Benefits and Risks of Using Health Data for Research

Ségolène Aymé, R Choquet, L Devillers, M Gilard, M Kelly-Irving, A Livartowski, B Lukacs

Health Data Hub
Scientific Advisory Board

Benefits and Risks of Using Health Data for Research

Report – October 2023

Authors:

Ségolène Aymé, (Chair of the Health Data Hub Scientific Advisory Board, Inserm); Rémy Choquet, Roche SAS); Laurence Devillers (Sorbonne University); Martine Gilard (Brest University Hospital); Michelle Kelly-Irving (Inserm); Alain Livartowski, (Unicancer Paris); Bertrand Lukacs (Vice-Chair of the Health Data Hub Scientific Advisory Board, French National Academy of Surgery).

To cite this paper:

S. Aymé, R. Choquet, L. Devillers, M. Gilard, M. Kelly-Irving, A. Livartowski, B. Lukacs, "Bénéfices et risques de l'utilisation des données de santé à des fins de recherche", Conseil scientifique consultatif du Health Data Hub, October 2023.

Acknowledgements:

The Working Group would like to thank Pierre Lombrail (Honorary Professor of Public Health, Education and Health Promotion Laboratory, Sorbonne Paris Nord University, and a member of the Inserm Ethics Committee) for his valuable comments, which enriched the discussion and assisted with the drafting of this document.

Executive summary of the paper

As the digitalisation of the healthcare system becomes a reality in developed countries, the availability of large volumes of health data is raising stakeholders' expectations, sometimes excessively, and concerns, sometimes groundlessly. A review of the benefits that can legitimately be expected from exploiting the available data is proposed, while considering the adverse effects that have already been documented or can legitimately be anticipated, before determining where the benefit/risk balance stands. While the current criteria place the emphasis on the absolute protection of privacy, which hinders or delays many highly relevant studies, we propose that the benefit/risk balance should henceforth form the basis for decisions to authorise the use of health data for research purposes.

Many **benefits** are expected from providing access to this data :
(1) Facilitating the epidemiological surveillance of major diseases, upon which many public health decisions are based.
(2) Enabling the measurement of healthcare efficacy and the detection of local or regional situations that deviate significantly from the mean.
(3) Facilitating the monitoring of changes in practices over time and their impact.
(4) Helping to document the positive and negative effects of innovations introduced as a result of changes in practices, be they medicinal, technical or organisational.
(5) Providing information about patients' access to healthcare services, allowing rates and delays to be calculated and cross-referenced with geographical variables and socio-economic data in order to identify inequalities, effective and ineffective strategies, and determine which areas should be prioritised for improvement.
(6) Contributing directly to improving knowledge – an essential step towards improving practices and designing effective and protective public policies.
(7) Playing an essential role in developing the uses of artificial intelligence, based on new, high-quality algorithms which can be trained and tested on representative populations, which is not always the case at present.
(8) Enabling the assessment of the effectiveness of prevention or screening campaigns, and of the extent to which best practice recommendations are followed.

However, the use of such data by multiple stakeholders is not without **risks**.
(1) The risk of revealing personal data is a major risk for personal data, but very minor for pseudonymised data and non-existent for anonymised data.
(2) The risk of uncontrolled use, beyond the need to inform people, depends on operators complying with the rules for use, and is very closely monitored;
(3) The risk of data theft and espionage by industrial operators or governments is often mentioned by opponents, but it is extremely low because non-identifying data has no commercial value, and this type of theft has never occurred.
(4) The risk of infringement of intellectual property rights and of financial loss is often mentioned as an argument against the valorisation of this data but has not been documented;
(5) Healthcare professionals are concerned about the risk of normative surveillance, although such surveillance is legitimate from the perspective of healthcare system funders, and tensions can be eased through dialogue.

The potential risks to individuals and the community are minimal if the protective measures adopted are effectively implemented and none of the risks are liable to have a significant impact (on health), either individually or collectively. However, the use of such data can have substantial benefits for society (e.g. appropriateness of care, prevention of health scandals, evaluation of innovations) and for the individual (such as the development of new

treatments), and may also have an impact on individuals themselves. The benefit/risk balance therefore leans strongly towards the benefits.

This leaves the question of how to **remove obstacles** to the provision of access to data collections and promote a culture of the common good represented by health data throughout all levels of society. These obstacles are numerous and include:

(1) a ban on the use of US data-storage tools in the absence of a European cloud, even though there have never been any complaints in the healthcare sector against US manufacturers for failing to comply with the GDPR in the healthcare field, nor any documented attempts to re-identify anonymised health research data;

(2) a ban on using the social security numbers recorded in the French National Social Security Register to match databases that need to be cross-referenced, despite this practice being possible in many European countries that are subject to the same European regulations as France;

(3) waiting times running into years for access to the main French National Health Data System (Système National des Données de Santé – SNDS) database, whereas this queue could be reduced if a copy of the main SNDS database were included in the HDH catalogue;

(4) an ecosystem that is overly complex for historical reasons, and which needs to be simplified in order to move towards a Single Gateway, as recommended in the European directive in preparation;

(5) public communications focusing solely on the risks and not presenting the benefits, despite well-informed citizens being in favour of the use of their data for research.

This is likely to have a major impact on the feasibility of many studies. Auditing the barriers to research within the current system could help to convince the doubters.

Introduction

The availability of large volumes of health data as a result of the growing digitalisation of healthcare systems worldwide is raising stakeholders' expectations, sometimes excessively, and concerns, sometimes groundlessly. Therefore, this is the right time to review the benefits that can legitimately be expected from exploiting the available data, while considering the adverse effects that have already been documented or can reasonably be anticipated, before determining where the benefit/risk balance stands.

This approach is in line with several recently adopted public policies, notably the French law on the organisation and transformation of the healthcare system (1), and the "Towards Open Science" plan (2), transposing a European directive. Certain theoretical arguments advocate making better use of the data available for public health decision-making and accelerating research and development with a view to improving people's quality of life (3). All of the surveys conducted to date show that citizens expect health data to be used to improve our public health and our healthcare system (4).

Data sharing refers to the circulation of data among many different actors, on the basis of multiple dispersed databases that are little known to the outside world and difficult to use by anyone not involved in their creation. The procedures for accessing health data may be perceived as complex due to the sensitivity of such data, which are subject to different governance systems, sometimes on a discretionary basis. The tools and expertise required to process data securely, which is a legitimate requirement, are expensive and often inaccessible, particularly for small research teams, institutions whose main activity is not research, and start-ups.

France is already well on the way to pooling the available health data, notably through its 2019 Health Law, which has extended the national health data system (Système national des données de santé – SNDS) and created a federative data access platform: the Health Data Hub (5). These choices give rise to questions, and sometimes create tensions, often linked to the complexity of the subject, which is why it is so important to provide the additional information needed to fully understand the merits of the path taken (6,7).

The French National Advisory Committee for Life Sciences and Health (Comité Consultatif National d'Ethique – CCNE), in its opinion No. 130 (8), helps to provide some answers to these questions in its analysis covering all ethical aspects of the digitalisation of data, which extend far beyond the scope of this paper devoted solely to the ethical issues raised by the use of health data for research, but not for healthcare.

On 9 May 2023, the CCNE and the French National Pilot Committee for Digital Ethics (Comité national pilote d'éthique du numérique – CNPEN) also published a joint opinion on health data platforms (9). These two committees made 21 recommendations in all: on data sharing and anonymisation, but also on sovereignty and environmental impact. The CCNE and the CNPEN are paying particular attention to the control of health data from private clinics, nursing homes for dependent elderly people, and more recently, biology platforms, all of which have been largely acquired by financial groups linked to international investment funds.

The aim of this paper is to present the implications of the open science policy for health data, and thereby help to provide the information required to hold an informed debate, and especially to assess the benefit/risk balance, which is key to informed decision-making.

It also discusses the institutional and regulatory ecosystem that guarantees the protection of privacy and public confidence, an ecosystem that should not hinder research by prioritising

individual rights over collective rights, and by only considering the potential risks without assessing the expected benefits.

It also provides a reminder that the research ecosystem is international and that national initiatives must be compatible with other research regulation systems in order to facilitate transnational partnerships, which are essential to research and innovation.

The ultimate purpose of this study is to propose a review of the benefits and risks generated by the reuse of health data for research, and to suggest changes to the health data provision ecosystem for research, in the best interests of public health and the quality of the healthcare system.

1. Health data and its organisation

"Health data" means any information relating to a person's physical or mental state of health (10). This may include information relating to the identification of a person for health purposes (number, symbol, etc.), information about tests or examinations including genetic and biological data, information relating to illnesses, symptoms, treatments, disabilities, history and lifestyle data if it has a potential impact on health, or information about the consumption of healthcare or services, particularly via medical devices or the healthcare reimbursement process.

