Chapter 1
Electronic Health Data Categories

The evolving landscape of healthcare technology has given rise to various types of data, each serving a unique purpose in practice. The definitions of the most commonly used categories of health data are presented in Table 1.1, while the descriptions available throughout the chapter show their diversity and significance, mainly in advancing healthcare research and practice.

1.1 Real-World Data

1.1.1 Real-World Data: Background

Real-world data (RWD) are data related to patient health status and/or the delivery of healthcare and are routinely collected from a variety of sources. Examples of RWD include data derived from electronic health records, medical claims data, data from product or disease registries, and data gathered from other sources (such as digital health technologies) that can inform health status [1]. The NICE further extends the definition of RWD in that most data sources are observational (or non-interventional), meaning that if any interventions (or exposures) are applied, they are not determined by a study protocol but are a result of decisions made by patients and healthcare professionals. Real-world evidence, on the other hand, is evidence generated from the analysis of RWD that provides information about the usage and associated benefits and risks of a medical product [1]. The NICE further elaborates that RWE can include a large variety of evidence types such as disease epidemiology, health service research, or causal estimation. In RWE studies, routinely collected data, customized data collection, or a combination of the two can be used. Single-arm trials that use RWE sources to create an external control are also considered RWE studies [2].

F. Cascini, *Secondary Use of Electronic Health Data*, SpringerBriefs in Public Health, https://doi.org/10.1007/978-3-031-88497-9_1

Table 1.1 Common health data categories

Electronic health data categories	Definition
Real-World Data (RWD) and Real-World Evidence (RWE)—Sect. 1.1	Real-world data (RWD) refer to patient-related health data (namely data representing patient health status and/or the delivery of healthcare) obtained from a variety of sources such as electronic health records, health insurance claims, patient surveys, product and disease registries, e-health services, etc. real-world evidence (RWE) is derived from RWD and represents the clinical evidence about a medical product's safety and risks or benefits
Administrative health data—Sect. 1.2	Synonyms are "health care utilization data", "administrative health care billing records", "administrative claims data", or simply "claims data". These are data that patients generate by encountering the healthcare system, either while visiting the physician's office, undergoing a diagnostic procedure, being admitted to a hospital, or receiving a receipt of a prescription at a community pharmacy. Their collection is primarily intended for administrative or billing purposes; however, they are also used to study healthcare delivery, benefits, harms, and costs
Biobanking data—Sect. 1.3	Data collected in biobanks which represent large collections of biospecimens connected/linked to relevant personal and health information (such as health records, family history, lifestyle, and genetic information) and their main purpose is for use in health and medical research
Clinical trial data—Sect. 1.4	Refers to data collected during a trial including various data types that are later transformed into analysable datasets to answer specific research questions and generate various publications/reports
Electronic Health Records (EHRs)—Sect. 1.5	An electronic health record (EHR) is a digital patient paper chart that contains health information (such as medical and treatment histories of patients) but also medications, treatment plans, immunization dates, allergies, radiology images, and laboratory and test results
Genomic and proteomic data—Sect. 1.6	These data are derived from genomics and proteomic studies. Genomics performs a systematic study of genes, their functions, and their interactions, while proteomics is focused on studying proteins, protein complexes, their localization, their interactions, and posttranslational modifications
National Registries—Sect. 1.7	National health registries represent health information systems, which are primarily intended to collect, process, and analyse data on diagnosed diseases. National registries are usually focused on specific diseases and include only, for example, diabetic or cancer patients, among others
Person-generated (electronic) health data (PGHD)—Sect. 1.8	Person-generated health data (PGHD) includes health-related data that patients (or family members or other caregivers) create, record, or collect, aiming to help address a health concern. This data type often sits outside the traditional clinical setting but can provide useful insights and give a more comprehensive picture of a patient's health. PGHD data sources mainly today include technology-driven solutions such as wearable devices, mobile apps, and home monitoring systems. Such data include results from home glucose/blood pressure monitoring wearable, fitness tracker data, among others
Social and Behavioural determinants of health—Sect. 1.9	Social and behavioural determinants of health (SBDH) include environmental conditions in which people are born, live, learn, work, play, worship, and age that are found to impact a wide range of outcomes and risks related to health, functioning, and quality of life. These variables can be divided into five main categories: Economic stability; education access and quality; social and community context; neighbourhood and built environment; and healthcare access and quality

In recent years, the generation of RWD has grown rapidly as a result of the immersion of new technologies in healthcare sectors in different scenarios. Because RWD is by nature observational, unstructured, or incomplete, and voluminous, it can be viewed as a suboptimal data type that faces both opportunities and challenges. Different research methods are being used and developed to make use of RWD, including clinical trials, target trial emulation, and applications of machine learning (ML) and artificial intelligence (AI) techniques [3]. Organizations such as the Big Data Steering Group (BDSG), the European Medicines Agency (EMA), and the European Medicines Regulatory Network (EMRN) are working to establish a framework that will facilitate the use of and establish values for RWE [4].

1.1.2 Real-World Data: Advantages and Challenges

RWD offers many advantages since it has a wide range of applications, including research (epidemiology, disease burden, and treatment patterns) and the evaluation of the economic value of medical products [5]. The benefits of reusing these data depend on the context and the type of application; however, there are many advantages, the most important of which include the following outlined below.

Enabling time and cost efficiency in new investigations, for example the implementation of RWE in early drug development stages, can decrease both time and expenses in conducting those clinical trials.

Facilitating data collection, as these data are, in many cases, routinely collected either for primary healthcare purposes or, in other stances, are readily available for use.

Enabling longitudinal analyses of large samples, the collection of these data is occurring at a very fast pace and thus generates data that are high in volume and longitudinal in nature.

Providing information and knowledge about outcomes that are otherwise hard to obtain, RWE can complement randomized-controlled trial (RCT) evidence, since these are not sufficient to provide a complete picture of medical products, as adverse events are more routinely tracked in clinical practice [6]. Therefore, regulatory bodies require that manufacturers collect information about safety in the post-marketing period (known as post-marketing studies), which are RWE studies. Thus, when complementing RCT evidence, RWE addresses the gaps in clinical knowledge [7].

- The use of RWD for secondary use in health research poses several challenges that stem from the inherent nature of these observational data. Unlike data collected in controlled settings, RWD is gathered in real-world scenarios, introducing variability and complexity. Moreover, a significant obstacle arises from the unstructured nature of RWD, which encompasses diverse formats such as texts, imaging, and networks. Inconsistencies abound, as data entry is carried out by different healthcare providers and systems, contributing to the messiness of the information.

- The voluminous and dynamic nature of RWDs, characterized by high-frequency collection, further complicates their management and analysis. Incompleteness is another issue, with critical end points often missing, as RWD was originally not intended for analytical purposes. For example, registry data frequently have limited follow-up points, hindering comprehensive analysis. Biases and measurement errors also afflict RWD, exemplified by selection bias in Internet and mobile device data. In essence, RWDs are characterized by their messy, incomplete, and heterogeneous nature and are compounded by various measurement errors and biases.
- It is well recognized that RWD quality is suboptimal and inconsistent [8–11]; as such, assessing and ensuring data quality becomes a challenging endeavour because of the complexity and heterogeneity of the information. Consequently, compromised data quality jeopardizes result validation, reproducibility, and replicability. Despite these challenges, the imperative for establishing best practices in data quality assessment remains crucial, especially considering that the evidence generated from RWD can form the basis for regulatory decisions impacting millions of lives.
- Furthermore, ethical and privacy concerns are inherent in the use of RWD, as the information often includes sensitive details such as medical histories, disease status, and financial information. The risks to privacy are amplified when linking data from different databases, a common practice in RWD analysis. These challenges underscore the need for rigorous ethical considerations and privacy safeguards in the utilization of RWD, as the implications of data analysis extend beyond scientific research to impact individuals and communities [3].

1.1.3 Real-World Data: Examples of Reuse

1.1.3.1 The Use of RWD for Research Purposes

The RWD plays a pivotal role in the secondary use of health data and has been shown to significantly impact clinical research. One notable application is in supporting randomized-controlled trials (RCTs), especially those focused on rare diseases. RCTs face challenges such as methodological issues, high costs, enrolment difficulties, and prolonged durations [12]. To address these challenges, a solution involves implementing a single-arm experimental or synthetic control design, utilizing historical information or data from EHRs and other sources for the control arm. While this study design is not without controversy, it becomes a viable option when investigating diseases with a well-understood course and predictable, fast, and significant treatment effects [13–15]. Additionally, RWE, particularly that obtained from EHR data, informs and optimizes clinical study designs by identifying unmet clinical needs and facilitating the recruitment of cohorts that would benefit the most from new treatments. It also aids in revising study inclusion criteria and identifying suitable study sites, ultimately supporting patient enrolment and retention [7, 13].

Furthermore, RWE facilitates the identification of the most important variables, streamlining data collection to be more costly and time efficient [13]. In the context of patient enrolment, RWE present in registries enhances the recruitment process, which is particularly crucial in rare disease studies where it can be challenging [16]. A notable example is the ADAPTABLE trial, the first large-scale EHR-enabled clinical trial utilizing EHR data to identify and enrol patients [3]. These examples underscore the multifaceted contributions of RWE in enhancing the efficiency and effectiveness of clinical research methodologies [3].

1.1.3.2 The Use of RWD in Clinical and Regulatory Decision-Making Processes

The secondary use of health data has ushered in a transformative era in supporting critical aspects of clinical and regulatory decision-making, particularly in the realm of medical product approvals. Traditionally, the foundation for new drug approvals rested on data derived from traditional RCTs. However, a noteworthy shift has occurred, with RWD now playing an instrumental role in informing decisions related to new drug approvals, supplementary approvals, and label revisions [17]. A compelling illustration of this paradigm shift can be seen in the case of palbociclib, a CDK4/6 inhibitor initially approved exclusively for the treatment of women with ER+/HER2—breast cancer. In 2019, the approval scope broadened to include men, a decision supported by EHR data documenting its off-label use among male patients [18].

1.1.3.3 The Use of RWD in Supporting Regulation Processes

Moreover, RWD serves as a crucial pillar for supporting European Union (EU) regulations. Current pathways for RWE generation for the EMA encompass diverse sources, including EU countries' in-house databases, the Data Analysis and Real-World Interrogation Network (DARWIN EU) [19], and investigations through the agency's research framework contracts. The results from the utilization of these diverse sources of information have proven invaluable in addressing a spectrum of questions and providing support for decision-making processes across various contexts and procedures. This includes addressing the research needs of key committees such as the Pharmacovigilance Risk Assessment Committee (PRAC), Paediatric Committee (PDCO), Committee for Orphan Medicinal Products (COMP), and the Scientific Advice Working Party (SAWP) [4]. RWE contributes significantly in contexts ranging from safety signals and periodic safety update reports to applications for paediatric investigation plans and waivers, maintenance of orphan designations, and scientific advice [4]. This dual application of RWE underscores its versatile and pivotal role in advancing regulatory frameworks and enhancing decision-making processes in healthcare.

1.2 Administrative Health Data

1.2.1 Administrative Databases: Background

The creation of administrative healthcare databases results from the utilization of healthcare services and payments made for payer or hospital billing purposes. These databases are a collection of large amounts of information from thousands or even millions of patients related to their diagnoses, procedures, resource utilization, and costs or charges [20]. These databases are the result of continuing gathering processes, which are often carried out by healthcare providers, healthcare maintenance organizations, health insurance organizations, and other relevant entities such as the civil registry.

Typically, administrative health databases are categorized as "big data", mainly because of certain characteristics of this type of data, including a large volume of information, high speed of data generation, and wide application fields [21].

In general, data gathering in such databases is not explicitly planned for research purposes. Administrative databases focus on preserving financial and administrative details to serve the interests of medical insurers and providers. In contrast, for example, EHRs are predominantly employed by clinicians to record the clinical conditions of patients. This implies key advantages and disadvantages of administrative databases compared with other data sources; for example, owing to the massive amount of data collection, these data allow for large epidemiologic surveys and precise epidemiologic surveillance, but they may suffer from various forms of information bias and a lack of generalizability of the results [22].

1.2.2 Administrative Databases: Advantages and Challenges

Administrative health data have become a powerful research tool for epidemiologists. Owing to its own nature, strengths and limitations are present within this type of data. A summary is presented in Table 1.2.

Current administrative databases still present issues within several types of studies. This can cause a flaw in the analysis resulting from these data or make the collected data inefficient for certain analysis purposes, limiting their secondary use. For example, in the Lombardia region (Italy), abnormal mortality rates were found due to a change in coding status during a study [25]. In general, the use of billing and coding data in research offers large sample sizes and increased statistical power, enabling multivariate adjustments for risk factors. However, potential limitations arise from the data source (i.e. nonbillable clinical events may not be registered), coding issues, linkages between multiple sources, lack of validation, and confounders not available for assessment. Additionally, variations in the definition of diseases across different hospitals or healthcare structures may introduce observer bias, emphasizing the need for careful consideration of these factors in the various applications of data [26].

Table 1.2 Advantages and limitations of administrative databases

Strengths	Limitations
Very large sample size with a diverse population basis	Lack of additional clinical information due to a bias of use for administrative and billing purposes
Large number of variables and good heterogeneity and accessibility	Selection bias may be present due to consent requirements, etc.
Enabling faster and less expensive research	Potential data gaps due to heterogeneous or changing coding practices in the field
Patients are followed over extended periods of time	Clinical significance of administrative data is not always evident
More comprehensive dataset when properly linked	Misclassification of data may turn into jeopardized study results [24]
Easier identification of rare or reduced patient populations, i.e. rare diseases, etc. [23]	

1.2.3 Administrative Databases: Examples of Reuse

1.2.3.1 The Use of Administrative Health Data in Research

In the realm of secondary use of electronic health data, administrative databases offer valuable resources for diverse research endeavours. Population-based epidemiological studies benefit significantly from these databases, enabling the evaluation of crucial indicators such as incidence, prevalence, and temporal trends in specific diseases or health conditions, along with associated mortality rates. For example, a comprehensive epidemiological study on heart failure was conducted, encompassing 2.1 million inhabitants of Sweden, using data collected posthospital admission [27]. Similarly, trends in diabetes incidence, prevalence, and mortality rates over a decade were explored by employing hospital discharge abstracts and physician service claims [28].

1.2.3.2 The Use of Administrative Health Data in Quality-of-Care Studies

Outcome studies provide a platform to research demographic characteristics and major medical conditions over an extended period, offering a representative sample for the general population, as exemplified in a ten-year study on the incidence and short-term outcomes of acute myocardial infarction in the Netherlands [29]. Administrative health data also provide benefits in quality-of-care studies, facilitating the evaluation of healthcare services and systems. Organizations such as the Agency for Healthcare Research and Quality (AHRQ) have introduced quality indicators on the basis of in-hospital discharge abstracts, which serve as tools for highlighting quality concerns and tracking changes over time [30]. Predictive models can contribute to quality-of-care improvement by identifying high-risk patients for targeted interventions, thus reducing hospitalizations [31].

1.2.3.3 The Use of Administrative Health Data in Detecting Adverse Events

Moreover, administrative databases offer a unique opportunity for adverse drug event studies, given their capacity to study large sample sizes over extended observation periods. Although advantageous, challenges arise from the lack of information on covariates and their potential impact on outcomes, as illustrated by the use of ICD-9-CM codes to detect adverse drug reactions (ADRs), which results in a chart with a positive predictive value ranging between 60% and 100% [32]. These examples underscore the versatility of administrative databases in supporting diverse research domains within the secondary use of electronic health data.

1.3 Biobanking Data

1.3.1 Biobanking Data: Background

The Declaration of Taipei defined "a collection of biological materials and associated data" as a biobanking term was first used by Loft and Poulsen [33]. Since then, the application of these data has expanded due to the advancement of omics science and the possibility of utilizing large databases in the health field. The term "biobanking" is often improperly applied to any collection of human biospecimens, without considering the ethical and legal requirements or the standardization of a variety of processes related to tissue collection. Biobanks represent large biospecimen collections that are linked to relevant personal and health information, such as health records, family history, lifestyle, and genetic information, whose primary aim is utilization in medical and health research [34]. This is aligned with the definition provided by the Organisation for Economic Co-operation and Development (OECD), which defines biobanks as organized resources intended for genetic research, encompassing human biological materials, data derived from genetic analysis, and related information [35]. According to the Declaration of Taipei, a biobank is "a collection of biological materials and associated data", whereas a health database is "a system for collecting, organizing and storing health information" [36].

