non-pathological states, the distinction between functional and dysfunctional biological systems, normative accounts that have looked at disease as an instance of harmful and naturally bad states, etc. (Murphy 2015).

The issues discussed in this debate are significant for the epistemology and ethics of health sciences research and health care, as distinctions between health and disease have an impact on what can, should and is studied in medicine. At the same time, however, I think that a follow-up project based on the approach used in this dissertation, looking at the conceptual level as situated in material and social dimensions, would push against some of the limitations of this debate. Firstly, a significant part of the debate does not seem to take into account scientific practice as other areas of philosophy of science do and is instead mostly focused on conceptual analysis. It could be argued that is a sensible choice, especially if the aim is to build a normative account for philosophical audiences, but an analysis of scientific practices around the conceptualisation and operationalisation of disease states could be relevant to the debate, as I have argued is the case when looking at the exposome. In addition, in the debate the perspective of a discipline such as epidemiology is rarely discussed. This seems rather odd, considering that epidemiology is usually defined as the discipline that studies health and disease in populations (Broadbent 2013) and provides significant evidence for policy-making on public health, as we have seen in the dissertation. If, for instance, we compare the health and disease debate with the ways in which disease in epidemiology and exposome research is framed, studied and discussed, there seem to be some disparities. The philosophical debate seems to be mostly focused on 'internalist' perspectives, and the role of environmental factors is usually not taken into account significantly.

However, we have seen that the exposome is pitched to counter gene-centric, internalist and biological approaches to disease because 70-80% of disease risks are due to changes in the environment (Rappaport and Smith 2010). Similar remarks can be extended to a lack of attention given to the social and socio-economic dimension of disease, which rarely seems to appear in the philosophical debate on these issues (see Ghiara and Russo under review).

By connecting my case to these other debates and issues, I hope to show the merits of a philosophical study centred on real-life, scientific practices of data collection, integration and use. In Chapter 1, I have defined this as the line of research in philosophy of science that is concerned with questions about the relations between data and cognitive and material elements of scientific practice and epistemology; focuses on data and data practices broadly construed, with units of analysis including data collection, processing, structuring, ordering and use as evidence; and aims at the documentation, critical engagement and philosophical interpretation of data. The dissertation is a contribution to this line of research, as I have been concerned with a number of issues discussed in data-centric philosophical studies of the sciences, including: innovation and novelty in science, in relation to the epistemic role of data in shaping the conceptual, material and technological context; the production of evidence in material and immaterial contexts, in relation to the ways in which data is considered and used as evidence; and the classification of evidence in different types of evidence, in relation of the assumptions, practices and uses that are involved in evidence classification. In addition, I have connected these discussions with topics of more general interest in philosophy of science, such as: approaches to and conceptualisations of scientific change, in the context of Kuhnian and post-Kuhnian discussions; causality, causal claims and causal knowledge in biomedical research; and evidential reasoning in the sciences, in relation to philosophical analyses of empirical knowledge. I hope that this dissertation contributes to our understanding of data, as an epistemically salient component of the sciences.

Appendices

Appendix A: Ethics Consent Form

Title of Research Project

The role of data and technology in contemporary biomedicine: Integrating evidence to study the relation between disease and the environment

Details of Project

The project aims to study: the theoretical approach to disease developed in EXPOsOMICS; the ways in which the different sources of evidence and data are used and integrated; the ways in which the project was set up and the level of interdisciplinarity involved; and the use and influence of technology on research.

Contact Details

For further information about the research and interview data, please contact:

Stefano Canali, Institute of Philosophy, Leibniz University Hannover, Am Klagesmarkt 14-17, 30159 Hannover, Germany.

Tel: +49 0511 762 14505. Email: stefano.canali@philos.uni-hannover.de.

If you have concerns/questions about the research you would like to discuss with someone else, please contact:

Professor Sabina Leonelli, Co-Director of Exeter Centre for the Study of the Life Sciences, Department of Sociology, Philosophy & Anthropology, Exeter University, Exeter, UK.

Tel: +44 01392 725137. Email: s.leonelli@exeter.ac.uk.

Participation in the study

I will ask you about your everyday research practices, your experiences with data integration and technology, the organisation of the project and its theoretical and methodological aspects. The interview will range from thirty minutes to two hours in duration. I would like to record this interview with your permission.

Data Protection Notice

Your contact details will be kept separated from your interview data. It will be kept for up to 10 years so that, if necessary, I can contact you during my research.

No voice or video data will be published or shared in any way. Interviews will be transcribed in an anonymised form. Anonymised data (transcripts) may be stored indefinitely.

Other than me, the only persons with temporary access to the recordings will be my supervisors and eventual academic co-authors.

The information you provide will be used for research purposes and your personal data will be processed in accordance with current data protection legislation and the University's notification lodged at the Information Commissioner's Office. Your personal data will be treated in the strictest confidence and will not be disclosed to any unauthorised third parties. The results of the research will be published in anonymised form.

Open access

After the interview, I will also ask you whether you are happy for me to make the eventual interview transcript available as open access research data. This is optional and disconnected from the question of anonymity. You can choose to be identifiable and at the same time for the transcripts not to be shared, or conversely, to not be identifiable but for anonymised versions of the transcripts to be made available. In any case, your contact details will be kept separately from your interview data.

Research findings

I will be happy to send you the transcription to review as well as any publications resulting from this study.

