Context and Health

From “Digital Ethology: Human Behavior in Geospatial Context,”
edited by Tomáš Paus and Hye-Chung Kum. Strüngmann Forum Reports, vol. 33,
Julia R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 978026254813

From “Digital Ethology: Human Behavior in Geospatial Context,”
edited by Tomáš Paus and Hye-Chung Kum. Strüngmann Forum Reports, vol. 33,
Julia R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 978026254813

11

Integrating Knowledge from Individual- and Aggregate-Level Data

Sven Sandin

Abstract

Modern technologies and societal changes have generated vast amounts of data, personal and individual or aggregated in clusters or geographic regions. Even though this development has stimulated a wealth of research aimed at understanding disease etiologies and promoting lifestyle changes, opportunities remain, and the integration of data is underutilized.

This chapter describes how geographic and aggregate-level data, with information about environmental and social exposures, can be combined with individual-level health data to increase our understanding of disease etiologies. With an emphasis on data primarily available in Nordic countries, it provides a summary of data sources, references for further reading, approaches and methods for analyses, legal aspects, and limitations.

Compared with data at the individual level, analysis of data at the aggregate level has many advantages in terms of access and privacy. Nonetheless, because the availability of individual-level data is the main strength of data from the Nordic countries, the summary starts with a description of these data and ends with aggregate and geographical (area-level) data. Note that in the Nordic countries, all register-based individual-level data can be linked to geographic regions (e.g., hospital, city, county) associated, for example, with place of birth or current residence. The information provided here should be helpful for anyone interested in disease-specific research and public health work to understand better underlying risks and causal paths.

Introduction

In 1943, the national Danish cancer register was created as a national research register, and in the 1950s, the other Nordic countries followed suit; reporting of malignant cancers became mandatory by law. National population registers

From "Digital Ethology: Human Behavior in Geospatial Context,"
edited by Tomáš Paus and Hye-Chung Kum. Strüngmann Forum Reports, vol. 33,
Julia R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 978026254813

and registers for vital statistics combined with unique personal numbers opened the door for population-based epidemiology (Pukkala et al. 2018). In the wake of cancer research and cancer epidemiology, a multitude of different registers and data sources have since been developed and become available for research purposes in the Nordic countries (Laugesen et al. 2021). Today, the Nordic countries (Denmark, Finland, Iceland, Norway, and Sweden) comprise a total population of approximately 27 million. The countries provide unique opportunities for joint health register-based research in large populations with long and complete follow-up, facilitated by shared features such as tax-funded public health-care systems, similar population-based registers, and the personal identity number as a unique identifier of all citizens (Laugesen et al. 2021). Notwithstanding these similarities, joint Nordic data resources remain underutilized in health research, and it should be possible to combine a wider array of data sources and apply modern methods to address research questions with better precision and accuracy. Examples of such data sources include weather data (temperature, rain, humidity, sun hours), pollution and air quality, road traffic and population density in different geographic regions as well as socially informative data (education, income, occupation, work). Furthermore, multigeneration and twin registers can provide information about inherited risks, opening the door for statistical analyses strengthening causal interpretation of results. In all Nordic countries, repositories of official statistics act as hubs linking different data sources through the personal identification number, which is in turn linked to tax records that provide information about geographic location.

The national population data sources available in the Nordic registers are not universally available to any citizen. For behavioral and lifestyle data or phenotype information not provided by national registers, cohort studies can be linked.

Whereas national registers and population-based cohorts are unique in their ability to generate unbiased estimates thanks to the complete (or almost complete) subject selection, other data sources offer methodological challenges, such as case-control studies, case cohorts, and self-selected samples. Consequently, the landscape of data sources has grown exponentially and includes a large variety of different designs, as well as data collected with no a priori design, or a lack of design. And whereas in the past national registers, cohort studies, and special case-control studies have provided undisputed information and knowledge useful for development of health measures, an efficient mapping and utilization of new data sources is required to keep up the pace of discovery. The development of statistical and computational methods, such as artificial intelligence, machine learning, and modern computer processing capabilities, provide useful tools.

With the goal of facilitating the creation of data informative for human health, the purpose of this chapter is to provide an overview of the different data sources available (see also Appendix 11.1), to demonstrate how to find, combine, and share the data, and to identify analytical challenges associated with their use.

Nordic National Registers

In each Nordic country, every citizen has a unique national identification number provided at birth or at immigration. The authorities use this number in all correspondence or registrations to ensure that citizens can be uniquely identified. Tax offices in the Nordic countries keep records on date of birth, emigration, or immigration. In addition, each country has a range of nationwide registers on health-related and other topics relevant for the authorities to monitor. Some of these were established decades ago, whereas others are more recent. As detailed below, all Nordic countries administer medical birth registers where information related to all births, preceding pregnancies, and maternal and perinatal conditions are recorded; patient registers record diagnoses by clinical specialists and vital statistics registers provide information about date of birth, death, immigration, and emigration.

Reporting to many registers (e.g., patient register) is mandatory by law and with few exceptions does not require consent (e.g., smoking during pregnancy in the Medical Birth Registry of Norway). Outside of the national cancer registers, the main purpose is not research but administration, monitoring, and quality assurance. Since personal ID numbers are used in all registrations, information from one register can be linked to information from others. This is permitted for research purposes under special circumstances (see below for description for each country). There are also requirements as to how data can be stored, used, and shared. When those circumstances are met, the researchers can apply for data from the register-keeping authorities. Health registers, for instance, are usually administered by different institutions than registers containing social information. When applications are approved, researchers receive data files containing copies of data they requested. These data usually require a lot of reorganization and cleaning before they can be used for statistical analyses. In addition, when combining data from two or more countries, extensive harmonization work is needed before analyses can be conducted in a similar (or as similar as possible) manner.

