Interim Memo on the Health Data Hub, health data warehouses, and the ethical questions raised by the collection and processing of Health Big Data

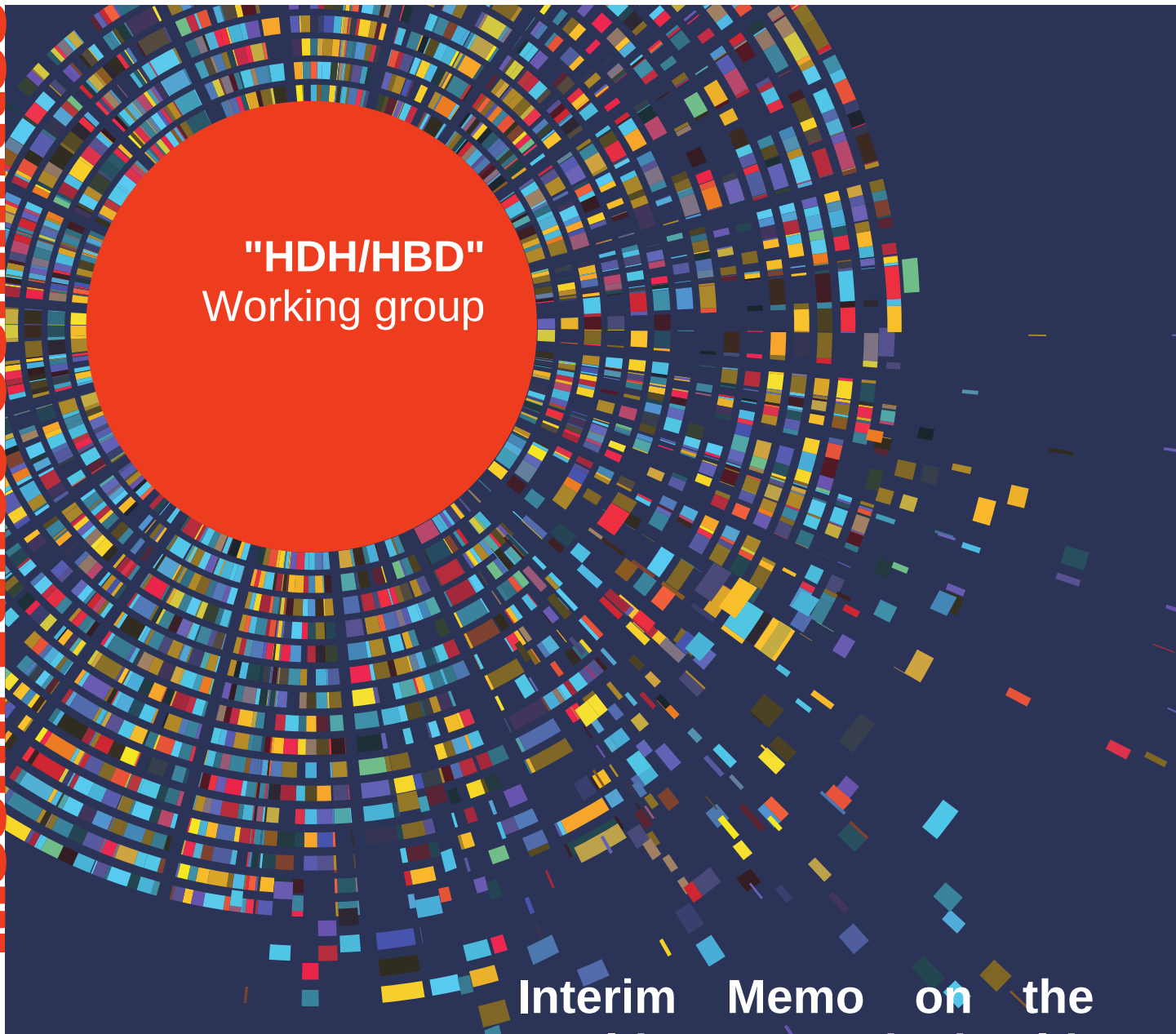
Pierre Lombrail, Israël Nisand, Christine Dosquet, Frédérique Lesaulnier, Catherine Bourgain, Bernard Baertschi, Anne Buisson, Catherine Cornu, François Hirsch, Christine Lemaitre, et al.

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

Inserm
La science pour la santé
From science to health

Inserm Ethics Committee

"HDH/HBD"
Working group



Interim Memo on the Health Data Hub, health data warehouses, and the ethical questions raised by the collection and processing of Health Big Data.

