Stud. Hist. Phil. Biol. & Biomed. Sci. 39 (2008) 415–428

Contents lists available at ScienceDirect

Stud. Hist. Phil. Biol. & Biomed. Sci.

journal homepage: www.elsevier.com/locate/shpsc

Mining data, gathering variables and recombining information: the flexible architecture of epidemiological studies

Susanne Bauer

Medical Museion, University of Copenhagen, Fredericiagade 18, DK-1310 Copenhagen, Denmark

ARTICLE INFO

Article history:
Received 12 May 2007
Received in revised form 20 May 2008

Keywords:
Risk factor epidemiology
Breast cancer aetiology
Collecting
Data mining
Biobanks
Population registries

ABSTRACT

Since the second half of the twentieth century, biomedical research has made increasing use of epidemiological methods to establish empirical evidence on a population level. This paper is about practices with data in epidemiological research, based on a case study in Denmark. I propose an epistemology of record linkage that invites exploration of epidemiological studies as heterogeneous assemblages. Focusing on data collecting, sampling and linkage, I examine how data organisation and processing become productive beyond the context of their collection. The case study looks at how a local population database established in 1976 to investigate possibilities for the prevention of cardiovascular disease is used thirty years later to test hypotheses on the aetiology of breast cancer. For two breast cancer investigations based on the same core data set, I follow the underlying record linkage practice and describe how research objects such as molecular markers become relevant with respect to public health through information networking. Epidemiological association studies function as tools that performatively enrol different contexts into statistical risk estimation, thereby configuring options for research as well as for clinical testing and public health policy.

When citing this paper, please use the full journal title *Studies in History and Philosophy of Biological and Biomedical Sciences*

1. Introduction

Biomedical research in the late twentieth century has made increasing use of epidemiological methods to establish empirical evidence about health risks on a population level. This paper is about data collecting and about the data work that underlies epidemiological assessments. It describes a local database established in the 1970s to investigate prevention possibilities for cardiovascular disease and how it came to inform the molecular aetiology of breast cancer three decades later.

A threefold increase in breast cancer risk associated with a mutation in the 'CHEK2 gene' was reported on 31 July 2006 by a research group from Copenhagen.[1] Local media accounts embraced the study as confirmation of a 'third breast cancer gene' and announced significant implications for treatment and testing.[2] The findings from epidemiological research resulted in a broad media response that echoed biomedical promises. In an earlier breast cancer study, the same local data had been used to investigate the aetiological contribution to breast cancer of the environmental exposure to organochlorines (Høyer et al., 1998); a reanalysis a few years later additionally took a molecular marker (p53) into account (Høyer et al., 2002). Upon first publication of the data in 1998, the association of organochlorines and breast cancer was broadly discussed in the local media with

E-mail address: sbau@sund.ku.dk

[1] See Weischer et al. (2007, published electronically on 31 July 2006). The CHEK2 gene has been discussed in relation to cancer risk since 1999. See Bell et al. (1999) for the first description of CHEK2 (then chk2) in people with Li Fraumeni syndrome. More detail on the scientific debate is provided below.

[2] After the scientific publication in the *Journal of Clinical Oncology*, the research result immediately made its way into Danish media: on 1 August 2006, among other local newspapers such as *Jyllands-Posten* and *The Copenhagen Post*, *Berlingske Tidende* reported that a new 'breast cancer gene' had been 'discovered' by Danish researchers. The principal investigator was quoted on the promise of 'a real possibility for prevention' with 'a simple test which now takes only one day' (Nielsen, 2006).

1369-8486/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.shpsc.2008.09.008

respect to food safety and regulation of pesticides (Bech-Danielsen, 1998).

A closer look into the original scientific papers reveals a complex architecture[3] behind seemingly straightforward statements on breast cancer: the data networks of Danish population registries, a long-term follow-up study on cardiovascular risk factors and, as part of the latter, a local biobank. In tracing the 'assemblages'[4] of epidemiological data and samples, this article explores how data work is instrumental in bringing molecular markers, originally products of laboratory research, into clinical and public health relevance. Focusing on record linkage, this is an attempt to contextualise aetiologic claims pertaining to molecular markers—such as CHEK2 or p53—within the history of a population study. Put differently, this paper traces back linkage procedures of two studies on breast cancer aetiology both based on the same local database and biobank to the context of data collection, that is the Copenhagen City Heart Study (CCHS), a prospective epidemiological study initiated in 1975. Such large-scale inventories of the general population to study cardiovascular disease were set up in Scandinavia during the 1960s and 1970s and have been instrumental to the rise of preventive medicine. In this context, the quantitative techniques of epidemiology have developed into the basic science of public health. This paper asks for the specific contribution of epidemiological techniques that use inferential statistics[5] and multivariate risk modelling; it tracks the datawork of two aetiological studies of breast cancer that heavily build on existing data and sample repositories.

'Modern epidemiology' has been described as 'population health aetiology' (and even as 'science of causation') by Alfredo Morabia (2005, p. 4). No longer focusing primarily on infectious disease 'epidemics', but broadly reformulated as a set of methods and designs for population studies, epidemiological techniques were applied to cardiovascular disease and cancer in the twentieth century.[6] The universe of chance and probabilities of today's numerical epistemologies in epidemiology and preventive medicine can be traced back to demographic statistics (Porter, 1995) and, more generally, to the nineteenth century rise of 'mechanical objectivity' (Daston & Galison, 1992). In Ian Hacking's terms, the quantitative statistical inventory of society in the nineteenth century brought about an 'avalanche of numbers, the erosion of determinism and the invention of normalism' (Hacking, 1990, p. 5). A number of studies document how numerical methods and mathematical statistics transformed medical research in the nineteenth and early-twentieth century (Matthews, 1995; Magnello, 2002; Magnello & Hardy, 2005).

Recently, historians have also begun to address the post World War II history of epidemiology, that is the generation of evidence in therapeutic research or the risk factor concept in preventive medicine (Marks, 1997; Aronowitz, 1998; Berridge, 2005; Weisz, 2005). Luc Berlivet (1999, 2005a) has reviewed the understanding of objectivity relative to health risks and the controversy on whether epidemiological results could prove 'causation' or just 'association' in the context of smoking and lung cancer in the 1950s, a topic that has also been covered by Mark Parascandola (2005), in particular in relation to the research conducted at the National Cancer Institute in the US.[7] Addressing the issue of causality[8], cancer epidemiologist Paolo Vineis (1997) has compared the predominant 'empiricist' tradition of observational medicine to the 'realist' tradition for instance in physics. He holds that epidemiology, as an 'intermediate science' should combine both approaches and attribute more value to mechanistic considerations instead of exclusively relying on formal statistical concepts of evidence. Olga Amsterdamska (2005) describes how epidemiologists spelled out the scientificity of their non-laboratory discipline against other biomedical sciences. Exploring epidemiology in a broader cultural framework, James Trostle (2005) stresses the aspect of social exchange in data collection and called for more collaboration between epidemiologists and medical anthropologists. Focusing on the history of epidemiological methods, a recent anthology also presents aspects of research practice, such as the development of study designs (Morabia, 2005).

While many of the contributions to the most recent history of the discipline were written by epidemiologists, few science studies scholars have focused on epidemiological knowledge production of population studies. This paper looks at the role of data collecting, storing and record linkage in epidemiologic research. Case studies from the Nordic countries are of particular interest here due to their central population registries with the capacity to link individual records. As an information resource on a population level, these nationwide registries have been instrumental to epidemiologic research in Denmark.

Since its start in 1976, the Copenhagen City Heart Study[9]—locally referred to as the Østerbro study—has been used to address a wide range of research questions from prevention, aetiology, treatment evaluation to health economy. Most of the studied topics were related to cardiovascular disease, but research soon extended to other outcomes. Among the diverse studies based on the Østerbro dataset,[10] this paper will focus in more detail on the re-use of these data for two lines of investigation on breast cancer aetiology: the first example deals with a study on exposure to industrial chemicals as an aetiological component for breast cancer (Høyer et al., 2002).[11] The second example relates to the investigation of CHEK2 as a 'candidate gene' for breast cancer (Weischer et al., 2007). Whereas the latter study is targeted at investigating a susceptibility gene, Høyer et al. (2002) take on an environmental health framework and focus on the influence of organochlorine exposure[12] on breast cancer under consideration of the molecular marker p53. I will argue that, based on the same data but each in a different research setup, these studies have brought forward different articulations within the discursive space of genomic epidemiology.

[3] With 'architecture' I take up an actor's category, used to discuss the set up of epidemiological research. See for example *Clinical epidemiology: The architecture of clinical research* (Feinstein, 1985).

[4] The use of the term 'assemblage' is inspired by Rabinow (2003) and Deleuze and Guattari (1988). For the concept of 'assemblage' with respect to the study of biotechnology, see Rabinow (2003, p. 56).

[5] In using inferential statistics, epidemiology draws on the biometrics tradition: the techniques of contingency tables go back to Karl Pearson; the concept of randomisation was introduced by Ronald A. Fisher in the 1920s (Fisher, 1925; Porter, 1995; Desrosières, 1998).

[6] On the shift from the bacterial agent of infectious disease to more holistic concepts as to chronic disease and multifactorial causation, see Mendelsohn (1998).

[7] Most of the research on recent epidemiology has concentrated on North America and the UK; see Berlivet (2005b) for the development of epidemiology in France in the second half of the twentieth century.

[8] It is beyond the scope of this paper to cover the concepts of causation; pragmatic criteria for causation have been formulated starting from Austin Bradford Hill (1937, 1965), to Mervyn Susser (1991), Nathaniel Rothman and Sander Greenland (1998), Paolo Vineis (1997, 2005) and many others.

[9] For reports on the Copenhagen City Heart Study's examination series, see Appleyard (1989) and Schnohr et al. (2001).

[10] The main areas of research conducted with these data that were originally compiled for the study of cardiovascular prevention include ageing, cancer, diabetes, epilepsy, genetics, lifestyle, obesity, psychosocial factors, respiratory diseases, and stroke.

[11] Høyer et al. published a series of articles investigating breast cancer and organochlorine exposures in biomedical journals between 1998 and 2002 (Høyer et al., 1998, 2000a, 2000b, 2000c, 2001, 2002).

[12] The summary term 'organochlorines' refers to a large number of chloro-organic compounds, of which ~15,000 were in wide use as pesticides and industrial chemicals (including DDT, dieldrin, PCB).

Science studies scholars have described the pathways of molecules from the laboratory to the clinic (Chadarevian, 1998; Löwy, 1998) and mapped out 'drug trajectories' (Gaudillière, 2005) in biomedical research. Similar to the science studies approaches that follow material objects and things through history (Heesen & Spary, 2001; Daston, 2005), this paper follows 'digital things' (Pálsson, 2005) together with the material samples collected for epidemiology. It examines the data networks of epidemiological studies as heterogeneous digital assemblages.[13] To follow the different data trajectories across context, I draw primarily on published documents—in this case rich resources in themselves, since they comprise detailed descriptions of the study set up as well as full tabulation of the data set; these were published by the Copenhagen City Heart Study office in 1989 to meet the needs of secondary data users. These published sources have been complemented with archival documents, media accounts as well as expert interviews.

The sections of this paper expose various practices with data: the first section introduces the contested field of breast cancer aetiology as a background to the actual case study (2), followed by more conceptual considerations on data assemblages in epidemiology (3). The second part zooms into the data collecting practices in the Østerbro study (4) and into related data mining through central registries (5). The third part (6) focuses on the recombination of these data in genomic epidemiology, tracing how genes and environment as breast cancer determinants are renegotiated in molecular epidemiology. The concluding paragraph (7) revisits the effects of these practices as to the configuration of knowledge in epidemiology; here, I will argue that following the pathways of digital data over time and across context can shed light on the ways epidemiological techniques co-shape biomedical knowledge.

2. Breast cancer epidemiology: sorting out genetic and environmental determinants

The aetiology of breast cancer has been subject to sustained controversy across many disciplines of the life sciences. In epidemiology, carcinogenesis is conceived as a multifactorial stochastic process with a plethora of 'probabilistic determinants'[14] at different levels; the variables of aetiologic studies can range from genetic and molecular to lifestyle, environmental and socio-economic factors.[15] It was in particular the Framingham Heart Study[16] that set the stage for 'risk factor epidemiology', a framework which became widely adopted in epidemiology as the 'risk factor approach' (Aronowitz, 1998). The concept of multiple causation[17] was methodologically developed in multivariate statistical modelling and remained largely unchallenged in most domains of epidemiology; the risk factor approach was perceived as a neutral framework for the study of chronic diseases. While developed within a biomedical concept of disease causation,[18] its success can be attributed to its capacity to provide a conceptual framework that also allowed inclusion of environmental and socio-economic variables as risk factors. As Robert Aronowitz (ibid., p. 144) put it, it is the openness and a 'purposeful ambiguity' in the risk factor concept that made it uniting and attractive to many disciplines and agendas, whilst representing precision, specificity, quantification, and individualism. In the 1990s, aetiological concepts themselves came under scrutiny, for example in explicit controversies over breast cancer; social epidemiologist Nancy Krieger (1994) has spelled out biomedical and socio-political models of causation as conflicting explanatory frameworks. More generally, the question of whether epidemiologists should adhere to the social sciences or to molecular biology has been subject to a vivid debate.[19]

These different conceptual traditions that played out in the epidemiological investigation of breast cancer in the later twentieth century are also reflected in the studies conducted with the Østerbro data in Copenhagen. In what follows I briefly explicate two such research lines in the study of breast cancer, relevant for my case study here: first the research conducted in an environmental health framework, with its focus on organochlorine exposure, is introduced; this is followed by a description of the epidemiological study of candidate genes for breast cancer (here a CHEK2 variant).