Biobanking plays a primary role in the new era of precision medicine and has enormous potential but also poses potential dangers, especially in terms of data sensitivity. Biobanks also support both public health and patient care efforts by providing resources for scientific and medical research [37, 38]. As biobanking data are ultimately collected from populations and individuals, concerns about crucial aspects related to human rights, such as dignity, autonomy, privacy, confidentiality, and non-discrimination, among others, increase the importance of promoting advancement in the biobanking field while individual rights are protected [39].

1.3.2 Biobanking Data: Advantages and Challenges

In the realm of secondary use of health data, the utilization of biobanking data offers a myriad of advantages. One significant benefit lies in the potential for a larger sample size and greater data diversity, enabling a more comprehensive understanding of various health conditions. Moreover, this trend is likely to continue with the existence of efficient recruitment methods that facilitate ongoing enrolment; for example, the recruitment target of the Kaiser Permanente and Million Veterans Program (MVP) is over 85,000 participants per year [40]. Furthermore, biobanking offers the ability to include a greater degree of race/ethnic diversity than many clinical research studies do, thereby overcoming the bias introduced by restricting studies to a single race/ethnic group. An example of this effort was made by the MVP, which recruited between 15% and 20% of their sample to be of African ancestry [41]. The benefit of merging these two data sources also lies in the fact that by genotyping the participants only once, there is the possibility of defining multiple case groups, which emphasizes the potential for efficient use of biobank data. Furthermore, given that the EHR data are gathered continuously during each repeated healthcare encounter, there are vast longitudinal data that can facilitate longitudinal follow-up studies focusing on incident events, recurrent events, response to treatments, and transitions in health over time [42]. Although the burden of rare diseases is high (e.g. in the United States), rare disease studies are often less supported than those of common diseases. The obstacle also lies in the need to collect large sample sizes for the genetic discovery of rare diseases. Therefore, linking electronic health records (EHRs) with biobanking increases the potential to identify and recruit rare disease cases and their family members, facilitating the conduct of diagnostic and therapeutic clinical studies. Such large-scale efforts have been made in the United States [43] and the UK [44]. Such large cohorts of genotyped individuals also enable the characterization of a sufficient number of rare genetic variants, which also require a large sample size to obtain accurate results [40]. Moreover, compared with traditional prospective cohort studies, which usually ascertain only a narrow range of phenotypes and assess risk factors once at baseline, EHRs linked with biobanks prove to be cost effective for collecting data on all clinically significant outcomes [40]. Finally, linking EHRs with biobanks enhances the opportunities for utilizing genetic data for personalized medicine, since it can both facilitate the discovery of genetic determinants of disease by conducting observational studies and perform genotype-based interventional studies at scale in the same population [40, 45].

However, their advantages come with their set of challenges. One of the challenges in analysing massive amounts of genetic data is combining biobanks around the globe, which are the basis for conducting comprehensive studies and improving the equity of obtaining genetic data in human genome research. Given that they aim to draw conclusions that should enhance population health globally (such as those related to genetic testing, disease diagnosis, and therapeutic solutions), they should not be based on a biased sample [46, 47]. Thus, to achieve this data diversity at the

geographical level, there is a need for a collective effort from researchers involved in genetic studies that use biobanks around the world. This goal is emphasized by initiatives such as the H3Africa Consortium [48], GenomeAsia100K Consortium [49], and the work of Robine and Varmus [50]. These initiatives underscore the importance of diverse datasets for producing robust and generalizable results.

Despite their great potential, biobanks are also considered dangerous to human rights since they are a source of sensitive data that can be accessed. In this context, the World Medical Association (WMA) published the Declaration of Taipei, which serves as a guideline for collecting, storing, and using identifiable data and biological material that are not intended only for their primary use [36]. Given that health databases and biobanks contain data from individuals and populations, concerns related to dignity, autonomy, privacy, confidentiality, and discrimination have arisen. This declaration outlines the need for harmonization between scientific advancements and the protection of individuals' rights, especially regarding the data obtained from their biospecimens. This is the first international guideline that addressed ethical issues and provided associated directions related to activities associated with human databanks and biobanks [39].

Since biospecimen data are linked with associated personal health data such as those contained in EHRs, issues associated with data quality are also present, such as data (in)completeness, missing data, inaccurate phenotype classification, and selection bias [51, 52]. Furthermore, there is inherent population bias, since people are either willing or not willing to share their data through biobanks, and those who are willing are more likely to self-report white race, higher educational attainment, and lower religiosity [53].

Appropriate informed consent (IC) is highly discussed as a crucial topic for biobanking data [54]. Despite the fact that the GDPR-2018 provides important indications for informed consent guidelines, at the moment, unfortunately, no international consensus on informed consent has been produced, leading to differences and intense debates in the interpretation and implementation of informed consent practices. To date, most biobanks have adopted a "broad consent model", basically meaning that biobanks have the right to use samples collected within a specified timeframe without the need to contact patients. However, other more ethical formats, such as "dynamic consent", also arise as a result of the possibility of using information technology tools for this purpose [55].

Effective data management is imperative for maximizing the utility of biobanking data. The Findability, Accessibility, Interoperability, and Reusability (FAIR) principles have recently been introduced to address issues of interoperability and data harmonization in biobanks [56]. These principles serve as a framework for ensuring that data are easily discoverable, accessible, and reusable across different research endeavours. Further attempts have been made by different authors to extend the FAIR principles and provide further requirements on data quality, specifically in the field of biobanking [57–59].

The international harmonization of biobanks has been a concerted effort over the past two decades. Many countries have established specific legislation or guidelines for the collection and processing of human tissues [60]. The best practices

guidelines for human and microorganism biological resources were published by international entities such as the Organization for Economic Co-operation and Development (OECD) [61], the International Society for Biological and Environmental Repositories (ISBER) [62], the National Cancer Institute (NCI) [63], the International Agency for Research on Cancer (IARC) [64], the Council for International Organizations of Medical Sciences (CIOMS) [65], the Nuffield Council on bioethics [66], and the European Society of Human Genetics [67]. Additionally, many countries have developed legislation and guidelines for human tissue collection and processing at the national and regional levels. Nevertheless, there is a recognized need for internationally harmonized criteria for sample collection, aiming to merge major requirements into a unified global standard for biobanking. This endeavour seeks to standardize information on the collection, access, and research activities in human data and biological resources, fostering collaboration and advancing the secondary use of health data on a global scale. To address this requirement, a working group was organized that comprises more than 70 experts from 29 countries at the ISO Technical Committee "Biotechnologies" and works on an international standard for biobanking (ISO 20387) that was published in 2018; this group relies on 14 national and international guidelines and standards [68]. This overarching standard is applicable to all organizations that are involved in biobanking of human, animal, fungus, and plant specimens, as well as microorganisms for research and development [69].

1.3.3 Biobanking Data: Examples of Reuse

The United Kingdom A prospective cohort study of 500,000 participants aged 40–69 years was produced starting with baseline assessments from 2006 to 2010 with the aim of researching the lifestyle, environmental and genomic determinants of life-threatening and disabling diseases of middle and old age. The data from participants include blood and urine biomarkers, as well as a series of web-based questionnaires and ongoing assessments of multimodal imaging and cardiac monitoring performed on target populations [70]. Participants have now been followed up for over a decade, with more than 52,000 incident cancer cases recorded and over 26,000 researchers worldwide currently using the data, positioning the UK Biobank to transform the understanding of cancer [69].

In the UK, a smaller study performed using data from 4509 tests explored the contributions of demographic, social, health, and other medical and environmental factors to COVID-19 risk [71]. These different examples highlight the potential adaptability and scalability of biobanking data to suit multiple population samples, objectives, etc.

Canada The Statistics Canada Biobank is an integral part of the Canadian Health Measures Survey. It collects and stores information from questionnaires; physical measures; and whole blood, plasma, serum, buffy coat, and DNA from consenting

Canadians between the ages of 3 and 79 years on an ongoing basis. Canadian researchers can apply to the programme to use biospecimens for research purposes [72].

China The China Kadoorie Biobank survey took place from 2004 to 2008 in ten geographically defined regions. Over 500,000 adults aged 30–79 years were recruited for the collection of data (mean blood pressure, body mass index, diabetes status, etc.) as well as blood samples (99.98% of the population). The final population had food reproducibility but large heterogeneity by age, sex, and study area [73]. Currently, this study represents one of the world's largest prospective cohort studies, with resurveys produced in 2008, 2013, and 2020 [74].

Japan The BioBank Japan (BBJ) project was launched in 2003 as a registry of patients diagnosed with any of the 47 common diseases selected for this purpose. Participants were enrolled from 2003 to 2008 and followed up until 2013. A total of 200,000 participants were registered in the study [75]. Since 2018, BBJ has been operated and managed as a project aiming to promote registered data and samples, with more than 500 papers based on BBJ samples and information published at the time [76].

Estonia The Estonian Biobank cohort is a volunteer-based sample of the Estonian resident adult population (>18 years), representing the largest epidemiological cohort of the Baltic region. The resulting database, which was collected through a network of medical professionals from Estonia (1.3 M population), allows for wide epidemiological and genomic research [77]. Currently, it comprises a cohort of more than 200,000 individuals and contains more than 700,000 single-nucleotide polymorphisms (SNPs), the most common type of genetic variation among people [78].

1.4 Clinical Trial Data

1.4.1 Clinical Trial Data: Background

The progress in transitioning from paper-based to electronic health data systems has also led researchers to increasingly consider different types of electronic data, including clinical trial registries, as important tools for investigation. According to the WHO, clinical trials and clinical intervention studies include any study that evaluates the health-related effects of interventions (such as drugs or medicines, cells and other biological products, surgical procedures, radiological procedures, devices, behavioural treatments, changes to the care pathway, preventive care, or other treatments) on human subjects [79]. The US National Institutes of Health (NIH) defines a clinical trial as a research study that prospectively assigns one or

more human subjects to one or more interventions (which may consist of placebo or other control) to elucidate their impact on health-related biomedical or behavioural outcomes [80].

Pharmaceutical companies and the research community are encouraged to share their data for secondary purposes and thus accelerate scientific discoveries and informed decision-making practices. Since 2013, international organizations such as the US National Institutes of Health (NIH), the Food and Drug Administration (FDA), and the European Medicines Agency (EMA) have developed policies and recommendations for clinical trial data sharing practices [81–83]. Other organizations have also called for improvements, specifically in terms of increased data sharing, with a focus on those generated by publicly funded research. Organizations such as the Organisation for Economic Co-operation and Development (OECD), the European Commission (EC), and the NIH have advocated that "publicly funded research is a public good produced for public interest", which therefore should translate into a responsible but open sharing of information, including clinical research [84]. The sharing of data from clinical research, therefore, can be justified from multiple grounds, including scientific, economic, ethical, etc. [85]. Furthermore, access to data to improve health is strongly related to the right to health and healthcare access. Institutions such as the Medical Research Council (MRC) have promoted principles that aim to determine principles for better use of clinical trial data in the future. Within these principles, transparency is presented as one of the key pillars to aim for research findings that will be as open, understandable, and reproducible as possible [86]. One way to ensure transparency in conducting clinical trials is prospective registrations of study protocols, which can be found in clinical trial registries, such as ClinicalTrials.gov. These registries represent another relevant and important source of clinical trial data and contain a vast number of metadata, including data characteristics from trials, protocols, funding, scientific leadership, etc. [87].

The European Health Data Space (EHDS) Regulation[1] includes data from clinical trials, clinical studies, and clinical investigations, which are subject to Regulation (EU) 2014/536, Regulation (EU) 2024/1938, Regulation (EU) 2017/745, and Regulation (EU) 2017/746, respectively, among the minimum categories of electronic health data for secondary use (Article 51). According to the EHDS regulation, to protect intellectual property rights or trade secrets (using legal, organizational, and technical measures), these data should be included when the clinical trial or clinical investigation has ended, without affecting any voluntary data sharing by the sponsors of ongoing trials and investigations. The health data access body (HDAB) can assess how to preserve this protection while also enabling access to such data for health data users to the extent possible.

[1] https://health.ec.europa.eu/publications/proposal-regulation-european-health-data-space_en

1.4.2 Clinical Trial Data: Advantages and Challenges

Within the realm of the secondary use of health data, clinical trial data have emerged as a valuable resource, highlighting both advantages and challenges. Leveraging existing data stands as a primary advantage, mitigating the need for redundant research efforts and fostering a more efficient use of resources [85, 86]. The ability to utilize larger sample sizes by aggregating data from multiple trials also enhances the statistical power and reliability of the findings. As already mentioned, larger sample sizes can facilitate the study of rare diseases or conditions that may not be adequately represented in individual trials. Furthermore, pooled data can help identify rare adverse effects or long-term safety concerns that single studies might miss. The reuse of clinical trial data enables the aggregation of information for meta-analyses, facilitating the re-examination and validation, or correction, if necessary, of existing results. Compared with conducting original clinical trials, this process not only enhances the validity of the data but also contributes to cost efficiency and time savings. Furthermore, the opportunity to test new hypotheses emerges as a noteworthy advantage in the exploration of clinical trial data [85, 86]. Finally, clinical trial data have been applied in regulatory and policy decisions by supporting regulatory submissions for new indications or approvals of treatments in different populations. It is also used in healthcare policy decision-making, as pharmacoeconomic analyses and comparative effectiveness research based on reused data can inform healthcare policy and reimbursement decisions. Finally, reusing existing clinical trial data can be applied more broadly to different populations, improving public health outcomes [88–90].

However, challenges persist in the secondary use of clinical trial data, necessitating further refinement of current practices. The complexities inherent in trial design and data sharing demand continuous improvement, posing questions about achieving full optimization. Standardization issues across various recording practices in trials create complexities for secondary analysts, potentially hindering a comprehensive understanding of the generated data. Additionally, a significant time investment is required for data to become fully available for analysis, impacting research efficiency and efficacy. Another important barrier to the secondary use of trial data is the lack of clarity around the availability of data. Notably, trial repositories do not provide sufficient descriptions of the trial setting, objectives, and design, whereas data dictionaries and schemas describing the dataset‘s content are often incomplete [91]. For example, discrepancies were found in reporting death events between trials, with variations in reporting deaths during or after the trial. Such differences may impede the ability to make meaningful comparisons between trials, and the delayed discovery of such information can affect the final interpretation of results [92].

Moreover, challenges specific to ClinicalTrials.gov further complicate the landscape of clinical trial data reuse. Issues such as missing or underspecified data, including outcome measures, study design, and conditions not denoted by Medical Subject Heading (MeSH) terms, pose hurdles for effective information retrieval and utilization [87]. The recognition of these challenges underscores the importance of

ongoing efforts to increase the accessibility, standardization, and completeness of clinical trial data for their successful secondary use in advancing medical research and knowledge.

Further complications are posed by the complexity of study designs, and tackling this issue might be challenging, and this is exaggerated even more when pooling and reusing of data is performed from multiple sources. Pooled analyses are prone to biases due to sample heterogeneity since even slight differences between datasets can introduce such problems. Addressing the inherent problem of heterogeneity when pooling data from multiple studies can rely on carefully designed protocols (very similar to those required for original clinical trials), which again need very detailed information from the original studies that are used for data pooling. Another way to address this problem is by encouraging close collaboration between secondary users and primary clinical trialists to prevent misunderstanding of trial complexities and nuances by secondary users [93]; however, this approach is not always easy to implement. Overcoming these complexities is crucial if secondary analysis is considered an important outcome itself, for which case, trial protocols should therefore improve their standards for clinical trial documentation and design with secondary analysis in mind [92].

1.4.3 Clinical Trial Data: Examples of Reuse

Patient-Level Data Were Used to Analyse Standard-of-Care Medications from Eight Prostate Cancer Clinical Trials. Data from eight prostate cancer clinical trials obtained from the Project Data Sphere® portal were analysed, with a total of 4127 subjects being combined, integrated, and studied. Among the participants, different heterogeneous demographic data, as well as different cancer stages, were present. By combining comparator arms from different trials, investigators were able to identify possible adverse survival outcomes for surgical castration therapy in prostate cancer patients [94].

Using an Integrated Clinical Trials Dataset for the Identification of Endophenotypes of Alzheimer's Disease Progression. Longitudinal patient-level data for 1160 Alzheimer's disease patients receiving placebo or no treatment with a follow-up of 18 months were investigated. At least three subgroups of Alzheimer's disease patients were identified by the results "learned" from clinical data, each of which is distinguished by a deterioration in cognitive function [94].