January 2022

Interim Memo on the Health Data Hub, health data warehouses, and the ethical questions raised by the collection and processing of Health Big Data

Pierre Lombrail, Israël Nisand, co-leaders of the Inserm Ethics Committee (CEI) HDH/HBD Working Group, Christine Dosquet, Frédérique Lesaulnier, Catherine Bourgain and the group members: Bernard Baertschi, Anne Buisson, Catherine Cornu, François Hirsch, Christine Lemaitre, Sylvie Ledoux, Flavie Mathieu, Isabelle Rémy-Jouet, Yamina Sadani.

In October 2020, the Inserm Ethics Committee (CEI) set up a Working Group in response to the questions raised by the decision to entrust the hosting of French National Health Data System (SNDS) data gathered by the Health Data Hub (HDH) to Microsoft through its cloud computing platform Azure. However, the group very quickly had to expand its reflection to include a much broader set of ethical questions raised by the collection and processing of Big Data relating to health data however remotely.

Since the group began its work, Decree no. 2021-848 of June 29, 2021 was published in France on personal data processing through the National Health Data System. This decree envisages the modes of governance and operation of the SNDS, whose scope was extended to include new categories of data[i] by French law no. 2019-774 of July 24, 2019. These categories include personal data from surveys in the health domain, matched with data from the SNDS and governed by the provisions of law no. 51-711 of June 7, 1951 on the obligation, coordination and confidentiality of statistics and it appears to be expected that the HDH catalog – which will be defined by decree – will include databases set up by Inserm teams (in addition to the registry of medical causes of death already integrated within the scope of the "historical" SNDS). Furthermore, in addition to occasional access to the SNDS – subject to the prior completion of a formality with the French Data Protection Authority (CNIL) – permanent access to the SNDS has been extended to include many public organizations or organizations that are tasked with a public service mission, for the needs of their missions, which include

“the Inserm research teams” (just like [French Public Health Code (CSP), art. R. 1461-13] "the research teams of university hospitals and comprehensive cancer centers (CLCCs), to which their members often belong).

It is becoming all the more necessary to reflect on the nature and conditions of compliance with the resulting obligations, i) for the Inserm institution (general policy for, and formalized governance of, access to the data, training of everyone involved, sufficient and competent support for the examination of research projects and the internal documentation of their regulatory compliance by teams with enhanced resources); and ii) for the research teams in terms of protecting the rights of the research participants and ensuring the scientific validity of the research conducted. It is about ethical and responsible research and, more generally, health democracy.

In this Interim Memo, we set out the issues identified following a first series of interviews and outline certain avenues for improving policies and procedures to ensure the greatest possible respect for the rights of research participants as well as the scientific validity of the research itself. At a time when some are declaring an urgent need to facilitate access to health data (*Les données de santé servent l’intérêt public, il y a urgence à en faciliter l’accès*. *Tribune, Collectif, Le Monde*, October 20, 2021), it seems to us that far from hampering research, these avenues are likely to ensure the trust of the research participants who contribute to scientific progress and as such the sustainability of quality research. The group's reflection will continue in order to better understand the uses in different contexts, with particular attention to the implementation of Artificial Intelligence techniques.

I. From initial concerns about the HDH hosting on the Azure platform to questions about the protection of people's rights and scientific integrity related to the creation of large data warehouses

The creation of the Health Data Hub (HDH) is the culmination of a long history which has made France the custodian of one of the largest health data warehouses in the world[ii]. This is a unique heritage in terms of the capacity to produce knowledge and in terms of the potential for optimizing the functioning of the health care system[iii] whose very existence raises questions that the CEI Chair formulated in a 2016 publication devoted to Big Data[iv]:

> *"The use of health or biomedical Big Data illustrates the ethical tension generated on the one hand by the huge potential of the method to advance knowledge of the origins, diagnosis, treatment and prevention of diseases, and on the other by the sensitivity of health information, protected in principle by medical confidentiality because of that, and the implicit vulnerability generated by their accessibility, partly linked to the lack of public knowledge of the meaning of these data and the opacity of the algorithms used to obtain them. "*

The hosting of the national Health Data Hub by Microsoft Azure's cloud computing platform is of concern to the IEC for at least three reasons:

- private operator (leading to fears of the influence of financial considerations on the choices of how the infrastructure is organized and how value is created from it that potentially go against scientific integrity and the collective interest; fears reinforced by questions about the economic model of the HDH in the long term should it be held hostage by these issues),
- U.S. law (fears about the protection of people's rights due to the Cloud Act and FISA regulations with the possibility of transferring data outside the territory for transmission to North American requesting authorities),
- and choice of a centralized form of architecture that is vulnerable to hacking (whereas the HDH preparatory dossier had suggested the choice of a distributed-storage configuration limiting the risks of intrusion). This last point has since been raised in the deliberations of the CNIL[v] which follow the publication of the decree[vi] relating to HDH[vii]. It is all the more salient as the CNIL is led to note that, "according to the details provided by the ministry, the HDH will

have a copy of the principal database, currently hosted by the National Health Insurance Fund (CNAM) and that the database catalog will be only hosted by the HDH".

The CEI is aware of the scientific, economic, and industrial – particularly pharmaceutical – stakes that govern the deployment of the HDH, but is vigilant in at least two respects:

- in a context of heightened competition between service providers claiming to contribute to the optimization of health systems operations[viii] (see the establishment of the Nvidia consortium in the UK[ix] or the recent creation of the Alliance for Real-World Data in our country[x]), the fragility of the economic model of the platforms that are being set up raises concerns about the balance between the necessary ultimate return on investment (one of the partners of Agoria, "a platform for the collection and analysis of health data serving better patient care" involving DocaPoste in particular, refers to a "trading platform"[xi,xii]) and the quality of the research projects (and services); furthermore, following the proliferation of scandals regarding fraudulent access/processing for commercial purposes, the CEI is concerned about the respect of people's rights (even if it is not strictly speaking about informed consent, the efficacy of the right to oppose the reuse of personal data does not appear to be systematically guaranteed[xiii]; finally, it is concerned about the lack of recognition of the work done for the "production of data", both by the volunteers who contribute to the collection of those data and by the researchers whose expertise gives them meaning);
- a clash of scientific paradigms is emerging which must be analyzed carefully: the predominant knowledge-production model in the biomedical sphere is hypothetico-deductive; it starts from the hypothesis to conduct refutation work, whether in the laboratory or in epidemiological research; the platform model is the opposite, with data collection first, the cross-referencing of multiple sources, and the identification of potentially significant associations through the use of Artificial Intelligence methods[xiv]. This calls into question first of all the very definition of what can be considered as health data. This then leads to arduous epistemological debates and an ethical question consists in being able to guarantee the validity of the data processing and the interpretation of the results produced. Assuming that this is still lawful in terms of the respect of people's rights once again. The questions relating to the implementation of Artificial Intelligence techniques do not fall within the scope of this initial Memo, they will be studied in a second phase and the subject of algorithms has been identified as being key[xv]. Without going so far as to denounce, like Eric Sadin, "radical antihumanism[xvi]," we are attentive to the comments of the French National Centre for Scientific Research (CNRS) Ethics Committee Chair, Jean-Gabriel Ganascia, when he declares that "thorough reflection is needed on the limits to impose on AI[xvii]". Let us note for the moment that France's bioethics law of August 2, 2021 (art. 17) governs the algorithmic processing of Big Data and medical decisions by creating

new obligations in terms of informing people and in terms of “explainability” to overcome the dangers of loss of control and disempowerment of professional users (CSP, new art. L.4001-3).

An initial discussion within the committee at the end of 2020 raised some of the challenges: “beyond the HDH, it is all of the health data collections that are concerned, including those from “data warehouses”[xviii] and cohorts[xix]; data quality and integrity are central, as are issues of confidentiality (does pseudonymization offer sufficient protection) and consent (to what, for how long, based on what information) or non-opposition by the data subjects; is centralization justified if it permits the implementation of security procedures as massive as the risks of forcing[xx] that it runs when methods for the “distributed” analysis of multi-source data exist[xxi]; the increasing complexity of access to HDH handicaps public research when it does not have the same capacity for investment as large private operators, and we should be just as concerned about "excess" processing as about the lack of processing which would be a source of progress; without forgetting the ethical issue of the quality of research projects which are submitted to the committee for opinion prior to authorizing access to the HDH, the Ethics and Scientific Committee for Research, Studies and Evaluation in Health (CESREES); how to articulate the local scientific committees, essential for the producers of data and the role devolved to CESREES; finally, if any of us here in Europe do not share the American vision of the good life, the General Data Protection Regulation (GDPR, *see box*) provides a valuable basis for reflection/protection.