It has been pointed out that the few established risk factors account for only about half of breast cancer risk (Snedeker, 2001, p. 35); therefore epidemiologists have repeatedly called for more studies on the role of modifiable environmental and social conditions in breast cancer aetiology. One major line of environmental breast cancer research has focused on organochlorines, which are understood as hormonally active when incorporated into the human body. The assessment of organochlorine exposures is considered a complex and difficult task, given the many congeners of organochlorines and metabolite compounds with different degrees of toxicity. The difficulty is that often only summary indicators of the overall exposures can be measured and therefore the interpretation of results is not straightforward, in particular when there are no exposure histories but data on biomarkers only (Snedeker, 2001). Epidemiological studies on the health effects of organochlorines have been conducted for more than two decades with inconsistent results (Brody & Rudel, 2003).[20] While it would be worthwhile to examine each of these studies 'in the making' as well as in their socio-cultural aspects, this paper focuses in particular on how the architecture of long-term epidemiological studies is set up through work with data, as exemplified by the data linkage and recombination for a specific case study.

Høyer et al. pursued an environmental health framework when they made use of the dataset of the Østerbro study and stored serum samples: their studies report increases in breast cancer risk with dieldrin exposures (but not DDT or DDE) and decreases in survival time of breast cancer patients (Høyer et al., 1998, 2000a). In particular since the 1990s with the emerging sequencing technol-

[13] Drawing on his studies on the Icelandic genome project, Gísli Pálsson describes 'digital assemblies', the byprodcuts of experimental biomedicine, as 'machines' to drill into the past and to generate novel connections (Pálsson, 2005, p. 250).

[14] The notion of 'probabilistic determinants' is often used alternatively to 'risk factors' in epidemiology; it relates to the framework of multiple causation and multivariate statistics, in which potential determinants of disease are tested within a population dataset. The probabilistic understanding of the term 'determinants' is key to 'modern epidemiology', defined as the 'study of the distributions and determinants [...] in populations' (Last,1983, pp. 32–33).

[15] For a summary overview on the field of cancer epidemiology, see for example Henderson et al. (1996).

[16] The Framingham Study is a large-scale epidemiological study begun in 1943; the study established what became known as cardiovascular risk factors (Kannel, 2000).

[17] Conceptualised as 'web of causation' (McMahon et al., 1960, p. 18), the modelling process can integrate knowledge from multiple levels of investigations and disciplines, such as molecular sciences, mechanistic models, environmental and socio-economic variables. Among other criteria, 'biological plausibility' makes up an important causality criterion in the interpretation of associations detected by epidemiologic studies (Hill, 1965).

[18] The biomedical proponents of the risk factor approach in the context of the Framingham Study used epidemiological methods to broaden pragmatically the traditional focus on disease mechanisms; thus the risk factor approach served as an extension of the 'ontological model of disease' in medicine (Aronowitz, 1998, p. 125).

[19] See for example Vineis (2005) and the debate on molecular and social epidemiology in 1999 in the *International Journal of Epidemiology*, *28*(5), in particular: 'Should the epidemiologist be a social scientist or a molecular biologist?' (Susser, 1999). Critical of 'reductionist biomedical concepts' and endorsing social epidemiology frameworks, Nancy Krieger proposed an 'ecosocial concept' as a model that considers structural factors as determinants in processes of 'embodiment' in a 'co-mingled social and biological world'. (Krieger, 2001, p. 674).

[20] Among the first studies in this field were the studies by Unger and Olsen (1980) and Unger et al. (1982, 1984), who conducted organochlorine measurements in tissue and blood; results of subsequent epidemiological studies remained inconclusive.

ogies, research into gene–environment interaction and into genetic polymorphisms as markers of susceptibility have gained prominence. The promise attributed to biomarker studies in this context is to clarify inconsistent results; negative studies often state the need to better account for individual susceptibility and underlying biological mechanisms (Krieger et al., 1994). It is in this sense that, again based on the CCHS dataset, the latest study by Høyer et al. (2002) investigated the influence of p53 as a genomic marker for the association between environmental exposure and breast cancer. Possible changes in epidemiological risk assessments due to the inclusion of biomarkers can implicate revisions in toxicity classification and, as a consequence, in the regulation of chemicals.

In the aftermath of the Human Genome Project, research into genetic determinants of breast cancer has become a highly funded domain. Clinical applications were envisioned and developed; in Denmark, BRCA1 and BRCA2 testing entered clinical practice in 1997 and is offered to breast cancer patients and ovarian cancer patients (Ejlertsen et al., 2007).[21] However, testing for breast cancer genes has remained controversial in view of limited predictive capacities[22] and related ethical considerations. Furthermore, only 25% of the familial aggregation has been found to be due to BRCA1 and BRCA2 (Antoniou & Easton, 2006), which drew the attention of genetic epidemiologists to other candidate genes for breast cancer, among them the CHEK2 polymorphism (Ahmed & Rahman, 2006; Nevannlina & Bartek, 2006). As shown in molecular genetics studies and described first in Li Fraumeni patients (Bell et al., 1999), CHEK2, encoding checkpoint kinase, is attributed an important role in DNA damage repair, that is preventing cellular entry into mitosis (CHEK2-Breast Cancer Consortium, 2002, p. 55). Due to its role in DNA repair, this polymorphism was soon studied in relation to cancer; a large international consortium was initiated to study the association of a specific mutation (CHEK2*1100delC) with cancer risk in a pooled study.[23] Nevanlinna & Bartek (2006) refer to CHEK2 as a 'candidate tumour suppressor' gene that contributes both to sporadic and hereditary breast cancer. Contrary to the stated 'simplicity' of the genetic test itself, the interpretation of so-called 'low penetrance' alleles is difficult; it may play some role in carcinogenesis but in complex interaction with many other factors.

Given the conceptual stakes as to genes, environment and their interactions, it is precisely epidemiology and biostatistical methods that are evoked to sort out multiple potential determinants by drawing on the risk factor approach. Delineated as a formal set of study designs and biostatistical methods, epidemiological techniques have become the tools to evaluate hypotheses across disciplines in the health sciences. Yet the specific setup and data networking practices depend on many aspects, from data infrastructures, cultures of research and collaboration between disciplines to the broader societal conditions. While in the age of genomics, breast cancer genetics, susceptibility and individual lifestyle research have become major research areas in contemporary biomedicine, environmental health research has been considered a rather marginalised topic.[24] Brown et al. (2006) evoke a Kuhnian model of paradigm shift in order to argue for environmental breast cancer research as an innovative yet under-funded approach. Going beyond a conceptual account of conflicting aetiologies and political research agendas, my take on epidemiology here differs in that it aims at exposing the role of data networks and research repositories. Thus before tracing back the data assemblages of the two breast cancer studies by Høyer et al. (2002) and Weischer et al. (2007) to the Østerbro study, I develop a more conceptual approach to databases and biobanks. In doing so, I draw inspiration from studies of material practices,[25] yet here these approaches are extended to following 'data trajectories' in order to explore the epistemology of record linkage in population studies.

3. On databases and biobanks: research collections as resources for knowledge production

Databases and biobanks are the research collections of much of the contemporary biomedical sciences, including epidemiology. Epidemiology is about collecting and storing information—notably, the discipline is commonly defined in terms of statistical 'distributions and determinants'.[26] Data collection can entail gathering new information through specific questionnaires and targeted screenings or making use of already existing data, for example from population registries. Through data linkage between different levels of investigation—from molecular to economic—epidemiological studies perform statistical investigations at a population level. In this sense, I examine epidemiological studies as complex biopolitical assemblages, where samples, data and techniques from different contexts are temporarily brought together in particular configurations. I use the term 'assemblage' to point out the heterogeneous, temporal and at times arbitrary character of these configurations that may change with novel measurement techniques, the constellations of access to data and samples, and funding opportunities. From such assemblages, the specific problematisations, experimental systems[27] and regulated research platforms[28] of epidemiology emerge. The further development of such 'quasi-experimental'[29] research systems in observational epidemiology can be followed, by tracing the different procedures of record linkage.

[21] It is beyond the scope of this paper to discuss the different regulatory regimes and practices of cancer susceptibility testing in clinical settings. For studies on the sociopolitical shaping of clinical practices of susceptibility testing in France and in the US and the UK, see Bourret (2005) and Parthasarathy (2005), respectively.

[22] Among those tested positively for BRCA1 and BRCA2, 46% and 43% respectively developed breast cancer before the age of 70 in a large sample in the US (Chen et al., 2006; Antoniou et al., 2006).

[23] For case series (families, individuals and controls) from clinical genetics centres in the UK, the Netherlands, North America and Germany, this association was demonstrated among non-carriers of BRCA1 or BRCA2 mutations (The CHEK2-Breast Cancer Consortium, 2002); other studies supported the model of CHEK2 as a susceptibility gene for breast cancer as well (Vahteristo et al., 2002). Subsequently Oldenburg et al. (2003) studied 'multiple case early onset breast cancer families' from the Netherlands in which these mutations were excluded and found that carrier patients developed breast cancer earlier than noncarriers; they concluded that CHEK2 may interact with unknown other genes to increase breast cancer risk. In a pooled analysis of ten case-control studies, the CHEK2-Breast Cancer Case-Control Consortium (2004) reported an increased risk of breast cancer in women unselected for family history. The Danish study (Weischer et al., 2007) added knowledge at the level of a general population.

[24] It has also been argued that this is due to the fact that ubiquitous chemicals are most difficult to study in epidemiology due to the lack of unexposed comparison groups (Brown et al., 2005, p. 515). See Snedeker (2001) for a review of epidemiological research on pesticides and breast cancer. A PubMed search (9 November 2007) showed 24,276 hits for 'gene & cancer & breast', while 'environment* & cancer & breast' retrieved 3,593 hits.

[25] Here, I refer to the concept of 'following the thing in itself' (Appadurai, 1988) and the notion of 'material agency' and 'performativity', as described by Andrew Pickering (1995) in *The mangle of practice*. See Latour (1988, 1993) and Rheinberger (1997) for research into the material practices in the life sciences and epistemic things, and Daston (2005) for 'thing studies' within history of science and history of art.

[26] 'Modern epidemiology' has been defined in statistical terms, i.e. as the 'study of the distributions and determinants of health-related states and events in populations, and the application of this study to control of health problems' (Last, 1983, pp. 32–33).

[27] See Rheinberger (1997) for the notion of 'experimental systems'.

[28] For the notion of 'biomedical platforms', see Keating and Cambrosio (2000, 2003).

[29] Referring to study designs that guide data collection and analysis, epidemiologists distinguish between 'experimental' studies (clinical trials) and, 'quasi-experimental' or 'semi-experimental' studies (observational studies) (Hill, 1971, p. 320). For an analysis of clinical trials and clinical experimentation in medical research, in particular clinical epidemiology, see also *The progress of experiment* (Marks, 1997).

When discussing the epistemological status of data and sample repositories in epidemiology, it is helpful to frame these practices within a more general history of collecting in the life sciences. While many historiographies of the life sciences view natural history collecting as an early feature that became subsequently replaced by experimental approaches, Bruno Strasser (2006) reminds us that natural history modes of collecting as a bioscientific practice go together with the rise of experimentalism, rather than being disrupted by the latter. John Pickstone (2007, p. 513) emphasises the significance of these 'other working knowledges [...] before and besides [synthetic experiments]',[30] such as information banks, mathematical models, interactive simulations, or electronic catalogues of molecular biology. While in its classic natural history tradition, collecting involves taking things away from their original context and moving them to a central location or to another context which enhances their value (Peirce, 1995), collecting can also imply the production of representations and the generation of data by taking measurements or by describing observations (Parry, 2004). The very practices with scientific collections constitute specific ways of ordering and knowing the world (Heesen & Spary, 2001). In material archives organised in a particular way, collections convey knowledge through their arrangement and classification. Overviews achieved through displays of arranged collections are closely related to the emergence of statistical tabulation: storing specimens in a certain mode allows overviews at a glance while performing classification, a similar function can be attributed to statistical techniques of data categorisation and analyses.