Consent

I have been fully informed about the aims and purposes of the project.

I understand that:

- there is no compulsion for me to participate in this research project and, if I do choose to participate, I may withdraw for a month after the interview;
- I can stop the interview at any time;
- I do not need to answer any questions that I do not wish to answer;
- any information which I give will be used solely for the purposes of this research project, which may include publications or academic conference presentations;
- if applicable, the information, which I give, may be shared with the supervisors of the project and potential co-authors in an anonymised form;
- all information I give will be treated as confidential;
- the researcher will make every effort to preserve my anonymity.

I agree to be audio-recorded: Yes ____ No ____

Name of participant: _____

Signature of participant: _____ Date: _____

Signature of researcher: _____ Date: _____

[2 copies to be signed by both interviewee and researcher, one kept by each]

To be completed after the interview, only if desired by the participant:

I _____ give permission to be identified and have my contributions attributed to me. Yes ____ No ____

I agree for the interview transcripts to be edited and shared in an open access format that will make it freely accessible online. Yes ____ No ____

Signature of participant: _____ Date: _____

Signature of researcher: _____ Date: _____

[2 copies to be signed by both interviewee and researcher, one kept by each]

Bibliography

Anderson, C. (2008). The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. *Wired Magazine*, 16(7), June 23. Available at: https://www.wired.com/2008/06/pb-theory/ (accessed January 2019).

Anderson, E. (2017). Feminist Epistemology and Philosophy of Science. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2017 Edition). Available at: https://plato.stanford.edu/entries/feminism-epistemology/ (accessed February 2019).

Ankeny, R. A., & Leonelli, S. (2015). Valuing Data in Postgenomic Biology. How Data Donation and Curation Practices Challenge the Scientific Publication System. In S. S. Richardson & H. Stevens (eds.) *Postgenomics: Perspectives on Biology after the Genome* (pp. 126-149). Durham, NA: Duke University Press.

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.

Ankeny, R. A., & Leonelli, S. (forthcoming). Using Repertoires to Explore Changing Practices in Recent Coral Research. In K. Matlin, J. Maienschein, & R. Ankeny (Eds.), *From the Beach to the Bench: Why Marine Biological Studies?* Chicago, IL: University of Chicago Press.

Ankeny, R. A., Chang, H., Boumans, M. & Boon, M. (2011). Introduction: Philosophy of Science in Practice. *European Journal for the Philosophy of Science* 1, 303-307.

Bailer-Jones, D. (2009). *Scientific Models in Philosophy of Science*. Pittsburgh: Pittsburgh University Press.

Bell, G., Hey, T., & Szalay, A. (2009). Beyond the Data Deluge. *Science*, 323(5919), 1297–1298.

Bertolaso, M., & Sterpetti, F. (2017). Evidence Amalgamation, Plausibility, and Cancer research. *Synthese* (online first), special issue "Evidence Amalgamation in the Sciences" edited by Samuel C. Fletcher, Jürgen Landes and Roland Poellinger.

Bluhm, R. (2005). From Hierarchy to Network: a Richer View of Evidence for Evidence-based Medicine. *Perspectives in Biology and Medicine*, 48(4), 535–547.

Bogen, J., & Woodward, J. (1988). Saving the Phenomena. *The Philosophical Review*, 97(3), 303–352.

Bonassi, S., & Au, W. W. (2002). Biomarkers in Molecular Epidemiology Studies for Health Risk Prediction. *Mutation Research/Reviews in Mutation Research*, 511(1), 73–86.

Boniolo, G., & Nathan, M. J. (2016). Introduction. In G. Boniolo & M.J. Nathan (Eds.) *Philosophy of molecular medicine: Foundational issues in research and practice* (pp. 1-12). London: Routledge.

Boniolo, G., & Campaner, R. (2018). Molecular Pathways and the Contextual Explanation of Molecular Functions. *Biology & Philosophy*, 33(3–4), 24.

Boumans, M., & Leonelli, S. (2013). Introduction: On the Philosophy of Science in Practice. *Journal for General Philosophy of Science*, 44(2), 259–261.

Bowker, G.C. (2005). *Memory Practices in the Sciences*. Cambridge, MA: MIT Press.

Bowker, G.C., & Star, S.L. (1999). *Sorting Things Out. Classification and Its Consequences*. Cambridge, MA: MIT Press.

boyd, D., & Crawford, K. (2012). Critical Questions for Big Data. *Information, Communication & Society*, 15(5), 662–679.

Brigandt, I. (2013). Integration in Biology: Philosophical Perspectives on the Dynamics of Interdisciplinarity. *Studies in History and Philosophy of Science*

Part C: Studies in History and Philosophy of Biological and Biomedical Sciences, 44(4), 461–465.

Broadbent, A. (2009). Causation and Models of Disease in Epidemiology. *Studies in History and Philosophy of Biological and Biomedical Sciences*, *40*(4), 302–311.

Broadbent, A. (2011). Inferring Causation in Epidemiology: Mechanisms, Black Boxes, and Contrasts. In: P. Illari, F. Russo F & J. Williamson (eds.) *Causality in the Sciences*, (pp. 45–69). Oxford: Oxford University Press.

Broadbent, A. (2013). *Philosophy of Epidemiology*. London: Palgrave Macmillan.