Optimizing Trial Design by Integrating Model-Based Clinical Trial Simulation, Pharmacoeconomics, and Value of Information. For example, whether clinical trial simulations based on a pharmacometric model can be utilized to generate prior distributions of treatment effects for Bayesian trial design when treatment effects estimated from previous studies are not suitable has been investigated. Several steps involved in simulating trial outcomes for different trial designs that can be used to optimize trial design while maximizing the return on investment were identified. A case study of febuxostat versus allopurinol for treating

hyperuricaemia in gout patients was performed. Different trial design scenarios vary in terms of alternative treatment doses, inclusion criteria, input uncertainty, and sample size. The optimal trial sample sizes varied depending on the uncertainty of the model inputs, trial inclusion criteria, and treatment doses. This interdisciplinary framework for trial design and sample size calculation can be valuable for supporting decisions in later stages of drug development. It can also help identify costly sources of uncertainty, thereby informing future research and development strategies [95].

Using Clinical Trial Data to Conduct Meta-analyses Revealed That Widely Used Treatments Are Not Effective or Safe. The safety of rosiglitazone (a widely used drug in the treatment of type 2 diabetes mellitus) was investigated through a meta-analysis to determine its impact on unexplored outcomes such as cardiovascular morbidity and mortality. After pooling data from 42 randomized-controlled trials, the study revealed an association between the use of rosiglitazone and an increased risk of myocardial infarction (with an odds ratio (OR) of 1.43, 95% confidence interval [CI], 1.03 to 1.98; $P = 0.03$) and an increased risk of death from cardiovascular causes, although with borderline significance (OR = 1.64, 95% CI, 0.98 to 2.74; $P = 0.06$) [96]. The safety of incorporating immune checkpoint blockade into perioperative cancer therapy as a therapy that has proven to be successful in clinical practice is another result. An association between the addition of immune checkpoint blockade to perioperative therapy and an increase in grade 3–4 treatment-related adverse events and adverse events that lead to treatment discontinuation was observed. These results emphasize the need for close safety monitoring in future clinical trials that assess neoadjuvant or adjuvant immune checkpoint blockade therapy [97].

1.5 Electronic Health Records

1.5.1 Electronic Health Records: Background

EHRs represent electronic records of patient health information generated over time whenever the patient has received health or care services. The information included in these records is patient demographics, progress notes, problems, medications, vital signs, past medical history, immunizations, laboratory data, and radiology reports [98]. In addition to containing primary information, data from other hospital information systems are included in an EHR (e.g. imaging data from the radiology department and genomics data from the genetic department). In addition to containment, EHRs can generate reminders for routine screenings and disease reporting or generate graphical trends against various parameters such as blood pressure and blood glucose level monitoring [99].

Since the incentivization schemes of different governments, including the US government, took place, a high adoption rate of these technologies has taken place

in hospitals and health offices, with countries such as the United States experiencing an 80% adoption rate at hospitals, naturally leading to further use of these data for research purposes [100].

Despite this improvement in the digital recording of health data and technology used for this purpose, some factors are still missing, especially when data are considered for secondary use. These include the fact that priorities at the time of registration will be aligned with the clinician's or administrative needs, as well as other factors related to the patient's behaviour during visits, including loss of follow-up, missing information, data inconsistencies, and others [99].

1.5.2 Electronic Health Records: Advantages and Challenges

The reuse of EHRs has emerged as a transformative practice with multifaceted advantages, influencing the clinical, organizational, and societal dimensions of healthcare. At the clinical level, the repurposing of EHRs has substantially improved the quality of care and contributed to the reduction of medical errors, showcasing tangible benefits for patients. From an organizational standpoint, EHR reuse has positively impacted financial and operational performance, underscoring its potential to streamline healthcare processes. Specifically, the use of EHR data was also found to expedite healthcare quality measurements. Unlike standard quality measures, electronic aggregation of patient data can accelerate analysis and lead to better results among standard quality measures. More than 250 studies that were included in a qualitative synthesis reported that the utilization of EHRs led to improved protocol adherence, large-scale clinical monitoring, and decreased adverse medication events [101]. Together with standardized quality measures, EHRs can also facilitate the formulation of new quality measures such as those that are more continuous (e.g. actual blood pressure measures) and not only binary (e.g. when blood pressure checking is performed or when it is below/above a specific threshold) [102]. Moreover, the secondary use of EHRs has facilitated superior research, promising advancements in population health [99]. EHR data reuse can facilitate the design of retrospective studies by using existing data pools. It will also accelerate patient enrolment in prospective studies, thereby substantially reducing the costs associated with patient recruitment [103]. Since EHRs include a wide array of data (such as laboratory data, vital signs, and other clinical parameters), they can also enable patient monitoring by examining a wide array of health outcomes [104]. Compared with traditional means of data collection, the use of EHR data also offers public health benefits such as individual telephone- or paper-based reporting. This enables public health officials to perform rapid monitoring of the spread and emergence of diseases [102]. Finally, EHR data can support decision-making processes related to medical products by facilitating safety monitoring. For example, in the United States, the FDA relies on aggregated EHR data to perform safety evaluations of FDA-regulated products—namely drugs and medical devices [102].

However, navigating the landscape of EHR reuse is not without its challenges. The issue of representativeness poses a significant concern, particularly when the geographical areas providing data fail to accurately represent the targeted population for inference. Furthermore, inaccuracies or limitations in data pertaining to socioeconomic status, race, ethnicity, and other crucial factors may impede the reliability of the findings. The original design of EHRs for clinical and administrative purposes introduces challenges related to data availability and interpretation. Insufficient recording of factors relevant for addressing social and behavioural determinants of health (SBDH) and reducing inequalities diminishes the data's potential for research purposes, as it may be hindered by various forms of bias. Additionally, the presence of unstructured data, such as adverse events and symptoms, necessitates advancements in technologies such as natural language processing for more effective analysis [105].

A critical gap exists in current EHRs concerning key research factors, which are designed primarily for billing and scheduling rather than for epidemiological research. Variables crucial for understanding socioeconomic status and SBDH factors are often absent, limiting the comprehensive use of EHR data for research purposes. Visit censoring, where visits are not seamlessly connected between different health facilities, introduces challenges such as right or left censorship, impacting the continuity of patient records, for example, when patients are referred to a new centre without previous clinical information [100].

Security and privacy concerns surround the reuse of EHRs. The vulnerability of EHRs to cyber threats has raised alarms, necessitating ongoing efforts to implement cryptographic, noncryptographic, and hybrid access control models and blockchain technology for data protection. Furthermore, cases of ambiguity in the question of data ownership may cause privacy issues, with patients consenting not only for the use of their data but also for the existence of a wide spectrum of data uses that may be included in this consent without the patient truly acknowledging it. In fact, obtaining explicit consent for every secondary use has been catalogued as time-consuming, costly, and exhausting. Additionally, the existence of previous events, which have led to data leaks, illicit or illegal data selling or loss, makes the trust relationship between patients and data primary users or collectors difficult [99].

Moreover, the discrepancies between the General Data Protection Regulation (GDPR) principles and the objectives of EHRs pose a regulatory challenge in the EU, as the principles may limit data collection and use, conflicting with the broader goals of secondary health data utilization. The GDPR defines six main data protection principles: (1) lawfulness, fairness, and transparency; (2) purpose limitations; (3) data minimization; (4) accuracy; (5) storage limitations; and (6) integrity and confidentiality. In the case of EHR, these principles can contradict the characteristics and objectives of EHR, as if followed to the rule, these principles generally mean that healthcare data collection should be limited, together with its use, and deleted immediately after its purpose has been achieved. In the case of secondary health data, almost all antagonist principles are promoted, as more data would mean more potential for epidemiological research and other purposes, which would increase its potential. Its secondary use is also encouraged for institutions and

individuals to serve as tools for healthcare improvement, and ultimately, secondary use is necessary for this purpose [99].

Dilatory regulations, such as the Health Insurance Portability and Accountability Act (HIPAA) and Health Information Technology for Economic and Clinical Health (HITECH) Act applied in the United States, face a widening gap with rapidly advancing technologies. Patient expectations for immediate data availability clash with regulations that do not meet present-day requirements. Emerging scenarios, such as the integration of social media platforms into health data, challenge outdated regulations, emphasizing the need for flexible and adaptive regulatory frameworks to keep pace with evolving health data and EHR usage [99]. In essence, while the advantages of EHR reuse are substantial, addressing these challenges is crucial for unlocking its full potential in the dynamic landscape of secondary health data utilization.

The recent EHDS regulation [106] establishes that in the European Union, certain categories of health data, such as EHRs, are registered in electronic format systematically and according to specific data quality requirements (defined by means of implementing acts adopted by the European Commission) to contribute to the high quality and continuity of healthcare. The European electronic health record exchange format forms the basis for specifications related to the registration and exchange of electronic health data, which will be accessible for secondary use across the Union through a common mechanism for the purpose of creating scientific, innovative, and societal value.

1.5.3 Electronic Health Records: Examples of Reuse

1.5.3.1 The Use of EHR Data in Research

EHRs have become invaluable in various realms of healthcare, highlighting their versatility in the secondary use of health data. In the domain of clinical research, EHRs play a pivotal role in designing and conducting clinical trials for new medicines, addressing the shortage of trained medical staff in healthcare organizations and facilitating the exploration of new or enhanced drugs to meet evolving healthcare needs [107]. Leveraging existing patient data, EHRs contribute to predicting diseases, investigating drug behaviours across different diseases or patient profiles, and even aiding in the development of vaccines [108–110]. EHR data can also be utilized to identify new uses for existing medicine, with the aim of enabling a rapid and cost-effective approach to drug (re)discovery. For example, a genetically informed drug-repurposing pipeline that can be used in diabetes management has been developed and investigated. Among the 20 candidate drug–gene pairs validated, angiotensin-converting enzyme inhibitors and calcium channel blockers (CCBs) showed evidence of glycaemic regulation involvement. CCBs were found to be strong candidates for both reducing blood glucose levels and managing cardiovascular disease [111]. Furthermore, retrospective analyses of EHR data can reveal

potential contributing factors that might cause certain adverse health outcomes such as prolonged postoperative opioid use in opioid-naïve patients. Several independent predictors of prolonged postoperative opioid use, such as alcohol abuse, black race, Medicaid insurance, diabetes, mood disorders, hypertension, and chronic kidney disease, can be investigated via perioperative patient screening, providing education to patients and clinicians, and performing close postoperative follow-up, among other methods [112].

1.5.3.2 The Use of EHR Data in Public Health

The utility of EHR data extends to public health surveillance (PHS), where it automates surveillance processes to identify potential outbreaks efficiently and prevent their escalation [113]. This application has proven effective not only in developed countries such as the UK, the United States, France, Norway, Canada, and Australia [114] but also in developing nations when necessary [115]. Public health agencies, including the US Centers for Disease Control and Prevention (CDC), have started utilizing EHR data. In 2003, the CDC launched the BioSense system to quickly identify bioterrorism-related illnesses [116]. Since then, BioSense's network has grown to include 1700 hospitals, and its mission has expanded beyond bioterrorism. Currently, BioSense is the only nationwide surveillance system in the United States that integrates local, regional, and national data to monitor all health-related threats. Consequently, through initiatives such as BioSense, the CDC can use EHR data to track and address various public health issues, from the emergence of influenza H1N1 to an anthrax attack [116]. The global impact of the COVID-19 outbreak further underscored the importance of digital data recording systems, such as EHRs, in enhancing public health responses and preparedness [117].

1.5.3.3 The Use of EHR Data in Quality Management

In the realms of clinical audit and quality assurance, EHRs provide a wealth of accurate and detailed information that is essential for systematically setting standards, analysing data, and monitoring performance to maintain high-quality standards. The use of EHR data in clinical audits not only streamlines the auditing process for greater convenience but also contributes to improving the quality of care delivered to patients [118]. For example, by implementing computer-based standing orders, the Regenstrief Institute in Indianapolis increased the pneumococcal vaccination rate, a standard quality measure, from 31% to 51% [119]. Timely and effective quality measurement via EHR data will allow institutions to identify and address shortcomings in real time rather than respond retrospectively. For example, the rate of deep venous thrombosis/pulmonary embolism (from 8.2% to 4.9%) in Partners' healthcare in Boston nearly halved when computerized alerts were used for anticoagulation prophylaxis [120]. Another example is the Geisinger Health System in Danville, Pennsylvania, which uses its EHR system to monitor

performance continuously on the basis of various standard and internal quality metrics. This approach has elevated nearly all its quality indicators above national averages and enhanced the quality of care across 32 performance measures evaluated in Medicare's Physician Group Practice Demonstration Initiative. These measures include diabetes, congestive heart failure, coronary artery disease, and preventive care [121–123].

The examples of EHR reuse presented in clinical research, public health surveillance, and clinical audit and quality assurance collectively highlight the transformative impact of EHRs in enhancing healthcare delivery, research, and response capabilities.

1.6 Human Genomic, Proteomic Data

1.6.1 Genomic and Proteomic Data: Background

Two decades ago, the invention of spotting oligonucleotide probes at a high density to glass or nylon arrays transformed the study of gene expression and accelerated the adoption of bioinformatics as an important element of biological studies [124]. The term "proteomics" was coined in the mid-1990s to describe the comprehensive study of a proteome, which encompasses the proteins expressed by a genome. This field involves large-scale analysis of proteins, namely examination of protein expression, structure, modifications, functions, and interactions [125]. Proteomics is a crucial postgenomic approach for enhancing the understanding of gene function [126]. On the other hand, human genomic data, according to the FDA, are data that can be obtained from germline sources (inherited from parents), somatic sources (e.g. mutations in tumour tissues), or mitochondrial sources (e.g. for tracing maternal lineages). Biological samples from humans may also contain nonhuman genomic molecules, such as microbial DNA or other potentially infectious agents. The type of genomic data produced depends on the analytes and the technology platforms used [127].

Currently, new research into how genetic variants affect health and disease has been conducted through the mapping of the human genome. Diseases such as diabetes, Alzheimer's disease, cardiovascular disease, and others are not being studied in relation to genetic variants for what has become part of a "personalized medicine" approach aimed at improving patient care. Institutions have been developed for this purpose. Community efforts include the Microarray Gene Expression Data (MGED) Society in 1999 and the Gene Ontology (GO) standard vocabulary. Other more recent efforts include the eMERGE network (2007), which is a consortium that integrates this innovation pathway, explores the utility of DNA repositories and electronic medical records (EMRs) to promote advancements in genomic science, and offers an initial step in the creation of data-driven strategies to integrate genomic data into standard healthcare delivery [128]. Gene Expression Omnibus (GEO) and

ArrayExpress are also part of this attempt to increase innovation in the field. Another EU initiative is very important: "1+ Million Genomes" (1+MG) seeks to provide secure access to genomic data and associated clinical information across Europe. It aims to advance cutting-edge research, inform health policy decisions, and promote personalized healthcare treatments that have the potential to enhance disease prevention. As one of the largest global genomic projects, 1+MG plays a crucial role in establishing international standards in the field of genomics [129].

1.6.2 Genomic and Proteomic Data: Advantages and Challenges

The utilization of publicly available genomic and proteomic data holds substantial advantages for both the scientific community and individual researchers [130]. Sharing research data with the public and storing them in dedicated databases with backup mechanisms can prevent information loss, addressing the heightened risk associated with storing data on private computers and servers. Notable repositories, such as the Sequence Read Archive (SRA)/European Nucleotide Archive (ENA) and Gene Expression Omnibus (GEO), offer secure platforms for genomic and gene expression data. Beyond safeguarding against loss, the reuse of such data fosters scientific progress and discovery [131]. For example, integrating diverse data sources, such as exRNA metadata and biomedical ontologies, via linked data technologies can enhance the generation and interpretation of hypotheses within independent biological contexts [132]. The transformative potential of reusing data in medicine extends to reshaping medical service practices, making the benefits of data reuse outweigh the associated risks [133, 134]. Additionally, the reuse of genomic and proteomic data has proven to be a strategic approach for addressing gaps in biomedical research and establishing starting points for experimental medical investigations [133]. Furthermore, this approach maximizes efficiency in terms of time, labour, and cost, offering economical alternatives to overcoming methodological constraints in conducting new experiments [135]. Critically, the sharing of datasets not only benefits the scientific community and society but also provides advantages for authors themselves [136–138]. Researchers can build reputations by publishing high-quality datasets while sharing these datasets enhances research visibility and increases the likelihood of additional citations. The surge in data reuse also underscores the growing need for larger databases, exemplified by the exponential growth of repositories such as the SRA and GenBank, which have expanded rapidly in size over just a few years [139–141].