In health statistics, collecting, measuring and correlating population data can be traced to the beginnings of mortality statistics and classifications of causes of death,[31] to censuses, insurance statistics and actuary risk calculation practices (Porter, 2000). Entrenched in governmental and administrative regimes, early vital statistics data sets were part of state practices that co-determined their collection, reporting and use. Their categories and routines of classification brought about certain modes of knowing which continue to be at work in recent biomedicine. Similar to the management of material samples in a biobank[32], data are arranged in the database in a specific format with particular variable descriptions and definitions.

For the case of a genomics tool, Christine Hine describes databases as 'emergent structures' (Hine, 2006, p. 269)[33] within a set of work practices—as both embodying and being embedded in orders of knowledge. With respect to collecting and databasing in epidemiological research, I am interested in how such emergent structures are brought forward by data and sample arrangement, by linkage opportunities and data processing practices. Yet, different from Hine's study on the development of a more 'instrumental' database tool over its lifespan, this paper deals with epidemiological data repositories and their temporal re-assemblages of data from different domains. Whereas databases store recorded 'inscriptions'[34] as digital data, biobanks contain the 'raw' material (from which more and novel inscriptions can be gained in the future) in a yet unknown potentiality. In particular since the 1990s, many epidemiological studies in Denmark have retained blood samples in biobanks; so did the Østerbro study. In 2006, the CCHS biobank was praised as a 'once again invaluable' resource, unforeseeably forming 'the basis for a new Danish research result, which first in the world can tell us something about the significance of the third known breast cancer gene, the CHEK2 gene, for the general population' (Kamph, 2006). What renders the biobanks of epidemiological studies into invaluable resources is that they are embedded in a defined sample with a detailed information infrastructure and high linkage capacity.

In aetiological epidemiology data and sample repositories are used to investigate connections between exposure and outcome. Outcome research, risk estimation and statistical hypothesis testing are enabled through analytical designs in which empirical data are brought together, for example in 'cohort studies'[35] or 'case-control studies'[36]. While registry-based cohort studies were the most frequently used study design in the Scandinavian context, the case-control design has increasingly been used since the 1960s: different from 'following entire cohorts', this type of study compares disease cases to a set of controls. Viewed as a means to optimise costs in expensive cohort studies, the case-control study became the most common design in late-twentieth century risk factor epidemiology (Paneth et al., 2005). Subsequently, the case-control technique has been applied to many cohort datasets—often using so-called 'nested' or 'embedded' designs within an existing study—this technique retrospectively enrols additional data, for example biomarkers for a subset of cases and controls in a 'cohort-nested case-control study' (Høyer et al., 2002).

In order to further explore how population databasing and biobanking functions epistemologically for this case study, the next section turns to the set up of the epidemiological database that underlies the two aetiological studies on breast cancer. As is common to the studies of material practices, I follow the digital data as inscriptions from a questionnaire or a biobank—starting from the very data recording. The Østerbro study is examined as a project of collecting that generates an inventory of life habits of a general population and, as the following sections will show, subsequently makes use of the Danish population registries' system as an infrastructure for data mining. Thus, after going back to the examination series of the Østerbro study, I will follow the data

[30] Pickstone distinguishes between four *ways of working and knowing*: reading and rhetorics, natural history and craft, analysis and rationalisation, and synthetic experimentation and systematic intervention, which he sees at work in different historical periods, albeit often simultaneously present in changing constellations and varying significance. Pickstone suggests these ways of working and knowing as analytical categories in the history of STM (Pickstone, 2000, 2007).

[31] Similar to several other European countries, 'vital statistics' in Denmark can be traced back to seventeenth century parish registries. Until present these data are used in historical demography (for Denmark, see e.g. Johansen, 2003) as well as in population genetics.

[32] I use the term 'biobank' in the sense of storage of any biological material with linkage to individual data. This corresponds to the definition used in the legislation of many countries, including the Danish regulations on biobanks (Loiborg et al., 2002).

[33] In her study of a database tool for mouse genomics, Hine follows the database from development through to its fading into irrelevance (Hine, 2006).

[34] With the notion of 'inscriptions' for the signals produced in the laboratory by 'inscription devices', I refer to Latour and Woolgar (1979) and Latour (1987). For contemporary epidemiology, the term 'inscriptions' can be extended beyond the laboratory to the digital sphere; 'inscription devices' then include questionnaires and registration routines of the health care system and of governments.

[35] In demography, the term 'cohort' denotes a group of persons with a common statistical characteristic. In epidemiological 'cohort studies', different exposure groups of otherwise 'homogeneous populations' are compared. Last (1983), p. 20, gave a broad definition of the term 'cohort study' (also known as concurrent, follow-up, longitudinal or prospective study): 'The method of epidemiologic study in which subsets of a defined population can be identified who are, have been or in the future may be exposed or not exposed, or exposed in different degrees, to a factor or factors hypothesized to influence the probability of occurrence of a given disease or other outcome'.

[36] In the 'case-control' design the study 'starts with the identification of persons with the disease (or other outcome variable) of interest, and a suitable control (comparison, reference) group of persons without the disease' (Last, 1983, p. 15), in order to compile odds ratios as measures of association, which for large numbers correspond to the relative risk in cohort studies. Janet Lane-Claypon's 1926 examination of breast cancer is generally referred to as the first modern case-control study. This design was used in particular in the 1950s in studies on smoking and lung cancer in the US and Britain (Paneth et al., 2005). These broadly applicable research designs in turn stimulated the development of multivariate techniques and statistical software convenient to epidemiologists. A set of mathematical modelling techniques were established and promoted in textbooks, for instance the logistic function to calculate the 'odds ratios' as risk estimates from case-control studies. See Breslow and Day (1980, 1987).

trajectories over time, until they contribute to epidemiologic statements on the role of environment and genes in breast cancer aetiology.

4. Epidemiology as a practice of collecting: the Copenhagen City Heart Study

In the second half of the twentieth century, large prospective cohort studies on cardiovascular disease were established in many European countries.[37] It was in the context of an emerging epidemiology of chronic disease that local studies of a new type—'follow-up studies' or 'cohort studies'—were initiated in Denmark. Large cohort studies on cardiovascular diseases and 'lifestyle risk' were started in several regions of Scandinavia in the 1960s and 1970s. These were inspired by the Framingham Study, set up in 1948 in the city of Framingham, Massachusetts, which had coined the term 'risk factor' (Kannel, 2000) and became a prototype for cardiovascular epidemiology.[38]

In comparative surveys—begun in 1956 as part of the Seven Countries Study—'lifestyle' (understood as smoking, physical activity and dietary habits) had been established as a risk factor for cardiovascular disease, yet the international comparison suggested considerable regional variation in cardiovascular risk (Mariotti et al., 1982; Keys et al., 1984; Tunstall-Pedoe, 2003). This prompted epidemiologists to initiate more comprehensive local studies to investigate these regional differences. As for the Nordic countries, a cardiovascular risk survey had been carried out in North Karelia (Finland) in the 1950s as part of the Seven Countries Study; early observational studies in Gothenburg (Sweden) and Rejkjavik (Iceland) were established in 1963 and 1967 respectively (Tunstall-Pedoe, 2003).

As a first large scale prospective cohort study in Denmark, a small group of clinicians initiated the Glostrup Population Studies in 1964 and began collecting data for a 50 years age group, successively including more 'birth cohorts' (Hagerup et al., 1981; Jørgensen, 2004). A decade later, in 1975, clinicians, mostly cardiologists, at the National University Hospital (Rigshospitalet) in Copenhagen initiated the Østerbro study, comprising a 'general adult population' of all ages.[39] Similar to Framingham, the study in Copenhagen was set up as a long-term observational study in a sample of the healthy population.

The loose network of clinicians involved in the early phase of the Østerbro study perceived themselves as pioneers, believing in prevention as an important contribution to medicine, beyond diagnostics and therapy.[40] Marginalised at the time, the population studies were carried out under tight funding conditions throughout the 1970s and 1980s. In order to conduct quantitative comparisons, standardised data for the general population to be observed prospectively were needed. Aiming at 'reference values for an unselected population' (Appleyard, 1989, p.7), researchers drew a random sample of 20,000 people aged >20 years among the ~90,000 inhabitants registered in ten electoral districts in the vicinity of Rigshospitalet (in Østerbro and Nørrebro). Letters of invitation to the examination were mailed out with a system of reminders in order to maximise response; the resulting participation rate of 74% in the first examination was high in international comparison.

Reducing costs through effective logistics has been an important issue throughout the study. While the variables collected varied between examinations, the basic setup of the first examination has been kept.

> At *the first station*, a blood sample for examination of lipid- and glucose level was drawn, presence of xantelasmata and earlobe-crease was noted, and a lung function test performed. At *the second station*, height and weight were measured, a 12 lead resting ECG recorded, and presence of arcus senilis and evaluation of signs of aging (i.e. grey hair, baldness and wrinkles), was noted according to a code [. . .]. At *the third and last station*, the questionnaire was checked and the blood pressure was measured. The results of the examination were explained to the participant, and were later mailed to the general practitioner. (Appleyard, 1989, p. 9)

The nurses who were in charge of examining participants changed between the different stations every two hours. The screenings themselves were set up as a 'conveyor belt' to allow the throughput of as many participants as possible.[41]

While participants waited for the examinations, they completed a questionnaire, which contained items on symptoms and diseases, on family history and, for women, on birth control and hormone treatment, as well as detailed questions on the new 'lifestyle risk factors'. These were items on 'smoking and drinking habits' (amount and frequencies of alcohol and tobacco consumption) as well as items on 'physical activity at work and during leisure time'[42] (Fig. 1). The data in the forms were checked for completeness by the nurses at the third station, while they explained the test results to participants.[43]

At the end of each day, blood samples were sent to the lab for analyses of total and HDL-cholesterol, triglycerides and glucose. All samples from the first investigations were retained in a freezer (Fig. 2). Storing samples in freezers for future use, later referred to as biobanking, was practiced in the Østerbro study from early on. In relation to the first examinations in 1976–1978, this was rather a backup routine, whereas from the third round of examinations in 1991–1994, researchers have been fully aware of the future importance of these resources for genetic epidemiology[44] and serum, plasma, full blood and DNA have been stored.

Collected data were documented in standardised sheets; questionnaires, lab results and electrocardiograms—literally 'inscriptions' of a signal on a piece of paper recorded by an inscription device—were archived and coded. For the first examinations, a commercial bureau verified the final datasheets and punched the information in punch cards. These were read into a computer, data were tabulated and checked for implausible values; these descriptive analyses were done at the central computer of Rigshospitalet.

[37] For an example of large-scale international networks of local studies, see the WHO coordinated study 'Monitoring trends in cardiovascular disease' ('MONICA') (Tunstall-Pedoe, 2003).

[38] For an early report on the set-up of the Framingham Study, see Dawber et al. (1951). See Oppenheimer (2005) for a recent account of this influential study, termed 'the epitome of successful epidemiological research' and 'prototype and model of the cohort study' by epidemiologist Mervyn Susser (Oppenheimer, 2005, p. 602). See Aronowitz (1998) for a historical and sociological study of the shift from angina pectoris to cardiovascular disease with the Framingham study.

[39] While the Glostrup population studies joined the international MONICA project of cardiovascular studies coordinated by the WHO, the CCHS remained a local study, administrated by a local executive scientific committee. International coordination and standardisation of epidemiological studies became a major goal of the WHO: the WHO ERICA studies initiated in 1964 comprised data from thirty-seven countries (ERICA Research Group, 1988; Lamm et al., 1989).

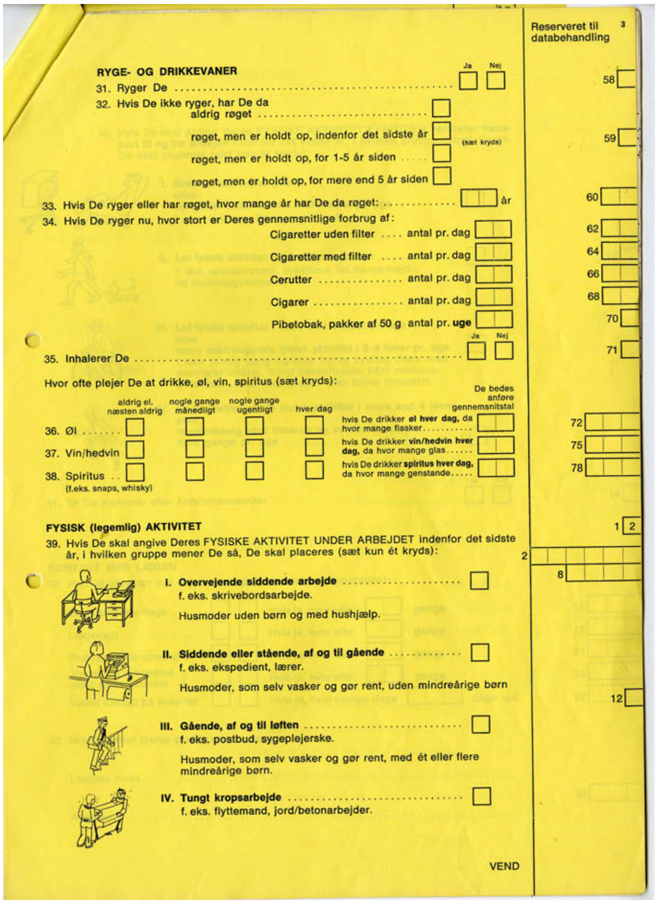
[40] Peter Schnohr, personal communication, 30 March 2006.