Medical Birth Registers

All Nordic countries have nationwide birth registers with complete coverage of live and stillbirths (Table 11.1). This register contains information on infant and maternal characteristics as well as on the pregnancy and delivery. The Swedish and Norwegian registers also collect information on fertility treatments, their indications, and procedures. The midwife or physician overseeing the delivery collects the following data at the hospital or home in the case of

Table 11.1 Overview of Nordic registers, showing the starting year that social and health-care data began to be collected in Finland, Denmark, Norway, and Sweden.

Type of data	Finland	Denmark	Norway	Sweden
Unique personal identifier of all residents	1968	1968	1967	1961
Medical birth register	1987	1973	1967	1973
Cause of death	1971	1970	1951	1961
Inpatient specialist care diagnoses	1969	1977	2008	1987/1973[1]
Outpatient specialist care diagnoses	1998	1995	2008	2001
Primary care diagnoses	2011	—	2006	—
Detailed neonatal specialty care	2005[2]	—	2009[3]	2001[3]
Cancer	1953	1943	1953	1958
Prescribed medicine/drugs	1964	1995	2004	2005
Medical pension and sickness leave (date, diagnosis)	1962/1999[4]	1976	1992	1990
Unemployment and social welfare	1970	1976	1992	1990
Taxable income	1970	1970	1993	1990
Educational attainment	1970	1973	1974	1970s
Occupation	1970	1981	—	1960s
Military draft cognition tests[5]	1982	1957	1970	1951–2010[6]

[1] Nationally, all psychiatric diagnoses from 1973 and somatic diseases from 1987

[2] Birth weight under 1500 gram or born before 32 weeks of gestation

[3] All children admitted to neonatal care

[4] Sickness leave from 1994

[5] Finnish data include personality; Finnish/Norwegian data include physical fitness

[6] Also from 2017 but with very low number summoned and tested

planned home deliveries: maternal height and weight, smoking status, parity and complications during pregnancy or delivery, infant gestational age, weight, length, head circumference, live/dead-born, and malformations and complications at birth.

Patient Registers

The national patient registers (NPR) are similar in the Nordic countries (Table 11.1). Since each Nordic country has a publicly financed health system with equal access, this ensures complete coverage of the population. NPRs include information about a patient's geographic location; the hospital, department, and clinical specialty needed; admission and discharge date; whether the visit was acute, planned, in- or outpatient; the type of diagnosis according to the International Classification of Diseases (ICD) diagnosis as well as surgical and medical procedure codes. Currently, ICD-10 diagnostic codes are used.

NPRs have evolved over the years. In Sweden, for example, its NPR was founded in 1964 but national coverage began only 1987 (except for psychiatric diagnoses when national coverage began in 1973 for inpatient specialist care). From the beginning, only inpatient visits with diagnoses from specialist care were included; diagnoses from outpatient specialist care were added sequentially, county-by-county, between 1999 and 2005. Extensive validation efforts had been made for different diseases with good results. Coverage and reliability vary, however, depending on the type of condition. Acute conditions requiring inpatient care (e.g., myocardial infarction) have better coverage than conditions such as obesity, type 2 diabetes, or subclinical depression and mood disorders which are typically treated by general practitioners.

Drug Prescription Registers

All Nordic countries have nationwide prescription registers that contain information about prescribed and collected drugs coded using the Anatomical Therapeutic Chemical (ATC) classification system. ATC has five levels: The first level indicates anatomical main group and contains 14 codes (e.g., N = nervous system and C = cardiovascular). The second level indicates the therapeutic subgroup. Levels three to five indicate finer details that describe chemical and pharmacological subgroups. The last level contains 5,067 codes. Even though there is information about drug dosage, the dose information is entered as free text and is therefore difficult to use. Limitations include the lack of information about drugs dispensed in hospitals and over-the-counter drugs. One practical limitation is the lack of data on why the drug was dispensed, which may provide information that helps to avoid biases due to confounding by indication (Catalog of Bias 2018; Greenland and Neutra 1980).

Sweden’s Multigeneration Register

The multigeneration register (Ekbom 2011) is a register administered by Statistics Sweden (SCB) and is comprised of persons who have been registered in Sweden after 1961 as well as those born in 1932 or later. These people are referred to as index persons. The register contains connections between index persons and their biological parents. In 2016, there were about ten million index persons in the register. Information is also collected for certain index persons from older national registration material. For index persons who were adopted, there is also information on their adoptive parents. Currently, there are about 150,000 index persons with information on adoptive mother or adoptive father. Thus, pedigree information on a child, mother, father, maternal, and paternal grandparents is available, and information about siblings (full, maternal and paternal half siblings), cousins (of different types), and aunts and uncles can be derived. This information on pedigrees has allowed family studies separating inherited risk from the environment without the need for genetic data. It has also the additional strength of capturing the entire inherited genetic information, whereas genome-wide association studies (GWAS) capture only a fraction (Bai et al. 2019, 2020). Another important use of this data source is analyses adjusting for family confounding; that is, factors related to the family cluster (including genes) and not the individual per se (also unobserved factors). For example, one study estimated the relative risk for individuals in the lowest Swedish income quintile of being convicted of violent criminality, compared with the highest quintile, to be a sevenfold increased risk. When adjusted for (unmeasured) family risk factors, the risk difference disappeared (Sariaslan et al. 2014). In another study, offspring exposed to higher levels of smoking during pregnancy had greater rates of severe mental illness rates than did unexposed offspring. This study failed, however, to find support for a causal effect of smoking when adjusting for (unobserved) family risk factors (Quinn et al. 2017).