General Data Protection Regulation (GDPR)

Since May 25, 2018, the GDPR has governed the processing of personal data within the European Union – and even outside of it from the moment that European residents are targeted[xxii]. While its material scope of application excludes personal data that are irreversibly rendered anonymous by a process ensuring that the data subject cannot then be re-identified, it does include pseudonymized data (see below). These remain attached to the data subject – for example, by means of an identifier – even if they can "no longer be attributed to a specific data subject without the use of additional information". The geographical scope of application of the GDPR extends to organizations established in the European Union, whether or not the processing takes place there – but not only, it is from the moment that the processing targets European residents. In addition, France has chosen to adopt local particularities in its adaptation of its *Loi Informatique et Libertés* (Data Protection Act), which requires that national rules "apply from the moment that the data subjects reside in France, including when the Data Controller is not established in France".

The fundamental principles of data protection are of particular interest to us (the quote marks indicate the quotations from a booklet by Aurélie Banck, already mentioned):

- The objective pursued by the Data Controller must be for purposes that are "specified" (which excludes any collection of data at random or for preventive purposes), "explicit" (that is to say communicated to the data subject), and "legitimate" in relation to the organization implementing the processing;

- The data must be "processed lawfully, fairly and in a transparent manner in relation to the data subject". The data subject must consent to the processing, or this must be required by particular specified conditions (the requirement of fairness and transparency refers to informing the data subjects and aims to avoid concealed or hidden data processing);

- The data must be "adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed" (this principle of data minimization, also referred to as the principle of proportionality, aims to ensure that the data collected are all strictly necessary for the purpose in question and to exclude any collection carried out in the event that these data should prove to be useful at a later date);

- The data must be "accurate" and, where necessary, kept up to date (which is for example critical when processing data from the hospital Medicalized Information System Program (PMSI) whose nomenclatures and algorithms for classifying stays change regularly, which requires context data to correctly interpret the results of processing over the long term);

- The data must be "processed in a manner that ensures appropriate security of the personal data, including protection against unauthorised or unlawful processing and against accidental loss,

destruction or damage, using appropriate technical or organizational measures" and appropriate to the risks (principles of integrity and confidentiality). Particularly in the case of "sensitive" data or relating to so-called "special" categories ("racial or ethnic origin, opinions, etc., and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation shall be prohibited") except to be able to take advantage of one of the exceptions exhaustively listed by the texts, one of which concerns scientific research. This prohibition in principle does not forbid the processing of these sensitive data but, in order to be able to process these data, it is necessary to be able to justify one of the legal exceptions to the prohibition and surround the processing of these sensitive data with appropriate guarantees (substantive and procedural).

- These considerations are developed, particularly regarding the characteristics of the consent of the data subjects, which must be free, specific (given for a specific precise purpose and in a granular manner) and informed (and especially collected in clear, accessible, and understandable language). Among the sensitive data, special considerations apply to the NIR (registration number in the national directory for the identification of natural persons) commonly referred to as the social security number because of its significant nature and the interconnection risks associated with it.

- "Accountability": the GDPR introduces a major paradigm shift with this principle of accountability, which switches from a static regime of prior formalities (declaration/authorization of processing) to a dynamic regime of global compliance. This principle is reflected in the definition of data protection and information system security policies, a record of processing activities and data breaches, and by taking into account the principles of *Privacy by design* and *Privacy by default*.

- We will end this non-exhaustive recap with one of the major new features introduced by the GDPR: the data protection impact assessment. This is an analysis of the risks to the privacy of the data subjects concerned resulting from the processing of personal data when this risk can be considered high, which is the case with human health research projects (based on nine specified criteria). Carrying out this assessment requires the close collaboration of all the operators concerned by the processing [xxiii]. It will be systematically required and must imperatively be provided in support of the initial authorization request in the event of:

- medical research on patients and/or minors that includes the processing of their genetic data;
- creation of a registry or database ("data or biological samples warehouse") intended to be open to the research community.

I.1 Protection of people's rights in terms of personal data processing

The CNIL works to enforce respect for data protection and the rights of individuals in France according to the principles set out in the French Data Protection Act of 1978, revised in 2019 to incorporate into French law the changes introduced by European law through the GDPR. Two main

types of requirements can be distinguished: data integrity and confidentiality on the one hand, and transparency of processing on the other. Within the research framework, we are adding the right to recognition of the work carried out to enable the constitution of databases, that of voluntary contributors who take their time to regularly answer what are often lengthy questionnaires, that of the researchers who designed the protocols and will conduct the analyses that will produce new knowledge. All of which as part of a two-pronged movement of open science and participatory research.