[41] Merete Appleyard, personal communication, 3 May 2006.

[42] The range of data collected was wider in the third examination and included additional items, for example on lifestyle and stress (1991–1994).

[43] Whereas the laboratory results were sent to participants' general practitioners in the 1970s, the results were later sent directly to the participants; this reflects changes in doctor–patient relationships towards patient autonomy and individualised responsibility.

[44] Gorm Jensen, Presentation at BioLogue symposium, University of Copenhagen, 27 November 2006.



Reserveret til databehandling 3

RYGE- OG DRIKKEVANER

31. Ryger De .... Ja □ Nej □ 58□

32. Hvis De ikke ryger, har De da
aldrig røget .... □
røget, men er holdt op, indenfor det sidste år □ (sæt kryds) 59□
røget, men er holdt op, for 1-5 år siden ..... □
røget, men er holdt op, for mere end 5 år siden □

33. Hvis De ryger eller har røget, hvor mange år har De da røget: .... □□ år 60□□

34. Hvis De ryger nu, hvor stort er Deres gennemsnitlige forbrug af:
Cigaretter uden filter .... antal pr. dag □□ 62□□
Cigaretter med filter .... antal pr. dag □□ 64□□
Cerutter .... antal pr. dag □□ 66□□
Cigarer .... antal pr. dag □□ 68□□
Pibetobak, pakker af 50 g antal pr. **uge** □□ 70□

35. Inhalerer De .... Ja □ Nej □ 71□

Hvor ofte plejer De at drikke, øl, vin, spiritus (sæt kryds):

	aldrig el. næsten aldrig	nogle gange månedligt	nogle gange ugentligt	hver dag	De bedes anføre gennemsnitstal	
36. Øl	□	□	□	□	hvis De drikker **øl hver dag,** da hvor mange flasker.......... □□	72□□□
37. Vin/hedvin	□	□	□	□	hvis De drikker **vin/hedvin hver dag,** da hvor mange glas...... □□	75□□□
38. Spiritus .. (f.eks. snaps, whisky)	□	□	□	□	hvis De drikker **spiritus hver dag,** da hvor mange genstande..... □□	78□□□

FYSISK (legemlig) AKTIVITET 1 2

39. Hvis De skal angive Deres FYSISKE AKTIVITET UNDER ARBEJDET indenfor det sidste år, i hvilken gruppe mener De så, De skal placeres (sæt kun ét kryds): 2 □□□□□□ 8 □□□□

I. **Overvejende siddende arbejde** .... □
f. eks. skrivebordsarbejde.
Husmoder uden børn og med hushjælp.

II. **Siddende eller stående, af og til gående** .... □
f. eks. ekspedient, lærer.
Husmoder, som selv vasker og gør rent, uden mindreårige børn

III. **Gående, af og til løften** .... □ 12□
f. eks. postbud, sygeplejerske.
Husmoder, som selv vasker og gør rent, med ét eller flere mindreårige børn.

IV. **Tungt kropsarbejde** .... □
f. eks. flyttemand, jord/betonarbejder.

VEND

Fig. 1. Questionnaire used in the first examination (1976–1978) of the Copenhagen City Heart Study. Medical Museion registration no. MM 58:2008 G a, detail, p. 3.

In the 1970s, the tabulation of variables for a large study like the Østerbro study was only possible at few central computers; in Copenhagen these were central computers at Denmark's Technical University and at Rigshospitalet, where time slots needed to be booked in advance.[45] Rapid advances in information processing led to a gradual decentralisation of computer capacity and the same computations could be performed on a standard PC two decades later.

In order to generate a database of reference values for the general population, a set of objectification strategies had to be brought to work: nurses had been trained according to the WHO manual for cardiovascular epidemiological studies. In the evaluation of electrocardiograms, indices were manually measured from the curves of the electrocardiograms (ECGs), following the 'Minnesota code', a pragmatic standard to read indices from the paper curves. Although 'often arbitrary and arrived at by compromises [...] [the Minnesota classification] provides a framework for reporting ECG items in uniform and clearly defined terms' (Appleyard, 1989, p. 155). Standardised sampling and coding secured that the results qualified as reference data for comparison purposes. The data that resulted from these large-scale examinations of a healthy population were frequently drawn upon by other researchers; due to the numerous requests, the investigators of the Østerbro study published their data in full length as reference values for an 'unselected population in Denmark' (ibid., p. 7).

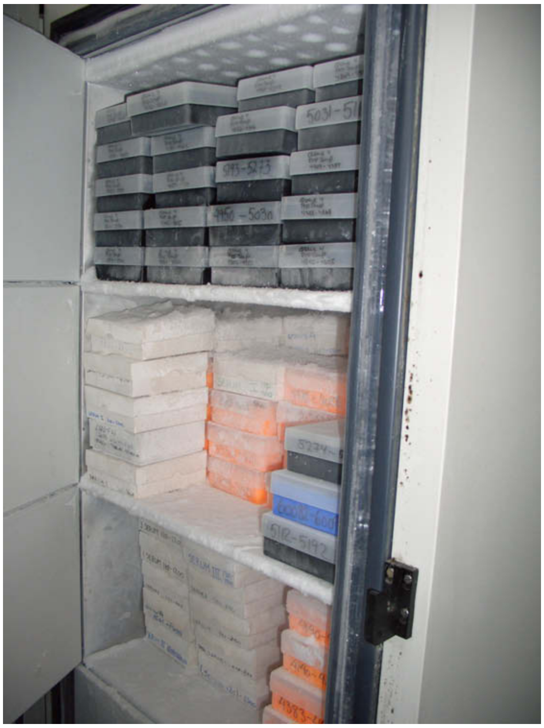


Fig. 2. Frozen samples of the Copenhagen City Heart Study (photograph by the author). Medical Museion registration no. MM 93:2008 B g.

To follow up participants over time, examinations were repeated at 5–10 year intervals (1981–1983, 1991–1994 and 2001–2003) to secure prospective data. Although participation rates declined over time—from 74% in the first examination (1976–1978), to 61% and 58% in the 1991–1994 and 2001–2003 examination, respectively—the follow-up through population registries was complete for study participants. Over the decades, data preservation became a critical issue with changing hardware and software systems; in addition to electronic storage, questionnaires, original schemes, and electrocardiograms were preserved and kept for reference.

The report on the first and second examination series of the Østerbro study introduced the epidemiological framework as 'the study of occurrence of disease in populations and of factors influencing the disease distribution within the population' and as 'essential in prevention and in health care' (ibid.). Laying out a whole agenda of chronic disease epidemiology, the authors of the 1989 report problematised in particular the previous lack of reference values for diagnostic variables and lifestyle risk factors for a general population.

Data inventories of cardiovascular reference values and lifestyle variables were established through a research system that applied diagnostics techniques to the general population, subjecting healthy citizens to an extended clinical gaze.[46] Studies such as the Østerbro study were envisioned to record and statistically relate detailed data on lifestyle habits to disease events later in life in order to calculate risk estimates that would inform prevention at a population level, and guide health policy in prioritizing and targeting future interventions. With respect to the range of potential determinants of disease, data collecting and analysis focused mainly on 'lifestyle' variables that were perceived as modifiable and as

[45] Merete Appleyard, personal communication, 7 June 2007.

[46] With epidemiological studies, the clinical gaze (Foucault, 1973) is extended to the general population to an unprecedented degree, enrolling data of whole populations statistically in the generation of biomedical knowledge.

under the control of the individual. In line with other studies, notably the Framingham Study, it was the individual level variables that gained priority over social indicators or occupational factors in what became understood as 'risk factor epidemiology'. Albeit frequently challenged in particular by social epidemiologists,[47] the conceptualisation of individual lifestyle risk continues to be influential in shaping health research and policies at the present.[48]

During the decades following the first examinations, the scope of data use was extended far beyond lifestyle risk and cardiovascular disease. The next section will discuss how these observational data gathered in the Østerbro study were connected to other research questions beyond the immediate context of the data collection.

5. Data mining: 'when an entire country is a cohort'[49]

From the early 1980s onwards, the data from the Østerbro study were linked to the Danish registry system, for example to the Causes of Death Registry[50] for mortality data. Epidemiologists working with data from the Østerbro study have made use of the system of central registries and connected the data to a wide range of further outcomes. Research ranged from the quantitative descriptions of disease incidence in relation to physical activity, nutrition, hypertension, and 'signs of aging' in the early years, to the more recent study of molecular markers. The linkage capacity of the dataset and application of methods to generate risk estimates for exposure variables in relation to outcomes provided a host of research opportunities and resulted in a considerable publications record.[51]

The system of population registries, which are specific to the Nordic countries, provided linkage opportunities to, at the time, more than sixty health registries and fifteen administrative registries, which were explicitly promoted for use in health research in Denmark (Meulengracht & Madsen, 1982). For instance, the National Danish Hospital Registry has stored data on hospital admissions from 1977 onwards and outpatient and emergency data have been included since 1995 (Andersen et al., 1999). A system of individual identification numbers (CPR-no.) introduced by the Central Population Registry on April 1, 1968 makes it possible to connect data within different registries. Although not primarily intended for research but for administration purposes, the CPR-no. became routinely used in epidemiology, both in local studies and for nationwide investigations. Computerised in the 1970s, population registries are seen as instrumental in generating internationally relevant research output (Mortensen, 2004, p. 122). The following paragraphs exemplify the use of these linkage opportunities for the two breast cancer studies that were conducted based on the data and biobanks of the Østerbro study.

When investigating the aetiology of breast cancer with data from a cardiovascular study, data on breast cancer needed to be retrieved from the registries. In both studies (Høyer et al., 2002; Weischer et al., 2007) the records of the Østerbro study were connected to the Danish Cancer Registry[52] and the breast cancer treatment registry (operated by the Danish Breast Cancer Cooperative Group (DBCG))—the two main sources for breast cancer incidence data.[53] The DBCG registry had been established for the nationwide clinical breast cancer trial that started in 1977 and stores treatment data, including data on histology and tumour stage (Fisherman et al., 1977, 1988).

Data mining was based on record linkage through CPR-nos.; for the study by Høyer et al. (2000a, 2000b, 2000c, 2001, 2002) the embedded 'nested case-control study' was described as follows:

> The aim of this cohort nested case-control study is to evaluate if mutations in p53 interact with organochlorines possible endocrine disrupting effect on breast cancer risk and survival [...]
>
> The cohort was followed with regard to development of invasive breast cancer to the end of 1992 by linkage to the Danish Cancer Registry, which is regarded virtually complete. [...] From the remainder of the CCHS cohort, two women were selected randomly to act as controls after matching for age and vital status at the time of diagnosis of the corresponding case. Information on vital status, cause, and date of death until July 31, 1997 was obtained by linkage to the Civil Registration System and the Causes of Death Registry, Danish National Board of Health. (Høyer et al., 2002, p. 60)

For studies conducted on the top of existing cohort studies yet under inclusion of new variables, the technique of 'nesting' plays a fundamental role in the mining and re-use of data from cohort studies and in the re-assemblages with new data. In 'nested designs' additional variables are included for a subset of cases and controls selected from the cohort for a specific study that is superimposed on the existing dataset. This bears the advantage of still benefiting from the infrastructure of the larger study, which makes sample and data access easier.

In order to investigate the influence of organochlorine exposure on cancer incidence based on the study's dataset, Høyer et al. brought together data from the Østerbro study with records of the central registries; as the authors affirm: 'The considerabl[e] amount of information obtained in the CCHS together with the Danish tradition of systematically registering cancer cases, and the unique ID number assures complete follow-up' (ibid., p. 63). In their new context, the data from the original questionnaire of the Østerbro study were mined for 'known and suspected risk factors for breast cancer' (ibid., p. 60).

Høyer et al. (2000b, 2002) re-circulated data and biological samples with a 'nested case-control design'. They matched two 'controls' to each breast cancer 'case' of the Østerbro data in order to compute odds ratios as estimates of breast cancer risk in dependence of the exposure.

To assess the influence of other factors, the variables—as available from the Østerbro study—were used for risk computation (Fig. 3). The regression model to calculate age-adjusted odds ratios can be performed for all other variables that are available. Depending on whether they were found statistically significant or not, they were included or excluded in the final statistical models (Høyer et al., 2000b, p. 180; 2002, p. 61).