In the other Nordic countries, the mother–child information from medical birth registers and information about the father can be used to derive similar information (Bai et al. 2019).

Cause of Death Register

All Nordic countries have cause of death registers, which include information about date and place of death, cause of death, and whether the death was natural, an accident, or suicide (Brooke et al. 2017; Helweg-Larsen 2011; Norwegian Institute of Public Health 2022; Statistics Finland 2021; Tolonen et al. 2007). All registers were founded before 1970.

Registers Informative for Social Exposures

All Nordic countries administer national registers for education, work and unemployment, occupation, income and taxation, housing, and other social factors. One register example is LISA (Longitudinal Integrated Database for Health Insurance and Labour Market Studies) in Sweden, with similar databases available in the other countries.

Created by SCB (Ludvigsson et al. 2019), LISA integrates existing data from the labor, education, and social sectors with the goal of enabling analysis and evaluation in the field of health/illness. LISA currently comprises 28 vintages and covers the period from 1990 to 2017. The database is expanded with a new vintage every year, with a delay of about 15 months, and is longitudinal: data for the same person can be linked for all years the person is in the population. Between 1965 and 1990, an extensive survey was sent out every five years to all inhabitants of Sweden, and this information is also linked to LISA. This detailed questionnaire, completed by all citizens, provided information about work and type of occupation as well as information on the conditions and standards of living. LISA includes data on yearly income and taxation, the highest level of education attained, occupation, number of days unemployed, income due to unemployment, early retirement, marital status, disposable income, number of children of different ages in a household, and the European Socioeconomic index created from the International Standard Classification of Occupations (ISCO).

Country-Specific Procedures for Data Access

Norway

In Norway, the use of register data for medical research is regulated by the General Data Protection Regulation (GDPR), the Health Research Act, the Health Registry Act, and the Statistics Act. In addition, most health registers have their own specific regulations.

In general, the use of health-related information for research purposes requires informed consent from the participant, yet information reported to the national health registries is confidential and reported without consent requirements. Therefore, the use of individual-level health-related information for research requires the approval and exemption from confidentiality from a Regional Committee for Medical and Health Research Ethics. Application to the ethics committee must include a project description that specifies the project aims and justifies the need for new knowledge, along with details on the planned data linkages and reasons why this information is needed to conduct the project. The application must also describe who will have access to data and how data will be stored. After acceptance, if someone not mentioned in the original application needs to have access to data, an amendment must be submitted.

Anonymized data (i.e., data which cannot be traced back to an individual living person) from the health registers (even linked between registers) can be used freely without applying for ethical approval. In such cases, the registry-keeping authorities are responsible for ensuring that the data provided to the researcher are "truly anonymized" (i.e., the data are indeed impossible to trace back to an individual) as judged by the responsible Norwegian authorities.

Statistics Norway administers data on education, income, social, and work-related information. The Statistics Act forbids any individual-level data from Statistics Norway to be stored in countries other than Norway. This severely constrains the use of Norwegian data in international research.

In practice, analyses involving such data must be carried out in Norway, and only the results can be shared. Researchers at an approved research institution or body within the EU/EEA may, as an exception, be granted access to indirectly identifiable data (pseudo anonymized) from the health registers. In its assessment, Statistics Norway places importance on measures to address the increased risk of data processed outside Norway's jurisdiction. In such cases, requirements are generally set for a specially adapted agreement with the foreign research institution/authority to ensure that Norwegian rules of law are applied and that a Norwegian legal venue is established (Statistics Norway 2022).

In practice, data may be shared in a common repository if there is no possibility to extract raw data on individuals and there is strict control of access to data. This "human restriction" security level includes contracts with register-keeping authorities and usually involves very few analysts (ideally only one for each study). This person is known and selected by the data processor who also ensures the competency level for the data processing. Together with the technical solution (SSH tunnel and time-limited certificates), this guarantees data protection.

Finland

In Finland, register data can be used for secondary purposes, including medical research, according to the Act on the Secondary Use of Health and Social Data (552/2019), the Personal Data Act, and the Act on the Openness of Government Activities. Other associated laws include the Statistics Act, the Act on National Personal Records Kept under the Health Care System, and the Medical Research Act. The Data Protection Ombudsman guides and controls the processing of personal data and provides related consultation.

The general principle regarding medical research is that whenever possible, non-individual-level data is preferred by the authorities (as stated in the Personal Data Act). If individual-level information is needed for research, informed consent is requested from the participants whenever possible. If getting consent is not possible, for example, due to a high number of individuals in the dataset (as is often the case in register studies) or because historical data is needed, a permission for research can be requested from the Health and Social Data Permit Authority (Findata) or, in some cases, directly from the authority keeping the register. Consent is always needed if register data are linked, for example, with survey data. If there is a need to combine data from the registers of multiple owners or obtain data from private social welfare and health-care service providers, the permits are issued by the Findata authority. If data are needed from a single register owner, the authority that oversees that register takes final responsibility for all research use of their data.