I.1.1 Data integrity and confidentiality

According to the GDPR: personal data shall be "processed in a manner that ensures appropriate security of the personal data, including protection against unauthorised or unlawful processing and against accidental loss, destruction or damage, using appropriate technical or organisational measures". We only reiterate the fears raised by the choice of hosting made by the HDH. Fears we consider to be all the more justified given the existence of French hosting alternatives, including a public offering, the Secure Data Access Centre (CASD). It seems that the government is gradually returning to the question of "sovereignty" with the obligation for administrations to only use trusted Cloud providers for services processing the data of citizens, companies, and public officials. This is an extremely fast-moving field, both technically and legally, for which the conduct of intelligence is essential. We will also insist on the fact that no procedure can guarantee the absolute security and confidentiality of personal data, whether it is pseudonymization or even anonymization (see box *Pseudonymized and anonymized data*).

Pseudonymized and anonymized data

1. What is pseudonymized data?

Data are referred to as pseudonyms when attribution to a data subject requires the use of additional information (correspondence table, encryption key, etc.). Data resulting from pseudonymization are therefore considered personal data and their processing remains subject to data protection principles.

In practice, pseudonymization consists of replacing the directly identifying data (surname, first name, etc.) of a dataset with indirectly identifying data (alias, sequential number, etc.). This makes it possible to process people's data without being able to identify them directly.

This definition covers various techniques commonly used in health research:

- use of a table of correspondence between the pseudonymous (coded) dataset required for the analyses and the identity data stored separately, conventionally used in clinical trials;

- encryption of directly identifying data to make them incomprehensible with a secret making it possible to chain data relating to an individual and follow his or her journey without the potential for identification.

Pseudonymization is a way to process personal data that ensures their security whilst fully preserving their utility. Unlike anonymization, it is reversible. In practice, it is possible to find the identity of people whose data have been pseudonymized.

The texts encourage the use of pseudonymization within the framework of scientific research (GDPR, art. 89). Pseudonymization does reduce the risk of a dataset being correlated with the original identity of a data subject and as such helps minimize risks to the data subjects.

Whatever the pseudonymization technique applied, the information used to link the pseudonyms generated with the directly identifying data is highly sensitive. Care must be taken to ensure that these elements are kept confidential by the use of the appropriate technical and organizational measures. This information should therefore only be accessible by authorized persons and under previously specified conditions.

2. Anonymous or anonymized data?

Anonymous or anonymized data are not subject to data protection legislation, whether they are anonymous to begin with or following anonymization by processing to ensure that the data subject cannot be re-identified subsequently (anonymized data). Using anonymous or anonymized data therefore makes it possible to bypass the regulations because their dissemination or reuse does not impact the privacy of the data subjects.

The impossibility of identifying individuals requires a case-by-case risk assessment. This is ascertained in relation to the means reasonably likely to be used by the Data Controller or by any other person. In practice, it involves taking into account the cost of identification, the time required for it, the technologies available – current but also future. If there is any doubt about the identifying nature of the data, it is recommended to consider them as identifying, until proven otherwise.

How should an anonymization process be evaluated?

In their 2014 opinion, the European data protection authorities define three criteria used to ensure that a dataset is truly anonymous:

- Non-individualization: it must not be possible to isolate an individual within the dataset;
- Non-correlation: it must not be possible to link separate datasets concerning the same individual;
- Non-inference: it must not be possible to deduce with near certainty new information about an individual.

If these three criteria are not fully met, it must be demonstrated, through an in-depth assessment of the risks of identification, that the risk of re-identification with reasonable means is zero.

Given that anonymization and re-identification techniques evolve regularly, it is essential to conduct regular intelligence in order to preserve, over time, the anonymous nature of the data produced.

In practice, anonymization therefore implies:

- An assessment, on a case-by-case basis, taking into account both the context and the risk, of the anonymization technique or combination of techniques, it being understood that no technique is infallible, as indicated by the Article 29 Working Party and confirmed by research in the field,
- Regular reassessment in relation to the evolution of the techniques (v),
- The irreversible destruction of the initial data (vi).

Still according to the GDPR, data must be “processed lawfully, fairly and in a transparent manner in relation to the data subject”. This requirement is complex to meet for several reasons:

- the first reason stems from the very definition of what can be considered as health data and which in part influences the level of information given to the citizen regarding certain uses;
- the second concerns the quality of the systems used to inform people, which is the basis of their ability either to give their consent to the use and reuse of their data, or to apply their right to oppose or even erase certain personal data.