Reusing biological data presents formidable challenges, as outlined by several documented hurdles. First, a substantial investment in infrastructure and software development remains imperative for effective data reuse. Scientists currently grapple with the need to navigate diverse proteomics resources to gather comprehensive information about a given protein. Despite facilitating the connection process

between omics datasets through sample identifiers, traditional repositories are often field specific such as genomics or proteomics. This proves to be limiting in the face of growing interdisciplinary research that integrates various omics sciences [142]. Major institutions such as the European Bioinformatics Institute (EBI) and National Center for Biotechnology Information (NCBI) have established specific databases for BioSamples [143] to facilitate the linking of different studies conducted using the same sample. Additionally, proteomics data are complex per se because of alternative splicing, PTMs, and protein degradation events, as well as the interconnectivity of proteins into complexes and signalling networks that are prone to changes in time and space [144]. Addressing this complexity requires new analytical and bioinformatics methodologies on an annual basis [145, 146], introducing further complications to data standardization and deposition. Furthermore, since the proteomics audience includes a variety of different roles (e.g. biologists who investigate protein mechanisms or computational biologists who develop new software for facilitating analysis and data interpretation) [147], data sharing in proteomics thus necessitates substantial investment and infrastructure. Because of this, there is a growing consensus on the need for public dissemination of proteomics data. In addition, in comparison with genomics, proteomic public data remain less common, thus still requiring substantial investment and infrastructure. Current projects in the proteomics field include the recent ProteomeXchange (PX) consortium, the Global Proteome Machine Database (GPBDB) [148], PeptideAtlas [149], and the Proteomics Identification Database (PRIDE) [150]. Other initiatives, such as Peptidome [151] and Tranche [152], are now discontinued due to a lack of funding, which negatively impacts data sharing in this field. The sheer multitude of data types and formats adds another layer of complexity to the task at hand. Ideally, repositories should not only house data but also ensure that it is linked to quality control (QC) metrics. Moreover, the generation of these metrics should ideally occur simultaneously with data acquisition in the laboratory. A critical challenge lies in the absence of experimental and technical metadata, particularly concerning proteomics, hindering effective data reuse in scientific endeavours. Finally, the identification of false positives in large datasets and when combining different datasets remains a problem [149, 153]. Therefore, protein expression resources should conduct the most thorough statistical analyses possible.

1.6.3 Genomic and Proteomic Data: Examples of Reuse

The Use of Genomic and Proteomic Data for Research

One avenue involves the reuse of raw genomic and proteomic data for novel biological investigations. For example, a dataset comprising 81 samples from white blood cells and 1463 from various organs, measured on the Affymetrix HG U133A platform, facilitated a study examining the correspondence in their expression profiles [154]. The results revealed significant overlap in the transcriptomes of white blood cells and other organs, which aligns with the understanding that most genes are

expressed across various tissues [155]. Expression data were also employed to pinpoint the tissue of origin for metastatic cancers lacking a known primary site, leveraging classifiers trained on a vast dataset encompassing 5577 samples from 56 cancer types and 1667 normal samples from 44 tissues [156]. These examples underscore how large and diverse datasets can provide profound insights, particularly in studies where global transcriptomic profiles enable comparisons across different cellular states.

The Use of Genomic and Proteomic Data for Conducting Meta-Analyses
Data reuse also involves the utilization of public data for conducting meta-analyses of summary data. Meta-analyses, which use summary-level data, offer a popular and flexible approach to repurposing existing data from various array platforms. Gene expression data have been subjected to meta-analysis for an array of biological inquiries, with emerging studies employing this technique on public datasets to discern signals that are not readily apparent in individual datasets [157–159]. Noteworthy examples include a study analysing 324 differentially expressed genes in individuals with Down syndrome, drawn from case–control data across 45 studies [160]. In the context of cancer research, microarray data from ArrayExpress, GEO, and Oncomine datasets were amalgamated to identify urinary biomarkers specific to prostate cancer [161].

The Integration with Other Omics Datasets
The integration of proteomics datasets with other public omics datasets opens new avenues for data scientists. Proteogenomics, a methodology employed in cancer studies, focuses on cancer-specific peptides with diagnostic and therapeutic potential. The Clinical Proteomic Tumor Analysis Consortium (CPTAC) of the US National Cancer Institute (NCI) has published impactful studies on various tumours, including colorectal, breast, and ovarian cancers [162–164]. This integration of proteomics with other omics datasets underscores the potential for cross-disciplinary insights and advancements in understanding complex biological phenomena.

1.7 National Registries

1.7.1 National Registries: Background

The term “patient registry” typically refers to registries focused on health information, distinguishing them from other types of record sets, although there is no universally accepted definition. E.M. Brooke, in a 1974 World Health Organization publication, described registries in health information systems as “a file of documents containing uniform information about individual persons, collected in a systematic and comprehensive way, in order to serve a predetermined purpose” [165]. The US National Committee on Vital and Health Statistics describes registries used for various purposes in public health and medicine as “an organized system for

collecting, storing, retrieving, analysing, and disseminating information on individuals who either have a particular disease, a condition (e.g., a risk factor) that predisposes them to a health-related event, or prior exposure to substances (or circumstances) known or suspected to cause adverse health effects" [166].

In the case of national registries, these are created with the objective of collecting data to measure progress, drive action, prevent disease, improve health, and ultimately improve health for all people [167]. Registries tend to be composed of data from multiple sources that are linked by a specific methodology, i.e. direct chart abstraction by trained data abstractors, who tend to work within a coordination centre at the registry. This data collection tends to be guided by a manual that defines the data elements, which are usually categorized as demographics, operative data, hospitalizations, follow-up, etc. [168].

Multiple national registries exist worldwide, focusing on different diseases, patient groups, and demographic characteristics of patients. Key infrastructure is necessary for national registries to promote interoperability; as an example, the rare disease network for care and research in France, established as a part of the first national plan for rare diseases (2005–2008), is based on a three-pillar architecture composed of (1) the definition of a national patient ID, (2) the definition of a common data format compatible with EHR standards, and (3) security [169].

1.7.2 National Registries Worldwide: Advantages and Challenges

National registries can collect large amounts of data (e.g. the registry provided by the Michigan Arthroplasty Registry Collaborative Quality Initiative [MARCQI] contains more than 375,000 cases), which can be collected from multiple sources and supplemented with other types of data (e.g. administrative billing data) [168]. National registries enable state-level collaboration. While in other cases, surgeons and hospitals compete in between, here, they engage in open data sharing and cooperate on quality improvement projects. The value of registries can be exemplified by the fact that they usually collect data from multiple institutions, making EHR and clinical databases the only components of a population registry [170]. In contrast to clinical databases, which do not require strict definitions of common features, this is mandatory for registries [170]. Furthermore, registries aim to collect the highest number of cases (if possible, in total), which makes them a representative data source for a defined population [165, 170, 171]. Registries also perform a constant update of the patient status by following a predefined schedule [170]. In addition, medical registries collect data in a way that allows comparisons of the results at the national (or even higher) level. Given that national medical registries are usually focused on a disease of interest, they fit their predefined purpose and collect specified outcomes [171], which makes them more uniform than EHR registries, for example. Taking the domain of cancer as an example, the primary goal of the cancer

national registry is to record all new cancer cases that arise within a defined population by focusing on epidemiology and public health practices [172]. These registries are rich sources of information on the cancer burden that can be used for healthcare planning and evaluation purposes and for conducting studies on prevention, early detection/screening, and cancer-related healthcare, which are used by epidemiologists and public health planners [16]. These concepts can be applied equally to other disease paradigms.

Registry data pose several challenges, encompassing issues related to data completeness, accuracy, and scope. First, ensuring data completeness is a persistent challenge, as demonstrated by the MARCQI, where reported cases are compared with validated statewide administrative billing data to confirm completeness. The extensive time and labour required for data abstraction on a large scale can result in substantial lag times between case performance and result reporting, prompting some registries to set specific time limits for data abstraction. Additionally, the scope and definition of data elements at the registry level may impose limitations. Another limitation arises from the fact that data are captured only for patients treated within institutions affiliated with the specific registry. Furthermore, the accuracy of registry data is compromised by potential inaccuracies in the coding process, as well as errors caused by programming and inaccurate transcription [173]. One of the pitfalls of registry data is nonstandardized and inactive follow-up, which may lead to incomplete and inaccurate outcomes. In certain cases, the follow-up is performed in a passive manner by using linkages with administrative databases. However, since the primary aim of these datasets is not for research purposes, these methods may introduce errors due to misclassification of outcomes resulting from coding errors and different coding practices. Additionally, the passive nature of data collection can lead to missing data, which can hinder further analysis in research practice. Finally, one should bear in mind that data entry into a registry is not strictly monitored, which may create the possibility that ineligible patients enter the registry and thus introduce bias in the findings obtained from registry data analysis. Therefore, registry data, like all other RWD, require careful monitoring, quality assessments, and validations so that they provide reliable results that can be applied to the population of interest [174]. Collectively, these challenges contribute to analytical limitations and hinder the ability to establish causality via registry data [168].

1.7.3 National Registries Worldwide: Examples of Reuse

1.7.3.1 The Use of National Registry Data in Public Health Research at the National and International Levels

The secondary use of health data, particularly those harnessed from national registries, represents a transformative force in addressing diverse health-related questions and guiding public health initiatives. National registries worldwide have become powerful reservoirs of health data, offering a wealth of information for

diverse research endeavours. In 2020, data from a National Cancer Centre in Japan were harnessed to explore the correlation between airborne pollen levels and cancer incidence. A significant association between elevated pollen levels measured two years prior and the incidence of breast cancer and other forms of cancer was revealed, highlighting the potential of national registries in uncovering environmental factors impacting health [175].

Further examples underscore the versatility of national registry data. The Lithuanian Compulsory Health Insurance Information System database, SVEIDRA, was used to support the hypothesis that patients with rheumatic diseases face a higher risk of mortality and lower life expectancy than the general population does. A total of 11,636 patients and a follow-up period of 43,064.34 person-years were included, and 950 registered deaths were identified, emphasizing the utility of national registry data for exploring disease-specific mortality risks [176]. Additionally, the National Arthroplasty Registry served as a valuable resource for a comparative observational study involving 581,818 procedures. Compared with procedures employing technology-assisted instrumentation, there was a noteworthy association between conventional instrumentation during total knee arthroplasty and early postoperative death, highlighting the potential of registry data to inform best practices in healthcare procedures [177]. A comprehensive population-based nationwide cohort study conducted in Finland, which utilized data from two national health registries, examined 933,823 children born between 1996 and 2011. This study investigated potential infections with *C. trachomatis*, revealing a rare occurrence rate of 0.22 per 1000 births. These findings challenge previous assumptions, shedding light on transmission dynamics during pregnancy and illustrating the ability of national registries to provide statistically significant answers to common clinical questions [178]. The impact of different dialysis modalities was addressed by drawing on data from national and regional registries of ten countries (Australia, New Zealand, Japan, China, Taiwan, Korea, Thailand, Hong Kong, Malaysia, and India). Despite challenges such as varying coding techniques and data limitations, data from 201,590 patients were useful for identifying significant influencing factors between dialysis modalities and outcomes [179].

1.7.3.2 The Use of National Registry Data for Worldwide Collaboration

The National Cardiovascular Data Registry (NCDR), established by the American College of Cardiology, stands as an exemplary source of standardized, evidence-based data on clinical topics and cardiovascular procedures [180, 181]. Another illustrative example arises from the multinational network of diabetic care centres known as SWEET. A 2021 study within this network utilized linear and logistic regression models to analyse participant data, revealing that lower HbA1c levels and fewer diabetic ketoacidosis (DKA) episodes were observed in participants utilizing insulin pumps, continuous glucose monitoring, or a combination of both [182]. The collaborative nature of registries, exemplified by initiatives such as the SWEET, fosters international collaboration, enabling the evaluation of information from small national populations on a global scale, the establishment of

benchmarking practices, and the synergistic effects of shared data [183]. Other examples of this collaboration of data between registries include the collaboration between ten national registries to evaluate the outcome of renal replacement therapy for diabetic neuropathy, suggesting the flexibility of research options that could be derived from international data use [184]. Finally, a collaborative project involving 7 well-known registries and the SWEET initiative provided a comparative perspective, combining data from over 900 facilities and 100,000 paediatric patients. This collaboration not only grants access to updated and representative data but also fosters open benchmarking collaboration between registries [183]. Collaborative initiatives among registries worldwide further amplify the potential of data management, allowing for the analysis of large datasets that were previously impractical for researchers.

1.7.3.3 The Use of National Registries for Supporting Regulatory Assessments

When RCT data are unavailable or are not ethical or feasible (as in the case of many rare diseases), patient registry data may support regulatory decision-making. For haemophilia, for example, the updated guideline for factor VIII products published by the EMA removes the obligation to conduct clinical trials in treatment-naïve patients but requires postlicensing trials on the basis of key data elements collected in patient registries [185]. Additionally, as in other European Medicines Agency (EMA) reviews, for conditional marketing authorization products, registration studies can provide postlicensing data that meet specific regulatory obligations to confirm safety and/or efficacy, such as for the recently approved chimeric antigen receptor (CAR) T-cell product thiagenlecleucel and axicabtagene in siloleucel [186, 187]. Certain registries may be particularly valuable in terms of the size and representativeness of the patient population, the length of follow-up of medically exposed patients, and the availability of data not collected by other real-world repositories.

The wealth of insights derived from these examples collectively underscores the pivotal role of national registries in advancing our understanding of health dynamics and in shaping evidence-based interventions, highlighting the substantial impact and diverse applications of national registries in advancing our understanding of health-related phenomena at both the national scale and the international scale.

1.8 Person-Generated Electronic Health Data

1.8.1 Person-Generated Health Data: Background

Information on health generated by patients is not a new concept. Instead, oral medical history has always served as the basis of medical consultation. Recently, this data collection process has also required the storage of records of medical events,

i.e. symptom frequency, blood glucose concentration records, to provide a longitudinal record that will assist in clinical diagnosis and management.

The Office of the National Coordinator for Health Information Technology (ONC) from the United States defined person-generated health data (PGHD) as "health-related data including health history, symptoms, biometric data, treatment history, lifestyle choices, and other information-created, recorded, gathered or inferred by or from patients or their designees" [188]. PGHD differs from data produced in clinical settings and provider encounters in two important ways. First, patients, not providers, are primarily responsible for collecting or maintaining this information. Second, patients direct the sharing of this information with healthcare providers and other stakeholders. In this way, PGHD complements the provider's collection and flow of health information in the health system. PGHD includes various data types, such as the following:

- Vital signs measured by the patient (or proxy) (e.g. temperature, blood pressure, blood glucose, weight), which can be captured manually, by reading a mechanical or electronic device, or automatically via a monitoring device.
- Self-reported lifestyle data (e.g. caloric intake, diet, exercise, hydration, and medication adherence) that are recorded by the patient or family member, and data collection usually occurs manually.
- Self-reported as perceived quality of life data (e.g. mood, sleep quality, level of pain, social contacts), which are mainly captured manually by the patient or a patient proxy.
- Other data, apart from health-related data, are known to the provider by the patient on a personalized individual basis [188].

PGHD collection can be supported by different tools, including smartphone apps, wearables, sensors, smart home appliances, medical record apps, etc. [189]. Although diverse organizations, such as the Veterans Health Administration, Kaiser Permanente, and others, are promoting a more systematic use of PGHD in its early stage, further work is still needed on many aspects of the PHGD to achieve its full potential.

1.8.2 Person-Generated Health Data: Advantages and Challenges

Person-generated health data (PGHD) have emerged as a dynamic source with notable advantages and challenges in the realm of secondary health data use. One key advantage lies in its capacity to be recorded outside of clinical settings, significantly extending potential data collection timeframes [190]. PGHD is valuable because it can be recorded outside of a clinical or laboratory setting in places where people spend most of their time (e.g. at home, in the office, or in their community). PGHD may also be measured more frequently than conventional medical data. PGHD can

be combined with medical and other data (such as geographic location, weather, and community events) to provide a more detailed picture of an individual's health over time [190].

By using digital tools to collect PGHD, researchers can communicate more widely and directly with potential study participants, allowing them to collect more data and improve their workflow. Special techniques include the use of scientific data platforms and remote sensing. Researchers can expand recruitment and enrolment in their studies by incorporating digital tools that incorporate PGHD into their study designs and protocols, such as mobile health devices, online discussion forums, and health information platforms. Researchers do not need to rely as much on relationships with physicians or health systems to identify potential study participants. Instead, they could provide promotional information about their studies directly to a broader and perhaps more diverse patient population through online postings on social media and email lists [191]. The use of PGHD technologies for electronic data collection and transmission helps streamline the research flow. Instead of manually entering patient data, data can flow electronically into the research database. Tools capable of cleaning data can simplify the process of ensuring that the data are complete and ready for analysis. These electronic processes reduce the amount of work required by researchers and reduce the potential for human error in data entry [192].