Such data therefore includes any information produced during consultations, hospitalisations and radiological or biological investigations, computerised and paper-based surveys, and administrative records. In addition, data that does not appear to be health data is considered as such if – due to its cross-referencing with other data – it enables a conclusion to be drawn about a person's state of health or a health risk, such as the cross-referencing of a weight measurement with other data (e.g. number of steps or a calorie intake measurement). Similarly, data may become health data because of its intended purpose, i.e. its use for medical purposes.

Only digitised data is considered in this paper, as this is the only directly usable data for research purposes.

We will now examine the different sources of health data that can be used for research purposes.

1.1. Administrative data

Many government bodies use health data when reimbursing healthcare costs, managing the health system and ensuring the quality of services provided to individuals. The main collection of medico-administrative data in France is the main database of the National Health Data System (Système national des données de santé – SNDS), created in 2016. This database compiles and provides access to health information collected by public bodies. This data is pseudonymised, i.e. the social security number recorded in the French National Social Security Register (numéro d'inscription au répertoire – NIR) is replaced by a randomly generated pseudonym. It includes information from three databases: the National Health Insurance Information System (SNIIRAM); the Medicalisation of Information Systems Programme (PMSI), which centralises billing information from healthcare establishments; and the Epidemiological Centre for the Medical Causes of Death Database (CépiDC), which is populated by data from death certificates. A single person has the same pseudonym in all three databases, and information about him or her can be compared. The Law of 2016 provided for the enrichment of this main SNDS database with two other sources of

medico-administrative data: "medico-social" data from departmental centres for disabled persons, and a representative sample of reimbursement data from complementary voluntary health insurance bodies.

French data assets have also been enriched with other sources of data, such as Santé Publique France data on visits to hospital accident and emergency departments and notifiable disease-reporting data.

Other administrative sources of data managed by different bodies, such as INSEE (French National Institute for Statistics and Economic Studies), are of great interest for research. These sources include the permanent demographic sample (Échantillon démographique permanent – EDP), a large-scale socio-demographic panel created in 1967, which centralises information of an essentially demographic nature from five sources: civil status records of births, marriages and deaths since 1968, data from censuses between 1968 and 1999 and from annual census surveys since 2004, data from the electoral register since 1967, data from the "all employees" panel since 1967, and socio-fiscal data since 2011. The EDP can be accessed by researchers or organisations outside the French Official Statistical Service via the Secure Data Access Centre, after consulting the Statistical Confidentiality Committee.

1.2. Healthcare data and hospital health data warehouses

Healthcare institutions also record the healthcare activity relating to each user. Patient records are increasingly computerised to improve the traceability of care and the sharing of information between professionals involved in healthcare. The data collected is generally reduced to the essentials so as to avoid adding to the workload of healthcare professionals. This data remains of great interest for research purposes, although it is not always adapted to this activity. Hospitals previously lacked the resources required to exploit this data for research purposes, but this situation is now changing: hospital data warehouses are being created and rolled out nationwide (11) and will enable secondary uses of this data. These data warehouses will indeed contain the proportion of healthcare data considered usable for research purposes and of interest, which will be pseudonymised.

Similarly, data on healthcare provided by private practitioners is a valuable source of information on the population's state of health, as more and more medical practices become computerised. However, a wide variety of practice-management software is currently used. The Platform for Data in Primary Care (P4DP) project (12), a consortium led by the National Board of Teachers in General Practice (CNGE), aims to create a national data warehouse for non-hospital healthcare data over the next three years, which will finally provide access to real-life non-hospital healthcare data.

Administrative data and healthcare data has the advantage of being collected routinely and exhaustively. For the most part, such data covers all citizens with health insurance, i.e. over 95% of the population. Its use in research enables it to be exploited at lower cost for significant collective benefit. However, the regulatory and technical measures implemented to protect such sensitive data mean that significant efforts must be made to make this data accessible to researchers. As we shall see, the data subjects must not have objected to the use of their data for research purposes, and the data itself must exist in a format that meets national or international technical and semantic standards.

1.3. Cohorts, registers and collections of data for research

Over the years, health researchers have built up substantial databases in order to obtain the data they need to validate or invalidate research questions. These databases encompass a variety of formats, ranging from registers collecting health events on a cross-disciplinary basis to cohorts (13) whose aim is to follow up, on a longitudinal basis, people who have

agreed to provide their data for research purposes. Clinical trial data, which is used to assess the benefits and risks of new treatments, is also a valuable source of data. Another category consists of survey data used to investigate the medical, social and psychological dimensions, in specific volunteer populations or in the general population. Connected objects have also recently begun to generate data, which is intended for monitoring purposes but can also be reused for other types of research. All of this data is formatted for research purposes and collected within a framework requiring individual personal notification, which is not always possible. This makes the data directly usable for purposes other than the original purpose, provided that the people in question have been notified and can exercise their rights. However, these data collections are expensive to maintain and limited in their geographical scope. A recent analysis by the French High Council for Public Health (Haut Conseil de la Santé Publique – HCSP) highlights these limitations and proposes solutions, which include matching these collections with the main SNDS database (14).

1.4. Environmental and social data that can affect health

A wide range of environmental data can be used to qualify and describe the impact of environmental factors on health, including sources of nuisance, environmental contaminants and land use. Examples of environmental data relevant to the field of health and the environment include air-quality data from the National Institute on Industrial Environment and risks (Institut national de l'environnement industriel et des risques – INERIS) (Géod'air, PREV'AIR, Cartothèque) or collected by approved air-quality agencies (AASQA), water-quality data from the regional Health Agencies (Agence régionale de santé – ARS) (SISE-Eaux), meteorological data from Météo France, data on polluted sites and soils (Infosols), noise mapping by the Studies and Expertise Centre on Risks, Mobility and Management (Centre d'études et d'expertise sur les risques, la mobilité et l'aménagement – Cerema) (plaMADE) and data on the sale of plant protection products managed by the French Office of Biodiversity (Office français de la biodiversité – OFB) (BNV-D).

Human environments are characterised by a social hierarchy, which gives rise to a social health gradient, i.e. the relationship between a person's social status and state of health. Indeed, the higher one's status in the social hierarchy, the better one's health, and vice versa.[1] To fully understand how social inequalities in health develop throughout life and according to different social categories (gender, socio-economic status, ethnic group, place of residence, etc.), it is vital for high-quality social data to be chained and analysed in relation to the various types of health data (diagnosis, incidence, follow-up, treatment, biomarkers). This will provide evidence for intervention research aiming to prevent and reduce social inequalities in health.

1.5. The SNDS is intended to represent the diversity of health data

Initially composed of healthcare claims data such as healthcare reimbursement forms, hospital billing data and medical causes of death data, the French National Health Data System (SNDS) was extended by the Law on the organisation and transformation of the healthcare system of 24 July 2019 to include all healthcare data that receives public funding, with the aim of expanding the pool of available data and thereby contributing to the broader use of data.

[1] Over the 2012-2016 period, life expectancy at birth for the richest 5% of the population was 84.4 years for men and 88.3 years for women, compared to 71.7 years and 80 years respectively among the poorest 5%, corresponding to a socio-economic gap of thirteen years for men and eight years for women (Blanpain, 2018) (15). Moreover, the most disadvantaged social categories are doubly penalised, facing a shorter lifespan and a poorer quality of life than the more affluent social categories. The most disadvantaged populations in France also live in areas in which they are more exposed to myriad forms of environmental pollution that are harmful to human health (16).

Therefore, the SNDS now includes data from registers, research cohorts, hospital data warehouses, etc.

By law, the HDH is responsible for compiling, organising and ensuring the availability of data from the SNDS and promoting innovation in the use of health data. The decree published in June 2021 stipulates that the HDH is jointly responsible for processing the main database with the French National Health Insurance Fund (Conseil national de l'Assurance maladie – CNAM) and is the data controller for the SNDS catalogue. These two subsets of the SNDS are of major interest to the ecosystem.

The main database is therefore a compilation of data covering the entire population, from the French Health Insurance System (Assurance Maladie) (SNIIRAM database), establishments (PMSI database), medical causes of death (Inserm CépiDC database), and disability-related data (MDPH - CNSA data), as a priority. It is intended to be continuously enriched by integrating other national databases, including, under this first decree, databases relating to the epidemic: on COVID vaccinations (Vaccin-covid) and the screening and information system (SI-DEP).

The catalogue is an evolving collection of databases. Created iteratively, it can be adapted to the challenges and needs of the ecosystem, and its content is defined by government order, as are the main database flows. The first version of this order was published on 12 May 2022 and provides for 10 databases that make up the first version of the SNDS catalogue. These include the National Rare Disease Databank (Banque nationale de données maladies rares – BNDMR), which centralises the computerised patient records created by the reference centres, the database covering visits to accident and emergency departments (OSCOUR), and the database compiling surveillance data on 33 notifiable diseases (MDO), which aims to prevent the risk of epidemics. A second version of the SNDS catalogue is currently being developed.