The definition of personal data (*see box Health data: a very broad definition*) appears to be legally clear[xxiv], just like that of processing[xxv], but that of personal health data lends itself to variable interpretations (for example, the SantéNathon collective disagrees with the Council of State on the status of the medical appointment data recorded by the appointment scheduling website Doctolib; although in our opinion using "medical" as a descriptor for these data leaves little room for doubt). Above all, this Memo only concerns data explicitly collected for research or health care/health administration purposes, which still leaves those that are collected and transmitted by all kinds of connected objects which people use out of a concern for their well-being[xxvi] (Internet of Things) or for explicit medical monitoring purposes (with a need to clarify the conditions for respecting data confidentiality when private providers provide individualized medical monitoring advice using data collected during treatment, for example)[xxvii].

Health data: a very broad definition

"Data concerning health" are defined as "personal data related to the physical or mental health of a natural person, including the provision of health care services, which reveal information about his or her health status".

Health status refers to **present, past, or future physical, mental, spiritual, and social health** and the data very broadly include:

- any information on the identification of the patient in the health care system,
- any health care services provided,
- information derived from the testing or examination of a body part or bodily substance, including from genetic data and biological samples (health data "by intended use");
- any information on, for example, a disease, disability, disease risk, medical history, clinical treatment or the physiological or biomedical state of the data subject independent of its source (recital 35).

Are concerned:

- the data that make it possible to indicate the pathology from which a person may be affected (health data "**by intended use**"); medical history, illnesses, health care provided, results of examinations, treatments, disability, etc.
- the cross-referencing of data that allow conclusions to be drawn about a person's state of health or risk to his or her health (e.g. cross-referencing of a weight measurement with age, height, etc.) (health data "**by cross-referencing**");
- data that become health data as a result of the medical use made of them, including in the context of health research involving human biological samples (health data "**by intended use**").

The concept of personal health data must be assessed on a case-by-case basis depending on the nature of the data collected.

Genetic data is also defined as "personal data relating to the inherited or acquired genetic characteristics of a natural person which give unique information about the physiology or the health of that natural person and which result, in particular, from an analysis of a biological sample from the natural person in question" (art. 4.13).

From a health research perspective, two types of health data appear to be distinguished at first glance: data produced for the primary purposes of research, whether or not it involves human subjects (RIPH or non-RIPH), and those that may be of interest for research whilst being primarily collected either within the framework of health care or for primary "medical-administrative" purposes, also called "routine data" (Medicalized Information System Programs (PMSIs), national medical

databases (SNIIRAM)/SNDS). According to the law, “excluded from the scope of personal data processing in the health domain is the processing (of data) within the scope of individual health care, mandatory or complementary state health insurance benefits, or the management of medical information in healthcare facilities”. But once these data are reused in the health research context, they fall within the scope of the Processing of personal data in the health domain (art. 64 et seq. of the French Data Protection Act) and are therefore subject to related formalities.