PGHD has also proven invaluable in the context of clinical trials, especially in the reporting of adverse events. Patient reported Outcomes (PROs) derived from PGHD offer a more accurate portrayal of adverse events, addressing the significant underreporting—up to 76%—observed in traditional clinical trials. This clarity in reporting, whether through the support of data or digital twins during clinical trials or the utilization of other sources of information post-product release based on PGHD, enhances the understanding of patient experiences and contributes to the overall assessment of interventions [193].

The large volume of health data generated on a continuous basis provides a plethora of opportunities for secondary uses and for gathering new scientific evidence that can impact the population.

The integration of PGHD into healthcare practices also introduces a set of challenges that necessitate careful consideration in the secondary use of health data. One primary challenge is associated with data governance, given that PGHD involves data generated directly by patients. Clear definitions regarding data holders and providers, the process of data collection, and establishing agreements between stakeholders become imperative for effective and ethical information collection [189]. The incorporation of new sources of data collection, such as the utilization of social media for pharmaco-epidemiologic research, presents an additional ethical consideration that requires thoughtful examination [189]. Motivating patients to actively engage as data providers poses another challenge, with concerns arising regarding unclear motivations and potential reluctance [189]. Additionally, patients can be inconsistent with using their apps and wearables, which introduces a new source of bias to the data [194].

Moreover, the storage and management of PGHD currently demand substantial resources, posing a risk of deteriorating the doctor–patient relationship and rendering the data difficult to interpret [104]. The delineation of clear responsibilities and liabilities is crucial to address potential issues arising from the use of inaccurate PGHD in diagnosis or treatment. This includes establishing accountability for errors, whether they originate from the patient, the app used for data collection, or other contributing factors [195]. Apps and wearables may use different tools and methods for data collection, thereby affecting the reliability and accuracy of the data [190]. Further complications to data accuracy and reliability are introduced by the participants' lack of knowledge. For example, they may incorrectly capture self-reported data (such as that related to anxiety) if they do not receive proper education upfront [196]. In addition to patients, practitioners may also misinterpret PGHD, and in some cases, PGHD lacks context (e.g. a spike in heart rate may be caused by exercise or stress); in both cases, the interpretation of the results may be misleading [197]. Additionally, the diverse characteristics of PGHD, encompassing various items, tools, communication channels, and workflows, necessitate the development of new standards to meet the demands of multiple stakeholders, including patients, app developers, EHR suppliers, and physicians [195]. Finally, data governance is an essential part of the effective use of PGHD and is required at every stage of the process, from data collection to dissemination of results. Agreement between study organizers and participants on the principles of information sharing is central to ensuring trust and includes consideration of confidentiality and legal liability. The development of patient privacy standards and guidelines and/or the updating of existing guidelines [198, 199] is important as the use of PGHD increases and new issues emerge. Legal aspects such as informed consent and data security are a concern in PGHD research, as are standards for electronic consent (consent is evolving) [200]. As the healthcare landscape evolves with the integration of PGHD, addressing these challenges becomes imperative to ensure the responsible and effective use of PGHD for improving healthcare outcomes.

1.8.3 PGHD: Examples of Reuse

The Use of PGHD in Research Endeavours

The secondary use of person-generated health data (PGHD) occurs in various domains, demonstrating its versatility and impact on biomedical research and healthcare practices. In the realm of scientific initiatives, the US Precision Medicine Initiative (PMI), specifically the All of Us Research Program [201], underscores the importance of collecting primary PGHD through participatory technologies such as web-based surveys, wearable devices, and smartphone apps. This ambitious programme aims to build a vast data repository with over one million participants, collaborating with projects such as the health US-based eHeart Study [202] and the Health Data Exploration Network [203] to collect primary PGHD reporting

environmental risk factors for biomedical research across diverse populations of both healthy individuals and patients.

Biometric data, encompassing wireless scales, blood pressure cuffs, thermometers, pulse oximeters, heart rate monitors, and blood glucose metres offer valuable insights when integrated with smartphone apps [204–206]. Some smartwatches even include electrocardiogram readings, which are more accurate than ambulatory ECGs and are useful for screening for atrial fibrillation and cardiac issues. The reuse of biometric data highlights the potential of PGHD in monitoring and managing various health parameters. Physical activity monitoring has evolved from accelerometers in the 1980s [207] to modern wrist-worn activity trackers, which have gained acceptance in research due to enhanced data accuracy [208–210]. Smartphones equipped with accelerometers can distinguish between different physical activities, offering a convenient and acceptable alternative to wrist-based trackers [211, 212]. Given the established benefits of regular physical activity, the use of PGHD from these devices holds promise for future studies.

The Use of Substandard Real-Time PGHD Data for Detecting and Predicting Health-Related Outcomes

Location data, derived from smartphones via GPS, Wi-Fi, and Google Location History, provide insights into patients' community mobility and its correlation with mental and physical well-being [213–216]. Such data can offer valuable information about daily activities outside the home, contributing to the assessment of quality of life. In mental health studies, GPS measurements of movement have been employed to detect depression with a remarkable accuracy of 87% [217].

The Use of PGHD in Decision-Making Processes

The use of technological innovations for PGHD has the potential to improve the individualized management of disease care (such as cancer) across the continuum from primary prevention to treatment and survival [218]. For example, a PGHD analysis focused on the detection of mental distress facilitated the identification of individuals in need of mental health treatment, help, and support [219]. Mobile technology applications that allow patients to identify and report symptoms or side effects of treatment between routine clinic visits can improve adherence to treatment plans [220].

The Use of PGHD to Improve the Collection of Certain Data Types

The assessment of dietary habits, a historically challenging endeavour, has improved with PGHD. Traditional retrospective questionnaires suffer from recall and social desirability biases. Web- or app-based dietary measurements provide a more accurate and less burdensome alternative [221, 222]. Future innovations may involve the use of smartphone cameras and machine learning algorithms to identify foods and portions, reducing participant burden and enhancing the collection of dietary information in studies [223, 224]. These examples illustrate the expansive potential of PGHD in transforming healthcare research and practices across diverse domains.

1.9 Social and Behavioural Determinants of Health

1.9.1 Social and Behavioural Determinants of Health: Background

The Healthy People programme, which is part of the US Department of Health and Human Services Office of Disease Prevention and Health Promotion, defines social determinants of health (SDOHs) as "the conditions in the environments where people are born, live, learn, work, play, worship, and age that affect a wide range of health, functioning, and quality-of-life outcomes and risks" [225]. SDOHs are also known as social and behavioural determinants of health (SBDHs) [226]. As stated by the WHO, the SDOH has important implications for health inequalities, namely unfair and avoidable differences in health status within and between countries. In countries of all income levels, health and disease follow a social gradient: the worse the socioeconomic status is, the worse the health. It is estimated that social factors may be more important than healthcare or lifestyle choices in influencing health. For example, many studies have shown that SDOH is responsible for 30–55% of health effects. Another estimation shows that sectors outside health contribute to population health outcomes to a greater extent than does the health sector itself [227].

With respect to five key areas: (1) economic stability, (2) education access and quality, (3) social and community context, (4) neighbourhood and environment, and (5) healthcare access and quality, research has demonstrated that SBDH factors account for more than one-third of the estimated annual deaths in the United States. In contrast, integrating SBDH factors such as education, lifestyle, physical activity, and diet could improve the management of diseases such as Alzheimer's disease or diabetes. Integrating SBDH into the EHR system could help address such a factor's key role in healthcare and further improve patient care outcomes in ways that would not be possible otherwise [226].

1.9.2 SBDH: Advantages and Challenges

The secondary use of social determinants of health (SBDH) data presents notable advantages and challenges, as recognized in the literature. Reusing SBDH data has the potential to significantly enhance care management by incorporating economic, social, and behavioural factors associated with a patient's health. This holistic approach to decision-making improves the efficiency of care plans and better addresses the needs of vulnerable populations, ensuring a more comprehensive and tailored healthcare strategy [228]. Moreover, addressing SBDH factors has the potential to reduce healthcare utilization, as exemplified by programmes initiated by the Centres for Medicare and Medicaid. These initiatives aim to systematically identify and address health-related social needs, such as food insecurity and inadequate housing, with the goal of lowering the overall demand for healthcare services

among community-dwelling Medicare and Medicaid beneficiaries [229]. Such programmes signify a transformative shift in how clinical and community service providers collaborate, ensuring that patients receive services and support aligned with their specific needs.

Furthermore, the integration of SBDH data holds promise for improving risk adjustment payment models and enhancing the design of performance-based incentives. Notably, the first US payment model incorporating SBDH variables, MassHealth, introduced variables such as a neighbourhood stress score and a measure of unstable housing. The inclusion of these variables, alongside additional data such as individuals' use of services from the Departments of Mental Health or Developmental Services, in the diagnosis-based model resulted in a more accurate alignment of payments to costs for several vulnerable categories. This example highlights the potential of SBDH data to refine payment models and incentivize performance in a manner that considers the diverse social determinants impacting healthcare outcomes [230]. In essence, the reuse of SBDH data offers a multifaceted approach to healthcare management, with the potential to optimize care plans, reduce healthcare utilization, and enhance the alignment of payments to actual costs, particularly for vulnerable populations.

Reusing SBDH data presents several challenges that warrant attention and thoughtful solutions. First and foremost, data collection and algorithm bias pose significant risks. Currently, the representation of SBDH in EHRs is limited, with only some aspects registered in clinical text. The integration of SBDH data with technologies such as AI to support clinical decisions holds promise but is susceptible to bias, potentially resulting in unfair decision scenarios. Addressing bias in data collection and interpretation is crucial to ensure the responsible integration of SBDH and PGHD to improve patient outcomes and avoid potential disparities [226].

Translating SBDH data into meaningful results is a complex task, particularly in the context of EHR incentive programmes. These programmes focus on capturing data, optimizing the clinical workflow, and promoting continuous quality improvement to eliminate healthcare inequalities. The third stage emphasizes the importance of addressing the relationship between SBDH and health outcomes. Recent research underscores the positive impact of incorporating SBDH into EHR, leading to improved outcomes such as increased performance, higher medication adherence rates, reduced hospitalization risk, and other positive indicators [231].

Furthermore, there are also barriers to the widespread implementation of SBDH screening in healthcare encounters. Key SBDH measures are not routinely collected, and when collected, rarely validated instruments are employed, limiting the secondary use of this valuable information. The obstacles include providers' unwillingness to address patients' needs without clear resources, concerns about the impact on patient–provider relationships, a lack of training, an absence of incentives to record findings, and a lack of agreement on which SBDH to screen for and the most reliable standardized tools [232].

Context challenges further complicate SBDH data collection, requiring special considerations for sensitive situations. Factors such as unstable housing situations

when caring for a child or an elder or when dealing with a drug-dependent family member introduce complexities in the selection of appropriate scenarios and types of questions for each patient. Similar to the approach of patient-centred and personalized care, SBDH data collection practices need to be tailored for each patient scenario, whether conducted in person or via data collection technologies [228]. In navigating these challenges, the responsible and ethical use of SBDH data becomes imperative for fostering equitable healthcare practices and outcomes.

1.9.3 SBDH: Examples of Reuse

The Use of SBDH in Disease Prediction
The secondary use of SBDH data has demonstrated its potential in various impactful applications. Prediction models have been developed to identify patients at risk on the basis of socioeconomic factors. Research has shown that low socioeconomic status and socioeconomic distress are associated with the risk of comorbid substance use disorders and opioid misuse [233, 234]. Furthermore, studies have explored the use of SBDH in predicting suicide risk, with factors such as education level, employment status, and income significantly correlated with suicidal behaviour in mentally ill patients [235–237]. Tools such as the OxMIS have been designed to predict suicide in severely mentally ill patients, incorporating SBDH factors such as education and substance abuse [238].

The Use of SBDH in Disease Management
SBDH data integration has also proven valuable in designing interventions for improved disease management. Semantic extract, transform, and load (ETL) services and dashboard systems such as the MOSAIC have contributed to better managing conditions such as obesity and diabetes. For example, patient education on lifestyle interventions associated with improved glycaemic control in diabetes patients highlights the potential for the use of SBDH data to enhance disease management [239–241].

The Use of SBDH in the Prediction of Disease Progress and Outcome
Moreover, the reuse of SBDH data plays a pivotal role in predicting the progression and outcomes of various disease states. Factors such as housing status, including homelessness, and geographic location have been identified as influencing mental, behavioural, and neurodevelopmental disorders, as well as circulatory system diseases [242]. Geographic location, which acts as a surrogate for socioeconomic characteristics, has been associated with multiple diseases, health conditions, and high mortality rates [243]. These examples underscore the versatility and significance of SBDH data in informing healthcare decision-making, designing interventions, and predicting health outcomes.

The Use of SBDH in Improving the Design of Risk-Adjusted Payment Models and Performance-Based Incentives

The first US payment model, which includes SDBH variables, was described previously [230]. In October 2016, MassHealth, which includes Massachusetts' Medicaid and Children's Health Insurance programmes, added two SDBH variables: the neighbourhood stress score and the amount of unstable housing. Adding the available SDBH and other data (e.g. whether subjects used the services of the Departments of Mental Health or Developmental Services) to the diagnosis-based model led to improved matching of payments to costs for several vulnerable member categories. However, improving such SDBH risk adjustment models requires fairly complete information on SDBH factors in medical encounters and claims data, which is a problem noted by Torres and colleagues [228].

References

1. FDA (2023) Real-world evidence. https://www.fda.gov/science-research/science-and-research-special-topics/real-world-evidence. Accessed 25 June 2024
2. National Institute for Health and Care Excellence (NICE) (2024) Glossary. https://www.nice.org.uk/Glossary?letter=R#Real-world%20evidence. Accessed 25 June 2024
3. Liu F, Demosthenes P (2022) Real-world data: a brief review of the methods, applications, challenges and opportunities. BMC Med Res Methodol 22:287. https://doi.org/10.1186/S12874-022-01768-6
4. European Medicines Agency (2023) RWE_framework_to_support_EU_regulatory_decision_making_1688023939.pdf. Open Science Framework
5. Dang A (2023) Real-world evidence: a primer. Pharmaceut Med 37:25–36. https://doi.org/10.1007/s40290-022-00456-6
6. Dang A, Jagan Mohan Venkateswara Rao P, Kishore R, Vallish BN (2021) Real world safety of bevacizumab in cancer patients: a systematic literature review of case reports. Int J Risk Saf Med 32:163–173. https://doi.org/10.3233/JRS-194051
7. MarksMan Healthcare (2018) How RWE can impact clinical trial design and help in decision making? https://marksmanhealthcare.com/2018/10/24/how-rwe-can-impact-clinical-trial-design-and-help-in-decision-making/. Accessed 8 Nov 2023
8. Hernández MA, Stolfo SJ (1998) Real-world data is dirty: data cleansing and the merge/purge problem. Data Min Knowl Discov 2:9–37. https://doi.org/10.1023/A:1009761603038/METRICS
9. Corrigan-Curay J, Sacks L, Woodcock J (2018) Real-world evidence and real-world data for evaluating drug safety and effectiveness. JAMA 320:867–868. https://doi.org/10.1001/JAMA.2018.10136
10. Makady A, de Boer A, Hillege H et al (2017) What is real-world data? A review of definitions based on literature and stakeholder interviews. Value Health 20:858–865. https://doi.org/10.1016/J.JVAL.2017.03.008
11. Franklin JM, Schneeweiss S (2017) When and how can real world data analyses substitute for randomized controlled trials? Clin Pharmacol Ther 102:924–933. https://doi.org/10.1002/CPT.857
12. Djurisic S, Rath A, Gaber S et al (2017) Barriers to the conduct of randomized clinical trials within all disease areas. Trials 18:360. https://doi.org/10.1186/s13063-017-2099-9
13. Rudrapatna VA, Butte AJ (2020) Opportunities and challenges in using real-world data for health care. J Clin Invest 130:565–574. https://doi.org/10.1172/JCI129197