2. Expected benefits from its use

Many pre-existing collections of data can be linked together for cross-fertilisation of data and can make a significant contribution to advancing our knowledge. However, the benefits of this data pooling for research are not clearly identified by many members of the public, health professionals and decision-makers, because the communication strategy targeting these audiences focuses mainly on the risks rather than on the benefits.

A review of the potential benefits is therefore presented below and summarised in the appended table:

2.1. General benefits: critical mass, representativeness of the populations studied and complementarity

There are obvious benefits to using and re-using data that is already available. At the very least, it provides a return on the costs invested in data collection, and at the most, an opportunity to significantly improve knowledge and practices. This advantage is all the greater when the event under analysis is rare, as it requires a large number of operators, which is never feasible for a single research team or healthcare institution. The pooling of databases provides access to the number of operators required to draw conclusions about rare events and compare events between geographical regions or social groups. It increases the power of studies and the robustness of the results.

The cross-referencing of databases enhances the data contained in each one. For example, data on deaths in France can be used to ascertain whether the people listed in databases specialising in a specific disease are still alive or have died – information that is impossible to obtain by any other means. Variations in disease frequency between regions can be explained by correlations with environmental exposures obtained from databases for industrial sites, landfill sites, incinerators, or levels of air or water pollution, for example. This type of work can only be accomplished by cross-referencing databases.

The proper management of sensitive data and the optimal use of complex databases designed by others are difficult to implement because they require rare expertise in which few people have been trained. It makes sense to share existing competencies, because many institutions will be unable to recruit the data professionals they need. It also requires a technical infrastructure providing excellent computing capacity and a high level of security, requirements which are too complex and costly to be available to all institutions that develop data collections. The sharing of infrastructure by multiple actors is the only satisfactory solution from a financial perspective, and the only way to meet the needs for expertise and performance. This does not correspond to the pooling of databases in a single national data warehouse, but rather the pooling of a copy of each of these databases if they need to be chained, in order to make them available to researchers. Several specialist warehouses may be established to meet the needs of particular communities. All these databases simply need to be identifiable due to their inclusion in research data warehouses, and interoperable through their compliance with semantic and technical standards.

2.2. Benefits for epidemiological surveillance

Epidemiological surveillance is a public health activity involving the continuous collection of information about health events, its analysis with a view to creating quantified indicators, and the mapping of this information prior to its dissemination, in order to assist decision-makers in the health sector. Having largely emerged since the 1950s, over the decades and after a succession of health crises, epidemiological surveillance has since played an essential role in the development and implementation of all health policies.

Specific databases were traditionally created to generate the required information, notably the morbidity registers of interest to Santé Publique France (the French National Public Health Agency – SPF) and INCA (the French National Cancer Institute), for example. These registers have many limitations. For example, they cover only a fraction of the population, which limits their statistical power, and they are very costly to maintain as they require the participation of a large number of professionals to ensure their quality and comprehensiveness (17). Furthermore, they are not designed to detect weak or new signals, since they only collect data previously identified as being of interest.

Registers and cohorts do have limitations in their role as epidemiological surveillance tools, but they can become excellent tools if they are combined with healthcare data. They can benefit directly from administrative databases by collecting data on the healthcare pathways of the people included and ascertaining their vital status. They can also validate the comprehensiveness of records, which increases the robustness of the research produced. Registers and cohorts can also be highly useful for validating hypotheses generated by healthcare data and for creating additional surveys in the event of an alert. Surveillance based on healthcare data enables fluctuations in the incidence of health events to be detected and analysed per territory and for relevant sub-populations, thereby enabling the identification of social inequalities in health and in access to care, or differences in exposure to local risk factors. This satisfies the public's desire to monitor the incidence rates of infectious diseases such as cancers, respiratory and cardiovascular diseases, embryo-foetal development anomalies and any other disease whose determinants are largely

environmental. Such surveillance is technically feasible at reasonable cost and is highly effective, with no potentially adverse effects. Epidemiological surveillance requires registers and cohorts, but it also needs healthcare data. Only real-life data – in this case, healthcare data – can confirm the conclusions of research studies, which are always conducted on specific populations and under controlled conditions that are sometimes far removed from reality. Our knowledge would be optimal if the different data sources were interoperable and connected.

2.3. Benefits for improving practices

Health data collected in the context of healthcare, to assist with the individual medical treatment of patients, is also of great interest for improving treatment from a collective perspective. It can be used to measure healthcare efficacy and detect local or regional situations that deviate significantly from the mean. It also enables the monitoring of changes in practices over time and their impact. The main SNDS database can be used to analyse the care pathway for certain diseases: the information provided in this manner, reflecting real-life situations, is extremely important for professionals because it highlights realities that raise questions about the relevance of treatments. For example, the Observapur database (18), which analyses the care pathways of men treated for mictional disorders, shows that the use of surgery, all other things being equal, varies by a factor of one to three depending on the regions. Such geographical variations have been demonstrated for many surgical procedures: hip-fracture surgery, tonsillectomy, appendectomy, caesarean section, obesity surgery, etc. (19). SNDS data can also be used to highlight discrepancies between medical recommendations and actual practices, and then inform information and training initiatives for professionals.

2.4. Benefits for evaluating innovations

Data can be used to monitor the positive and negative effects of innovations introduced as a result of changes in practices, be they medicinal, technical or organisational. Of course, such innovations have already been studied before their adoption and marketing authorisation – in clinical trials, for example – but on a limited scale. When they are disseminated and adopted on a larger scale, a number of issues may arise, such as the transferability of the trial and the effects. Some, but not all, innovations are also subject to mandatory post-marketing monitoring, but such monitoring is hard to implement without access to the real-life data provided by administrative databases. This especially applies to medical devices, whose benefits are rarely documented by clinical trials, unlike new medicinal products.

2.5. Benefits for improving access to healthcare services

Healthcare data analysis can be used to document patient access to healthcare interventions: rates and delays can be calculated and cross-referenced with geographical variables and social and economic data in order to identify what is working and what is not, and which geographical areas should be prioritised for improvement, even if there may be a long way to go between identifying a problem and implementing corrective measures, particularly in the disability and mental health fields, for example.

2.6. Benefits for improving knowledge

Health data can also contribute to improving knowledge directly – an essential step towards improving practices and designing effective and protective public policies. The natural history of many uncommon, let alone rare, diseases is poorly understood. Registers or cohorts focusing on a single disease are very rarely recruited nationally, and it is extremely difficult to keep track of the patients they include, which complicates the understanding of long-term

trends (17). Administrative health data can fill gaps in the knowledge generated by registers and cohorts.

2.7. Benefits for developing tools and services to improve the population's health

The healthcare system uses increasingly powerful tools. Investigative and diagnostic techniques, in addition to care and treatment methods, are developed on the basis of fundamental and applied academic research, and are transformed into universally available products by companies that develop, produce and distribute them. The availability of healthcare data is a pivotal factor in the industrial innovation ecosystem, enabling manufacturers to gauge the size and nature of the needs to be covered, collaborate with healthcare professionals and academic researchers during the research and development phase, and evaluate their products once they are marketed.

Artificial intelligence is used in many of the tools being developed to improve the quality of medical care. The development of algorithms is dependent on the availability of huge volumes of data. The data used must be essential to the purpose of the processing and for nothing else. Like any other innovation, the usefulness of algorithms needs to be proven. Greater availability of data will enable the development of higher-quality algorithms because they will have been trained and tested on representative populations, which is not always the case at present.

2.8. Benefits for guiding health policy

Health data plays a crucial role in developing health policies and assessing whether they are effective. It is used to assess the effectiveness of prevention or screening campaigns and compliance with best practice recommendations issued by the French National High-Authority on Health (Haute autorité de santé – HAS), the French National Agency for Medical and Health Product Safety (Agence nationale de sécurité du médicament et des produits de santé – ANSM), and learned societies. Health data can be used to identify territorial, social or age-group inequalities, so that corrective measures can be envisaged. It is key to detecting upward trends in certain diseases, including epidemics. None of these activities would be possible without using administrative health data. The processing of data available to all researchers during the SARS-Cov 2 pandemic demonstrated the usefulness of providing sweeping access to databases and showed that access to French data was less open than in English-speaking and Northern European countries.

It is from this perspective that the Strategic Committee on Health Data (Comité stratégique aux données de santé) was established. This committee offers the Minister guidance and decision-making support for the implementation and development of the French National Health Data System (SNDS). As provided for in Article R. 1461-10 of the French Public Health Code, the Strategic Committee is responsible for proposing guidelines for developing the SNDS, issuing opinions on legislative and regulatory developments, identifying databases that should be included in the catalogue, and categories of missing data, in addition to convening thematic working groups and interviewing prospective database managers. It is chaired by the minister for Health who can delegate to the French Directorate for Research, Surveys, Evaluation and Statistics (Direction de la Recherche, des Études, de l'Évaluation et des Statistiques – DREES) and vice-chaired by the Directorate General for Research and Innovation (Direction générale de la recherche et de l'innovation – DGRI), with the Health Data Hub serving as its secretariat.