In principle, if a study uses only register-based information, an approval of an ethics committee is not required by law. In practice, however, research institutions where the study is conducted can require ethics committee approval for all studies conducted by that institution. Medical studies using register data usually apply for a statement from the regional ethics committee in the hospital district. In Finland, as in Norway, application to the ethics committee must include a project description/research plan specifying its aims and detailing planned data linkages, as well as an explanation as to why this information is needed to carry out the project.

As with the application to the ethics committee, the application for a data permission must include a data utilization plan, a list of individuals who will process the data, and a description of the requested data. If someone not mentioned in the original application needs to have access to data, an amendment must be submitted. Data from health registers can be shared with research collaborators in other countries if data security is sufficiently high. This applies primarily to collaborators in Europe (EU and EEA countries). Data sharing outside Europe is much more strictly regulated.

In most cases, remote access to pseudonymized data is granted. Identifiable data can be delivered to researchers in some restricted cases, if data security is sufficiently high; for instance, if the researcher already has the identification numbers (e.g., own cohort) or if the researcher will link additional data to the dataset (e.g., medical records from the hospitals). Permission and processing of the register data for research purposes is liable to charges.

Sweden

In Sweden, research using the Swedish registers requires affiliation with a university and approval from the Swedish Ethical Review Authority (2023). The registers are primarily administered by three government bodies: SCB, The National Board of Health and Welfare (*Socialstyrelsen*, or SOS), and the Swedish tax agency. As of 1947, all Swedish citizens are assigned a unique personal identification number at birth, which makes it technically possible to link all governmental registers. In research, the personal ID number is always replaced by a random identifier by the register holder for privacy reasons. To request data for research purposes from a national register, an ethics permission is needed. Approval is not, however, sufficient to enable access to the register data; each authority alone decides on what information can be provided to the applicant. After approval from the national ethics board, a lawyer at each register reviews and approves the use of data through a process that does not need to take research aims into consideration. Their goal is solely to protect the privacy of individual Swedish citizens, based on regulations to which the respective authorities are subject. When ordering data for research purposes, a main responsible person is usually assigned at either Statistics Sweden or the National Board of Health and Welfare to coordinate the activities linking the different registers and selecting the appropriate records. This work will be charged to the researcher ordering the data.

The National Board of Health and Welfare (SOS) is a government agency under the Ministry of Health and Social Affairs. The primary register for medical research is the NPR, which contains records of all visits to a clinical specialist; nationwide inpatient care since 1987 (1973 for psychiatric diagnoses). Outpatient specialist diagnoses are available in the patient register between 1999 and 2005 for different counties. All diagnoses are recorded using ICD 7, 8, 9, and 10. SOS is also responsible for the cause of death and the cancer registers. It is not the policy of SOS to provide individual-level data to researchers outside Sweden and the EU/EES. Instead, they advise researchers from other countries to cooperate with colleagues affiliated with a Swedish university, to whom SOS can provide data according to standard legal provisions and procedures. Over the last few years, the Swedish government has invested in health registers, which has resulted in the creation of the National Quality Registries. The National Quality Registries have been built up by dedicated health-care professionals with the aim of monitoring the outcome of specific health conditions (e.g., breast cancer, psychiatry, heart disease). The objective has been to generate valuable knowledge to improve health care and support research.

SCB is the Swedish government agency responsible for producing official statistics in Sweden. It is the holder for registers of vital statistics (date of birth, death, immigration and emigration), for education, as well as social measures. SCB collects, supports, and coordinates official statistics. It produces statistics from many subject areas with different kinds of geographic divisions, such as county, municipality, partial areas, and postal code areas. The products are developed by Statistics Sweden as commissioned work. In Sweden, data for individual respondents (microdata) are protected by the Secrecy Act. It is, however, possible for researchers to apply for access to microdata for use in specified research projects. The system for researchers' access to microdata stored at Statistics Sweden is called Microdata Online Access (MONA). Data are described through Statistics Sweden's standard system for documentation of microdata. Information about MONA and the documentation is published on the website in Swedish. The SCB longitudinal database LISA contains individual data on sickness, parental, and unemployment insurance.

Denmark

In Denmark, there are two main owners of data from national registers: Statistics Denmark and the Danish Health Data Authority. As public authorities and data processors, both are subject to Danish laws for treatment of personal data, including the Act on Processing of Personal Data and the Danish Act of Health.

Statistics Denmark manages data registers on the total population, including information on various demographic factors and social conditions. To obtain access to data from Statistics Denmark, a research project must be associated with a Danish public research unit. Furthermore, the Danish Data Protection Agency must approve the research project if data are linked to data from other authorities or registers. If data from Statistics Denmark are linked with data from the Danish Health Data Authority, approval from the Danish Health Data Authority is also required. Subsequently, Statistics Denmark extracts data from the registers and places all the data on a server at Statistics Denmark (EIT Health Scandinavia 2022).

The Danish Health Data Authority is the supreme authority of health care in Denmark and is part of the Ministry of Health. The Danish Health Data Authority is responsible for all health registers, including the medical birth register, the cause of death register, and NPR. To access data from the Danish Health Data Authority, the Danish Data Protection Agency must approve the research project; if the research project includes direct contact with humans or human biological material, approval must also be obtained from the National Committee on Health Research Ethics. Over the Scientific Service of the Danish Health Data Authority, researchers can obtain access to these data in a safe IT environment, known as "the Research Machine" (*Forskermaskinen*) (Sundhedsdatastyrelsen 2022). The Research Machine allows remote access to most health registers in a secure environment; it requires a personal user ID and two-factor login and no remote access. The user is allowed to use email to send out results from the Research Machine but may send individual-level data.