Broadbent, A. (2014). Disease as a Theoretical Concept: The Case of "HPV-itis". *Studies in History and Philosophy of Biological and Biomedical Sciences*, *48*, 250–257.

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.

Bschir, K., Lohse, S., & Chang, H. (2018). Introduction: Systematicity, the Nature of Science? *Synthese* (online first), special issue "Systematicity - The Nature of Science" edited by Karim Bschir, Simon Lohse and Hasok Chang.

Burian, R. (1997). Exploratory Experimentation and the Role of Histochemical Techniques in the Work of Jean Brachet, 1938-1952. *History and Philosophy of the Life Sciences*, 19(1), 27-45.

Burian, R. M. (2001). The Dilemma of Case Studies Resolved: The Virtues of Using Case Studies in the History and Philosophy of Science. *Perspectives on Science*, 9(4), 383–404.

Callebaut, W. (2012). Scientific Perspectivism: A Philosopher of Science's Response to the Challenge of Big Data Biology. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, 43(1), 69–80.

Campaner, R. (2011). Causality and Explanation: Issues from Epidemiology. In: D. Dieks, W.J. Gonzalez, S. Hartmann, T. Uebel & M. Weber (eds.) *Explanation, Prediction, and Confirmation* (pp 125–135). Dordrecht: Springer Netherlands.

Campaner, R., & Galavotti, M.C. (2012). Evidence and the Assessment of Causal Relations in the Health Sciences. *International Studies in the Philosophy of Science*, 26(1), 27–45.

Cartwright, N. (2011). A philosopher's view of the long road from RCTs to effectiveness. *The Lancet*, 377(9775), 1400–1401.

Chadeau-Hyam, M., Athersuch, T. J., Keun, H. C., De Iorio, M., Ebbels, T. M. D., Jenab, M., Sacerdote, C. et al. (2011). Meeting-in-the-middle Using Metabolic Profiling – a Strategy for the Identification of Intermediate Biomarkers in Cohort Studies. *Biomarkers*, 16(1), 83–88.

Chadeau-Hyam, M., Campanella, G., Jombart, T., Bottolo, L., Portengen, L., Vineis, P., Liquet, B. et al. (2013). Deciphering the Complex: Methodological Overview of Statistical Models to Derive Omics-based Biomarkers. *Environmental and Molecular Mutagenesis*, *54*(7), 542-557.

Chang, H. (2004). Inventing Temperature: Measurement and Scientific Progress. Oxford: Oxford University Press.

Chang, H. (2011). Integrating History and Philosophy of Science. In S. Mauskopf & T. Schmaltz (Eds.) *Integrating History and Philosophy of Science Problems and Prospects* (pp. 109-124). Dordrecht: Springer Netherlands.

Chapman, R., & Wylie, A. (2016). *Evidential Reasoning in Archaeology*. London: Bloomsbury.

Clarke, B., & Russo, F. (2017). Causation in Medicine. In: J. A. Marcum (Ed.) *The Bloomsbury Companion to Contemporary Philosophy of Medicine* (pp. 297–322). London: Bloomsbury.

Clarke, B., Gillies, D., Illari, P., Russo, F., & Williamson, J. (2013). The Evidence That Evidence-based Medicine Omits. *Preventive Medicine*, 57(6), 745–747.

Cooper, R. (2016). Health and Disease. In J. A. Marcum (Ed.) *The Bloomsbury Companion to Contemporary Philosophy of Medicine* (pp. 275-296). London: Bloomsbury.

Currie, A. (2015). Philosophy of Science and the Curse of the Case Study. In C. Daly (Ed.) *The Palgrave Handbook of Philosophical Methods* (pp. 553–572). London: Palgrave Macmillan UK.

Davies, G., Frow, E. K., & Leonelli, S. (2013). Bigger, Faster, Better? Rhetorics and Practices of Large-scale Research in Contemporary Bioscience. *BioSocieties*, 8(4), 386–396.

Dennis, K. K., Marder, E., Balshaw, D. M., Cui, Y., Lynes, M. A., Patti, G. J., Rappaport, S. M. et al. (2017). Biomonitoring in the Era of the Exposome. *Environmental Health Perspectives*, 125(4), 502–510.

De Vreese, L. (2009). Epidemiology and Causation. *Medicine, Health Care and Philosophy*, 12(3), 345–353.

Douglas, H. E. (2009). *Science, Policy, and the Value-Free Ideal*. Pittsburgh, PA: University of Pittsburgh Press.

Dupré, J. (1996). "Metaphysical Disorder and Scientific Disunity". In Peter Galison & David J. Stump (Eds.) *The Disunity of Science: Boundaries, Context, and Power* (101–117.). Stanford, CA: Stanford University Press.

Espín-Pérez, A., Font-Ribera, L., van Veldhoven, K., Krauskopf, J., Portengen, L., Chadeau-Hyam, M., Vermeulen, R. et al. (2018). Blood Transcriptional and microRNA Responses to Short-term Exposure to Disinfection Byproducts in a Swimming Pool. *Environment International*, *110*, 42-50.

European Commission (2019). The Human Exposome Project: A Toolbox for Assessing and Addressing the Impact of Environment on Health. Available at: https://ec.europa.eu/info/funding-tenders/opportunities/portal/screen/opportunities/topic-details/sc1-bhc-28-2019 (accessed February 2019).