The Strategic Committee on Health Data has recently established a number of working groups representing the ecosystem, with the aim of defining a common core of data for

hospital data warehouses and a health-data-access governance system, as well as sustainable financing arrangements for health databases and fee-paying mechanisms.

2.9. Benefits for participatory research

Participatory science and research are forms of scientific knowledge production in which actors in civil society purposefully and actively participate, either individually or collectively, alongside researchers. They are a means of involving citizens in scientific research by combining citizens' expertise with scientific expertise, and serve as a concrete means of addressing societal issues and benefiting from a wealth of experiential knowledge, which harmoniously complements the research community's expert knowledge. The availability of large datasets is one of the prerequisites for the development of participatory research. On the research side, users' views and experience complement those of experts. On the users' and citizens' side, their involvement enables them not only to dialogue with experts, but also to increase their knowledge in the research field and in project development, and to understand the use of health data and the associated regulatory framework. The aim is not only to promote efforts by public bodies to create appropriate citizen-led governance systems and forums for dialogue in line with their missions, but also to implement awareness-raising, information and training projects, or even workshops on the use of health data.

3. Potential deleterious effects and their risk of occurrence

After reviewing the expected benefits of using available health data, which would be impossible without this data, it is logical to closely examine the actual or potential risks, which are summarised in Table 2.

3.1. Risk of revealing personal data

Personal health data is extremely private information that must not be divulged beyond the very small circle of healthcare professionals chosen by an individual. This is the consensus view and there can be no exceptions. The identity of the people whose data will be used must therefore be protected as effectively as possible. Data can only be used for purposes other than healthcare, i.e. for research, if it has been anonymised or pseudonymised, with names being transformed into randomly assigned pseudonyms that make it impossible to trace the actual name from the pseudonym. Nobody wants to see their medical records being published on an open-access website.

The challenge is therefore to provide maximum technical security in order to combat data theft from doctors' surgeries, hospitals and clinics and administrative databases, in other words, from any collections containing personal data. And the stakes are high, because personal data is often hacked due to its high monetary value, either to obtain a ransom or to sell it to other economic players.

Fortunately, the data used for research purposes is never personal. It is either completely anonymised or pseudonymised.

Anonymisation is a technique that definitively separates personal data from the associated data. No links can be reconstituted subsequently in order to track down the identity of the persons concerned. This is a radical way of making health data usable by researchers, while fully protecting privacy. However, the system has several flaws. Some data, due to its uniqueness or rarity, is indirectly identifying. Genomic data, particularly that derived from

sequencing, is unique to each individual. Someone who happened to know the characteristics of a person's genome could find their data in all the genomic databases. This requires particular malicious intent and a great deal of technical effort, but it is now impossible to prevent since genome sequencing is so widely used for purposes other than medicine and science.

In theory, pseudonymisation is a technique that can protect the identity of people whose data has been collected, by replacing their first and last names with a randomly assigned pseudonym containing no information that can be traced back to the individual. It has the advantage of enabling the protection of people's identity without compromising any return to them or additional data concerning them if required by the research or its findings. The information required to match the pseudonym to the personal data is kept in a separate database. Researchers only have access to the data linked to the pseudonyms and never to the matching table. Pseudonymisation is a regulatory requirement in the event of the SNDS being used by researchers. However, this technique has a number of potential disadvantages, the most common of which is when the pseudonym contains elements that help with day-to-day data management. Doctors sometimes include a patient's initials, the year of recruitment, the title of the research project and the patient's inclusion number in the study. All these indications greatly facilitate the re-identification of people by someone working in the institution, for example, despite providing sufficient protection against identification by malicious outsiders. Such poor practices must be eradicated.

Other loopholes allow for the re-identification by malicious persons without any links to an institution, provided that they want to search for a particular person, whether the data is anonymised or pseudonymised. This can be done by cross-referencing databases with publicly available data. If files from several sources contain the year of birth, residential postcode, gender and blood group, it can be deduced that a given person suffers from a specific pathology because he or she is included in a file of cancer sufferers, for example. This actually occurred in 1997 when a journalist re-identified the medical file of a governor in the USA. Human genetic data, facial photographs, voice or video recordings and medical imaging can all be used to re-identify specific individuals in cases of malicious intent. To date, there have been no known cases of large-scale re-identification of pseudonymised data. The documented cases are the result of people trying to demonstrate that this was theoretically possible, based on the identification of a specific person.

Methods to reduce the probability of re-identification have been developed. One method is based on the principle of differential confidentiality, which entails the addition of statistical noise to diminish the accuracy of the information contained. Such methods are not unanimously approved.

No method can claim to be totally effective in preventing the re-identification of a person if large volumes of data are available and can be cross-referenced, but such re-identification requires a great deal of effort. In other words, the risk of invasion of privacy is not completely zero if data is used maliciously with the intent to cause harm, but it is drastically reduced when data use is supervised for research purposes if the pseudonymisation rules are properly respected. The risk is certainly not zero, but it remains low and cannot be considered an obstacle to the use of data of significant public interest, given the associated benefits.

3.2. Risks of uncontrolled use

The concerns expressed relate to respect for the rights of the people affected by the use of their data. The concept of systematic prior notification is linked to the fact that the data contains information about people's private lives, which is a highly protected constitutional

right. Data items are public, not private objects, and data custodians have no rights over them, but rather duties to protect them and use them only for the purposes intended when the data was collected or subsequently processed in accordance with the General Data Protection Regulation (GDPR). The focus should be placed on data collection rather than on individual data (20).

In healthcare, this field has undergone significant changes over time. In France, the general principle of consent established by the Public Health Code is that individuals must, in principle, consent to the collection of their data when research is conducted on their body. This principle has been extended to the use of invasive samples of parts of the human body for research purposes, with people having to give their consent in order to authorise samples that might not be used for their own medical treatment, but could further scientific knowledge. Participation in research is seen as an altruistic donation of one's body parts or time spent undergoing a test or filling in a questionnaire. Free and informed consent must remain the rule.

However, when personal data is collected in a healthcare context for research purposes, individuals play no additional role that could be assimilated to a donation, as this data is collected during the provision of healthcare for the purposes of individual treatment and for monitoring the quality of the service provided. This is therefore a different situation which, according to the GDPR and the French Data Protection Act, should not require explicit consent and perhaps not even systematic notification since the use of the data takes nothing away from individuals and does not require them to make any particular contribution. Nonetheless, certain groups need to be reassured, such as members of sexual minorities, people treated for a mental health problem or women who have undergone birth control procedures. To this end, healthcare institutions need to be irreproachable in their approach to protecting the data entrusted to them, which is clearly not yet entirely the case. The provision of information about data security therefore seems essential at the time of data collection, together with information about its secondary uses.

At present, the regulations require the systematic provision of individual notification, but a general notification system could legitimately be established in the knowledge that individuals have the right to withdraw their consent upon request, as has been envisaged for organ donation. This would vastly increase the volume of data that could be used to drive progress in public health, at no detriment to the beneficiaries of the healthcare system.

Research data is collected with the patient's active assistance. This data is generally collected after the collection of explicit consent following detailed notification about the objectives of the research and its protocol, as this concerns a donation for research purposes (opt-in scheme). At present, in France, this distinction is improperly understood and applied, which hampers many studies unnecessarily.

Another fear expressed about uncontrolled uses concerns the use of data for purposes other than those intended in the general interest, especially by industry, possibly for commercial purposes, but also by groups with partisan or malicious targeting intentions. This risk is theoretically possible, but it is stringently controlled by the current regulations, which require an independent committee’s opinion on the purpose of the research, with only research serving the common good being authorised. Furthermore, European regulations provide for heavy fines in the event of non-compliance with the European Union's General Data Protection Regulation.

3.3. Risk of data theft and industrial or state espionage

The third category of potential deleterious effects relates to the possible consequences of health data theft for the purpose of industrial or state espionage. This is a very prominent concern in the discourse of the opponents of open access to real-life data for research, on the pretext that many data warehouses store their data in clouds managed by North American companies. Indeed, the US Cloud Act stipulates that the US government can, if necessary, access data located anywhere in the world, in the event of a risk to US security. In fact, this clause only applies in certain specific circumstances, such as judicial enquiries authorising a judge to investigate in order to retrieve the health data of a person accused of terrorism. Therefore, it will potentially only apply to collections of personal data, which is not the type of data stored for research purposes.

Moreover, US companies operating in Europe must comply with the GDPR, and they do so because otherwise they would risk being fined 4% of their annual turnover for failing to comply with the GDPR. On top of that, such a precedent would cause them to lose all their European clients.