Aggregate-Level Data

While individual-level data provide the most precise information on individuals, aggregate-level data offers valuable insight. For instance, different *occupations* are often associated with different environmental exposures (e.g., the exposure of workers in sawmills and lumberyards to wood fiber dust). This information can be exploited after individuals are linked to occupational registers. Although individual exposures may vary depending on the exact job task and length of work, such classification can provide important information (Knight et al. 2010) and relate occupation to health outcomes. It is important, however, to adjust for confounding since occupation is strongly linked to education and other socioeconomic factors which are also generally associated with health.

Urbanicity, another type of information defined on an aggregate level, has often been proposed to influence psychiatric outcome and mental illness (e.g., schizophrenia) and is available from national registers and polls as well as from cohorts. For example, SCB offers information from demographic areas (DeSO), using unique codes to indicate nine positions. The first four positions indicate the county and municipality to which an area belongs, as it consists of the county and municipality code. The fifth position is a letter: A, B, or C, which groups the DeSO into three different categories: A is located primarily outside major population concentrations or urban areas; B is mostly located in a population concentration or urban area, but not in the central city of the municipality; C is located in the central part of the municipality (Figure 11.1). In each area, information about age, education, and living conditions is available.

Geographic variations in disease frequency, or exposure (e.g., air pollution), can be used in the search for underlying risk factors. Geographic variations in

From "Digital Ethology: Human Behavior in Geospatial Context,"
edited by Tomáš Paus and Hye-Chung Kum. Strüngmann Forum Reports, vol. 33,
Julia R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 978026254813

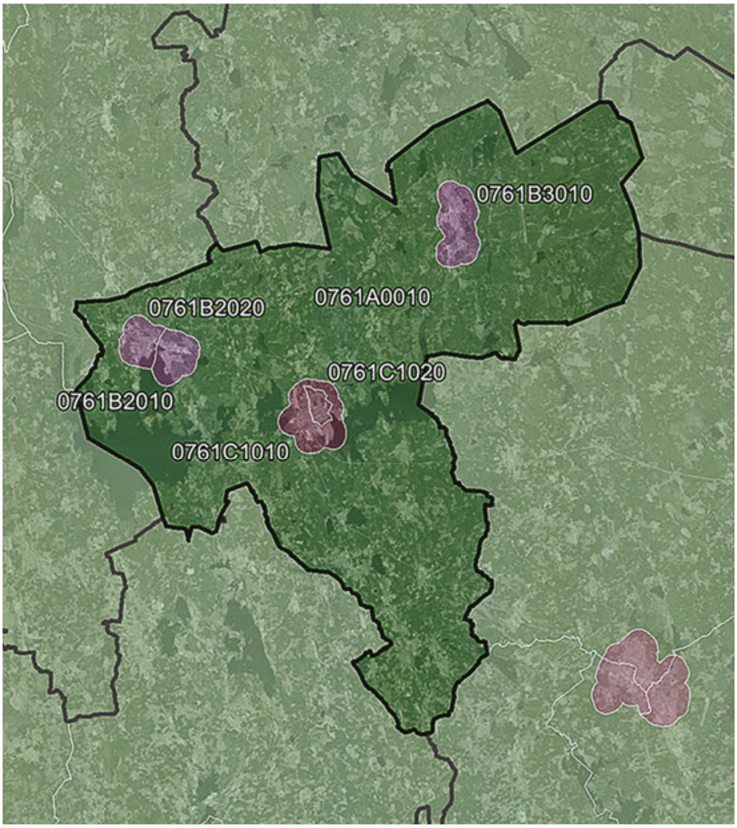

Figure 11.1 Demographic areas coding. Image source: Processing © SCB, other geodata © SCB, Lantmäteriet.

physical environment (e.g., temperature, humidity, wind, and sun exposure) may give insight in health and wellness and may be increasingly important with future changes in the climate (Beauté et al. 2016; Bhopal 1993). Such data is generally available on a geographic regional level from national meteorological institutes. These measures are available across the European Union using the *Nomenclature des Unités Territoriales Statistiques* (NUTS), the hierarchical geographical region classification system (Publications Office of the European Union 2003). The aim was to obtain comparable areas in terms of, for example, surface area and population size in the various EU member states. Introduced in 1988 by EUROSTAT, it is also used by the Nordic countries for area classification, which can be linked to the individual data in the national registers. NUTS can then, in a next step, be linked to national geographic areas such as postal codes (zip codes). Using different units or different definitions of geographic units can result in increased variations in health outcomes; see the study of Legionnaires disease (Beauté et al. 2016).

Combining Heterogeneous Data Sources

For a number of health outcomes, the immense gain in statistical power achieved by pooling research studies has allowed a detailed examination of various relationships, such as attempts to quit smoking, the duration of hormonal replacement therapy after menopause, and combined effects of maternal and paternal age in autism (Collaborative Group on Hormonal Factors in Breast Cancer 1997; Doll et al. 2004; Sandin et al. 2015; Sundström et al. 2019). Naturally, the pooling of studies is constrained by the available exposure data collected in different sites/countries, sometimes several decades after the original study was created. Perhaps more important, pooling studies often requires a reduction of the exposure level to a lowest common denominator. Aligning data measured in many different ways and for different purposes is a challenge and can result in severe oversimplification. When combining several, more or less heterogeneous, data sources, the following issues must be considered.