Thus, a set of variables collected for the context of cardiovascular disease is rendered into an aetiological matrix for breast cancer; data availability and possibilities to retrieve them are key to the

[47] There has been criticism of the focus on lifestyle, in particular from social epidemiologists, who emphasise the social and structural determinants of health, see for example Berkman & Kawachi (2000) and Krieger (2001); see also Section 2 of this paper on challenges to the biomedical model by social epidemiologists.

[48] If compared to other Scandinavian countries, the lifestyle focus has been particularly prominent in Denmark (Vallgårda, 2006).

[49] See Frank (2000).

[50] For a description of the Danish Causes of Death Registry, see Juel and Helweg-Larsen (1999).

[51] For a publication list of the Copenhagen City Heart Study from 1977 to the present, see Copenhagen City Heart Study (n.d.) The publication record reached >500 papers by 2008.

[52] The Danish Cancer Registry is the first nationwide registry of that kind and has been operating since 1 January 1943. It was managed by the Danish Cancer Society, before it became part of the Danish National Board of Health in 1997.

[53] See Storm (1991) for an account of the differences (and the implications for epidemiology) between the data of the Danish Cancer Registry and of the DBCG.

choice of modelling parameters. As additional risk factors, these variables were examined for modifying effects via statistical modelling and the parameters found significant in the analysis (parity, body weight and hormone replacement therapy) were kept in the final model. Moving across context, the data initially recorded as cardiovascular risk factors in the Østerbro study have become potential modifiers of breast cancer risk.

In a similar way, in their study on the CHEK2 gene, Weischer et al. (2007) performed data mining from multiple sources, yet in a differing configuration. To identify breast cancer cases (and other cancers) in the CCHS for the period between 1947 through 2002:

> Diagnoses of invasive cancer [...] were obtained from the Danish Cancer Registry, which identifies 98% of all cancers in Denmark. [...] Follow-up time for each participant began at the establishment of the Danish Civil Register System April 1, 1968, or the participant's 20th birthday, whichever came later, and ended at death, event, emigration, or on December 31st, 2002, whichever came earlier. [...] Follow-up was 100% complete. (Weischer et al., 2007, p. 58)

Thus, the first study reported was an analysis of incidence data for, among other cancer sites, breast cancer and the CHEK2*1100delC mutation status within the complete prospective cohort study. Data on cancer incidence were mined for the entire cohort from the cancer registry and full genotyping of blood samples from participants had been conducted for this part of the study. In addition to this cohort analysis, a case-control study was designed, recruiting the breast cancer cases treated at the hospital, while using the Østerbro data as a source for selecting 'population controls' from a database considered representative of the general population. The 'hospital cases' were interviewed, while for the 'controls' from the Østerbro study the questionnaire data on lifestyle and the results from the physical examinations in 1976–1978 and 1981–1983 were made available to the new study. In the 1990s, the epidemiologists used information on smoking habits, alcohol consumption, body mass index, parity, use of oral contraceptive, menopausal status and hormonal replacement therapy (ibid., pp. 58–59) from the data that were collected (via questionnaires) in the context of the investigation of cardiovascular disease. Again, in the new context, the variables of the Copenhagen City Heart Study acquire the status of potential risk factors for breast cancer—factors that interact with genetic predispositions and other factors in carcinogenesis.

The system of central population registries co-shaped the ways epidemiological research has been set up. For instance, this system also determined the beginning of the follow-up in the breast cancer studies analysed here. As Mortensen (2004, p. 121) has put it, 1968 is the 'year zero in Denmark' for most current registry research. It was the record linkage opportunities that made it possible that the data from the study of cardiovascular prevention could help investigate breast cancer. In this way, the Østerbro data serve post hoc as an empirical resource to study other outcome variables, made possible through data mining.

Whereas epidemiological studies elsewhere have to collect most variables using questionnaires, the regulations in the Nordic countries allow epidemiologists to directly access certain individual data from central registries:

> Denmark has a well-operating system of unique person-numbers, which has great practical significance in everyday life, and which makes it possible to follow the Dane 'from the cradle to the grave'. The basic data of the CPR can be linked to individual information about for example health, disease and occupational status. (Olsen et al., 2004, p. 1459)[54]

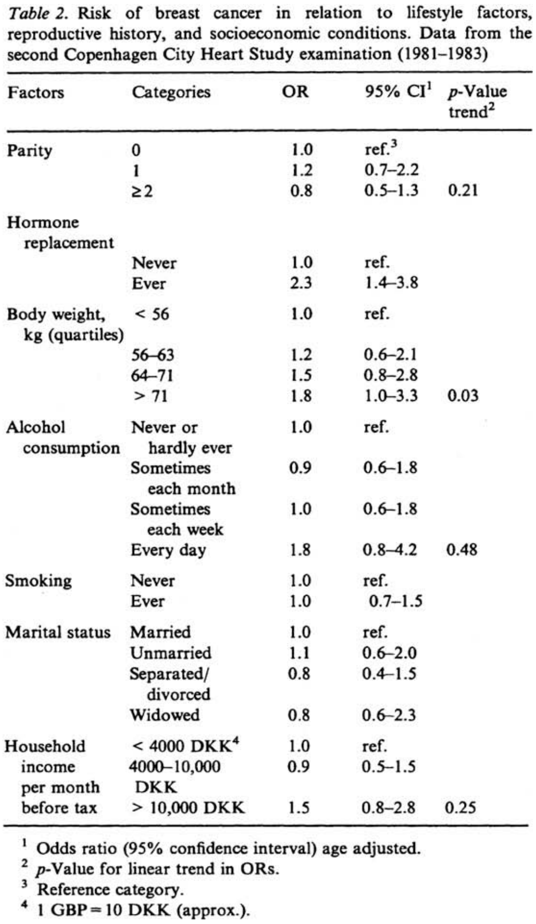


Table 2. Risk of breast cancer in relation to lifestyle factors, reproductive history, and socioeconomic conditions. Data from the second Copenhagen City Heart Study examination (1981–1983)

Factors	Categories	OR	95% CI[1]	*p*-Value trend[2]
Parity	0	1.0	ref.[3]	
	1	1.2	0.7–2.2	
	≥2	0.8	0.5–1.3	0.21
Hormone replacement				
	Never	1.0	ref.	
	Ever	2.3	1.4–3.8	
Body weight, kg (quartiles)	< 56	1.0	ref.	
	56–63	1.2	0.6–2.1	
	64–71	1.5	0.8–2.8	
	> 71	1.8	1.0–3.3	0.03
Alcohol consumption	Never or hardly ever	1.0	ref.	
	Sometimes each month	0.9	0.6–1.8	
	Sometimes each week	1.0	0.6–1.8	
	Every day	1.8	0.8–4.2	0.48
Smoking	Never	1.0	ref.	
	Ever	1.0	0.7–1.5	
Marital status	Married	1.0	ref.	
	Unmarried	1.1	0.6–2.0	
	Separated/ divorced	0.8	0.4–1.5	
	Widowed	0.8	0.6–2.3	
Household income per month before tax	< 4000 DKK[4]	1.0	ref.	
	4000–10,000 DKK	0.9	0.5–1.5	
	> 10,000 DKK	1.5	0.8–2.8	0.25

[1] Odds ratio (95% confidence interval) age adjusted.
[2] *p*-Value for linear trend in ORs.
[3] Reference category.
[4] 1 GBP = 10 DKK (approx.).

Fig. 3. Re-evaluation of data from the Copenhagen City Heart Study for breast cancer risk. Table originally published by Høyer et al., 2000b, Table 2, p. 180; reprinted with the kind permission of Springer Science and Business Media.

With epidemiological and biostatistical tools, the relationship between 'health outcomes' as held in these registries and a large array of potential determinants—as potential risk factors, indicators or surrogate variables—can be connected, tested for associations and the latter expressed as probabilistic risk estimates. In particular through the regulation of data access favourable to the needs of epidemiologists, registry research has become an academic subdiscipline in the health sciences (Mortensen, 2004). The specific conditions for record linkage in Denmark were launched as a competitive resource, when it comes to epidemiological research in the age of genomics (Melbye, 2001; Frank, 2003; Mortensen, 2004).

> Moreover, in Denmark in the last 30 years a number of biological banks storing blood samples, urine samples, biopsies (biobanks) have been established. Such biobanks are of extraordinary value in epidemiological research. To this picture of excellent research conditions belongs that the Danish population is to a great extent willing to contribute information to research, for example in the form of the donation of a blood sample and answering a questionnaire. (Olsen et al., 2004, p. 1459)

Combining data from different sources in new ways is particularly important in molecular and genetic epidemiology. The subdiscipline of genetic epidemiology has changed the mode in which research on cardiovascular disease and cancer was con-

[54] The translations from the review article 'From counting cancers to the cancer registry' in the *Ugeskrift for Læger* (*Danish Medical Weekly*) are my own.

ducted by the mid 1990s. The option to recombine existing datasets with new molecular variables renders epidemiologic databases into productive resources for novel hypothesis testing.

6. Recombining information in genomic epidemiology

Existing biobanks and data resources have increasingly been used to study the significance of genetic variability in 'complex diseases' with multifactorial aetiologies and population studies in this field are often referred to as 'public health genomics'.[55] To shed light on the specific productivity of epidemiological techniques in making sense of sequence information and molecular markers at a population level, this section continues to follow the data from the Østerbro study. The focus now moves from data mining to the re-use of both data and samples in the two breast cancer studies. In what ways were the data from the CCHS and population registries recombined with new measurements? What kinds of data rearrangements were made in the study designs to include molecular determinants?

6.1. Organochlorine exposures and breast cancer: testing new markers on already existing samples

To assess exposure to organochlorines as of several decades ago, Høyer et al. negotiated access to the samples that had been retained. Thus, the frozen serum samples were now, twenty years later, used and reanalysed to assess the exposure. The researchers secured detailed exposure reconstruction by collaboration with a specialised laboratory.[56] As the published paper reports, 0.5 ml from the original serum samples were retrieved and sent to the US Center for Disease Control (CDC) for comprehensive analyses. In this way, new exposure data were gained from the original samples; still they remain bound to the configuration of the original study. Like in data mining, the 'nested' design was the technique that made it possible to include further 'exposure' variables as determinants; it reduced the number of measurements and therefore costs, while keeping the representivity of the cohort as methodological benefit.

Høyer et al. (2002) not only re-used the retained serum samples from the Østerbro study's biobanks to measure exposure in detail, but also refined the classification of the outcome—breast cancer—with molecular techniques. They traced the original diagnostic biopsies in order to measure immunohistochemical and genetic biomarkers of the stored tumour tissue.

> Further linkage to the nation-wide clinical trial of the Danish Breast Cancer Cooperative Group provided information on breast cancer tumor characteristics and the reviewing pathology facility, from which the paraffin embedded tumor tissue specimen was obtained. Ten micrometer sections of the most viable part of the paraffin specimen were prepared for the detection of p53 mutations. (Ibid., p. 60)

As the authors further describe, DNA was extracted from the tissue sections and exons 4–11 were screened for the 'tumour suppressor gene' encoding the p53 protein. In this context, the p53 marker is understood as 'a fingerprint to implicate environmental exposure' (ibid.). Thus, the p53 mutation remains an ambiguous intermediary, situated between an exposure effect and a tumour characteristic, at the same time it becomes a potential 'signature of exposure'. By including this marker in the study, it is hoped that inconclusive results can be clarified and 'the causal associations are expected to emerge more clearly' (ibid.) after accounting for p53. In epidemiological terms, the interaction of the genetic polymorphism and organochlorine exposure as to cancer risk was to be determined.

Brought together in a nested case-control study, molecular variables, registry data and lifestyle variables from the questionnaires were enrolled into multivariate statistical modelling. Compared to the earlier studies by the same authors, the innovation was that the newly gained exposure variables from the samples stored in the biobank as well as the p53 marker measured in tumour tissue enter the risk calculation. The biostatistical analyses included comparisons in terms of risk differences and risk ratios between groups defined by organochlorine levels and 'mutant p53' or 'wildtype p53' status (ibid.), with adjustment for a set of 'potential confounders' (ibid., p. 61). All potential determinants (including parity, household income and education) are subsequently tested and those found statistically significant were kept in the final logistic regression model. Thus variables from different levels and contexts were brought into one model equation and then statistically tested for associations with breast cancer under multiple adjustments. The hypothesis testing takes place as biostatistical modelling, a retrospective quasi-experiment *in silico*[57] constructed post-hoc with the data at hand, i.e. from the Østerbro study, the registries and new variables gained from the stored tissue. When translated to public health—for the case of organochlorines this is of concern with respect to health risk assessment and regulation[58]—the resulting risk figures carry on categories of data collecting and decisions made in data modelling.