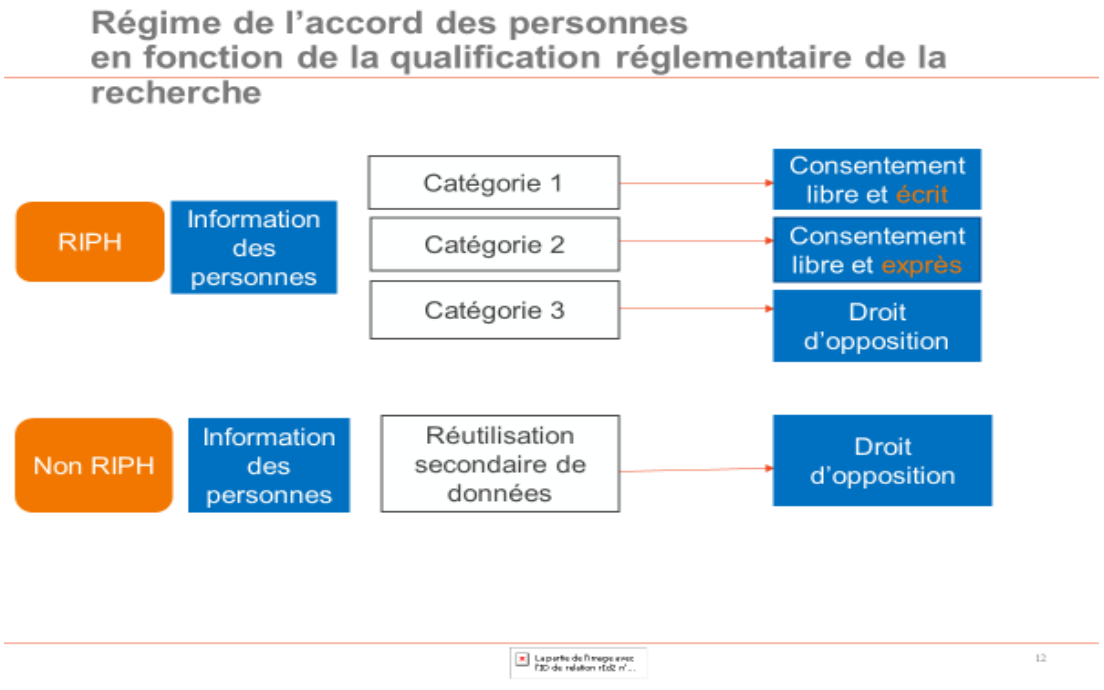
Hospital medical information is not limited to data belonging to the core of the SNDS, those of the various PMSIs. It also includes health care data that are stored in health data warehouses. These contain masses of highly sensitive identifying data (from orthogenics and psychiatry to genome sequencing) collected by personnel who often combine health care and research functions, requiring them to constantly juggle two registers: one being personalized health care within the framework of an interpersonal relationship of trust covered by medical confidentiality, the other being research that mobilizes these data for purposes other than those for which they were initially collected – except in the case of clinical research protocols that immediately come under the definition of research involving human subjects (RIPH) (Amiel and Dosquet, 2021[xxviii]). Particular vigilance is required in terms of respecting people’s rights, especially the right to be informed of processing based on the reuse of their health care data.

In an attempt to simplify, we can reason in terms of the “data cycle” by distinguishing, for convenience, between the emergence of the data (“initial production of primary data”) and its reuse (for purposes other than those for which it was originally collected).

I.1.2 Data cycle, legal rules, information to participants, and confidentiality

Initial production/emergence of primary health data within the framework of research and obligations to inform participants

The regulatory designation of the research project directly determines the formalities applicable to it (submission of an application for approval or commitment to comply with a Reference Methodology) and the nature of the rights afforded to the participants (*see below Participant agreement type according to the regulatory designation of the research*). While some health research projects are exempt from formalities or can be implemented immediately by Data Controllers if they comply with a Reference Methodology, other research projects must be approved by the CNIL.



Whatever the designation of the research and the formalities that apply as a result, it is essential to inform the participants. The GDPR guarantees greater transparency with regard to data subjects aimed at giving them better control over their personal data. In terms of health research, this transparency is an essential guarantee given to participants in return for the lifting of professional

secrecy. It enables them to understand the objectives of the research, the terms of their participation, the scope of the agreement they give, and it enables them to control the use that will be made of their data. It determines and facilitates the effective application of their rights. It helps to establish a relationship of trust with researchers.

As a matter of principle, the information must be provided individually to each person participating in the research, whether the data is collected from him or her or from third parties, and must be specific to each project (French Data Protection Act, art. 58). This must be done for each project in which the patient participates or for which the patient's data will be the subject of processing.

The individual information must be accompanied by general information in health care facilities that transmit personal data to enable research activities (display on the premises, inclusion in the welcome booklet, etc.)

The conditions are precisely governed by the French Public Health Code (CSP) with regard to research data (Amiel, Dosquet and CEEI, 2021[xxix]). In the case of research involving human subjects (RIPH), the studies can only be legally implemented once a favorable opinion (ethics opinion) has been issued by a committee for the protection of human subjects (CPP) and the latter examines with particular attention the methods of obtaining the subjects' consent – and therefore the clarity and the completeness of the information that will be given to them – in order to make its decision. Inserm's Ethics Evaluation Committee (CEEI-IRB) does the same within its field of competence, namely for research not designated as involving human subjects (non-RIPH), like the ethics committees created by universities. In the case of non-interventional research in the field of health, studies do not require the collection of consent but that of the manifestation of non-opposition. The information provided is just as governed by the texts and Reference Methodologies to which these protocols can be attached by precisely defining the requirements to be met. According to the CNIL, "Reference Methodology MR-003 governs processing comprising health data and of a public interest nature, carried out within the framework of research for which the data subject does not oppose participation after being informed. More specifically, it concerns non-interventional research and medicinal product cluster trials. Information of the patients on an individual basis is mandatory. The Data Controller undertakes to collect only that data which is strictly necessary and relevant with regard to the objectives of the research. "