Selection of Data Sources

Finding relevant data can be a challenge. It is clear that documenting various data collections and data samples in public access databases would facilitate such a task. For population-based studies, the Maelstrom project can serve as a good example (Bergeron et al. 2018). The Maelstrom project administers a database where studies can be registered, the study variables can be mapped to existing variables that facilitate cross-study comparisons, longitudinal measurements are displayed, and the contact information for study principal investigators is accessible.

The lack of generally accepted and utilized variable standard(s) hamper pooling efforts. Thus, harmonization work is repeated for each pooling project, which is a waste of sparse resources. Again, Maelstrom may offer a solution or a start.

To evaluate validity and reliability of collected data, local experts (e.g., clinicians) are usually needed, thus allowing human knowledge to be embedded. For example, a pooling study including longitudinal clinical diagnoses of type 1 diabetes and a second data sample using self-reports can make the overall results impossible to interpret. The clinical diagnoses may change over time as well as the coding system, and changes in the health system may affect ascertainment.

Study Design

Integrating knowledge from different data sources is influenced by the underlying study design. A cohort created as a random sample from a well-defined population has the advantage of allowing many different research questions to be addressed, but other designs may offer different advantages. While cohort studies are often considered easiest to combine, data from different designs should be considered complementary, not competing. Even though much of the criticism of case-control studies is valid, as in biased case selections or lack of a relevant control group, it is not a feature of the design per se. In the Nordic countries, there exists an infrastructure for designing and creating case-control studies with at least the same quality as a prospective cohort study (e.g., where the full population can be enumerated). Strategies for generating new knowledge should be open for inclusion of data from different designs, and data from most designs may be combined using statistical techniques.

Sharing Data

International collaboration as well as data pooling and sharing is key to modern research. Not all collaborators and data sources are positioned within the same legal system. Thus, ways of sharing and combining data must be considered. The most common and, from an analyst's perspective, best way to share data is by *sending the original data*. Encryption in combination with data transfer using secure protocols (e.g., https/TLS) ensure sharing of data with minimal risk of data theft during data transport. Combining all data onto one site optimizes the analytical choices. While it is now difficult to share data between the European Union and the United States (Hallinan et al. 2021), it is possible to share data within each region. In the European Union, data is shared by applying standard agreements for data transfer agreements.

When this is not always possible, more advanced and restricted ways of data sharing must be considered. If the safety concern is related to sharing of individual-level personal and sensitive data, sharing aggregated data may offer an alternative. This, however, comes with restrictions on the analytical tools that are available to analyze the data and will therefore not always fit. A simple example of *aggregated data* is the study of mortality between males and females. For a country of Germany's size, a table of 80 million rows and two columns would be needed yet to calculate the difference in proportion, a 2 × 2 table containing the number of rows where males and females die and survive will suffice. These information lossless measures are called *sufficient statistics* (e.g., for estimating the difference in proportions of dead males and females). Only statistical analyses where sufficient statistics can be derived from aggregate-level data can be performed without losing any information (Hallinan et al. 2021; Persson et al. 2020; Sandin et al. 2006). When the aggregated data is too crude (too high loss of information) for the intended analysis to be executed, simulation approaches may be used. Applying statistical simulation methods made possible by the power of modern computers allows us to "simulate" or *generate a synthetic database* with the same numeric properties as the original data, but where all links to original (individual-level) data have been eliminated (Nowok et al. 2016). Once the synthetic database has been shared and analyzed, the computer code can be sent back to the original data owner and applied to the original (real) data.

For data sharing in larger collaborations across several sites, data federation techniques offer a viable solution to this problem by permitting controlled access to datasets located and managed in disparate locations without the need for permanent storage at a single location (Haas et al. 2002). Under this scenario, each study site retains control of their own data in separate databases at their respective site (Figure 11.2). The GenomEUTwin project stored epidemiological data for around 600,000 twins from across Europe and Australia (Muilu et al. 2007). In the iCARE project—a collaboration of national registers for autism research between Sweden, Denmark, Norway, Finland, Israel, and West Australia—software was developed to share data as well as to analyze data in a central node using data aggregated at each site (Figure 11.2) (Carter et al. 2016).

Depending on legal requirements, an even more privacy protective approach may be applied, such as by using technologies offered by Datashield (Wilson et al. 2017b; Wolfson et al. 2010). Datashield implements a database federation but in combination with statistical computational techniques similar to the aggregated data (above). Here, only minimal statistics are shared to the

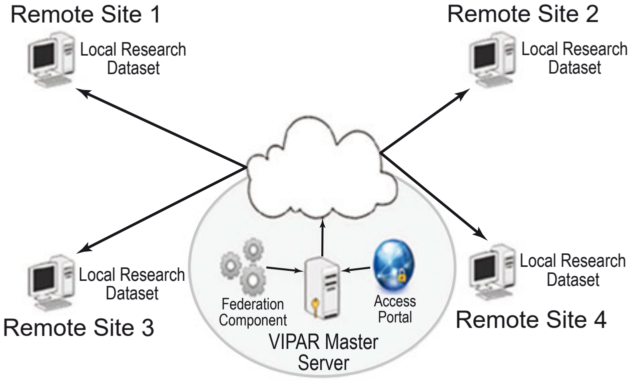

Figure 11.2 Topology of ViPAR (from Carter et al. 2016; https://creativecommons.org/licenses/by/4.0/deed.en). This database application is built around a master server, linked to remote sites. Each site maintains their own data. Analysts access the web-based portal where they run analyses. During analysis, the federation component retrieves data from the sites into computer RAM on the master server where they are analyzed and removed without ever permanently being stored.

central analytical server; individual-level data never leave the local site. As an example, for calculating a linear regression line, only measures (n, $\sum x$, $\sum x^2$) are needed from each site ("sufficient statistics").