Fiorito, G., Vlaanderen, J., Polidoro, S., Gulliver, J., Galassi, C., Ranzi, A. Krogh, V. et al. (2018). Oxidative Stress and Inflammation Mediate the Effect of Air Pollution on Cardio- and Cerebrovascular Disease: A Prospective Study in Non-smokers. *Environmental and Molecular Mutagenesis*, 59 (3), 234–246.

Fleming, L.E., Tempini, N., Gordon-Brown, H., Nichols, G., Sarran, C., Vineis, P., Leonardi, G. et al. (2017). *Big Data in environment and human health: Challenges and opportunities: Oxford encyclopaedia for environment and human health*. Oxford: Oxford University Press.

Flick, U. (2014). *An Introduction to Qualitative Research* (Fifth Edition). SAGE Publications.

Floridi, L. (2012). Big data and their Epistemological Challenge. *Philosophy and Technology*, 25(4), 435–437.

Forber, P. (2010). Confirmation and Explaining How Possible. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, 41(1), 32–40.

Fuller, J. 2018. Universal Etiology, Multifactorial Diseases and the Constitutive Model of Disease Classification. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *67*, 8–15.

Fuller, J., Broadbent, A., & Flores, L. J. (2015). Prediction in Epidemiology and Medicine. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *54*, 45–48.

Gannett, L. (2016). The Human Genome Project. In: E. N. Zalta (Ed.) *The Stanford Encyclopedia of Philosophy* (Summer 2016 Edition). Available at https://plato.stanford.edu/archives/sum2016/entries/human-genome/(accessed November 2018).

Ghiara, V., & Russo, F. (under review). Reconstructing the Mixed Mechanisms of Health: The Role of Bio- and Socio-markers.

Gibbon, S., Prainsack, B., Hilgartner, S. & Lamoreaux, J. (Eds.). 2018. *Routledge Handbook of Genomics, Health and Society. Second edition*. London: Routledge.

Gillies, D. (2011). The Russo-Williamson Thesis and the Question of Whether Smoking Causes Heart Disease. In: P. Illari, F. Russo F & J. Williamson (eds.) *Causality in the Sciences*, (pp. 110–125). Oxford: Oxford University Press.

Green, S. (2016). What Is the 'post' in Postgenomics? *Metascience*, 25: 83–86.

Green, S., & Vogt, H. (2016). Personalizing Medicine: Disease Prevention in Silico and in Socio. *Humana.Mente*, 30, 105–145.

Gross, F., Kranke, N., & Meunier, R. (2019). Pluralization Through Epistemic Competition: Scientific Change in times of Data-intensive Biology. *History and Philosophy of the Life Sciences*, 41(1), 1.

Gulliver, J., Morley, D., Dunster, C., McCrea, A., van Nunen, E., Tsai, M.-Y., Probst-Hensch, N. et al. (2018). Land Use Regression Models for the Oxidative Potential of Fine Particles (PM 2.5) in five European areas. *Environmental Research*, 160, 247-255.

Guttinger, S. & Dupré, J. Genomics and Postgenomics. In: E. N. Zalta (Ed.) *The Stanford Encyclopedia of Philosophy* (Winter 2016 Edition). Available at https://plato.stanford.edu/archives/win2016/entries/genomics/ (accessed November 2018).

Hacking, I. (1983). Representing and Intervening: Introductory Topics in the Philosophy of Natural Science. Cambridge: Cambridge University Press.

Haraway, D. (1988). Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective. *Feminist Studies*, 14(3), 575.

Hey, T., Tansley, S., & Tolle, K. (2009). *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Redmond, WA: Microsoft Research.

Hilgartner, S. (2017). *Reordering Life: Knowledge and Control in the Genomics Revolution*. Cambridge, MA: MIT Press.

Howick, J. (2011). Exposing the Vanities – and a Qualified Defense – of Mechanistic Reasoning in Health Care Decision Making. *Philosophy of Science*, 78(5), 926–940.

Hoyningen-Huene, P. (1993). *Reconstructing Scientific Revolutions: Thomas S. Kuhn's Philosophy*. Chicago, IL: University of Chicago Press.

Iliadis, A., & Russo, F. (2016). Critical Data Studies: An introduction. *Big Data & Society*, 3(2), 1–7.

Illari, P.M. (2011a). Mechanistic Evidence: Disambiguating the Russo-Williamson Thesis. *International Studies in the Philosophy of Science*, 25(2), 139–157.

Illari, P.M. (2011b). Why Theories of Causality Need Production: An Information Transmission Account. *Philosophy and Technology*, 24(2), 95–114.

Illari, P., & Russo, F. (2016b). Causality and Information. In L. Floridi (Ed.) *The Routledge Handbook of Philosophy of Information*. London: Routledge.

Illari P., & Russo F. (2016) Information Channels and Biomarkers of Disease. *Topoi* 35(1), 175–190.

Illari, P.M., & Williamson, J. (2012). What is a Mechanism? Thinking About Mechanisms *Across* the Sciences. *European Journal for Philosophy of Science*, 2(1), 119–135.

Jeong, A., Fiorito, G., Keski-Rahkonen, P., Imboden, M., Kiss, A., Robinot, N., Gmuender, H. et al. (2018). Perturbation of Metabolic Pathways Mediates the Association of Air Pollutants with Asthma and Cardiovascular Diseases. *Environment International*, 119(July), 334–345.