If many public institutions host their data in US clouds, it is because technological solutions developed by companies in the United States offer tools that meet the highest technical and security standards. These are the only operators capable of meeting the requirements of invitations to tender. There is no technological alternative that meets the stringent security standards, either in France or in Europe. It seems more important to prioritise IT security rather than protection against a hypothetical risk of the USA appropriating pseudonymised data, which would be unlikely to harm citizens either individually or collectively. Nevertheless, the development of a European cloud is highly desirable in order to establish our sovereignty in this field.

For all that, personal primary care data can be stolen by private hackers or hackers operating on behalf of foreign powers. An economic market already exists for such data. This is a genuine threat for primary data collections, but not for research data collections.

3.4. Risk of intellectual property right infringement and financial loss as an argument against exploitation

Professionals involved in the collection, management and analysis of health data, and the institutions that employ them, claim intellectual property rights over this data. Many even consider that they own it, which is misleading. Indeed, there is no property right to health data per se for its custodian, or even the patient, but an intellectual property right does exist for databases. Protection is also afforded against the plagiarism of works. However, data collectors face a number of obligations: they are responsible for the proper use of the data they collect and for generating value from this data. Effective data collection is necessarily difficult, requiring financial investment and a high degree of professionalism. Collectors may therefore be reluctant to share their data with third parties that have not participated in the data collection effort. This is compounded by this is the illusion that their collection has a very high value in use, due to the media coverage of the profits made by the major private operators from personal data, which can therefore be used for commercial targeting. Research data only has commercial value if it meets international quality standards and is collected in large volumes, enabling the performance of authoritative studies.

The exploitation of data collections for research purposes is clearly inadequate, without any open-access data provision for third parties – whether academic or industrial – and without

the chaining of this data with the other existing collections. The tensions in this field could be resolved by properly funding the teams that maintain collections of interest, which is not currently the case, and by developing sharing arrangements that would enable them to obtain scientific and financial benefits if third parties were granted open access to the data. Concrete ways to address stakeholders' concerns do exist, with no proven or potentially adverse effects.

Moreover, the value in use of large data collections complying with semantic and technical standards is indisputable, as shown by the experience of other countries, notably Israel, Singapore, Estonia and the USA via the Mayo clinics network. There is currently no clearly defined business model in France, where the financing is almost exclusively public. Efforts should therefore be made to agree on the definition of a public policy for the long-term financing of databases and warehouses capable of ensuring the collection of this data, covering the initial investments that cannot be covered by fees, and reassuring the stakeholders. If financed in this way, the effective collection, quality assurance and availability of health data would enable it to sustain the knowledge economy without depending on private players, while also meeting the expectations of the industrial sector, which needs access to this data in the R&D process, where speed is of the essence.

3.5. Risk of normative surveillance of healthcare professionals

Some actors view the use of databases of medical practices as an unacceptable form of control over their professional activity, which could be prejudicial to them by providing access to publicly available performance scores or professional sanctions in the event of a significant deviation from the mean. The actors who pay for these types of data exploitation consider them a legitimate means of ensuring the best use of public funds. Addressing these concerns requires the involvement of the professionals and the institutions concerned in the interpretation of the results, as was the case when external laboratory quality controls were introduced.

3.6. Risk of the profiling of healthcare system users

The exploitation of Health Insurance data opens up the possibility of identifying individuals or groups who abuse services or are carriers of transmissible diseases and do not comply with the health obligations in force, or who have atypical profiles considered to be at risk of contracting certain diseases. This is a genuine risk which may lead to the justification of discriminatory political proposals. Protection depends on maintaining a democratic governance of data use.

After this review of the possible benefits and potentially adverse effects, it is now time to describe the French and European data management ecosystem designed to protect health data, in both its technical and regulatory dimensions.

4. Current governance of health data

The collection, management and use of health data, like research data, are highly regulated in response to the public's wishes and concerns. The measures implemented are of a technical, legal and organisational nature, as presented below.

4.1. Technical data protection

The security of health data collections is a major organisational and technological issue, which is overseen by the French National Agency for the Security of Information Systems (Agence nationale de la sécurité des systèmes d'information – ANSSI) and the Senior Defence and Security Official for the Ministries of Health and Prevention. The following security measures for protecting access to health data are recommended: data must be encrypted; operating rights must be segmented between people to enable the containment of any malicious act; the management of accounts and permissions must be secured by authentication based on a combination of several authentication factors to prevent identity theft; users must have a secure workspace dedicated solely to their project; the technical security building blocks must be independent and use flow-filtering, malware-detection and encryption-key-generation solutions; all access operations must be traced; data may also be hosted in the European Union by a certified "Health Data Host" (certificat "Hébergement données de santé" – HDS).

If all these criteria are met, the data will be as secure as possible and only accessible via criminal acts, against which there can be no absolute protection.

A theoretical risk of data theft therefore exists in the event of a technical security flaw. This is a problem for health data hosts, as such data is personal and therefore has a market value, but this danger is mitigated in the context of health research by the pseudonymisation and anonymisation measures applied to the data. In addition, data collections for research purposes are generally merely copies of databases which are hosted by their managing body. There can therefore be no risk of data loss.
The main danger is linked to the high degree of security required to counter hacking. Many institutions, particularly healthcare institutions, have neither the know-how nor the budgets required to secure their IT systems.

4.2. Legal protection of personal data

Personal data protection is regulated by the General Data Protection Regulation (GDPR) and certain requirements concerning health data are laid down in the French Data Protection Act (Loi Informatique et Libertés – LIL). Research projects involving personal health data are carried out under the supervision of the French Data Protection Agency (Commission nationale de l'informatique et des libertés – CNIL): in some cases, an explicit request for authorisation must be sent to the CNIL, while in others they must comply with simplified standards[2] drawn up by the CNIL. This means that they are not required to obtain formal authorisation, provided that they meet all of the conditions listed in these standards. The CNIL is likely to verify this point.

The texts also set out the security requirements to be met for hosting, such as the SNDS security standards, transparency requirements for projects and their results, and the obligation to notify and support citizens who wish to exercise their right to withdraw or modify their data.

4.3. Evaluation of research objectives

It is consensually agreed that the use of health data for research projects is only warranted if the purpose of the research is legitimate and ethical and the methodology is appropriate. The project must also be in the public interest and the project coordinators must have genuine legitimacy.

[2] "Simplified standards" or "reference methodologies" refer to simplified health data access procedures, which facilitate the processing of personal health data since express authorisation from the CNIL is unlikely to be required for processing operations that conform to them. See the CNIL: https://www.cnil.fr/fr/les-referentiels-et-methodologies-de-reference-sante

These points are verified by the French Scientific and Ethical Committee for Research, Studies and Evaluations in the Health Sector (Comité éthique et scientifique pour les recherches, les études et les évaluations dans le domaine de la santé – CESREES), which replaced the Expert Committee for Health, Research, Studies and Evaluations (Comité d'Expertise pour les Recherches, les Études et les Évaluations dans le domaine de la Santé – CEREES). This committee was created following the publication of the implementing decree for the French Data Protection Law of 15 May 2020, and the official notice of the appointment of the committee members on 9 June 2020. The committee reports to the French Ministers for Research and for Health. To formulate its opinion, it ascertains whether the project is socially worthwhile and whether the project is serious and credible. The committee also makes sure that the project leader is not requesting access to more data than is necessary for the project, and that civil society will benefit from sharing its data. To this end, the committee is composed of some twenty members appointed by government order, including two representatives of healthcare users, and a network of around forty experts appointed by the chair on the basis of members' proposals. This composition ensures the availability of the expertise required to examine the applications submitted, i.e. clinical as well as methodological expertise.

The CESREES issues its opinion, which must then be examined by the French Data Protection Authority (CNIL), which remains the only institution empowered to authorise project coordinators to process data. The CNIL verifies that the project complies with all the security obligations required by the General Data Protection Regulation.

4.4. Legal framework for partnerships

When stakeholders' data is shared, the sharing arrangements may be governed by contracts between the data producer and the user, between the producer and the sharing platform, or between the user and the sharing platform.

These contracts set out the obligations in terms of compliance with the GDPR, in the context of data transfers and the receipt or correction of data, where applicable. The contracts also cover the measures implemented to raise the profile of the database (e.g. inclusion in a metadata catalogue, allocation of a DOI [Digital Object Identifier, a permanent digital element used to identify data]) and specify the terms and conditions for scientific exploitation by the parties involved, including the publication conditions. They may also include the arrangements for enriching the data with other data, providing access to certain findings and the licences chosen, the pricing arrangements, and even the funding associated with the performance of the work required to implement the sharing, where applicable. Contracts binding users also include general terms and conditions governing the use of the platforms and awareness-raising requirements. These contracts are drawn up by the parties, which must reach a consensus before the project can begin. Institutions and researchers therefore retain control over the use of their data by third parties. The main problem in this sector is the lack of mutual trust between public institutions, which is a source of major delays in the signing of contracts and may event prevent their signature (21).