6.2. A candidate for epidemiology: the CHEK2 gene

Different from the 'environmental' study by Høyer et al., the configuration of variables in the 'genetic' study by Weischer et al. is entirely targeted at the CHEK2 polymorphism: the investigators set up the study to test the influence of the CHEK2*1100delC mutation. This genetic marker was investigated as a potential determinant, whilst adjusting for environmental and lifestyle variables. At the University Hospital, the blood samples of the 1991–1994 examination of the Østerbro study were fully genotyped. In order to investigate the influence of this candidate gene, the generated sequence data were statistically modelled together with the variables from the Østerbro study. To interrogate the dataset for the significance of polymorphisms in the CHEK2 tumour suppressor gene as to breast cancer risk, Weischer et al. (2007) used two study designs in parallel.

In the first analysis, the data from the examinations of the Østerbro study were modelled following the cohort design; the data were divided into different groups, according to the 'CHEK2 status' of participants. Connecting the individual data, recorded over three decades, to the DNA sequence information, breast cancer risks were calculated in dependence of 'CHEK2 status'. As reported in the paper, statistical adjustments for lifestyle risk factors included:

> age, body mass index ($<25\ v \geqslant = 25\text{kg/m}^2$), alcohol consumption (0g/wk $v > 0$ g/wk), nulliparity (yes *v* no), current use of oral contraceptives (yes *v* no) menopausal status (pre- *v* post-menopausal), and current use of hormone replacement therapy (yes *v* no). (Weischer et al., 2007, p. 58)

[55] See for example the recent 'public health genomics' and 'public health genetics' initiatives in the health sciences—such as 'Genpub', University of Copenhagen (n.d.).

[56] Organochlorine analyses were conducted at the US Center for Disease Control, Atlanta; the measurements included eighteen chlororganic pesticides, twenty-eight PCB congeners and DDTs (Høyer et al., 2002, p. 60).

[57] For the notion of 'biology *in silico*' and biology as an information science, see for example Lenoir (1999).

[58] For the public debate on threshold values for pesticides in food which followed the 1998 publication of the study on organochlorines, see Dørge (1998).

Thus, most parameters, as available in the Østerbro study, were included to adjust for different lifestyle risks, while the aetiologic hypothesis tested was the association with the CHEK2 polymorphism. With this 'candidate gene' approach that investigates specific genes in 'association studies' at a population level, the breast cancer risk estimates were adjusted for differences in 'lifestyle'.

As a second design reported in the same publication, Weischer et al. used a case-control approach to compare breast cancer 'cases' diagnosed and treated at the University Hospital to 'controls' borrowed from the Østerbro study:

> The case-control study included 1,101 women with invasive breast cancer consecutively recruited at Herlev University Hospital between February 2001 and August 2004 [...] Controls were 4,665 general population (The Copenhagen City Heart Study) from the same age range as the patients and who had no history of breast cancer before the end of 2002. (Ibid.)

In this data recombination, both for 'cases' and 'controls', detailed lifestyle and exposure data were accessed, which allowed the computation of risk estimates ('odds ratios') for the different subgroups defined by CHEK2 status. Thus, in this mode of recombining data, the Østerbro study is used as a pool of population-based 'controls', in other words a reference population for comparison.

For candidate gene studies in cohorts with complete follow-up for thirty years as in the Østerbro study, even the null results are frequently cited, as they can serve to falsify hypotheses that were suggested by smaller case-control studies.[59] As its main and most cited result, the CHEK2-study showed a threefold increase in breast cancer risk and no increase in risk for other cancers with the inherited mutation in the CHEK2 gene. In this combined study, the dataset stabilised CHEK2 as a tumour suppressor gene and as a significant risk factor for breast cancer. The authors conclude that 'women with CHEK2*1100delC may benefit from breast cancer screening, preferably excluding the use of ionizing radiation' (ibid., p. 62). Further optimistic statements regarding clinical testing, given by the researchers to the media[60] raised a controversy among clinical practitioners on the relevance and ethics of testing for low penetrance risk markers such as CHEK2 (Gerdes, 2006; Weischer et al., 2006).

6.3. Re-mobilizing data for molecular epidemiology

Genetic epidemiology has had an important place in the recent research with the Østerbro study's data since the 1990s (Nordestgaard et al., 1996). What made long-term cohort studies such as this local database interesting to molecular epidemiologists in the mid 1990s was the biological material collected and retained combined with complete medical and lifestyle information at an individual level. Building on the CCHS data, breast cancer is investigated using the molecular markers available from the tissue and blood samples. With the broader availability of sequencing technologies, epidemiological studies have increasingly taken into account genetic variables. Epidemiological techniques have come to work as tools to make sense of genomic data; conceptually epidemiologists continue to draw on the notion of multicausality—the 'web of causation' (McMahon et al., 1960, p. 18)—to study multiple gene–environment interactions. As documented for the environmental breast cancer studies (Høyer et al., 2002), the shift to include molecular variables in epidemiology also holds for environmental health research. Whereas during the 1960s and 1970s, the matrix of risk factors extended above all to new 'lifestyle' variables, the fabric of epidemiological knowledge have come to look very different thirty years later, when risk factors include the molecular level.[61] The ways aetiological risk factors are addressed in epidemiological research increasingly move to the molecular level of biomarkers.

While the Copenhagen City Heart Study had initially been aimed at investigating cardiovascular disease, its database came to function as a productive empirical archive for the study of new markers in the context of genomic epidemiology. Reassembled with molecular variables, such as the CHEK2 polymorphism or p53, this database moved on to new research on molecular aetiologies. As exemplified for the Østerbro study's database, genomic epidemiology makes use of multiple levels, from micro to macro; from DNA sequence to lifestyle and socioeconomic variables such as household income. This mode of genomic epidemiology has implications not only with respect to candidate genes or molecular markers, but in a transformed matrix of disease determinants. Thus, with genomic epidemiology the ways social, lifestyle and environmental factors are addressed have changed to the molecular level of biomarkers and transformed the epistemic configurations of how aetiologies are and can be understood. Yet, as the two breast cancer studies illustrate, there is a range of different positions and conceptual traditions—environmental health versus genetics—that remain present within genomic epidemiology.

Establishing the databases, collecting and genotyping samples, extracting and connecting the information is laborious and involves well designed collecting logistics and meticulous data organisation as well as advanced methods in statistical modelling. In epidemiological practice, it may not always be feasible to obtain the variables of interest, then surrogates and proxies for instance for exposure history need to be developed and included in the model. Despite resistances encountered, the epidemiological techniques are highly flexible in drawing together data at multiple levels and from different contexts—registries, hospitals, biomedical laboratories, biobanks etc. Epidemiology follows pragmatic considerations of data access as much as conceptual frameworks, prior beliefs, methodological standards, mechanistic understandings and ideas on aetiology. It is the possibility and performativity of record linkage that enrols population databases with retained blood and tissue samples into 'biovalue'.[62] However, the more variables are included, the larger the datasets required to achieve sufficient statistical power, a limitation often faced in studies on gene-environment interactions. Therefore it is precisely the already existing large-scale studies and registries that are invaluable for genomic epidemiology. Together, databases and biobanks become unique and sustained resources for research, which are drawn upon to develop and test new aetiological hypotheses, made possible by data mining and data recombination. At the same time, such aetiological studies are contingent on their history of data collecting and biobanking, as well as on the registry infrastructure and the socio-cultural factors that shaped their categorisations, which present important epistemological implications.

[59] Børge G. Nordestgaard, personal communication, 11 October 2006.

[60] See Nielsen et al. (2006). While some genetic epidemiologists consider testing for the CHEK2 allele in principle ready for the clinic, many clinicians and epidemiologists hold that screening is not justified. This is due to problems in the interpretation of tests for 'low penetrance' mutations; furthermore, the risk estimations are considered still inconclusive: for instance, a population-based study from Sweden, published in late 2006, found no association of the CHEK2 with breast cancer (Einarsdottir et al., 2006).

[61] As the p53 study by Høyer et al. (2002) exemplifies, the field of environmental health research also incorporates molecular and genetic markers. This focus of research in turn modifies conceptual frameworks of risk assessment: in toxicogenomics, genetic susceptibility towards exposure moves centre stage (Shostak, 2005; Bauer, 2006).

[62] As for the term 'biovalue', I refer to Waldby (2002) and Rose & Novas (2005).

7. Conclusions

This paper has approached epidemiology as a practice of collecting and traced selected data trajectories of a large-scale cohort study. The analysis of two re-assemblages of data from the Østerbro study—in aetiological studies of breast cancer—has exposed the role of data mining and record linkage in stabilising biomedical knowledge at a population level. Various data strategies of epidemiological research practice can be described: active gathering of new variables and samples in a defined study design is key to large-scale prospective follow-up studies. Mining data from registries refers to the deployment of already existing data—data recorded in other contexts but used for epidemiological research—such as routine data, for example social statistics, cause of death data, or, in the case of the Nordic countries, data from central population registries. Recombining information entails a re-assemblage of data into novel constellations, which re-evaluate determinants and outcomes based on molecular techniques. In these re-arrangements, data travel over time, across levels and context of investigation, whilst they continue to carry on contextual categories.

In this sense, epidemiological knowledge generation follows a pragmatic approach, studying statistical associations by making use of the data at hand. Luc Berlivet (2005a, pp. 57–58) has pointed to Austin Bradford Hill's 'pragmatics of risk factor epidemiology', where the epistemological questions came to be regulated by a set of causality criteria and conventions. While this pragmatism is compatible with the 'regulatory objectivity' (Cambrosio et al., 2005) that is paramount to highly regulated rigorous methodologies, epidemiological practices with data also entail flexibility in the multiple data re-assemblages. As to the Østerbro data, the original cardiovascular prevention study is reshuffled into aetiological studies on breast cancer by data mining and recombination. Record linkage and new measurements on tissue samples can be brought to work in different conceptual frameworks and contexts: for instance in their different data re-assemblages, the two epidemiological investigations based on the Østerbro data establish different articulations as to breast cancer aetiology.

Epidemiological research such as the Østerbro study is multidisciplinary and brings together a range of techniques from field sciences and the laboratory; the database functions as a local platform for hypothesis testing, modelling and simulation. While in many ways, the epidemiological investigation of disease rates and risks is similar to quantitative social science and demography, Paolo Vineis has stressed that, compared to the social sciences, 'epidemiology makes use more often of study designs that simulate experiments, than of surveys in the general population' (Vineis, 2005, p. 349). The experimental configuration applied to data and sample collections is paramount to the epistemic value epidemiologists attribute to their studies.

Computerised record linkage and multivariate modelling are key epistemic techniques instrumental to the simulation of experimental approaches in epidemiology. In data (re)arrangements, population-based 'quasi-experimental' comparisons are projected back in time and retrospectively embedded in historical cohort studies, such as in the nested case-control studies. Mediated by the architecture of a nested case-control study, questionnaire data, tissue samples and genetic or environmental aetiologies are recombined in a new constellation (conceived as retrospectively enabled comparison) and navigated with regression modelling; in this way epidemiological techniques construct retrospective data experiments *in silico*. Thus, epidemiology works through large-scale data generation and data networking combined with the simulation of experiments using observational data. As much as about designing original studies, successful epidemiology is about securing long-term access to data and samples as future resources, in this case to the biobank and the database of a long-term follow-up study and to population registries. Collecting and observing as well as experimenting and simulating are deeply entangled in epidemiological research.

Studying epidemiology's practices of record linkage focuses on the data organisation and configuration, thereby shedding light on the epistemic status of databases. Inspired by the science studies' mode of following things and molecules, I propose to also track the less tangible 'digital trajectories' of data and data collections over time and across contexts. As an analytic strategy in the study of epidemiological knowledge, this can help describe how local data and samples are enrolled into contingent yet highly mobile risk estimates, eventually becoming significant to public health. Moreover, this approach can allow insight into the flexible architectures of epidemiological research which, despite their significance in biomedical science, in evidence-based medicine and in health policy, have remained largely unexplored.

Acknowledgements

Research for this paper has been conducted as part of the project 'Danish Biomedicine, 1955–2005: Integrating Medical Museology and the Historiography of Recent Biomedicine' (funded by the Novo Nordisk Foundation) at the Medical Museion, University of Copenhagen. I am grateful to the researchers of the Copenhagen City Heart Study who shared their knowledge and experience with me. For the exploration of collecting practices in epidemiology, the Ischia Summer School on the History of the Life Sciences 2005 provided much inspiration at the early stage of this work. I would like to thank Sniff Andersen Nexø, Jan Eric Olsén, Hanne Jessen, Adam Bencard, Søren Bak-Jensen, Thomas Söderqvist and all participants in the Medical Museion seminars; I thank Signild Vallgårda and Christine Hanke for critical reading of earlier drafts, and Cecily Palmer for help with language revision. Many thanks also to the two anonymous reviewers of this journal for their insightful comments and valuable suggestions.