Juarez, P. D., Matthews-Juarez, P., Hood, D. B., Im, W., Levine, R. S., Kilbourne, B. J., Kilbourne, B. J. et al. (2014). The Public Health Exposome: A Population-based, Exposure Science Approach to Health Disparities Research. *International Journal of Environmental Research and Public Health*, 11(12), 12866–12895.

Karaca, K. (2013). The Strong and Weak Senses of Theory-Ladenness of Experimentation: Theory-Driven versus Exploratory Experiments in the History of High-Energy Particle Physics. *Science in Context*, 26(01), 93–136.

Karaca, K. (2017). A Case Study in Experimental Exploration: Exploratory Data Selection at the Large Hadron Collider. *Synthese*, 194(2), 333–354.

Keller, E.F. (2015). The Postgenomic Genome. In S. S. Richardson & H. Stevens (eds.) *Postgenomics: Perspectives on Biology after the Genome* (pp. 9-31). Durham, NC: Duke University Press.

Kelly, S. E. (2010). Qualitative Interviewing Techniques and Styles. In I.L. Bourgeault, R. DeVries & R. Dingwall (Eds.) *Handbook on Qualitative Health Research* (307-326). Sage Publications.

Kelly, T. (2016). Evidence. In: E.N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (winter 2016 edition). Available at: https://plato.stanford.edu/entries/evidence/ (accessed July 2018).

Kitchin, R. (2013). Big Data and Human Geography: Opportunities, challenges and risks. *Dialogues in Human Geography*, 3(3), 262–267.

Kitchin, R. (2014a). The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences. London: Sage.

Kitchin, R. (2014b). Big Data, New Epistemologies and Paradigm Shifts. *Big Data & Society*, 1(1), 1-12.

Kitchin, R., & McArdle, G. (2016). What Makes Big Data, Big Data? Exploring the Ontological Characteristics of 26 Datasets. *Big Data & Society*, 3(1), 1–10.

Kitchin, R., & Lauriault, T. (2018). Towards Critical Data Studies: Charting and Unpacking Data Assemblages and Their Work. In J. Thatcher, J. Eckert & A. Shears (Eds.) *Thinking Big Data in Geography. New Regimes, New Research* (pp. 3-20). Lincoln, NE: University of Nebraska Press.

Kuhn, T. S. (1962). *The Structure of Scientific Revolutions*. Chicago, IL: University of Chicago Press.

Kuhn, T. S. (1977). The Essential Tension: Selected Studies in Scientific Tradition and Change. Chicago, IL: University of Chicago Press.

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).

Latour, B. & Woolgar, S. (1979). Laboratory Life: The Social Construction of Scientific Facts. Beverly Hills, CA: Sage.

Laudel, G., & Gläser, J. (2007). Interviewing Scientists. *Science, Technology & Innovation Studies*, 3(2), 91–111.

Lee, T.A., & Pickard, A.S. (2013). Exposure Definition and Measurement. In P. Velentgas, N.A. Dreyer & P. Nourjah (eds.) *Developing a Protocol for Observational Comparative Effectiveness Research: A User's Guide*. Rockville, MD: Agency for Healthcare Research and Quality.

Leonelli, S. (2009). On the Locality of Data and Claims about Phenomena. *Philosophy of Science*, 76(5), 737–749.

Leonelli, S. (2010a). Documenting the Emergence of Bio-ontologies: Or, Why Researching Bioinformatics Requires HPSSB. *History and Philosophy of the Life Sciences*, 32(1), 105–25.

Leonelli, S. (2010b). Packaging Data for Re-Use: Databases in Model Organism Biology. In P. Howlett & M. S. Morgan (eds.) *How Well Do Facts Travel? The Dissemination of Reliable Knowledge* (pp. 325–348). Cambridge: Cambridge University Press.

Leonelli, S. (2012a). Introduction: Making Sense of Data-driven Research in the Biological and Biomedical Sciences. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, 43(1), 1–3.

Leonelli, S. (2012b). When Humans Are the Exception: Cross-species Databases at the Inter-face of Biological and Clinical Research. *Social Studies of Science*, 42(2), 214–236.

Leonelli, S. (2013a). Data Interpretation in the Digital Age. *Perspectives on Science*, 22(3), 397–417.

Leonelli, S. (2013b). Why the Current Insistence on Open Access to Scientific Data? Big Data, Knowledge Production, and the Political Economy of Contemporary Biology. *Bulletin of Science, Technology & Society*, 33(1–2), 6–11.

Leonelli, S. (2013c). 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.

Leonelli, S. (2014). What Difference Does Quantity Make? On the Epistemology of Big Data in Biology. *Big Data & Society*, 1(1), 1–11.

Leonelli, S. (2015) What Counts as Scientific Data? A Relational Framework. *Philosophy of Science*, 82(5), 810–821.

Leonelli, S. (2016a) *Data-Centric Biology: A Philosophical Study*. Chicago, IL: University of Chicago Press.

Leonelli, S. (2016b). Locating Ethics in Data Science: Responsibility and Accountability in Global and Distributed Knowledge Production Systems. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2083), 20160122.

Leonelli, S. (2017). Robert Chapman and Alison Wylie // Evidential Reasoning in Archaeology, Reviewed by Sabina Leonelli. In *British Journal for the Philosophy of Science*. Available at: https://bjpsbooks.wordpress.com/2017/08/08/robert-chapman-and-alison-wylie-evidential-reasoning-in-archaeology/ (accessed May 2018).