14. US Food & Drug Administration (2017) New drug therapy approvals 2017. https://www.academia.edu/36341647/New_Drug_Therapy_Approvals_2017_FDA_report. Accessed 8 Nov 2023
15. Chatterjee A, Chilukuri S, Fleming E, et al (2018) Real-world evidence: driving a new drug-development paradigm in oncology | McKinsey. https://www.mckinsey.com/industries/life-sciences/our-insights/real-world-evidence-driving-a-new-drug-development-paradigm-in-oncology. Accessed 8 Nov 2023
16. Li G, Sajobi TT, Menon BK et al (2016) Registry-based randomized controlled trials- what are the advantages, challenges, and areas for future research? J Clin Epidemiol 80:16–24. https://doi.org/10.1016/j.jclinepi.2016.08.003
17. Purpura CA, Garry EM, Honig N et al (2022) The role of real-world evidence in FDA-approved new drug and biologics licence applications. Clin Pharmacol Ther 111:135–144. https://doi.org/10.1002/CPT.2474
18. US Food & Drug Administration (2018) Framework for FDA's real-world evidence program. https://www.fda.gov/media/120060/download. Accessed 12 Nov 2023
19. DARWIN EU (2023) Data analysis and real world interrogation network. https://darwin-eu.org/. Accessed 11 Nov 2023
20. Harbaugh CM, Cooper JN (2018) Administrative databases. Semin Pediatr Surg 27:353–360. https://doi.org/10.1053/J.SEMPEDSURG.2018.10.001
21. Sutherland SM (2020) Big data and pediatric acute kidney injury: the promise of electronic health record systems. Front Pediatr 7:536. https://doi.org/10.3389/FPED.2019.00536
22. Gavrielov-Yusim N, Friger M (2014) Use of administrative medical databases in population-based research. J Epidemiol Community Health 68:283–287. https://doi.org/10.1136/JECH-2013-202744
23. Ulrich EH, So G, Zappitelli M, Chanchlani R (2021) A review on the application and limitations of administrative health care data for the study of acute kidney injury epidemiology and outcomes in children. Front Pediatr 9:742888. https://doi.org/10.3389/fped.2021.742888
24. Shrestha S, Dave AJ, Losina E, Katz JN (2016) Diagnostic accuracy of administrative data algorithms in the diagnosis of osteoarthritis: a systematic review. BMC Med Inform Decis Mak 16:82. https://doi.org/10.1186/S12911-016-0319-Y
25. Mazzali C, Maistriello M, Ieva F, Barbieri P (2015) Methodological issues in the use of administrative databases to study heart failure. In: Paganoni AM, Secchi P (eds) Advances in complex data modelling and computational methods in statistics, Contributions to statistics. Springer, pp 149–160
26. Parker SL, McGirt MJ (2015) Editorial: administrative database research. J Neurosurg 122:441–442. https://doi.org/10.3171/2014.4.JNS14689
27. Zarrinkoub R, Wettermark B, Wändell P et al (2013) The epidemiology of heart failure, based on data for 2.1 million inhabitants in Sweden. Eur J Heart Fail 15:995–1002. https://doi.org/10.1093/EURJHF/HFT064
28. Lipscombe LL, Hux JE (2007) Trends in diabetes incidence, incidence, and mortality in Ontario, Canada 1995-2005: a population-based study. Lancet 369:750–756. https://doi.org/10.1016/S0140-6736(07)60361-4
29. Koopman C, Bots ML, Van Oeffelen AAM et al (2013) Population trends and inequalities in incidence and short-term outcome of acute myocardial infarction between 1998 and 2007. Int J Cardiol 168:993–998. https://doi.org/10.1016/J.IJCARD.2012.10.036
30. Mazzali C, Duca P (2015) Use of administrative data in healthcare research. Intern Emerg Med 10:517–524. https://doi.org/10.1007/S11739-015-1213-9
31. Gao J, Moran E, Li YF, Almenoff PL (2014) Predicting potentially avoidable hospitalizations. Med Care 52:164–171. https://doi.org/10.1097/MLR.0000000000000041
32. Miguel A, Marques B, Freitas A et al (2013) Detection of adverse drug reactions using hospital databases-a nationwide study in Portugal. Pharmacoepidemiol Drug Saf 22:907–913. https://doi.org/10.1002/PDS.3468

33. Loft S, Poulsen HE (1996) Cancer risk and oxidative DNA damage in man. J Mol Med (Berl) 74:297–312. https://doi.org/10.1007/BF00207507
34. Annaratone L, De Palma G, Bonizzi G et al (2021) Basic principles of biobanking: from biological samples to precision medicine for patients. Virchows Arch 479:233. https://doi.org/10.1007/S00428-021-03151-0
35. Hewitt R, Watson P (2013) Defining biobank. Biopreserv Biobank 11:309–315. https://doi.org/10.1089/BIO.2013.0042
36. Ballantyne A (2019) Adjusting the focus: a public health ethics approach to data research. Bioethics 33:357–366. https://doi.org/10.1111/BIOE.12551
37. Harris JR, Burton P, Knoppers BM et al (2012) Toward a roadmap in global biobanking for health. Eur J Hum Genet 20:1105–1111. https://doi.org/10.1038/EJHG.2012.96
38. Jacobs G, Wolf A, Krawczak M, Lieb W (2018) Biobanks in the era of digital medicine. Clin Pharmacol Ther 103:761–762. https://doi.org/10.1002/CPT.968
39. Dhai A (2016) The WMA declaration of Taipei: human databases and biobanks for the common good. S Afr J Bioeth Law 9:50–51. https://doi.org/10.7196/SAJBL.2016.v9i2.270
40. Small AM, O'Donnell CJ, Damrauer SM (2018) Large-scale genomic biobanks and cardiovascular disease. Curr Cardiol Rep 20. https://doi.org/10.1007/S11886-018-0969-8
41. Gaziano JM, Concato J, Brophy M et al (2016) Million veteran program: a mega-biobank to study genetic influences on health and disease. J Clin Epidemiol 70:214–223. https://doi.org/10.1016/J.JCLINEPI.2015.09.016
42. Celis-Morales CA, Lyall DM, Welsh P et al (2017) Association between active commuting and incident cardiovascular disease, cancer, and mortality: prospective cohort study. BMJ 357:j1456. https://doi.org/10.1136/bmj.j1456
43. National Center for Advancing Translational Sciences (2024) Rare Diseases Clinical Research Network (RDCRN). https://ncats.nih.gov/research/research-activities/rdcrn. Accessed 25 June 2024
44. Genomics England (2024) 100,000 genomes project. https://www.genomicsengland.co.uk/initiatives/100000-genomes-project. Accessed 25 June 2024
45. Kalia SS, Adelman K, Bale SJ et al (2017) Recommendations for reporting of secondary findings in clinical exome and genome sequencing, 2016 update (ACMG SF v2.0): a policy statement of the American College of Medical Genetics and Genomics. Genet Med 19:249–255. https://doi.org/10.1038/GIM.2016.190
46. Harris J, Sulston J (2004) Genetic equity. Nat Rev Genet 5:796–800. https://doi.org/10.1038/NRG1454
47. Editorial (2021) Precision medicine needs an equity agenda. Nat Med 27:737. https://doi.org/10.1038/S41591-021-01373-Y
48. Rotimi C, Abayomi A, Abimiku A et al (2014) Enabling the genomic revolution in Africa: H3Africa is developing capacity for health-related genomics research in Africa. Science 344:1346. https://doi.org/10.1126/SCIENCE.1251546
49. Wall JD, Stawiski EW, Ratan A et al (2019) The GenomeAsia 100K project enables genetic discoveries across Asia. Nature 576:106–111. https://doi.org/10.1038/S41586-019-1793-Z
50. Robine N, Varmus H (2022) New York's Polyethnic-1000: a regional initiative to understand how diverse ancestries influence the risk, progression, and treatment of cancers. Trends Cancer 8:269–272. https://doi.org/10.1016/J.TRECAN.2021.11.005
51. Gange SJ, Golub ET (2016) From smallpox to big data: the next 100 years of epidemiologic methods. Am J Epidemiol 183:423–426. https://doi.org/10.1093/AJE/KWV150
52. Ehrenstein V, Nielsen H, Pedersen AB et al (2017) Clinical epidemiology in the era of big data: new opportunities, familiar challenges. Clin Epidemiol 9:245–250. https://doi.org/10.2147/CLEP.S129779
53. Sanderson SC, Brothers KB, Mercaldo ND et al (2017) Public attitudes toward consent and data sharing in biobank research: a large multisite experimental survey in the US. Am J Hum Genet 100:414–427. https://doi.org/10.1016/J.AJHG.2017.01.021

54. D'Abramo F, Schildmann J, Vollmann J (2015) Research participants' perceptions and views on consent for biobank research: a review of empirical data and ethical analysis. BMC Med Ethics 16:60. https://doi.org/10.1186/S12910-015-0053-5
55. Coppola L, Cianflone A, Grimaldi AM et al (2019) Biobanking in health care: evolution and future directions. J Transl Med 17:172. https://doi.org/10.1186/S12967-019-1922-3
56. Boeckhout M, Zielhuis GA, Bredenoord AL (2018) The FAIR guiding principles for data stewardship: fair enough? Eur J Hum Genet 26:931–936. https://doi.org/10.1038/S41431-018-0160-0
57. Holub P, Kohlmayer F, Prasser F et al (2018) Enhancing reuse of data and biological material in medical research: from FAIR to FAIR-health. Biopreserv Biobank 16:97. https://doi.org/10.1089/BIO.2017.0110
58. Ranasinghe S, Pichler H, Eder J (2018) Report on data quality in biobanks: problems, issues, state-of-the-art. https://doi.org/10.48550/arXiv.1812.10423
59. Kodra Y, Weinbach J, Posada-De-La-Paz M et al (2018) Recommendations for improving the quality of rare disease registries. Int J Environ Res Public Health 15. https://doi.org/10.3390/IJERPH15081644
60. Moore HM, Compton CC, Alper J, Vaught JB (2011) International approaches to advancing biospecimen science. Cancer Epidemiol Biomarkers Prev 20:729. https://doi.org/10.1158/1055-9965.EPI-11-0021
61. OECD (2007) OECD best practice guidelines for biological resource centres. https://www.oecd.org/health/biotech/oecdbestpracticeguidelinesforbiologicalresourcecentres.htm. Accessed 26 June 2024
62. Campbell LD, Astrin JJ, DeSouza Y et al (2018) The 2018 revision of the ISBER best practices: summary of changes and the editorial team's development process. Biopreserv Biobank 16:3. https://doi.org/10.1089/BIO.2018.0001
63. National Cancer Institute (2016) NCI best practices for biospecimen resources
64. Mendy M, Caboux E, Lawlor RT et al (2017) Common minimum technical standards and protocols for biobanks dedicated to cancer research. International Agency for Research on Cancer
65. Council for International Organizations of Medical Sciences (CIOMS) in collaboration with the World Health Organization (WHO) (2016) International ethical guidelines for health-related research involving humans. www.cioms.ch. Accessed 26 June 2024
66. Nuffield Council on Bioethics (2015) The collection, linking and use of data in biomedical research and health care: ethical issues. A guide to the report. Accessed 26 June 2024
67. Recommendations of the European Society of Human Genetics (2003) Data storage and DNA banking for biomedical research: technical, social and ethical issues. Eur J Hum Genet 11:S8–S10. https://doi.org/10.1038/sj.ejhg.5201115
68. ISO (2018) ISO 20387:2018 - biotechnology — biobanking — general requirements for biobanking. https://www.iso.org/standard/67888.html. Accessed 26 June 2024
69. Conroy MC, Lacey B, Bešević J et al (2023) UK biobank: a globally important resource for cancer research. Br J Cancer 128:519–527. https://doi.org/10.1038/S41416-022-02053-5
70. Conroy M, Sellors J, Effingham M et al (2019) The advantages of UK Biobank's open-access strategy for health research. J Intern Med 286:389–397. https://doi.org/10.1111/JOIM.12955
71. Chadeau-Hyam M, Bodinier B, Elliott J et al (2020) Risk factors for positive and negative COVID-19 tests: a cautious and in-depth analysis of UK biobank data. Int J Epidemiol 49:1454–1467. https://doi.org/10.1093/IJE/DYAA134
72. Statistics Canada Biobank. https://www.statcan.gc.ca/en/microdata/biobank. Accessed 11 Nov 2023
73. Chen Z, Chen J, Collins R et al (2011) China Kadoorie Biobank of 0.5 million people: survey methods, baseline characteristics and long-term follow-up. Int J Epidemiol 40:1652–1666. https://doi.org/10.1093/IJE/DYR120
74. China Kadoorie Biobank (CKB). https://www.ckbiobank.org/. Accessed 11 Nov 2023

75. Nagai A, Hirata M, Kamatani Y et al (2017) Overview of the BioBank Japan project: study design and profile. J Epidemiol 27:S2–S8. https://doi.org/10.1016/J.JE.2016.12.005
76. BioBank Japan. https://biobankjp.org/en/index.html. Accessed 11 Nov 2023
77. Alver M, Palover M, Saar A et al (2019) Recall by genotype and cascade screening for familial hypercholesterolemia in a population-based biobank from Estonia. Genet Med 21:1173–1180. https://doi.org/10.1038/S41436-018-0311-2
78. Estonian Biobank. https://genomics.ut.ee/en/content/estonian-biobank. Accessed 11 Nov 2023
79. World Health Organization. Clinical trials. https://www.who.int/health-topics/clinical-trials/#tab=tab_1. Accessed 9 Nov 2023
80. NIH Grants & Funding NIH's. Definition of a clinical trial. https://grants.nih.gov/policy/clinical-trials/definition.htm. Accessed 26 June 2024
81. European Medicines Agency (2024) Clinical trials information system (CTIS): online training modules. https://www.ema.europa.eu/en/human-regulatory-overview/research-development/clinical-trials-human-medicines/clinical-trials-information-system-training-support/clinical-trials-information-system-ctis-online-training-modules. Accessed 26 June 2024
82. U.S. Food & Drug Administration (FDA) (2020) FDA continues to support transparency and collaboration in drug approval process as the clinical data summary pilot concludes. https://www.fda.gov/news-events/press-announcements/fda-continues-support-transparency-and-collaboration-drug-approval-process-clinical-data-summary. Accessed 26 June 2024
83. National Institutes of Health (NIH) (2022) NIH office of data science strategy announces new initiative to improve access to NIH-funded data. https://datascience.nih.gov/news/nih-office-of-data-science-strategy-announces-new-initiative-to-improve-data-access. Accessed 26 June 2024
84. Research Councils UK (2015) Guidance on best practice in the management of research data. http://www.oecd.org/sti/sci-tech/38500813.pdf. Accessed 26 June 2024
85. Ohmann C, Banzi R, Canham S et al (2017) Sharing and reuse of individual participant data from clinical trials: principles and recommendations. BMJ Open 7:18647. https://doi.org/10.1136/bmjopen-2017-018647
86. Medical Research Council (MRC) (2016) MRC policy on open research data from clinical trials and public health intervention studies. https://www.ukri.org/wp-content/uploads/2021/08/MRC-0208212-20200116_MRC-Policy-on-Open-Research-Data_FINAL-update.pdf. Accessed 12 Nov 2023
87. Miron L, Gonçalves RS, Musen MA (2020) Obstacles to the reuse of study metadata in ClinicalTrials.gov. Sci Data 7:443. https://doi.org/10.1038/S41597-020-00780-Z
88. Committee on Strategies for Responsible Sharing of Clinical Trial Data; Board on Health Sciences Policy; Institute of Medicine (2015) Guiding principles for sharing clinical trial data. National Academies Press (US)
89. Chan A-W, Song F, Vickers A et al (2014) Increasing value and reducing waste: addressing inaccessible research. Lancet 383:257–266. https://doi.org/10.1016/S0140-6736(13)62296-5
90. EFPIA, Barnes B (2019) Safeguards framework for secondary use of clinical trial data for scientific research. https://www.efpia.eu/media/413227/position-paper-safeguards-framework-for-secondary-use-of-clinical-trial-data-for-scientific-research-September-2019.pdf. Accessed 26 June 2024
91. Wilkinson T, Sinha S, Peek N, Geifman N (2019) Clinical trial data reuse—overcoming complexities in trial design and data sharing. Trials 20:513. https://doi.org/10.1186/s13063-019-3627-6
92. Wilkinson T, Sinha S, Peek N, Geifman N (2019) Clinical trial data reuse - overcoming complexities in trial design and data sharing. Trials 20:513. https://doi.org/10.1186/S13063-019-3627-6/FIGURES/1
93. O'Connor CM, van Veldhuisen DJ (2017) Data sharing from the editors' perspective. JACC Heart Fail 5:314–315. https://doi.org/10.1016/j.jchf.2017.02.007