Generating Knowledge

Internal and External Validity

While many associations and treatment contrasts can be reliably estimated within single studies, external validity may be less dependable. For instance, even if the relative risk of a health outcome is estimated close to the underlying truth internally in the study, the absolute measures may be biased. This needs to be considered and may be addressed by weighting (Wang et al. 2020).

Confounding needs to be considered as an important topic, both when designing new studies or gathering data from different data sources. Healthy worker effect is such an example. Originally observed in occupational cohort studies, healthy worker effect refers to a situation where people available for and willing to participate in a study tend to be healthier than the target population. This specific form of selection bias usually results in an underestimation of risks, such as for mortality caused by occupational exposures (Naimi et al. 2013).

Replication

One single study, no matter how well designed or implemented, is unable to provide irrefutable evidence regarding the correctness of an association. By using study replication design with independent data samples, the generalizability of results can be addressed as well as the increasing concern of bias and nonreproducibility of results from research studies (Ioannidis 2005; Moonesinghe et al. 2007). This is a current priority of the NIH (National Institute of Dental and Craniofacial Research 2018), which also calls for large population-based studies with contemporary and accurate clinical diagnoses and for studies that can adjust for individual and familial confounding as well as temporal trends.

Knowledge Embedding

Failing to embed properly human knowledge, experience, and empirical knowledge is wasteful. Immediate examples of this include the integration of clinical knowledge about case ascertainment and clinical exposure or known features of health system(s). An analytical example is when applying known genetic correlations in equations (Bai et al. 2019; Svensson et al. 2009) instead of estimating the correlations from the data itself. On the other hand, embedding human knowledge in the wrong way or embedding less solid knowledge could increase both bias and measurement errors.

Challenges

Combining data sources, especially large data with high statistical power, but sometimes limited validity, can lead to false alarms (e.g., warning for spurious association between diet or other environmental exposures and health outcomes). False alarms undermine the credibility of science, move the focus from more important and causally true associations, and increase the anxiety of consumers of the research literature. The reasons for false alarms include badly designed studies, nontransparent (or entirely lacking) analysis plans (often with extensive and ad hoc subgroup analyses), lack of adjustment for multiplicity of statistical tests, and findings uncritically promoted by the investigators. Media attention often worsens the problem when a potentially large proportion of the population may be concerned about a particular exposure. Examples of this include the fear that one might contract brain cancer from using a cell phone use (IARC Working Group on the Evaluation of Carcinigenic Risks to Humans 2013) or that vaccines increase the risk for autism. It took over a decade to ease public concern over media alarms regarding cellular phones (IARC Working Group on the Evaluation of Carcinigenic Risks to Humans 2013). The false claim of autism risk following vaccination has yet to be dismantled in the minds of a large proportion of society and has affected other health outcomes (Madsen et al. 2002). An important yet often neglected consequence of false alarms is that they can undermine efforts to promote healthy lifestyles based on well-established evidence. False alarms increase the risk that the general public will deny all evidence and leave them with a sense that nothing matters.

Given the many rare outcomes and sparse exposures, big data approaches are needed. Geographic disparities as well as temporal trends in disease risk and health markers may indicate the presence of environmental factors. Still, it is a challenge to bring in such innovation, which must be paired with funding, into these and related fields.

Summary and Notes about Future Needs

There are many key issues that need to be addressed in the future:

1. High-quality data do not occur automatically. As researchers, or users of data, we all have a responsibility to *generate new data*. Current research models give too little credit to these issues. After years of planning, generating funding, and collecting quality assurance data, analysts often expect to take first and last place in the list of authors

by arguing they made the scientific contribution. We need a new model to reward the creation and management of new data. New data in a new area should be designed, not as isolated islands, but with the aim and target to combine with other data sources upfront. All new studies want to perform new measures (e.g., new risk scores, new measures of physical activity, rating scales). Each new study, however, should not overlook the importance of reusing existing measures, which would allow the field to connect multiple related studies to facilitate pooling and replication.

2. Study documentation: Projects like Maelstrom should be supported.
3. Legal concern: To solve issues around data access and sharing, more conscious, *brave, and scientifically engaged lawyers are needed as collaborators* to move the field forward. Too often lawyers act in the role of guardians of a company, university, or database. As such, denial of data access is often the first level of defense. For the community, this may result in unethical procedures where accrual and generation of new knowledge is hampered—often contrary to the wish of patients or study participants (Dufva et al. 2021).
4. Privacy and data sharing: In a globalized world where collaborations are key for fast and efficient development, we need to *develop community-based agreements on how to use personal data*. Whereas the European Union has taken one standpoint in strengthening the rights of individual citizens to own and control their personal data, other countries do not agree and have instead adopted laws where governments have the right to all data. This situation seriously hampers collaborations.
5. Methods and competence: There is an urgent need for *advanced statistical methods*, analysts, and software tools to apply these methods to optimize the use of data, targeted for the research question at hand.
6. Data, method, and software: Publications in health research, and other work, should include not only a written description of the analytical approach. In Open Science publications, and for science funded by the NIH, there is often a requirement that data should be made available after publication. While this is a step forward, it is not sufficient. The analytical method should be documented through *publication of the software code* used, and comments on the different steps taken to reach the final conclusions.
7. Replication: We need models and approaches that *encourage replication* and verification of research results. Not only should "new" hypotheses be rewarded; more credit should be given to replication studies. This will allow studies that cannot be replicated to be downplayed and studies which are replicated, but where results cannot be replicated and verified, to be shamed.