Leonelli, S. (2019). What Distinguishes Data from Models? *European Journal for Philosophy of Science*, 9(2), 22.

Leonelli, S., & Ankeny, R. A. (2015). Repertoires: How to Transform a Project into a Research Community. *BioScience*, 65(7), 701–708.

Leonelli, S., & Tempini, N. (2018). Where Health and Environment Meet: The Use of Invariant Parameters in Big Data Analysis. *Synthese* (online first), special issue "Philosophy of Epidemiology" edited by Sean Valles and Jonathan Kaplan.

Leonelli, S., Rappert, B., & Davies, G. (2017). Data Shadows. *Science, Technology, & Human Values*, *42*(2), 191–202.

Lioy, P. J. (2010). Exposure Science: A View of the Past and Milestones for the Future. *Environmental Health Perspectives*, 118(8), 1081–1090.

Lloyd, S. & Raikhel, E. (2018). Environmental Epigenetics and Suicide Risk at a Molecular Scale. In S. Gibbon, B. Prainsack, S. Hilgartner & J. Lamoreaux (eds.) *Handbook of Genomics, Health and Society. Second edition* (pp. 254–262). London: Routledge.

Logan, A. C., Prescott, S. L., Haahtela, T., & Katz, D. L. (2018). The Importance of the Exposome and Allostatic Load in the Planetary Health Paradigm. *Journal of Physiological Anthropology*, 37(1), 15.

Lynch, M. (2014). From Normative to Descriptive and Back: Science and Technology Studies and the Practice Turn. In L. Soler, S. Zwart, M. Lynch & V. Israel-Jost (Eds.) *Science After the Practice Turn in the Philosophy, History, and Social Studies of Science* (93–113). New York and London: Routledge.

Maitre, L., de Bont, J., Casas, M., Robinson, O., Aasvang, G. M., Agier, L., Andrušaitytė, S. et al. (2018). Human Early Life Exposome (HELIX) Study: a European Population-based Exposome Cohort. *BMJ Open*, 8(9), e021311.

Maxson Jones, K., Ankeny, R. A., & Cook-Deegan, R. (2018). The Bermuda Triangle: The Pragmatics, Policies, and Principles for Data Sharing in the History of the Human Genome Project. *Journal of the History of Biology*, 51(4), 693–805.

Mayer-Schönberger, V., & Cukier, K. (2013) Big Data: A Revolution that Will Transform How We Live, Work and Think. London: John Murray.

Miller, G. W., & Jones, D. P. (2014). The Nature of Nurture: Refining the Definition of the Exposome. *Toxicological Sciences*, 137(1), 1–2.

Mitchell, S. D., & Dietrich, M. R. (2006). Integration without Unification: An Argument for Pluralism in the Biological Sciences. *The American Naturalist*, 168(S6), S73–S79.

Mitchell, S. D., & Gronenborn, A. M. (2015). After Fifty Years, Why Are Protein X-ray Crystallographers Still in Business? *The British Journal for the Philosophy of Science*, 68(3), 703–723.

Morabia, A. (2005). Epidemiology: an Epistemological Perspective. In: A. Morabia (Ed.) *A history of epidemiologic methods and concepts* (pp. 3-125). Berlin: Springer.

Morgan, M. S. (2012). Case Studies: One Observation or Many? Justification or Discovery? *Philosophy of Science*, 79(5), 667–677.

Morgan, M. S. (2014). Resituating Knowledge: Generic Strategies and Case Studies. *Philosophy of Science*, 81(5), 1012–1024.

Morgan, M. S. (2019). Exemplification and the Use-values of Cases and Case Studies. *Studies in History and Philosophy of Science Part A* (online first).

Morgan, M. S., & Morrison, M. (Eds.). (1999). *Models as Mediators: Perspectives on Natural and Social Science*. Cambridge: Cambridge University Press.

Murphy, D. (2015). Concepts of Disease and Health. In E. N. Zalta (Ed.) *The Stanford Encyclopedia of Philosophy* (Spring 2015 Edition). Available at https://plato.stanford.edu/archives/spr2015/entries/health-disease/ (accessed January 2019).

Müller-Wille, S., & Charmantier, I. (2012). Natural History and Information Overload: the Case of Linnaeus. Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences, 43(1), 4–15.

Nicholson, D. J., & Dupré, J. (eds.) (2018). *Everything Flows. Towards a Processual Philosophy of Biology*. Oxford: Oxford University Press.

O'Malley, M. A., & Soyer, O. S. (2012). The Roles of Integration in Molecular Systems Biology. Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences, 43(1), 58–68.

Parkkinen, V. P., Wallmann, C., Wilde, M., Clarke, B., Illari, P., Kelly, M.P., Norell, C. et al. (2018). *Evaluating Evidence of Mechanisms in Medicine. Principles and Procedures*. SpringerBriefs in Philosophy.

Pietsch, W. (2015a). The Causal Nature of Modeling with Big Data. *Philosophy and Technology*, 29(2), 137–171.

Pietsch, W. (2015b). Aspects of Theory-Ladenness in Data-Intensive Science. *Philosophy of Science*, 82(5), 905-916.

Pietsch, W. (2016). Two Modes of Reasoning with Case Studies. In T. Sauer & R. Scholl (Eds.) *The Philosophy of Historical Case Studies* (pp. 49–67). Boston Studies in the Philosophy and History of Science, Springer.