94. Geifman N, Kennedy RE, Schneider LS et al (2018) Data-driven identification of endophenotypes of Alzheimer's disease progression: implications for clinical trials and therapeutic interventions. Alzheimers Res Ther 10:4. https://doi.org/10.1186/S13195-017-0332-0
95. Hill-McManus D, Hughes DA (2021) Combining model-based clinical trial simulation, pharmacoeconomics, and value of information to optimize trial design. CPT Pharmacometrics Syst Pharmacol 10:75–83. https://doi.org/10.1002/PSP4.12579
96. Nissen SE, Wolski K (2007) Effect of rosiglitazone on the risk of myocardial infarction and death from cardiovascular causes. N Engl J Med 356:2457–2471. https://doi.org/10.1056/NEJMOA072761
97. Fujiwara Y, Horita N, Adib E et al (2024) Treatment-related adverse events, including fatal toxicities, in patients with solid tumours receiving neoadjuvant and adjuvant immune checkpoint blockade: a systematic review and meta-analysis of randomized controlled trials. Lancet Oncol 25:62–75. https://doi.org/10.1016/S1470-2045(23)00524-7
98. Ambinder EP (2005) Electronic health records. J Oncol Pract 1:57–63. https://doi.org/10.1200/JOP.2005.1.2.57
99. Shah SM, Khan RA (2020) Secondary use of electronic health record: opportunities and challenges. IEEE Access 8:136947–136965. https://doi.org/10.1109/ACCESS.2020.3011099
100. Gianfrancesco MA, Goldstein ND (2021) A narrative review on the validity of electronic health record-based research in epidemiology. BMC Med Res Methodol 21:234. https://doi.org/10.1186/S12874-021-01416-5
101. Chaudhry B, Wang J, Wu S et al (2006) Systematic review: impact of health information technology on quality, efficiency, and costs of medical care. Ann Intern Med 144:742–752. https://doi.org/10.7326/0003-4819-144-10-200605160-00125
102. Sandhu E, Weinstein S, McKethan A, Jain SH (2012) Secondary uses of electronic health record data: benefits and barriers. Jt Comm J Qual Patient Saf 38:34–40. https://doi.org/10.1016/S1553-7250(12)38005-7
103. DiMasi JA, Hansen RW, Grabowski HG (2003) The price of innovation: new estimates of drug development costs. J Health Econ 22:151–185. https://doi.org/10.1016/S0167-6296(02)00126-1
104. Elkin PL, Trusko BE, Koppel R et al (2010) Secondary use of clinical data. Stud Health Technol Inform 155:14–29
105. Yim W-W, Wheeler AJ, Curtin C et al (2018) Secondary use of electronic medical records for clinical research: challenges and opportunities. Converg Sci Phys Oncol 4:014001. https://doi.org/10.1088/2057-1739/AAA905
106. European Commission (2024) European health data space. https://health.ec.europa.eu/ehealth-digital-health-and-care/european-health-data-space_en. Accessed 12 Jan 2024
107. Coorevits P, Sundgren M, Klein GO et al (2013) Electronic health records: new opportunities for clinical research. J Intern Med 274:547–560. https://doi.org/10.1111/joim.12119
108. Xiao Y, Wu J, Lin Z, Zhao X (2018) A deep learning-based multimodel ensemble method for cancer prediction. Comput Methods Prog Biomed 153:1–9. https://doi.org/10.1016/J.CMPB.2017.09.005
109. Willyard C (2017) The drug-resistant bacteria that pose the greatest health threats. Nature 543:15. https://doi.org/10.1038/NATURE.2017.21550
110. Spicknall IH, Looker KJ, Gottlieb SL et al (2019) Review of mathematical models of HSV-2 vaccination: implications for vaccine development. Vaccine 37:7396–7407. https://doi.org/10.1016/J.VACCINE.2018.02.067
111. Shuey MM, Lee KM, Keaton J et al (2023) A genetically supported drug repurposing pipeline for diabetes treatment using electronic health records. EBioMedicine 94:104674. https://doi.org/10.1016/J.EBIOM.2023.104674
112. Lanzillotta-Rangeley J, Clark A, Christianson A, Kalarchian MA (2020) Association of prescription opioid exposure and patient factors with prolonged postoperative opioid use in opioid-Naïve patients. AANA J 88:18–26

113. Atreja A, Gordon SM, Pollock DA et al (2008) Opportunities and challenges in utilizing electronic health records for infection surveillance, prevention, and control. Am J Infect Control 36:S37–S46. https://doi.org/10.1016/J.AJIC.2008.01.002
114. Keck JW, Redd JT, Cheek JE et al (2014) Influenza surveillance using electronic health records in the American Indian and Alaska native population. J Am Med Inform Assoc 21:132–138. https://doi.org/10.1136/AMIAJNL-2012-001591
115. Odero W, Rotich J, Yiannoutsos CT et al (2007) Innovative approaches to application of information technology in disease surveillance and prevention in Western Kenya. J Biomed Inform 40:390–397. https://doi.org/10.1016/J.JBI.2006.12.007
116. Loonsk JW (2004) BioSense--a national initiative for early detection and quantification of public health emergencies. MMWR Suppl 53:53–55
117. Kapoor A, Guha S, Kanti Das M et al (2020) Digital healthcare: the only solution for better healthcare during COVID-19 pandemic? Indian Heart J 72:61–64. https://doi.org/10.1016/J.IHJ.2020.04.001
118. Esposito P, Dal Canton A (2014) Clinical audit, a valuable tool to improve quality of care: general methodology and applications in nephrology. World J Nephrol 3:249. https://doi.org/10.5527/WJN.V3.I4.249
119. Litzelman DK, Dittus RS, Miller ME, Tierney WM (1993) Requiring physicians to respond to computerized reminders improves their compliance with preventive care protocols. J Gen Intern Med 8:311–317. https://doi.org/10.1007/BF02600144
120. Kucher N, Koo S, Quiroz R et al (2005) Electronic alerts to prevent venous thromboembolism among hospitalized patients. N Engl J Med 352:969–977. https://doi.org/10.1056/NEJMoa041533
121. CMS (2011) Medicare physician group practice demonstration. https://www.cms.gov/priorities/innovation/innovation-models/medicare-demonstrations/medicare-physician-group-practice-demonstration. Accessed 26 June 2024
122. Lee TH, Bothe A, Steele GD (2012) How Geisinger structures its physicians' compensation to support improvements in quality, efficiency, and volume. Health Aff (Millwood) 31:2068–2073. https://doi.org/10.1377/HLTHAFF.2011.0940
123. Paulus RA (2009) ProvenCare: Geisinger's model for care transformation through innovative clinical initiatives and value creation. Am Health Drug Benefits 2:122–127
124. Rung J, Brazma A (2013) Reuse of public genome-wide gene expression data. Nat Rev Genet 14:89–99. https://doi.org/10.1038/NRG3394
125. Domon B, Aebersold R (2006) Mass spectrometry and protein analysis. Science 312:212–217. https://doi.org/10.1126/SCIENCE.1124619
126. Tezel G (2013) A proteomics view of the molecular mechanisms and biomarkers of glaucomatous neurodegeneration. Prog Retin Eye Res 35:18–43. https://doi.org/10.1016/J.PRETEYERES.2013.01.004
127. FDA, CDER, CDRH, CBER (2018) E18 genomic sampling and management of genomic data guidance for industry. https://www.fda.gov/files/drugs/published/E18-Genomic-Sampling-and-Management-of-Genomic-Data-Guidance-for-Industry.pdf. Accessed 26 June 2024
128. McCarty CA, Chisholm RL, Chute CG et al (2011) The eMERGE network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. BMC Med Genet 4:13. https://doi.org/10.1186/1755-8794-4-13
129. European Commission (2024) Shaping Europe's digital future. European "1+ Million Genomes" Initiative. https://digital-strategy.ec.europa.eu/en/policies/1million-genomes. Accessed 1 Oct 2024
130. Tenopir C, Rice NM, Allard S et al (2020) Data sharing, management, use, and reuse: practices and perceptions of scientists worldwide. PLoS One 15:e0229003. https://doi.org/10.1371/JOURNAL.PONE.0229003
131. Curty RG, Crowston K, Specht A et al (2017) Attitudes and norms affecting scientists' data reuse. PLoS One 12:e0189288. https://doi.org/10.1371/JOURNAL.PONE.0189288

132. Subramanian SL, Kitchen RR, Alexander R et al (2015) Integration of extracellular RNA profiling data using metadata, biomedical ontologies and linked data technologies. J Extracell Vesicles 4:27497. https://doi.org/10.3402/JEV.V4.27497
133. Wade TD (2014) Refining gold from existing data. Curr Opin Allergy Clin Immunol 14:181–185. https://doi.org/10.1097/ACI.0000000000000051
134. Safran C (2017) Update on data reuse in health care. Yearb Med Inform 26:24–27. https://doi.org/10.15265/IY-2017-013
135. Raju HB, Tsinoremas NF, Capobianco E (2016) Emerging putative associations between non-coding RNAs and protein-coding genes in neuropathic pain: added value from reusing microarray data. Front Neurol 7:168. https://doi.org/10.3389/FNEUR.2016.00168
136. McKiernan EC, Bourne PE, Brown CT et al (2016) How open science helps researchers succeed. Elife 5:e16800. https://doi.org/10.7554/ELIFE.16800
137. Leitner F, Bielza C, Hill SL, Larrañaga P (2016) Data publications correlate with citation impact. Front Neurosci 10:419. https://doi.org/10.3389/FNINS.2016.00419
138. Ali-Khan SE, Harris LW, Gold ER (2017) Motivating participation in open science by examining researcher incentives. Elife 6:e29319. https://doi.org/10.7554/ELIFE.29319
139. NCBI Resource Coordinators (2017) Database resources of the National Center for Biotechnology Information NCBI Resource Coordinators. Nucleic Acids Res 45:D12–D17. https://doi.org/10.1093/nar/gkw1071
140. Sayers EW, Cavanaugh M, Clark K et al (2019) GenBank. Nucleic Acids Res 47:D94–D99. https://doi.org/10.1093/nar/gky989
141. Sayers EW, Agarwala R, Bolton EE et al (2019) Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 47:D23–D28. https://doi.org/10.1093/nar/gky1069
142. Perez-Riverol Y, Alpi E, Wang R et al (2015) Making proteomics data accessible and reusable: current state of proteomics databases and repositories. Proteomics 15:930–949. https://doi.org/10.1002/PMIC.201400302
143. Gostev M, Faulconbridge A, Brandizi M et al (2012) The BioSample Database (BioSD) at the European Bioinformatics Institute. Nucleic Acids Res 40:D64–D70. https://doi.org/10.1093/NAR/GKR937
144. Altelaar AFM, Munoz J, Heck AJR (2013) Next-generation proteomics: towards an integrative view of proteome dynamics. Nat Rev Genet 14:35–48. https://doi.org/10.1038/NRG3356
145. Distler U, Kuharev J, Navarro P et al (2014) Drift time-specific collision energies enable deep-coverage data-independent acquisition proteomics. Nat Methods 11:167–170. https://doi.org/10.1038/NMETH.2767
146. Lanucara F, Holman SW, Gray CJ, Eyers CE (2014) The power of ion mobility-mass spectrometry for structural characterization and the study of conformational dynamics. Nat Chem 6:281–294. https://doi.org/10.1038/NCHEM.1889
147. Riffle M, Eng JK (2009) Proteomics data repositories. Proteomics 9:4653–4663. https://doi.org/10.1002/PMIC.200900216
148. Craig R, Cortens JP, Beavis RC (2004) Open source system for analysing, validating, and storing protein identification data. J Proteome Res 3:1234–1242. https://doi.org/10.1021/PR049882H
149. Farrah T, Deutsch EW, Omenn GS et al (2014) The state of the human proteome in 2013 as viewed through PeptideAtlas: comparing the kidney, urine, and plasma proteomes for the biology and disease-driven human proteome project. J Proteome Res 13:60–75. https://doi.org/10.1021/PR4010037
150. Vizcaíno JA, Côté RG, Csordas A et al (2013) The PRoteomics IDEntifications (PRIDE) database and associated tools: status in 2013. Nucleic Acids Res 41. https://doi.org/10.1093/NAR/GKS1262
151. Slotta DJ, Barrett T, Edgar R (2009) NCBI peptidome: a new public repository for mass spectrometry peptide identifications. Nat Biotechnol 27:600–601. https://doi.org/10.1038/NBT0709-600

152. Smith BE, Hill JA, Gjukich MA, Andrews PC (2011) Tranche distributed repository and ProteomeCommons.org. Methods Mol Biol 696:123–145. https://doi.org/10.1007/978-1-60761-987-1_8
153. Ezkurdia I, Vázquez J, Valencia A, Tress M (2014) Analysing the first drafts of the human proteome. J Proteome Res 13:3854. https://doi.org/10.1021/PR500572Z
154. Kohane IS, Valtchinov VI (2012) Quantifying the white blood cell transcriptome as an accessible window to the multiorgan transcriptome. Bioinformatics 28:538–545. https://doi.org/10.1093/BIOINFORMATICS/BTR713
155. Lukk M, Kapushesky M, Nikkilä J et al (2010) A global map of human gene expression. Nat Biotechnol 28:322–324. https://doi.org/10.1038/nbt0410-322
156. Ojala KA, Kilpinen SK, Kallioniemi OP (2011) Classification of unknown primary tumors with a data-driven method based on a large microarray reference database. Genome Med 3:63. https://doi.org/10.1186/GM279
157. Tseng GC, Ghosh D, Feingold E (2012) Comprehensive literature review and statistical considerations for microarray meta-analysis. Nucleic Acids Res 40:3785–3799. https://doi.org/10.1093/NAR/GKR1265
158. Kang DD, Sibille E, Kaminski N, Tseng GC (2012) MetaQC: objective quality control and inclusion/exclusion criteria for genomic meta-analysis. Nucleic Acids Res 40:e15. https://doi.org/10.1093/NAR/GKR1071
159. Ramasamy A, Mondry A, Holmes CC, Altman DG (2008) Key issues in conducting a meta-analysis of gene expression microarray datasets. PLoS Med 5:e184. https://doi.org/10.1371/JOURNAL.PMED.0050184
160. Vilardell M, Rasche A, Thormann A et al (2011) Meta-analysis of heterogeneous down syndrome data reveals consistent genome-wide dosage effects related to neurological processes. BMC Genomics 12:229. https://doi.org/10.1186/1471-2164-12-229
161. Chen M, Wang K, Zhang L et al (2011) The discovery of putative urine markers for the specific detection of prostate tumor by integrative mining of public genomic profiles. PLoS One 6:e28552. https://doi.org/10.1371/JOURNAL.PONE.0028552
162. Gundem G, Perez-Llamas C, Jene-Sanz A et al (2010) IntOGen: integration and data mining of multidimensional oncogenomic data. Nat Methods 7:92–93. https://doi.org/10.1038/NMETH0210-92
163. Halling-Brown MD, Bulusu KC, Patel M et al (2012) canSAR: an integrated cancer public translational research and drug discovery resource. Nucleic Acids Res 40:D947–D956. https://doi.org/10.1093/NAR/GKR881
164. Lamb J, Crawford ED, Peck D et al (2006) The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease. Science 313:1929–1935. https://doi.org/10.1126/science.1132939
165. Brooke EM, WHO (1974) The current and future use of registers in health information systems. https://iris.who.int/handle/10665/36936. Accessed 26 June 2024
166. National Committee on Vital and Health Statistics (2022) Frequently asked questions about medical and public health registries What is a registry? https://ncvhs.hhs.gov/wp-content/uploads/2022/11/FAQ-on-Public-Health-Registries.pdf. Accessed 26 June 2024
167. National Program of Cancer Registries (NPCRs) | CDC. https://www.cdc.gov/cancer/npcr/index.htm. Accessed 11 Nov 2023
168. Hallstrom BR, Hughes RE, Huddleston JI (2022) State-based and National U.S. Registries: the Michigan Arthroplasty Registry Collaborative Quality Initiative (MARCQI), California Joint Replacement Registry (CJRR), and American Joint Replacement Registry (AJRR). J Bone Joint Surg Am 104:18–22. https://doi.org/10.2106/JBJS.22.00564
169. Choquet R, Landais P (2014) The French national registry for rare diseases: an integrated model from care to epidemiology and research. Orphanet J Rare Dis 9:O7. https://doi.org/10.1186/1750-1172-9-S1-O7