References

Ahmed, M., & Rahman, N. (2006). ATM and breast cancer susceptibility. *Oncogene, 25*, 5906–5911.

Amsterdamska, O. (2005). Demarcating epidemiology. *Science, Technology, and Human Values, 30*(1), 17–51.

Andersen, T. F., Madsen, M., Jørgensen, J., Mellemkjoer, L., & Olsen, J. H. (1999). The Danish National Hospital Register: A valuable source of data for modern health sciences. *Danish Medical Bulletin, 46*(3), 263–268.

Antoniou, A. C., & Easton, D. F. (2006). Models of genetic susceptibility to breast cancer. *Oncogene, 25*, 5898–5906.

Antoniou, A. C., Pharoah, P. D. P., Easton, D. F., & Evans, D. G. (2006). BRCA1 and BRCA2 cancer risk. *Journal of Clinical Oncology, 24*(29), 3312–3313.

Appadurai, A. (1988). Introduction: Commodities and the politics of value. In A. Appadurai (Ed.), *The social life of things* (pp. 3–63). Cambridge: Cambridge University Press.

Appleyard, M. (1989). The Copenhagen City Heart Study. Østerbroundersøgelsen. A book of tables with data from the first examination (1976–1978) and a five year follow-up (1981–1983). *Scandinavian Journal of Social Medicine Suppl., 41*, 1–160.

Aronowitz, R. A. (1998). *Making sense of illness. Science, society, and disease*. Cambridge: Cambridge University Press.

Bauer, S. (2006). The genomics of environmental response: Re/visions in risk assessment? In A. Bammé, G. Getzinger, & B. Wieser (Eds.), *Yearbook 2006 of the Institute for Advanced Studies on Science, Technology and Society* (pp. 121–138). Munich/Vienna: Profil.

Bech-Danielsen, A. (1998). Pesticider gav brystkræft. *Politiken, 5 December*, Sect. 1, 5.

Bell, D. W., Varley, J. M., Szydlo, T. E., Kang, D. H., Wahrer, D. C., Shannon, K. E., Lubratovich, M., Verselis, S. J., Isselbacher, K. J., Fraumeni, J. F., Birch, J. M., Li, F. P., Garber, J. E., & Haber, D. A. (1999). Heterozygous germ line hCHK2 mutations in Li-Fraumeni syndrome. *Science, 286*, 2528–2531.

Berkman, L., & Kawachi, I. (Eds.). (2000). *Social epidemiology*. New York: Oxford University Press.

Berlivet, L. (1999). Argumentation scientifique et espace publique: La quête de l'objectivité dans les controversies autour de 'risque de santé'. In B. Francois & E. Neveu (Eds.), *Espaces publics mosäiques: Acteurs, arènes et rhétoriques des débats publics contemporains* (pp. 185–208). Rennes: Presses Universitaires de Rennes.

Berlivet, L. (2005a). 'Association or causation'? The debate on the scientific status of epidemiology, 1945–c.1965. In V. Berridge (Ed.), *Making health policy: Networks in research and policy after 1945* (pp. 43–74). Amsterdam: Rodopi.

Berlivet, L. (2005b). Exigence scientifique et isolement institutionnel: L'essor contrarié de l'épidémiologie française dans la seconde moitié du XXe siècle. In G. Jorland, A. Opinel, & G. Weisz (Eds.), *Body counts: Medical quantification in historical & sociological perspectives* (pp. 335–358). Montreal: McGill-Queen's University Press.

Berridge, V. (2005). *Making health policy: Networks in research and policy after 1945*. Amsterdam: Rodopi.

Bourret, P. (2005). BRCA patients and clinical collectives: New configurations of action in cancer genetics practices. *Social Studies of Science, 35*(1), 41–68.

Breslow, N. E., & Day, N. E. (1980). *Statistical methods in cancer research, Vol. I. The analysis of case-control studies*. IARC Scientific Publications, 32. Lyon: ARC.

Breslow, N. E., & Day, N. E. (1987). *Statistical methods in cancer research, Vol. II. The design and analysis of cohort studies*. IARC Scientific Publications, 82. Lyon.

Brody, J. G., & Rudel, R. A. (2003). Environmental pollutants and breast cancer. *Environmental Health Perspectives, 111*(8), 1007–1019.

Brown, P., McCormick, S., Mayer, B., Zavestocki, S., Morello-Frosch, R., Gasior Altmann, R., & Senier, L. (2006). 'A lab of our own': Environmental causation of breast cancer and challenges to the dominant epidemiological paradigm. *Science, Technology, and Human Values, 31*, 499–536.

Cambrosio, A., Keating, P., Schlich, T., & Weisz, G. (2006). Regulatory objectivity and the generation and management of evidence in medicine. *Social Science and Medicine, 63*(1), 189–199.

Chadarevian, S. de. (1998). Following molecules: Hemoglobin between the clinic and the laboratory. In S. deChadarevian & H. Kamminga (Eds.), *Molecularizing biology and medicine: New practices and alliances, 1910–1970s* (pp. 171–201). Amsterdam: Harwood.

Chen, S., Iversen, E. S., Friebel, T., Finkelstein, D., Webe, B. L., Eisen, A., Peterson, L. E., Schildkraut, J. M., Isaacs, C., Peshkin, B. N., Corio, C., Leondaridis, L., Tomlinson, G., Dutson, D., Kerber, R., Amos, C. I., Strong, L. C., Berry, D. A., Euhus, D. M., & Parmigiani, G. (2006). Characterization of BRCA1 and BRCA2 mutations in a large United States sample. *Journal of Clinical Oncology, 24*(6), 863–871.

Copenhagen City Heart Study. (n.d.). http://www.copenhagencityheartstudy.dk/sider/artikler.htm. (Accessed 15 May 2008)

Daston, L. (Ed.). (2005). *Things that talk*. Cambridge MA: Zone books.

Daston, L., & Galison, P. (1992). The image of objectivity. *Representations, 50*, 81–128.

Dawber, T. R., Meadors, G. F., & Moore, F. E. Jr., (1951). Epidemiological approaches to heart disease: The Framingham Study. *American Journal of Public Health, 4*(3), 279–281.

Deleuze, G., & Guattari, F. (1988). *A thousand plateaus*. London: Athlone.

Desrosières, A. (1998). *The politics of large numbers: A history of statistical reasoning*. Cambridge MA: Harvard University Press.

Dørge, H. (1998). En giftig sag. *Weekendavisen, 23 October*, section1, 4.

Einarsdottir, K., Rosenberg, L. U., Humphreys, K., Bonnard, C., Palmgren, J., Li, Y., Li, Y., Chia, K. S., Liu, E. T., Hall, P., Liu, J., & Wedren, S. (2006). Comprehensive analysis of the ATM, CHEK2 and ERBB2 genes in relation to breast tumour characteristics and survival: A population-based case-control and follow-up study. *Breast Cancer Research, 8*, R67.

Ejlertsen, B., & Gerdes, A.-M. (2007). Arvelig brystkræft: Behandling og forebyggelse. *Ugeskrift for Læger, 169*, 2972.

ERICA Research Group. (1988). The CHD Risk-Map of Europe: The 1st Report of the WHO-ERICA Project. *European Heart Journal*, *9*(*Suppl. I*), 1–36.

Feinstein, A. R. (1985). *Clinical epidemiology: The architecture of clinical research*. Philadelphia: Saunders.

Fischerman, K., & Mouridsen, H. T. (1988). Danish Breast Cancer Cooperative Group (DBCG). *Structure and results of the organization. Acta Oncologica, 27*, 593–596.

Fisher, R. A. (1925). *Statistical methods for research workers*. Edinburgh: Tweeddale Court.

Fisherman, K., & Mouridsen, H. T. (1977). Danish Breast Cancer Cooperative Group (DBCG). *Ugeskrift Læger, 139*(42), 2493–2494.

Foucault, M. (1973). *The birth of the clinic: An archaeology of medical perception*. London: Tavistock.

Frank, L. (2000). Epidemiology: When an entire country is a cohort. *Science, 287*(5462), 2398–2399.

Frank, L. (2003). Epidemiology: The epidemiologist's dream: Denmark. *Science, 301*(5630), 163.

Gaudillière, J.-P. (2005). Introduction: Drug trajectories. *Studies in History and Philosophy of Biological and Biomedical Sciences, 36*(4), 603–611.

'Genpub', University of Copenhagen. (n.d.). http://genpub.pubhealth.ku.dk/postgeno_en/Public_Health_Genomics/. (Accessed 15 May 2008)

Gerdes, A.-M. (2006). Presse, etik og genetisk screening. *Ugeskrift for Læger, 168*(40), 3448.

Hacking, I. (1990). *The taming of chance*. Cambridge: Cambridge University Press.

Hagerup, L., Eriksen, M., Schroll, M., Hollnagel, H., Agner, E., & Larsen, S. (1981). The Glostrup Population Studies collection of epidemiologic tables. Reference values for use in cardiovascular population studies. *Scandinavian Journal of Social Medicine Suppl., 20*, 1–112.

Heesen, A. te., & Spary, E. (Eds.). (2001). *Sammeln als Wissen*. Göttingen: Wallstein.

Henderson, B., Pike, M., Bernstein, L., & Ross, R. (1996). Breast cancer. In D. Schottenfeld & J. F. Fraumeni (Eds.), *Cancer epidemiology and prevention* (pp. 1022–1039). New York: Oxford University Press.

Hill, A. B. (1965). The environment and disease: Association or causation. *Proceedings of the Royal Society of Medicine, 58*, 295–300.

Hill, A. B. (1971). *Principles of medical statistics*. London: The Lancet Limited (First published 1937).

Hine, C. (2006). Databases as scientific instruments and their role in the ordering of scientific work. *Social Studies of Science, 36*(2), 269–298.

Høyer, A. P., Jørgensen, T., Grandjeau, P., Brock, J. W., & Hartvig, H. B. (1998). Organochlorine exposure and breast cancer. *Lancet, 352*, 1816–1820.

Høyer, A. P., Jørgensen, T., Grandjeau, P., & Brock, J. W. (2000a). Organochlorine exposure and breast cancer survival. *Journal of Clinical Epidemiology, 52*, 323–330.

Høyer, A. P., Jørgensen, T., Grandjeau, P., & Hartvig, H. B. (2000b). Repeated measurements of organochlorine exposure and breast cancer risk (Denmark). *Cancer Causes Control, 11*, 177–184.

Høyer, A. P., Grandjeau, P., Jørgensen, T., Brock, J. W., & Hartvig, H. B. (2000c). Organiske klorede forbindinger og brystkræft. *Ugeskrift Læger, 162*(7), 922–926.

Høyer, A. P., Jørgensen, T., Rank, F., & Grandjeau, P. (2001). Organochlorine exposures influence on breast cancer risk and survival according to estrogen receptor status: A Danish cohort-nested case-control study. *BMC Cancer, 1*, 8.

Høyer, A. P., Jørgensen, T., Rank, F., Grandjeau, P., & Hartvig, H. B. (2002). Organochlorines, p53 mutations in relation to breast cancer risk and survival: A Danish cohort-nested case-control study. *Breast Cancer Research and Treatment, 71*, 59–65.

Johansen, H. C. (2003). *Danish population history 1600–1939*. Odense: University of Southern Denmark Press.

Jørgensen, T. (2004). Epidemiologisk forskning gennem 40 år: Befolkningsundersøgelserne i Glostrup—Center for Sygdomsforebyggelse—Forskningscenter for Forebyggelse og Sundhed. *Ugeskrift Læger, 166*(15–16), 1425–1428.

Juel, K., & Helweg-Larsen, K. (1999). The Danish registers of causes of death. *Danish Medical Bulletin, 46*, 354–357.

Kamph, L. (2006). Dansk bioanalytikere, 'Østerbroundersøgelsens biobank atter uvurderlig' (The Copenhagen City Heart Study's biobank once again invaluable). http://www.dbio.dk/neobuilder.20060801103839309000056641.html. (Accessed 15 May 2008)

Kannel, W. B. (2000). The Framingham Study: Its 50-year legacy and future promise. *Journal of Atherosclerosis and Thrombosis, 6*(2), 60–66.

Keating, P., & Cambrosio, A. (2000). Biomedical platforms. *Configurations, 8*, 337–387.

Keating, P., & Cambrosio, A. (2003). *Biomedical platforms: Realigning the normal and the pathological in late-twentieth-century medicine*. Cambridge, MA: MIT Press.

Keys, A., Menotti, A., Aravanis, C., Blackburn, H., Djordevic, B. S., Buzina, R., Dontas, A. S., Fidanza, F., Karvonen, M. J., Kimura, N., et al. (1984). The Seven Countries Study: 2,289 deaths in 15 years. *Preventive Medicine, 13*(2), 141–154.

Krieger, N. (1994). Epidemiology and the web of causation: Has anyone seen the spider? *Social Science and Medicine, 39*, 887–903.