Pietsch, W. (in preparation). Big Data – The New Science of Complexity.

Prescott, S. L., & Logan, A. C. (2017). Each Meal Matters in the Exposome: Biological and Community Considerations in Fast-food-socioeconomic Associations. *Economics and Human Biology*, 27, 328–335.

Rappaport, S. M. (2011). Implications of the Exposome for Exposure Science. *Journal of Exposure Science & Environmental Epidemiology*, 21(1), 5-9.

Rappaport, S. M., & Smith, M. T. (2010). Environment and Disease Risks. *Science*, 330(6003), 460–461.

Reydon, T. A. C. (2012). How-possibly Explanations as Genuine Explanations and Helpful Heuristics: a Comment on Forber. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, 43, 302–310.

Rheinberger, H. J. (1997). Towards a History of Epistemic Things: Synthesizing Proteins in the Test Tube. Stanford: Stanford University Press.

Rheinberger, H. J. (2010). An Epistemology of the Concrete: Twentieth-Century Histories of Life. Durham, NC: Duke University Press.

Rheinberger, H. J. (2011). Infra-experimentality: From Taces to Data, from Data to Patterning Facts. *History of Science*, 49(3), 337–348.

Rheinberger, H. J., & Müller-Wille, S. (2017). *The Gene: From Genetics to Postgenomics*. Chicago, IL: The University of Chicago Press.

Reiss, J., & Ankeny, R. A. (2016) Philosophy of Medicine. In E. N. Zalta (Ed.) *The Stanford Encyclopedia of Philosophy* (Summer 2016 Edition). Available

at: https://plato.stanford.edu/archives/sum2016/entries/medicine/ (accessed January 2019).

Rieder, G., & Simon, J. (2016). Datatrust: or, the Political Quest for Numerical Evidence and the Epistemologies of Big Data. *Big Data and Society*, 3(1), 1–6.

Robinson, O., & Vrijheid, M. (2015). The Pregnancy Exposome. *Current Environmental Health Reports*, 2(2), 204–213.

Russo, F. (2009). Variational Causal Claims in Epidemiology. *Perspectives in Biology and Medicine*, 52(4), 540–554.

Russo, F. (2017). On the Poietic Character of Technology. *Humana.Mente Journal of Philosophical Studies*, 30, 147–174.

Russo F, & Vineis, P. (2016). Opportunities and Challenges of Molecular Epidemiology. In: G. Boniolo & M. J. Nathan (Eds.) *Philosophy of Molecular Medicine: Foundational Issues in Research and Practice* (pp. 252–281). London: Routledge.

Russo, F., & Williamson, J. (2007). Interpreting Causality in the Health Sciences. *International Studies in the Philosophy of Science*, 21(2), 157–170.

Russo, F., Williamson, J. (2012). EnviroGenomarkers: The Interplay Between Mechanisms and Difference Making in Establishing Causal Claims. *Medicine Studies*, 3(4), 249–262.

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* (pp .192-209). Durham, NC: Duke University Press.

Simon, J. (2015). Distributed Epistemic Responsibility in a Hyperconnected Era. In L. Floridi (Ed.) *The Onlife Manifesto: Being Human in a Hyperconnected Era* (pp. 145-159). Springer International Publishing.

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.

Sismondo, S. (2010). *An Introduction to Science and Technology Studies* (Second Edition). Malden, MA: Wiley.

Soler, L., Zwart, S., Lynch, M., & Israel-Jost, V. (2014). Introduction. In L. Soler, S. Zwart, M. Lynch & V. Israel-Jost (Eds.), *Science after the Practice Turn in the Philosophy, History, and Social Studies of Science* (pp. 1–43). London: Routledge.

Stegenga, J. (2014). Down with the Hierarchies. *Topoi*, 33(2), 313–322.

Stevens, H., & Richardson, S.S. (2015). Approaching Postgenomics. In S. S. Richardson & H. Stevens (eds.) *Postgenomics: Perspectives on Biology after the Genome* (pp. 1-8). Durham, NC: Duke University Press.

Stevens, H., & Richardson, S.S. (2015). Approaching Postgenomics. In S. S. Richardson & H. Stevens (eds.) *Postgenomics: Perspectives on Biology after the Genome* (pp. 1-8). Durham, NC: Duke University Press.

Stingone, J. A., Buck Louis, G. M., Nakayama, S. F., Vermeulen, R. C. H., Kwok, R. K., Cui, Y., Balshaw, D. M. et al. (2017). Toward Greater Implementation of the Exposome Research Paradigm within Environmental Epidemiology. *Annual Review of Public Health*, 38(1), 315–327.

Strasser, B. J. (2011). The Experimenter's Museum: GenBank, Natural History, and the Moral Economies of Biomedicine. *Isis*, 102(1).

Strasser, B.J. (2012). Data-Driven Sciences: From Wonder Cabinets to Electronic Databases. *Studies in History and Philosophy of Biological and Biomedical Sciences* 43 (1): 85–87.

Strimbu, K., Tavel, J. A. (2010). What are Biomarkers? *Current Opinion in HIV and AIDS*, 5(6), 463–466.

Suppes, P. (1976). What is a Scientific Theory? In S. Morgenbesser (Ed.) *Philosophy of Science Today* (pp. 55-67). New York, NY: Basic Books.

van Tongeren, M., & Cherrie, J. W. (2012). An Integrated Approach to the Exposome. *Environmental Health Perspectives*, 120(3), a103–a104.