170. Newton J, Garner S, University of Oxford. Institute of Health Sciences (2002) Disease registers in England: a report commissioned by the Department of Health Policy Research programme in support of the White Paper entitled Saving lives: our healthier nation
171. Gliklich RE, Dreyer NA, Leavy MB (2014) Registries for evaluating patient outcomes: a user's guide, 3rd edn. AHRQ Methods for Effective Health Care, Rockville, MD
172. Schmidt M, Schmidt SAJ, Sandegaard JL et al (2015) The Danish National Patient Registry: a review of content, data quality, and research potential. Clin Epidemiol 7:449–490. https://doi.org/10.2147/CLEP.S91125
173. Pop B, Fetica B, Blaga ML et al (2019) The role of medical registries, potential applications and limitations. Med Pharm Rep 92:7–14. https://doi.org/10.15386/CJMED-1015
174. Levine MN, Julian JA (2016) Registries that show efficacy: good, but not good enough. J Clin Oncol 26(33):5316–5319. https://doi.org/10.1200/JCO.2008.18.3996
175. Awaya A, Kuroiwa Y (2020) The relationship between annual airborne pollen levels and occurrence of all cancers, and lung, stomach, colorectal, pancreatic and breast cancers: a retrospective study from the National Registry Database of Cancer Incidence in Japan, 1975-2015. Int J Environ Res Public Health 17:3950. https://doi.org/10.3390/IJERPH17113950
176. Dadonienė J, Charukevič G, Jasionytė G et al (2021) Mortality in inflammatory rheumatic diseases: Lithuanian National Registry Data and systematic review. Int J Environ Res Public Health 18:12338. https://doi.org/10.3390/IJERPH182312338
177. Harris IA, Kirwan DP, Peng Y et al (2022) Increased early mortality after total knee arthroplasty using conventional instrumentation compared with technology-assisted surgery: an analysis of linked national registry data. BMJ Open 12:e055859. https://doi.org/10.1136/BMJOPEN-2021-055859
178. Honkila M, Wikström E, Renko M et al (2017) Probability of vertical transmission of *Chlamydia trachomatis* estimated from national registry data. Sex Transm Infect 93:416–420. https://doi.org/10.1136/sextrans-2016-052884
179. Johnston DS (2017) Digital maturity: are we ready to use technology in the NHS? Future Healthc J 4:189–192. https://doi.org/10.7861/FUTUREHOSP.4-3-189
180. The National Cardiovascular Data Registry (2023) https://cvquality.acc.org/NCDR-Home. Accessed 11 Nov 2023
181. Kutcher MA, Klein LW, Ou FS et al (2009) Percutaneous coronary interventions in facilities without cardiac surgery on site: a report from the National Cardiovascular Data Registry (NCDR). J Am Coll Cardiol 54:16–24. https://doi.org/10.1016/J.JACC.2009.03.038
182. Cardona-Hernandez R, Schwandt A, Alkandari H et al (2021) Glycemic outcome associated with insulin pump and glucose sensor use in children and adolescents with type 1 diabetes. Data from the international pediatric registry SWEET. Diabetes Care 44:1176–1184. https://doi.org/10.2337/DC20-1674
183. Lanzinger S, Zimmermann A, Ranjan AG et al (2022) A collaborative comparison of international pediatric diabetes registries. Pediatr Diabetes 23:627–640. https://doi.org/10.1111/PEDI.13362
184. Van Dijk PCW, Jager KJ, Stengel B et al (2005) Renal replacement therapy for diabetic end-stage renal disease: data from 10 registries in Europe (1991-2000). Kidney Int 67:1489–1499. https://doi.org/10.1111/J.1523-1755.2005.00227.X
185. European Medicines Agency (2011) Committee for medicinal products for human use (CHMP) Guideline on the clinical investigation of recombinant and human plasma-derived factor VIII products. www.ema.europa.ee. Accessed 26 June 2024
186. Ali S, Kjeken R, Niederlaender C et al (2020) The European medicines agency review of Kymriah (Tisagenlecleucel) for the treatment of acute lymphoblastic leukemia and diffuse large B-cell lymphoma. Oncologist 25:e321–e327. https://doi.org/10.1634/THEONCOLOGIST.2019-0233
187. Papadouli I, Mueller-Berghaus J, Beuneu C et al (2020) EMA review of Axicabtagene Ciloleucel (Yescarta) for the treatment of diffuse large B-cell lymphoma. Oncologist 25:894–902. https://doi.org/10.1634/THEONCOLOGIST.2019-0646

188. Paper W, Shapiro M, Johnston D, et al (2012) Patient-generated health data white paper. Accessed 26 June 2024
189. Bourke A, Dixon WG, Roddam A et al (2020) Incorporating patient generated health data into pharmacoepidemiological research. Pharmacoepidemiol Drug Saf 29:1540–1549. https://doi.org/10.1002/PDS.5169
190. Codella J, Partovian C, Chang HY, Chen CH (2018) Data quality challenges for person-generated health and wellness data. IBM J Res Dev 62(1):1–3. https://doi.org/10.1147/JRD.2017.2762218
191. FDA, Office of the Commissioner (1998) Recruiting study subjects. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/recruiting-study-subjects. Accessed 27 June 2024
192. Office of the National Coordinator for Health Information Technology (2018) Conceptualizing a data infrastructure for the capture, use, and sharing of patient-generated health data in care delivery and research through 2024. Accessed 27 June 2024
193. Jim HSL, Hoogland AI, Brownstein NC et al (2020) Innovations in research and clinical care using patient-generated health data. CA Cancer J Clin 70:182–199. https://doi.org/10.3322/CAAC.21608
194. Huang Y, Upadhyay U, Dhar E et al (2022) A scoping review to assess adherence to and clinical outcomes of wearable devices in the cancer population. Cancers (Basel) 14:4437. https://doi.org/10.3390/CANCERS14184437
195. Linwood SL (2022) Digital health. Exon Publications
196. Krist AH, Tong ST, Aycock RA, Longo DR (2017) Engaging patients in decision-making and behavior change to promote prevention. Inf Serv Use 37:105–122. https://doi.org/10.3233/ISU-170826
197. Abdolkhani R, Gray K, Borda A, DeSouza R (2019) Patient-generated health data management and quality challenges in remote patient monitoring. JAMIA Open 2:471–478. https://doi.org/10.1093/JAMIAOPEN/OOZ036
198. Health IT Buzz, DeSalvo KB, Samuel J (2016) Examining oversight of the privacy & security of health data collected by entities not regulated by HIPAA. https://www.healthit.gov/buzz-blog/privacy-and-security/examining-oversight-privacy-security-health-data-collected-entities-not-regulated-hipaa. Accessed 27 June 2024
199. European Commission (2018) Reform of EU data protection rules. https://commission.europa.eu/law/law-topic/data-protection/reform_en. Accessed 27 June 2024
200. FDA, OHRP, CDER, et al (2016) Use of electronic informed consent questions and answers guidance for institutional review boards, investigators, and sponsors. https://www.fda.gov/media/116850/download. Accessed 27 June 2024
201. Mooney KH, Beck SL, Friedman RH et al (2014) Automated monitoring of symptoms during ambulatory chemotherapy and oncology providers' use of the information: a randomized controlled clinical trial. Support Care Cancer 22:2343–2350. https://doi.org/10.1007/s00520-014-2216-1
202. Wheelock AE, Bock MA, Martin EL et al (2015) SIS.NET: a randomized controlled trial evaluating a web-based system for symptom management after treatment of breast cancer. Cancer 121:893–899. https://doi.org/10.1002/CNCR.29088
203. Kearney N, McCann L, Norrie J et al (2009) Evaluation of a mobile phone-based, advanced symptom management system (ASyMS) in the management of chemotherapy-related toxicity. Support Care Cancer 17:437–444. https://doi.org/10.1007/S00520-008-0515-0
204. Nelson BW, Allen NB (2019) Accuracy of consumer wearable heart rate measurement during an ecologically valid 24-hour period: intraindividual validation study. JMIR Mhealth Uhealth 7:e10828. https://doi.org/10.2196/10828
205. Thomson EA, Nuss K, Comstock A et al (2019) Heart rate measures from the Apple Watch, Fitbit Charge HR 2, and electrocardiogram across different exercise intensities. J Sports Sci 37:1411–1419. https://doi.org/10.1080/02640414.2018.1560644

206. Haberman ZC, Jahn RT, Bose R et al (2015) Wireless smartphone ECG enables large-scale screening in diverse populations. J Cardiovasc Electrophysiol 26:520–526. https://doi.org/10.1111/JCE.12634
207. Klesges RC, Klbsges LM, Swenson AM, Pheley AM (1985) A validation of two motion sensors in the prediction of child and adult physical activity levels. Am J Epidemiol 122:400–410. https://doi.org/10.1093/OXFORDJOURNALS.AJE.A114121
208. Breteler MJM, Janssen JH, Spiering W et al (2019) Measuring free-living physical activity with three commercially available activity monitors for telemonitoring purposes: validation study. JMIR Form Res 3:e11489. https://doi.org/10.2196/11489
209. Zhang P, Godin SD, Owens MV (2019) Measuring the validity and reliability of the Apple Watch as a physical activity monitor. J Sports Med Phys Fitness 59:784–790. https://doi.org/10.23736/S0022-4707.18.08339-1
210. Bai Y, Hibbing P, Mantis C, Welk GJ (2018) Comparative evaluation of heart rate-based monitors: Apple Watch vs Fitbit charge HR. J Sports Sci 36:1734–1741. https://doi.org/10.1080/02640414.2017.1412235
211. Nolan M, Mitchell JR, Doyle-Baker PK (2014) Validity of the Apple iPhone®/iPod Touch® as an accelerometer-based physical activity monitor: a proof-of-concept study. J Phys Act Health 11:759–769. https://doi.org/10.1123/JPAH.2011-0336
212. Wu W, Dasgupta S, Ramirez EE et al (2012) Classification accuracies of physical activities using smartphone motion sensors. J Med Internet Res 14:e130. https://doi.org/10.2196/JMIR.2208
213. Wind DK, Sapiezynski P, Furman MA, Lehmann S (2016) Inferring stop-locations from WiFi. PLoS One 11:e0149105. https://doi.org/10.1371/journal.pone.0149105
214. Geyer K, Ellis DA, Piwek L (2019) A simple location-tracking app for psychological research. Behav Res Methods 51:2840–2846. https://doi.org/10.3758/s13428-018-1164-y
215. Zhou X, Li D (2018) Quantifying multidimensional attributes of human activities at various geographic scales based on smartphone tracking. Int J Health Geogr 17:11. https://doi.org/10.1186/s12942-018-0130-3
216. Ruktanonchai NW, Ruktanonchai CW, Floyd JR, Tatem AJ (2018) Using Google location history data to quantify fine-scale human mobility. Int J Health Geogr 17:28. https://doi.org/10.1186/s12942-018-0150-z
217. Saeb S, Zhang M, Karr CJ et al (2015) Mobile phone sensor correlates of depressive symptom severity in daily life behavior: an exploratory study. J Med Internet Res 17:e175. https://doi.org/10.2196/jmir.4273
218. Clauser SB, Wagner EH, Aiello Bowles EJ et al (2011) Improving modern cancer care through information technology. Am J Prev Med 40:S198–S207. https://doi.org/10.1016/J.AMEPRE.2011.01.014
219. Place S, Blanch-Hartigan D, Rubin C et al (2017) Behavioral indicators on a mobile sensing platform predict clinically validated psychiatric symptoms of mood and anxiety disorders. J Med Internet Res 19:e75. https://doi.org/10.2196/JMIR.6678
220. Peterson SK, Shinn EH, Basen-engquist K et al (2013) Identifying early dehydration risk with home-based sensors during radiation treatment: a feasibility study on patients with head and neck cancer. J Natl Cancer Inst Monogr 2013:162–168. https://doi.org/10.1093/JNCIMONOGRAPHS/LGT016
221. Carter MC, Albar SA, Morris MA et al (2015) Development of a UK online 24-h dietary assessment tool: myfood24. Nutrients 7:4016–4032. https://doi.org/10.3390/nu7064016
222. Liu B, Young H, Crowe FL et al (2011) Development and evaluation of the Oxford WebQ, a low-cost, web-based method for assessment of previous 24 h dietary intakes in large-scale prospective studies. Public Health Nutr 14:1998–2005. https://doi.org/10.1017/S1368980011000942
223. Lo F, Sun Y, Qiu J, Lo B (2018) Food volume estimation based on deep learning view synthesis from a single depth map. Nutrients 10:2005. https://doi.org/10.3390/nu10122005

224. Resende e Silva BV, Rad MG, Cui J et al (2018) A mobile-based diet monitoring system for obesity management. J Health Med Inform 09:307. https://doi.org/10.4172/2157-7420.1000307
225. University of Pittsburgh, Health Sciences Library System Guides (2024) Social & behavioral determinants of health. https://hsls.libguides.com/social-behavioral-determinants-of-health-data. Accessed 27 June 2024
226. Bompelli A, Wang Y, Wan R et al (2021) Social and behavioral determinants of health in the era of artificial intelligence with electronic health records: a scoping review. Health data. Science 2021:9759016. https://doi.org/10.34133/2021/9759016
227. WHO (2024) Social determinants of health. https://www.who.int/health-topics/social-determinants-of-health#tab=tab_1. Accessed 27 June 2024
228. Clark MA, Gurewich D (2017) Integrating measures of social determinants of health into health care encounters: opportunities and challenges. Med Care 55:807–809. https://doi.org/10.1097/MLR.0000000000000788
229. Alley DE, Asomugha CN, Conway PH, Sanghavi DM (2016) Accountable health communities--addressing social needs through medicare and medicaid. N Engl J Med 374:8–11. https://doi.org/10.1056/NEJMP1512532
230. Ash AS, Mick EO, Ellis RP et al (2017) Social determinants of health in managed care payment formulas. JAMA Intern Med 177:1424–1430. https://doi.org/10.1001/JAMAINTERNMED.2017.3317
231. Chen M, Tan X, Padman R (2020) Social determinants of health in electronic health records and their impact on analysis and risk prediction: a systematic review. J Am Med Inform Assoc 27:1764–1773. https://doi.org/10.1093/JAMIA/OCAA143
232. Deferio JJ, Breitinger S, Khullar D et al (2019) Social determinants of health in mental health care and research: a case for greater inclusion. J Am Med Inform Assoc 26:895–899. https://doi.org/10.1093/JAMIA/OCZ049
233. Zhang-James Y, Chen Q, Kuja-Halkola R et al (2020) Machine-learning prediction of comorbid substance use disorders in ADHD youth using Swedish registry data. J Child Psychol Psychiatry 61:1370–1379. https://doi.org/10.1111/JCPP.13226
234. Afshar M, Joyce C, Dligach D et al (2019) Subtypes in patients with opioid misuse: a prognostic enrichment strategy using electronic health record data in hospitalized patients. PLoS One 14:e0219717. https://doi.org/10.1371/journal.pone.0219717
235. Zheng L, Wang O, Hao S et al (2020) Development of an early-warning system for high-risk patients for suicide attempt using deep learning and electronic health records. Transl Psychiatry 10:72. https://doi.org/10.1038/s41398-020-0684-2
236. Chen Q, Zhang-James Y, Barnett EJ et al (2020) Predicting suicide attempt or suicide death following a visit to psychiatric specialty care: a machine learning study using Swedish national registry data. PLoS Med 17:e1003416. https://doi.org/10.1371/journal.pmed.1003416
237. Walsh CG, Ribeiro JD, Franklin JC (2018) Predicting suicide attempts in adolescents with longitudinal clinical data and machine learning. J Child Psychol Psychiatry 59:1261–1270. https://doi.org/10.1111/JCPP.12916
238. Senior M, Burghart M, Yu R et al (2020) Identifying predictors of suicide in severe mental illness: a feasibility study of a clinical prediction rule (Oxford mental illness and suicide tool or OxMIS). Front Psych 11:268. https://doi.org/10.3389/fpsyt.2020.00268
239. Poulymenopoulou M, Papakonstantinou D, Malamateniou F, Vassilacopoulos G (2015) A health analytics semantic ETL service for obesity surveillance. Stud Health Technol Inform 210:840–844
240. Dagliati A, Sacchi L, Tibollo V et al (2018) A dashboard-based system for supporting diabetes care. J Am Med Inform Assoc 25:538–547. https://doi.org/10.1093/JAMIA/OCX159
241. Zhang H, Hosomura N, Shubina M et al (2016) Electronic documentation of lifestyle counselling in primary care is associated with lower risk of cardiovascular events in patients with diabetes. Diabetes 65:A363

242. Davoudi A, Ozrazgat-Baslanti T, Ebadi A et al (2017) Delirium prediction using machine learning models on predictive electronic health records data. In: Proceedings - 2017 IEEE 17th International Conference on Bioinformatics and Bioengineering, BIBE 2017. Washington, DC, USA, pp 568–573
243. Nau C, Ellis H, Huang H et al (2015) Exploring the forest instead of the trees: an innovative method for defining obesogenic and obesoprotective environments. Health Place 35:136–146. https://doi.org/10.1016/J.HEALTHPLACE.2015.08.002