Krieger, N. (2001). Theories for social epidemiology in the twenty-first century: An ecosocial perspective. *International Journal of Epidemiology, 30*(4), 668–677.

Krieger, N., Wolff, M. S., Hiatt, R. A., Rivera, M., Vogelman, J., & Orentreich, N. (1994). Breast cancer and serum organochlorines: A prospective study among white, black, and Asian women. *Journal of the National Cancer Institute, 86*(8), 589–599.

Lamm, G. on behalf of the WHO ERICA study group. (1989). The risk-map of Europe. *Annals of Medicine, 21*(3), 190–192.

Last, J. (1983). *dictionary of epidemiology*. New York: Oxford University Press.

Latour, B. (1987). *Science in action: How to follow scientists and engineers through society*. Milton Keynes: Open University Press.

Latour, B. (1988). *The pasteurization of France*. Cambridge, MA: Harvard University Press.

Latour, B. (1993). *We had never been modern*. Cambridge, MA: Harvard University Press.

Latour, B., & Woolgar, S. (1979). *Laboratory life: The social construction of scientific facts*. Los Angeles: Sage.

Lenoir, T. (1999). Shaping biomedicine as an information science. In T. Bellardo Hahn, M. E. Bowden, & R. V. Williams (Eds.), *Proceedings of the 1998 Conference on the History and Heritage of Science Information Systems. ASIS Monograph Series* (pp. 27–45). Medford, NJ: Information Today.

Loiborg, S., Brøndum-Nielsen, K., Daasnes, C., Melbye, M., Saugmann-Jensen, P., von Linstow, H., Aarø-Hansen, N., & Rasmussen, B. (2002). *Redegørelse om biobanker. Forslåg til retlig regulering af biobanker inden for sundhedsområdet*. Betænkning nr. 1414. Indenrigs- og Sundhedsministeriet. Copenhagen.

Löwy, I. (1998). Immunotherapy of cancer from Coley's toxins to interferons: Molecularization of a therapeutic practice. In S. deChadarevian & H. Kamminga (Eds.), *Molecularizing biology and medicine: New practices and alliances, 1910–1970s* (pp. 249–271). Amsterdam: Harwood.

Magnello, E. (2002). The introduction of mathematical statistics into medical research: The roles of Karl Pearson, Major Greenwood and Austin Bradford Hill. In E. Magnello & A. Hardy (Eds.), *The road to medical statistics* (pp. 95–123). Amsterdam: Rodopi.

Magnello, E., & Hardy, A. (Eds.). (2002). *The road to medical statistics*. Amsterdam: Rodopi.

Mariotti, S., Capocaccia, R., Farchi, G., Menotti, A., Verdecchia, A., & Keys, A. (1981). Differences in the incidence rate of coronary heart disease between North and

South European cohorts of the Seven Countries Study as partially explained by risk factors. *European Heart Journal, 3*, 481–487.

Marks, H. (1997). *The progress of experiment: Science and therapeutic reform in the United States, 1900–1990*. Cambridge: Cambridge University Press.

Matthews, J. R. (1995). *Quantification and the quest for medical certainty*. Princeton, NJ: Princeton University Press.

McMahon, B., Pugh, T., & Ipsen, J. (1960). *Epidemiologic methods*. Boston: Little, Brown & Company.

Melbye, M. (2001). *Registerforskning: Enestående danske muligheder*. Copenhagen: IT- og Forskningsministeriet.

Mendelsohn, A. (1998). From eradication to equilibrium: How epidemics became complex after World War I. In C. Lawrence & G. Weisz (Eds.), *Greater than the parts: Holism in biomedicine, 1920–1950* (pp. 303–331). Oxford: Oxford University Press.

Meulengracht, A., & Madsen, M. (1982). *Registre indenfor sundhedsområdet: En oversigt over registreringssystemer, der kan anvendes i epidemiologisk forskning og i sundhedsplanlægning*. København: Dansk Institut for Klinisk Epidemiologi.

Morabia, A. (2005). *A history of epidemiologic methods and concepts*. Basel: Birkhäuser.

Mortensen, P. B. (2004). Registerforskning i Danmark. *Norsk Epidemiologi, 14*(1), 121–124.

Nevanlinna, H., & Bartek, J. (2006). The *CHEK2* gene and inherited breast cancer susceptibility. *Oncogene, 25*, 5912–5919.

Nielsen, G. (2006). Et tredje brystkræftgen kortlagt. *Berlingske Tidende, 1 August*, Sect. 1, 3.

Nordestgaard, B., Agerholm-Larsen, B., Wittrup, H. H., & Tybjærg-Hansen, A. (1996). A prospective cardiovascular population study used in genetic epidemiology: The Copenhagen City Heart Study. *Scandinavian Journal of Clinical Laboratory Investigation, 56*(Suppl. 226), 65–71.

Oldenburg, R. A., Kroeze-Jansema, K., Kraan, J., Morreau, H., Klijn, J. G., Hoogerbrugge, N., Ligtenberg, M. J., van Asperen, C. J., Vasen, H. F., Meijers, C., Meijers-Heijboer, H., de Bock, T. H., Cornelisse, C. J., & Devilee, P. (2003). The CHEK2*1100delC variant acts as a breast cancer risk modifier in non-BRCA1/BRCA2 multiple-case families. *Cancer Research, 63*(23), 8153–8157.

Olsen, J. H., Mellemkjær, L., & Friis, S. (2004). Fra kræfttælling til cancerregister. *Ugeskrift Læger, 166*(15/16), 1458–1461.

Oppenheimer, G. M. (2005). Becoming the Framingham Study, 1947–1950. *American Journal of Public Health, 95*(4), 602–610.

Pálsson, G. (2005). Of Althings! In B. Latour & P. Weibel (Eds.), *Making things public* (pp. 250–257). Cambridge, MA: MIT Press.

Paneth, N., Susser, E., & Susser, M. (2005). Origins and early development of the case-control study. In A. Morabia (Ed.), *A history of epidemiologic methods and concepts* (pp. 291–311). Basel: Birkhäuser.

Parascandola, M. (2005). Epidemiology in transition: Tobacco and lung cancer in the 1950s. In G. Jorland, A. Opinel, & G. Weisz (Eds.), *Body counts: Medical quantification in historical and sociological perspectives* (pp. 226–248). Montreal: McGill-Queen's University Press.

Parry, B. (2004). *Trading the genome: Investigating the commodification of bio-information*. New York: Columbia University Press.

Parthasarathy, S. (2005). Architectures of genetic medicine: Comparing genetic testing for breast cancer in the USA and the UK. *Social Studies of Science, 35*(1), 5–40.

Pearce, S. (1995). Collecting as a medium and message. In E. Hooper-Greenhill (Ed.), *Museum, media, message* (pp. 15–23). London: Routledge.

Pickering, A. (1995). *The mangle of practice: Time, agency & science*. Chicago: University of Chicago Press.

Pickstone, J. (2000). *Ways of knowing*. Manchester: Manchester University Press.

Pickstone, J. (2007). Working knowledges before and after circa 1800: Practices and disciplines in the history of science, technology and medicine. *Isis, 98*, 489–516.

Porter, T. M. (1995). *Trust in numbers: The pursuit of objectivity in science and public life*. New Jersey: Princeton University Press.

Porter, T. M. (2000). Life insurance, medical testing, and the management of mortality. In L. Daston (Ed.), *Biographies of scientific objects* (pp. 226–246). Chicago: University of Chicago Press.

Rabinow, P. (2003). *Anthropos today: Reflections on modern equipment*. Princeton: Princeton University Press.

Rheinberger, H.-J. (1997). *Towards a theory of epistemic things: Synthesizing proteins in the test tube*. Stanford: Stanford University Press.

Rose, N., & Novas, C. (2005). Biological citizenship. In A. Ong & S. J. Collier (Eds.), *Global assemblages: Technologies, politics, and ethics as anthropological problems* (pp. 439–463). Oxford: Blackwell.

Rothman, K., & Greenland, S. (1998). *Modern epidemiology*. Raven: Lippincott.

Schnohr, P., Jensen, G., Lange, P., Scharling, H., & Appleyard, M. (2001). The Copenhagen City Heart Study, Østerbroundersøgelsen: Tables with data from the third examination 1991–1994. *European Heart Journal, 3*(Suppl. H), H1–H83.

Shostak, S. (2005). The emergence of toxicogenomics: A case study of molecularization. *Social Studies of Science, 35*(3), 367–403.

Snedeker, S. (2001). Pesticides and breast cancer risk: A review of DDT, DDE, and dieldrin. *Environmental Health Perspectives, 109*(Suppl. 1), 35–47.

Storm, H. (1991). The Danish Cancer Registry: A self-reporting national cancer registration system with elements of active data collection. In O. Jensen, D. Parkin, R. MacLennan, C. S. Muir, & R. G. Skeet (Eds.), *Cancer registration: Principles and methods* (pp. 220–236). IARC Monograph No. 95. Lyon: International Agency for Research on Cancer (IARC).

Strasser, B. (2006). Collecting and experimenting: The moral economies of biological research, 1960–1980s. In S. de Chadarevian, & H.-J. Rheinberger (Eds.), *History and epistemology of molecular biology and beyond: Problems and perspectives* (pp. 105–123). Preprint, 310. Berlin: Max Planck Institute for the History of Science.

Susser, M. (1991). What is a cause and how do we know one? *A grammar for pragmatic epidemiology. American Journal of Epidemiology, 133*(7), 635–648.

Susser, M. (1999). Should the epidemiologist be a social scientist or a molecular biologist? *International Journal of Epidemiology, 28*(5), S1019–S1022.

The CHEK2-Breast Cancer-Case Control Consortium. (2004). CHEK2*1100delC and susceptibility to breast Cancer: A collaborative analysis involving 10,860 breast cancer cases and 9,065 controls from 10 studies. *American Journal of Human Genetics, 74*, 1175–1182.

The CHEK2-Breast Cancer Consortium. (2002). Low penetrance susceptibility to breast cancer due to CHEK2*1100delC in noncarriers of BRCA1 or BRCA2 mutations. *Nature Genetics, 31*, 55–59.

Trostle, J. A. (2005). *Epidemiology and culture*. New York: Cambridge University Press.

Tunstall-Pedoe, H. for the WHO MONICA project (Ed.). (2003). *MONICA Monograph and multimedia sourcebook*. Geneva: World Health Organization.

Unger, M., & Olsen, J. (1980). Organochlorine compounds in the adipose tissue of deceased people with and without cancer. *Environmental Research, 23*(2), 257–263.

Unger, M., Kiaer, H., Blichert-Toft, M., Olsen, J., & Clausen, J. (1984). Organochlorine compounds in human breast fat from deceased with and without breast cancer and in a biopsy material from newly diagnosed patients undergoing breast surgery. *Environmental Research, 34*(1), 24–28.

Unger, M., Olsen, J., & Clausen, J. (1982). Organochlorine compounds in the adipose tissue of deceased persons with and without cancer: A statistical survey of some potential confounders. *Environmental Research, 29*(2), 371–376.

Vahteristo, P., Bartkova, J., Eerola, H., Syrjäkoski, K., Ojala, S., Kilpivaara, O., Tamminen, A., Kononen, J., Aittomäki, K., Heikkilä, P., Holli, K., Blomqvist, C., Bartek, J., Kallioniemi, O. P., & Nevanlinna, H. (2002). A CHEK2 genetic variant contributing to a substantial fraction of familial breast cancer. *American Journal of Human Genetics, 71*(2), 432–438.

Vallgårda, S. (2007). Public health policies: A Scandinavian model? *Scandinavian Journal of Public Health, 35*(2), 205–211.

Vineis, P. (1997). Proof in observational medicine. *Journal of Epidemiology and Community Health, 51*, 9–13.

Vineis, P. (2005). Causality in epidemiology. In A. Morabia (Ed.), *A history of epidemiologic methods and concepts* (pp. 337–349). Basel: Birkhäuser.

Waldby, C. (2002). Stem cells, tissue cultures and the production of biovalue. *Health: An Interdisciplinary Journal for the Social Study of Health. Illness and Medicine, 6*(3), 305–323.

Weischer, M., Bojesen, S. E., Tybærg-Hansen, A., Axelsson, C. K., & Nordestgaard, B. G. (2006). Svar (Debat). *Ugeskrift for Læger, 168*, 3448.

Weischer, M., Bojesen, S. E., Tybærg-Hansen, A., Axelsson, C. K., & Nordestgaard, B. G. (2007). Increased risk of breast cancer associated with CHEK2*1100delC. *Journal of Clinical Oncology, 25*(1), 57–63. (Epublished 31 July 2006)

Weisz, G. (2005). From clinical counting to evidence-based medicine. In G. Jorland, A. Opinel, & G. Weisz (Eds.), *Body counts: Medical quantification in historical & sociological perspectives* (pp. 377–393). Montreal: McGill-Queen's University Press.