Acknowledgments

Part of the country-specific information on accessing register data was taken from work done within the Research on European Children and Adults Born Preterm project (RECAP preterm 2022). The publication by former colleagues at the Karolinska Institutet, Hans-Olov Adami and Nyrén, provided inspiration for future requirements and the need to generate new data (Adami and Nyrén 2016).

Appendix 11.1: Useful Links

Europe

- Infrastructure for spatial information in Europe (INSPIRE): https://inspire-geoportal.ec.europa.eu/
- European Union official statistics (EUROSTAT): https://ec.europa.eu/eurostat
- Data at the World Health Organization: https://www.who.int/data

Sweden

- The National Board of Health and Welfare web page: http://www.socialstyrelsen.se/english
- The Swedish Medical Birth Register: http://www.socialstyrelsen.se/register/halsodataregister/medicinskafodelseregistret/inenglish
- National Patient Register: http://www.socialstyrelsen.se/register/halsodataregister/patientregistret/inenglish
- The Swedish Cancer Registry: http://www.socialstyrelsen.se/register/halsodataregister/cancerregistret/inenglish
- Swedish National Quality Registries, a unique research base: http://kvalitetsregister.se/englishpages/useregistrydatainyourresearch.2251.html
- Ethical aspects of registry-based research in the Nordic countries: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4664438/
- Statistics Sweden (SCB), government of Sweden's bureau for official statistics: https://www.scb.se
- Regional statistical products: https://www.scb.se/en_/Services/Regional-statistical-products/
- Guidance for researchers and universities: https://www.scb.se/en_/Services/Guidance-for-researchers-and-universities/
- Longitudinal integration database for health insurance and labor market studies (LISA): https://www.scb.se/en_/Services/Guidance-for-researchers-and-universities/SCB-Data/Longitudinal-integration-database-for-health-insurance-and-labour-market-studies-LISA-by-Swedish-acronym/
- MONA – *leveranssystemet för microdata*: https://www.scb.se/sv_/Vara-tjanster/Bestalla-mikrodata/MONA/

Denmark

- Statistics Denmark, government of Denmark's bureau for official statistics: https://www.dst.dk/en
- The Danish Health Data Authority: https://www.sst.dk/da
- Overview of Danish health data: https://www.danishhealthdata.com
- Publications for several Danish registers: https://journals.sagepub.com/toc/sjp/39/7_suppl

Finland

- The Act on the Secondary Use of Health and Social Data: https://stm.fi/en/secondary-use-of-health-and-social-data
- Personal Data Act (unofficial translation): http://www.finlex.fi/fi/laki/kaannokset/1999/en19990523.pdf
- Act on the Openness of Government Activities: http://www.finlex.fi/fi/laki/kaannokset/1999/en19990621
- Statistics Act: http://tilastokeskus.fi/meta/lait/tilastolaki_en.html & http://tilastokeskus.fi/meta/lait/2013_tilastolaki_en.pdf
- Medical Research Act: http://www.finlex.fi/fi/laki/kaannokset/1999/en19990488.pdf
- Data Protection Ombudsman: http://www.tietosuoja.fi/en/index.html
- Findata, Health and Social Data Permit Authority: https://www.findata.fi/en/
- Finnish Information Centre for Register Research: https://rekisteritutkimusen.wordpress.com/
- Institute for Health and Welfare (THL): https://www.thl.fi/en/web/thlfi-en (e.g., Medical Birth Register, Hospital Discharge Register, Care Register for Health Care, Register of Primary Health Care visits)
- Statistics Finland, government of Finland's bureau for official statistics: http://www.stat.fi/index_en.html
- Population Register Centre: http://vrk.fi/en/frontpage (data e.g., on address, nationality, mother tongue, and family relations)
- Social Insurance Institution of Finland: http://www.kela.fi/web/en (data e.g., on reimbursed prescription medication purchases and welfare benefits)
- Finnish Cancer Registry: http://www.cancer.fi/syoparekisteri/en/
- Finnish Centre for Pensions: http://www.etk.fi/en/ (data on all old-age and disability pensions)

Norway

- The Regional Committees for Medical and Health Research Ethics: https://helseforskning.etikkom.no/?_ikbLanguageCode=us
- A translated (unofficial) version of The Health Research Act: https://app.uio.no/ub/ujur/oversatte-lover/data/lov-20080620-044-eng.pdf
- Statistics Norway, government of Norway's bureau for official statistics: http://www.ssb.no/en/
- The Norwegian Institute of Public Health (NIPH), which administers the Medical Birth Registry of Norway, the Norwegian Cause of Death Registry,

the Norwegian Neonatal Network (a quality registry for neonatal medicine) and the Norwegian Prescription Database: https://www.fhi.no/en/

- The Norwegian Patient Registry administered by the Norwegian Directorate of Health: https://helsedirektoratet.no/english/norwegian-patient-registry
- The Cancer Registry of Norway: https://www.kreftregisteret.no/en/

From “Digital Ethology: Human Behavior in Geospatial Context,”
edited by Tomáš Paus and Hye-Chung Kum. Strüngmann Forum Reports, vol. 33,
Julia R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 978026254813