4.5. Citizens' involvement in choices

Several legal measures have been taken to involve citizens in the governance of access to health data. For example, the law stipulates that the Health Data Hub, a national operator bringing together 56 stakeholders in the health data ecosystem, shall be vice-chaired by the president of France Assos Santé, and that two representatives of health data users will sit on the CESREES committee and on the Strategic Committee on Health Data, a body created by law to set the broad guidelines for the SNDS.

4.6. General policy on health data use in France

The public authorities are aware of the benefits of re-using health data and paved the way for the provision of access to it in 2016 by enacting the law on the modernisation of the French healthcare system (Loi Modernisation de notre système de santé – LMSS).
This law marks a major step forward. By creating the SNDS, it has unified the governance of several large administrative databases that were previously managed by different bodies, each implementing their own specific data-access rules. The French government is now responsible for the strategic governance of these data assets and has clarified the doctrine and principles governing access to SNDS data: all actors, whether public or private, may access this data, under conditions ensuring its security and the protection of privacy, as long as the purposes for which this data is used do not run counter to the public interest. Access rules and procedures have been defined, and a Single Gateway – the Institut national des données de santé (INDS) – has been established to facilitate this access.
The law on the organisation and transformation of the healthcare system (Loi pour une Organisation et Transformation du Système de Santé – OTSS) of 24 July 2019 has strengthened this ambition to develop uses to serve innovation and research. By extending the French National Health Data System to all publicly-funded health data, it establishes the principle that such data should be used more widely. This law replaced the INDS with the Health Data Hub which, in addition to its role as a Single Gateway, has a much broader remit:

- providing a secure and state-of-the-art platform for storing and processing data;
- developing and progressively enriching a documented data catalogue to provide the scientific community with "priority" databases for furthering knowledge of health, such as cohorts, registers and hospital data, in addition to the "historical" SNDS. The Health Data Hub is jointly responsible for processing the main SNDS database with Assurance Maladie, the French health insurance system, and is the data controller for the catalogue;
- providing services and tools for users, and coordinating the ecosystem to accelerate innovation by promoting the sharing of experience and knowledge.
- ensuring the coordination of the Strategic Committee on Health Data's activities, which is responsible for proposing guidelines for developing the SNDS, issuing opinions on legislative and regulatory developments, identifying databases that should be included in the catalogue and categories of missing data, convening thematic working groups and hearings of prospective database managers.

4.7. General policy on health data use in Europe

To facilitate access to the different types of data available in the Member States, the European Commission has made the future European Health Data Space (EHDS) a priority of its health policy (22). This ambition is notably reflected in the proposal for a European regulation creating the European Health Data Space, covering the primary and secondary uses of health data. The draft regulation is currently being negotiated by the European Parliament and the Council of the European Union with a view to adoption in 2024. To prepare for the implementation of the text, a number of prefiguration instruments for the future European Health Data Space have been put in place since 2019, including TEHDaS (Towards a European Health Data Space) (23), a think-tank programme bringing together more than 26 Member States, in which the Health Data Hub has coordinated five French partners. Led by the HDH, the French contribution has been particularly active on the themes of civic engagement and governance. In July 2022, France led the winning consortium in the European Commission's call for applications to develop a test version of the future European Health Data Space. Work began in October 2022 and is scheduled to last two years. The aim is to provide an end-to-end user pathway for researchers wishing to

use health data from several European countries. Services offered in this user pathway include a European metadata catalogue, populated by national catalogues, and a single access-application form, to dispense with the need for researchers to submit multiple access requests at the same time.
The French HDH is held up as an example for other European countries to follow, and the European Commission has drawn inspiration from the French model by proposing the creation of bodies responsible for health data access in each country. The European Commission recently approved the project led by the Health Data Hub and its partners in preparation for the national implementation of the European Health Data Space ("French HealthData@EU"). The French consortium will receive funding to prepare for the implementation of the EHDS at national level, which includes 1) measures to reinforce the HDH's services; 2) funding to support a coordinated approach to the quality and standardisation of data with a view to its future re-use, and 3) the completion of the national metadata catalogue by the data-custodian partners, in compliance with the relevant European standard (Health DCAT-AP).

4.8. The data-access process and obstacles to it

Despite changes made in recent years, there are still obstacles to data sharing.

The access procedures are numerous and complex, and it is not always easy for stakeholders to know which steps to take between applications for authorisation, and whether their project corresponds to the simplified standards or reference methodologies. In addition, the latter change regularly, and lack a clear development strategy.

The division of roles between government ministries and the CNIL in the development of national data access governance is generally unclear. Problems persist concerning the use of the social security numbers allocated to each person at birth based on civil status data sent by local municipalities to the French National Institute for Statistics and Economic Studies (INSEE) and recorded in the French National Social Security Register (NIR), the anonymisation rules, database compliance procedures and the correct way to implement the requirements for notification and the exercise of rights. In some cases, the attainment of the various regulatory, contractual or technical milestones results in long and dissuasive access times that may run into years. This leads to absurd situations in which data only becomes available once the funding for the research project has ended and the researchers have left to work elsewhere. These situations genuinely amount to missed opportunities when worthwhile projects that have received funding are prevented from going ahead by unjustified administrative delays. No studies have been conducted to quantify the extent of these problems and their impact on research. However, the research community frequently expresses its frustration. The recent publication of a white paper by the CNRS Scientific Council on the administrative obstacles that researchers encounter on a daily basis reveals a deep distress (21) that needs to be resolved. A process audit seems essential, because the data protection system was developed progressively, at a time when the open science concept was not embraced as a value. The proliferation of actors is a problem in itself, and a source of unjustified delays. The authorisation system should therefore be redesigned, with a single body issuing processing authorisations, as provided for by the European directive on the European Health Data Space currently being drawn up.

The requirements concerning security, and more recently, sovereignty, are extremely stringent and disconnected from the realities in the field, both in terms of the resources that actors can marshal for sharing and accessing health data, and the level of services available on the private market. The CNIL regularly demands that research data should not be stored

in a cloud managed by a US company, even though no other French or European cloud meets the required security specifications.

In addition, the CNIL's refusal to authorise a unique identifier sometimes leads to more complex and less effective cross-referencing, while adding several months of additional work to the performance of statistical matching. Many other European countries routinely perform matching based on a unique identifier in compliance with the GDPR. This puts French research teams at a competitive disadvantage, greatly delays research projects and reduces the power of studies, since the statistical matching rates stand at around 80%, a rate that could rise to almost 100% if social security numbers were to be used. Such a ban seems hard to justify, as it does nothing to increase the protection of data.

The implementation of large-scale healthcare databases requires significant expenditure on technical infrastructure, IT development and business software development, for example. Institutions are also required to document, standardise, pseudonymise, match and improve the quality of data so that it can be used by third parties, while also implementing the measures for notifying individuals or enabling them to exercise their rights under the GDPR. They must also develop the organisational system required for contracting and the provision of access to data by third parties while respecting rights and providing a high level of security. These costs are often underestimated by the administrative bodies in charge, and strategic decisions are complicated by political decision-makers' reluctance to share health data. For this to happen, there needs to be strong support for producers of health data of interest, in addition to the methodological, technical and regulatory assistance currently provided by the Health Data Hub.

5. Services offered by the HDH in response to these ethical pressures

The Health Data Hub's missions and projects have been developed in accordance with the values of open science, scientific integrity, the creation of social value, personal data protection, the advancement of knowledge for the benefit of all, and the optimisation of ecological, financial and human resources.

The main need expressed by stakeholders in France is for the clarification of the available databases, their content and the rules governing access to the data. They also want to be able to access them quickly, failing which they may resort to data that can be obtained more quickly or more easily from other countries.

In 2016, France appointed a single national actor to assist project coordinators with these issues. Although its role was strengthened by the 2019 law, the HDH was never intended to be the only actor capable of effective data provision. Now more than ever, actors are expressing the need to have access to an interlocutor capable of assisting them with the regulatory process: from identifying the applicable access procedure to preparing the application form and liaising with the ethical and scientific committee.

In France, the HDH serves as the secretariat for the National Scientific and Ethics Committee and submits applications to the French Data Protection Authority (CNIL). Data custodians can always refer to an ethical and scientific committee at a more local level, but this committee's opinion will form part of the dossier submitted to the national governance body if authorisation from the CNIL is required. The CNIL is currently the only body empowered to authorise the processing of such data, ensure the fair treatment of all cases, and guarantee the transparency of the process and the opinions issued.

The HDH coordinates the community of data users, identifies their needs for support, produces educational documentation and promotes the open-source approach. It also puts data users in touch with the relevant institutional actors, according to the difficulties they encounter; it can speak to these institutional actors on behalf of data users in the event of any difficulties encountered, and can also provide practical assistance with overcoming obstacles wherever possible.