Thompson, R. Paul & Upshur, Ross (2017). *Philosophy of Medicine: An Introduction*. London: Routledge.

Turner, M. C., Nieuwenhuijsen, M., Anderson, K., Balshaw, D., Cui, Y., Dunton, G., Hoppin, J. A. Koutrakis, P., Jerrett, M. (2017). Assessing the Exposome with External Measures: Commentary on the State of the Science and Research Recommendations. *Annual Review of Public Health*, 38(1), 215–239.

US National Research Council. (2012). *Exposure Science in the 21st Century: a Vision and Strategy*. Washington: National Academies Press.

Vandenbroucke, J.P., Broadbent, A., & Pearce, N. (2016). Causality and Causal Inference in Epidemiology: The Need for a Pluralistic Approach. *International Journal of Epidemiology*, 45 (6), 1776–1786.

Verma, M., Hontecillas, R., Tubau-Juni, N. & Bassaganya-Riera, J. (2018). Challenges in Personalized Nutrition and Health. *Frontiers in Nutrition*, 5, 117.

Vineis, P. (2015). Exposomics: Mathematics meets Biology. *Mutagenesis*, *30*(6), 719–722.

Vineis, P. (2018). What Is the Exposome and How It Can Help Research on Air Pollution. *Emission Control Science and Technology*.

Vineis, P., & Perera, F. (2007). Molecular Epidemiology and Biomarkers in Etiologic Cancer Research: The New in Light of the Old. *Cancer Epidemiology Biomarkers and Prevention*, 16(10), 1954–1965.

Vineis, P., Chadeau-Hyam, M., Gmuender, H., Gulliver, J., Herceg, Z., Kleinjans, J., Kogevinas, M. et al. (2017a). The Exposome in Practice: Design of

the EXPOsOMICS Project. *International Journal of Hygiene and Environmental Health*, 220(2), 142–151.

Vineis, P., Illari, P., & Russo, F. (2017b). Causality in Cancer Research: A Journey Through Models in Molecular Epidemiology and Their Philosophical Interpretation. *Emerging Themes in Epidemiology*, 14(1), 7.

Vrijheid, M. (2014). The Exposome: A New Paradigm to Study the Impact of Environment on Health. *Thorax*, 69(9), 876–878.

Wagenknecht, S., Nersessian, N. J., & Andersen, H. (2015). Empirical Philosophy of Science: Introducing Qualitative Methods into Philosophy of Science. Heidelberg: Springer.

Wang, T. J., Larson, M. G., Vasan, R. S., Cheng, S., Rhee, E. P., McCabe, E., McCabe, E. et al. (2011). Metabolite Profiles and the Risk of Developing Diabetes. *Nature medicine*, 17(4), 448–453.

Wild, C.P. (2005). Complementing the Genome with an "exposome": The Outstanding Challenge of Environmental Exposure Measurement in Molecular Epi-demiology. *Cancer Epidemiology Biomarkers and Prevention*, 14(8), 1847–1850.

Wild, C. P. (2009). Environmental Exposure Measurement in Cancer Epidemiology. *Mutagenesis*, 24(2), 117–125.

Wild, C. P. (2011). Future Research Perspectives on Environment and Health: The Requirement for a More Expansive Concept of Translational Cancer Research. *Environmental Health: A Global Access Science Source*, 10(SUPPL. 1), S15.

Wild, C. P. (2012). The Exposome: from Concept to Utility. *International Journal of Epidemiology*, 41(1), 24–32.

Wilholt, T. (2013). Epistemic Trust in Science. *British Journal for the Philosophy of Science*, 64(2), 233–253.

Wimsatt, W.C. (2007). Re-engineering Philosophy for Limited Beings: Piecewise Approximations to Reality. Cambridge, MA: Harvard University Press.

Wimsatt, W. C. (2013). Articulating Babel: An Approach to Cultural Evolution. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *44*(4), 563–571.

Wimsatt, W. C., & Griesemer, J. R. (2007). Reproduction Entrenchments to Scaffold Culture: The Central Role of Development in Cultural Evolution. In R. Sansom & R. N. Brandon (Eds.), *Integrating Evolution and Development: From Theory to Practice* (pp. 227–324). MIT Press.

Winther, R. G. (2016). The Structure of Scientific Theories. In E. N. Zalta (Ed.) *The Stanford Encyclopedia of Philosophy* (Winter 2016 Edition). Available at: https://plato.stanford.edu/archives/win2016/entries/structure-scientific-theories/ (accessed January 2019).

Woodward, J. (2000). Data, Phenomena, and Reliability. *Philosophy of Science*, 67, S163–S179.

Worrall, J. (2002). What Evidence in Evidence-Based Medicine? *Philosophy of Science*, 69(S3), S316–S330.

Wylie, A. (2000). Questions of Evidence, Legitimacy, and the (Dis)Unity of Science. *American Antiquity*, *65*(02), 227–237.

Wylie, A. (2012). Feminist Philosophy of Science: Standpoint Matters: Presidential Address. *Proceedings and Addresses of the American Philosophical Association*, 84(4), 47–76.

Wylie, A. (2017). How Archaeological Evidence Bites Back. *Science, Technology, & Human Values*, 42(2), 203–225.