Above all, the HDH is the national operator for the implementation of the major strategic orientations identified and can conduct the activities prioritised by the Strategic Committee, in particular. The HDH facilitates activities at national level on harmonising access fees, identifying minimum data sets, introducing simplified access procedures, setting up health data information systems, etc.

The HDH also supports stakeholders and advises them on the implementation of measures linked to the provision of open access to data: in France, for example, a policy is currently being implemented to support the design and roll-out of hospital health data warehouses. The Health Data Hub operates as a partner in this initiative, from the design stage through to the provision of support for the winners.

The HDH centralises information to maximise transparency towards civil society (a project to produce a national digital form enabling citizens to submit requests to exercise their rights in a simplified and orderly manner is currently being developed). It contributes to raising healthcare professionals' awareness and providing training (initial and in-service) in secondary uses of data to enable them to promote the creation and sharing of health databases of interest for research and innovation.

It provides access to the different data custodians' data by acting as a trusted third party. The volumes of data transmitted through it can require major investments in technological infrastructure, which may be unrealistic at local level.

The HDH enables the storage of data of national interest in order to take full advantage of the efforts already made to improve the quality, linking and availability of data and thereby reduce the burden of data provision on data providers.

The HDH has also established a Citizens Department responsible for devising a work programme that will facilitate the performance of its statutory duties to ensure transparency and provide support for citizens. These actions involve listening to citizens' concerns (e-consultation, consensus conference), involving them in governance and discussions (dialogue groups), providing information and training (a citizen training programme has been developed with FAS, CNIL, DREES and CNAM), and offering support for the exercise of rights, including the creation of a national form for managing objections.

Conclusion

The use of health data collections for scientific, medical and public health research is fully justified, given their numerous potential uses for the public good, which remain unexploited in the current context.

The potential risks to individuals and the community are minimal if the protective measures adopted are effectively implemented and none of the risks are liable to pose a danger with a significant impact (on health), either individually or collectively. The benefit/risk balance is therefore clearly in favour of the benefits, and calls for rapid corrective action to address delays and blockages, which are unjustified in the light of this benefit/risk balance that goes

beyond the theoretical risks. One example is the inability of French expert centres for constitutional bone diseases to contribute to the international register of fibrodysplasia ossificans progressiva, a very severe and exceedingly rare disease which leads to the ossification of muscles – literally petrifying patients, on the pretext that the data in this register is stored in the United States, because the register was created by a US patients' association. French regulatory agencies consider data storage in the USA to be a major risk in itself, despite the implausibility of such a risk and even though its probability is nil, having never been observed after decades of transatlantic scientific collaboration. This ban is slowing down the development of knowledge concerning the natural history of the disease, which is vital for evaluating the innovative treatments currently being developed.

The purposes for which data is used in research are widely shared in our societies, and consistent with choices already enshrined in our legal and legislative framework. Facilitating access to data for research is essential with a view to upholding the values of open science enshrined in law, meeting citizens' expectations of receiving appropriate and properly evaluated care, improving practices by furthering knowledge of the determinants of overall health, ensuring territorial and social equity, and providing the knowledge required for sound decision-making at all levels and evaluating public policies.

These objectives will only be achieved if the obstacles to opening up access to data collections are removed and if the culture of the common good represented by data is disseminated throughout all levels of society. The large health data collections currently used by researchers (24) are hosted in countries such as the UK or the USA, and the institutions responsible for their dissemination have never reported any adverse effects attributable to open access, contradicting the warnings issued by the custodians of large data collections in France.

France can pride itself on having implemented a successful policy of health data provision for research, developed a national health data warehouse and created the HDH. French administrative databases are unrivalled worldwide in terms of their exhaustiveness and the wealth of information they contain. However, numerous obstacles hinder the performance of the wide variety of research required to effectively manage healthcare policy, improve the healthcare system and devise innovative solutions. Examples include the fact that each research project is individually assessed for its technical security aspects and compliance with regulatory rules, even though these issues depend on the institutions in which the researchers work. Common sense would advocate the evaluation of these data management services at the institutional level rather than at the project level. This would simplify procedures enormously, reduce the time taken to obtain authorisations by several months and make research institutes more accountable for their data management policies. This is a current demand of the University Hospital Institutes (25). Practical difficulties in accessing SNDS data are responsible for the non-performance, or significantly delayed performance, of many studies. This situation must change because a solution has been found, which resides in the HDH's capacity to host a copy of the SNDS. However, the implementation of this solution is currently being delayed by the debate about the risks of hosting data on a server in France using the technologies of a US company – Microsoft – when the HDH should instead be considered as a risk reducer.

Priority should be given to resolving the causes of the major delays encountered in obtaining processing authorisations, signing partnership agreements between data producers, and providing access to the available health insurance data. It is unethical to prevent useful research from being carried out by imposing security requirements that go far beyond those set out in the GDPR, which are sometimes impossible to meet and therefore obstructive (26).

We must therefore move away from a purely legalistic reasoning that only considers potential theoretical risks, without considering the expected benefits of health data use for research. This requires the development of a benefit/risk assessment grid to reach a consensus on the respective importance of each item. This approach is commonly used in decision-making processes for marketing authorisations for health products, and for the organisation of population screening, for example. These methods are transferable.

An audit of the current situation would be welcome with a view to adapting the French regulatory system, which should protect against consequential risks without hindering progress on the pretext of protecting against risks whose effects are negligible and seldom occur. This means considering the benefit/risk balance when granting authorisation or compliance status, and not merely the risks, as the CNIL currently does. The only body that takes the benefit/risk balance into account is the French Scientific and Ethical Committee for Research, Studies and Evaluations in the health sector (CESREES). In order to gain a better understanding of the public health risks and benefits of secondary research on health data, we propose the implementation of an objective benefit/risk assessment tool, in which the risks are measured, weighted by the probability of the risk occurring, and objectified by their impact on society.

Three other aspects need to be addressed: distinguishing between personal consent granted for health data and research data, using the national identification (social security) number for data chaining, and simplifying the authorisation system. The report by the French Senate's Social Affairs Committee adopts a similar position (27). It recommends facilitating the matching by researchers of the different current databases by easing the rules of the decree on the professionals authorised to use the national identifier, and clarifying the procedures for obtaining patients' consent. The report also recommends clarifying the roles of the CNIL and the HDH with regard to authorisations. Removing these stumbling blocks would transform the situation and finally enable the performance of studies that the country and its citizens are legitimately calling for.

The report on the European Joint Action entitled "Towards the European Health Data Space" (TEHDAS) (23) notes that "Despite the efforts made in some countries to facilitate the secondary use of health data, obstacles persist. As far as cross-border projects are concerned, the non-harmonised and largely uncoordinated national authorisation processes remain a major obstacle." It is therefore necessary to clarify, rationalise and simplify the steps involved in accessing health data in all European Union Member States, first and foremost in France, whose regulatory requirements are more stringent than the European requirements. To achieve this, the TEHDAS activities underline "the importance of a clear division of roles and responsibilities held by the various players at national level: data custodians, bodies responsible for data access, and coordinators of data access bodies".

In addition, the European initiative highlighted "the support of citizens for the re-use of health data", confirmed by a major consultation carried out in the United Kingdom, Belgium and France, with the support of France Assos Santé and the HDH, which collected more than 6,000 contributions"(23). Citizens expect health data to be used to improve our public health and our healthcare system. It is up to us to find solutions to make this happen. We need a coherent, reliable and effective framework for the secondary use of health data, and this can be developed. It is important to regulate without hindering (28) and refrain from protecting a handful of individuals at the expense of progress for millions of people.

Appendices

Table 1 Benefits of using health data for research

nature of benefits	benefits	impacts
general benefits	critical mass of data	decisive for rare events
		more powerful studies
	representation of the general population	more relevant results
	pooling of competencies	professionalisation of practices
	pooling of technical resources	improved security of data
		economies of scale
		facilitated participation of small institutions
epidemiological surveillance	exhaustive nationwide surveillance	faster trend detection
	improvement in the quantity and quality of data	better use of registers
	lower data acquisition costs	identification of inequalities
improvement of practices	measurement of the efficacy of healthcare	better allocation of resources
	detection of territorial inequalities	better management of public policies to correct inequalities in care
	relevance of medical care	improvement of professional practices
evaluation of innovations	measurement of the effects of innovations in real life	better allocation of resources
	(non-) justification for their financing	avoidance of health scandals
	early detection of adverse effects	support for public decision-making
improved access to healthcare services	detection of access failures and over-consumption	improvements to ensure an equitable, territorial and sociological service
	detection of territorial inequalities	
improvement of knowledge	better understanding of the natural history of diseases	better R&D
	identification of geographical, economic, social and environmental determinants	well-targeted innovations
		support for public decision-making
health-enhancing tools and services	more effective tools for diagnosis and treatment	improvement of medical care

	decision-making, management and monitoring algorithms	improvement of professional practices
guidance for policy development	tools for evaluating existing policies	better management of public policies with improved cost-effectiveness
	monitoring of the impact of recommendations	reduction of territorial inequalities
	measurement of territorial inequalities	agile adaptation to crises
participatory research	citizens' involvement in research	greater support for public policies
	consideration of patients' and carers' knowledge	improved services because they are better adapted to expectations
	democratic forums for democratic dialogue	confidence in services

Table 2 Risks of using health data for research

nature of risks	potential consequences	risk level
revelation of personal data	disclosure of private information	very low for pseudonymised data
		nil for anonymised data
	malicious re-identification	very low for pseudonymised data
		nil, especially for anonymised data
uncontrolled use	use without notifying people in the event of poor practices	very low due to strong regulation in place
data theft and espionage	very little impact for pseudonymised or anonymised data	extremely low, as never yet observed for research data
infringement of intellectual property rights	loss of the priority advantage	very low due to strong regulation in place
normative surveillance of healthcare professionals	sanctioning of professionals who deviate from standard practices	low, as differences will be dealt with through negotiation
	benefits for payers who ensure the proper allocation of resources	
profiling of healthcare users	stigmatisation of certain minorities	nil in democracies, but high in illiberal democracies

References

1- Stratégie de transformation du système de santé, Gouvernement, August 2019. link: https://www.gouvernement.fr/action/strategie-de-transformation-du-systeme-de-sante
2- Plan national pour la science ouverte, Ministère de l'enseignement supérieur et de la recherche, July 2018. link: https://www.enseignementsup-recherche.gouv.fr/fr/le-plan-national-pour-la-science-ouverte-les-resultats-de-la-recherche-scientifique-ouverts-tous-49241
3- "Big data en santé : des défis techniques et éthiques à relever", Inserm, June 2022. link: https://www.inserm.fr/dossier/big-data-en-sante/
4- "Recommendations on how to engage citizens in the European Health Data Space", Tehdas, 2023. link: https://tehdas.eu/app/uploads/2023/03/tehdas-recommendations-on-how-to-engage-citizens-in-the-european-health-data-space.pdf
5- Health Data Hub, link: https://www.health-data-hub.fr
6- Marcel Goldberg , Marie Zins, "Le Health Data Hub - Pourquoi ? Comment ?" Med Sci (Paris), March 2021 pp.271-276. link: doi: 10.1051/medsci/2021016.
7- Pierre Lombrail, Israël Nisand, Christine Dosquet, Frédérique Lesaulnier, Catherine Bourgain, et al, "Note d'étape sur le Health Data Hub, les entrepôts de données de santé et les questions éthiques posées par la collecte et le traitement de données de santé dites " massives "", Comité d'éthique de l'Inserm, January 2022. link: inserm-03533863
8- "Données massives et santé : une nouvelle approche des enjeux éthiques", Comité Consultatif National d'Ethique, Avis 130 du CCNE, May 2019. link: https://www.ccne-ethique.fr/sites/default/files/2021-02/avis_130.pdf
9- "Plateformes de données de santé : enjeux d'éthique". Comité Consultatif National d'Ethique, Avis commun du CCNE et du CNPEN, Avis 143 du CCNE, Avis 5 du CNPEN, February 2023. link: https://www.ccne-ethique.fr/sites/default/files/2023-05/CCNE-CNPEN_GT-PDS_avis_final27032023.pdf
10- "Qu'est-ce ce qu'une donnée de santé", Cnil, 2023. link: https://www.cnil.fr/fr/quest-ce-ce-quune-donnee-de-sante (consulted on 2 octobre 2023)
11- "Entrepôts de données de santé hospitaliers en France : Quel potentiel pour la Haute Autorité de santé ?", Haute autorité de santé, October 2022. link: https://www.has-sante.fr/jcms/p_3386123/fr/entrepots-de-donnees-de-sante-hospitaliers-en-france
12- "P4DP, un consortium pour créer le premier entrepôt de données de santé pour la médecine générale", Health Data Hub, 16 March 2023. link: https://www.health-data-hub.fr/actualites/p4dp-un-consortium-pour-creer-le-premier-entrepot-de-donnees-de-sante-pour-la-medecine (consulted on 2 October 2023)
13- "France Cohortes : comment pérenniser un outil de recherche exceptionnel", Inserm, septembre 2020. link: https://www.inserm.fr/actualite/france-cohortes-comment-perenniser-outil-recherche-exceptionnel/ (consulted on 2 October 2023).
14- "Registres et données de santé : utilité et perspectives en santé publique", Haut Conseil de la Santé Publique, September 2021. link: https://www.hcsp.fr/Explore.cgi/avisrapportsdomaine?clefr=1126
15- Nathalie Blanpain, "L'espérance de vie par niveau de vie : Méthode et principaux résultats", document de travail n°F1801, Insee, 2018. link: https://www.insee.fr/fr/statistiques/3322051

16- "Inégalités environnementales et sociales se superposent-elles ?", note d'analyse n°112, France stratégie, September 2022. link: https://www.strategie.gouv.fr/sites/strategie.gouv.fr/files/atoms/files/fs-2022-na-112-inegalites_environnementales-septembre_0.pdf
17- Marcel Goldberg, Mireille Coeuret-Pellicer, Céline Ribet et Marie Zins, "Cohortes épidémiologiques et bases de données d'origine administrative : Un rapprochement potentiellement fructueux", Med Sci (Paris), 2012 ; 28 pp. 430–434
18- Lukacs B, Cornu JN, Aout M, Tessier N, Hodée C, Haab F, Cussenot O, Merlière Y, Moysan V, Vicaut E., "Management of lower urinary tract symptoms related to benign prostatic hyperplasia in real-life practice in France: a comprehensive population study."Eur Urol, September 2013(3) pp.493-501. link: https://pubmed.ncbi.nlm.nih.gov/23465519/
19- Morgane Le Bail, Zeynep Or (dir.), "Atlas des variations de pratiques médicales. Recours à dix interventions chirurgicales, Edition 2016", Irdes, November 2016. link: https://www.irdes.fr/recherche/ouvrages/002-atlas-des-variations-de-pratiques-medicales-recours-a-dix-interventions-chirurgicales.pdf
20- Jan Piasecki, Phaik Yeong Cheah, "Ownership of individual-level health data, data sharing, and data governance" *BMC Med Ethics*, October 2022. link: doi: 10.1186/s12910-022-00848-y. PMID: 36309719
21- "Livre blanc préliminaire du conseil scientifique du CNRS sur les entraves administratives à la recherche", Conseil scientifique du CNRS, May 2023. link: https://www.cnrs.fr/comitenational/cs/recommandations/Rapport_Entraves_vf.pdf
22- "Espace européen des données de santé", Commission européenne, 2023. link: https://health.ec.europa.eu/ehealth-digital-health-and-care/european-health-data-space_fr (consulted on 2 October 2023).
23- "Advancing data sharing to improve health for all in europe, Main findings of joint action Towards the European Health Data Space 2021–2023", Sitra, Tehdas, Markus Kalliola, Elina Drakvik and Maria Nurmi (Ed.), 2023. link: https://tehdas.eu/results/eu-wide-collaboration-needed-to-optimise-health-data-use-for-research-and-innovation/
24- Baptiste Couvy-Duchesne, Simona Bottani, Etienne Camenen, Fang Fang, Mulusew Fikere, Juliana Gonzalez-Astudillo, Joshua Harvey, Ravi Hassanaly, Irfahan Kassam, Penelope A. Lind, Qianwei Liu, Yi Lu, Marta Nabais, Thibault Rolland, Julia Sidorenko, Lachlan Strike, Margie Wright, "Main existing datasets for open data research on humans", Machine Learning for Brain Disorders book, 2023, pp 753-806. link: https://link.springer.com/book/10.1007/978-1-0716-3195-9
25- " Il faut faire des données de santé un bien commun pour la recherche", *Le Monde*, 31 March 2023. link: https://www.lemonde.fr/idees/article/2023/03/31/intelligence-artificielle-il-faut-faire-des-donnees-de-sante-un-bien-commun-pour-la-recherche_6167803_3232.html
26- Kenneth P. Seastedt, Patrick Schwab, Zach O'Brien, Edith Wakida, Karen Herrera, Portia Grace F. Marcelo, Louis Agha-Mir-Salim, Xavier Borrat Frigola, Emily Boardman Ndulue, Alvin Marcelo, and Leo Anthony Celi, "Global healthcare fairness: We should be sharing more, not less, data", PLOS Digit Health, October 2022
27- "Données de santé : une réforme encore en cours de chargement", Rapport d'information n° 873, Commission des affaires sociales, Sénat, 12 July 2023. link: https://www.senat.fr/rap/r22-873/r22-8731.pdf
28- Yann Joly, Stephanie O.M. Dyke, Bartha M. Knoppers, Tomi Pastinen, "Are Data Sharing and Privacy Protection Mutually Exclusive?", Cell, n°167, 17 November 2016

HEALTH DATA HUB

Table of contents