"Reference Methodology MR-004 governs the processing of personal data for the purposes of study, evaluation or research that do not involve human subjects. More precisely, these are studies that do not meet the definition of research involving human subjects, in particular **studies focusing**

exclusively on the reuse of data." (CNIL deliberation no. 2018-155 of May 3, 2018 specifies that "The research must be of public interest. The Data Controller undertakes to collect only that data which is strictly necessary and relevant with regard to the objectives of the research."). Its requirements are clear and carry obligations to inform people on a collective and individual basis.

Reuse and sharing of data and information to data subjects

The answer to certain research questions may require – and these uses are increasing – the reuse of data (research or routine data), which can only be a good thing when it comes to recognizing the initial work put in in order to "produce" them. The reuse of data does not change the designation of the research. This is the case for cohorts, whose succession of studies most often falls within research in human subjects Category 3 (non-interventional research). The occasional secondary reuse of data constitutes processing of personal data that is distinct from the source processing. This new processing is subject to specific formalities and the provision of specific information to the data subjects.

The essential point being that any new processing fulfils a new purpose which therefore requires a new procedure for informing people. This is a very sensitive point, which was strongly emphasized to us by the board members of the Constances association as well as by the scientific managers of the cohort of the same name.

Data reuse also occurs when health care data are used secondarily for research purposes. They often fall within the scope of MR004, which primarily aims to regulate the exclusive reuse of these data.

It is essential to design flexible and dynamic methods for the exercising of rights, which are likely to enable infrastructures to address the challenges of a deeper understanding of pathological mechanisms, for example, whilst guaranteeing better control by people of their data, which contributes to the essential trust of the data subjects. As Georges Dagher underlines with regard to biobanks, *"(. . .) it is time to rethink the role of source persons more broadly in terms of participation in and contribution to research (. . .). The new paradigm developed by the use of biological collections and aimed at creating a resource for research invites an update of the regulatory and ethical framework that governs the issue of patient participation in research projects (. . .)* " (*Le Monde Sciences & Santé* supplement, Wednesday July 8, 2015). We must welcome the evolution of the CNIL's doctrine on this point, illustrated by draft Reference Methodology MR004 (*see box "Information to persons and respect*

of 'data protection act' rights within the scope of research under MR004 relating exclusively to data, biological samples, or both")

Each MR004-compliant project must be recorded in a public directory maintained by the HDH and which can be consulted on its website.

Information to persons and respect of 'data protection act' rights within the scope of research under MR004 relating exclusively to data, biological samples, or both

MR004 accepts that data and/or biological samples can be reused, and that people can be considered to be validly informed when:

1. "General information concerning the research activities in the establishment must be provided to the data subjects (displayed on the premises, mentioned in the welcome booklet, etc.).
2. Added to this general information is the individual information given to the patient who is included in the research. This must be done for each project in which the patient participates or for which the patient's data will be processed.
3. Data and/or biological samples that are not specifically collected for the research may be reused without any new individual information to the data subjects:

. When the data subject already has the information envisaged in articles 13 and 14 of the GDPR; this could, for example, concern several research projects, performed by the same Data Controller with identical purposes, identical data categories, and identical recipients;
. Or when the information provided when collecting the data and/or biological samples envisages the possibility of reusing the data and/or samples and refers the data subjects to a specific information mechanism that they can consult prior to the implementation of each new data processing (for example, a website presenting each research project carried out on the data and/or samples collected within the framework of the initial information). "

It can be a website, centralizing information relating to all the projects carried out, their characteristics and to which people can refer.
Otherwise, exemption from the obligation to inform will always be possible, but the absence of such information provision will render the research ineligible for the Reference Methodologies and a request for authorization will be necessary.

Recommendations:

Taking a transparent approach, it is essential to anticipate the fate of the personal data collected as part of an initial project by mentioning, in the participant information document, the future sharing of the data and their reuse (including in third countries) for one or more health research purposes, and by referring people to a specific information system, such as a "transparency portal", to which they can refer so as to not have to directly and personally contact them before implementing new processing.

The general information will not dispense with the need for prior individual information specific to each new research project requiring the secondary reuse of data already collected, but if project documentation on this website is envisaged, the data subjects will have the opportunity to inquire should they so wish.
If this is not done, an exemption from the obligation to inform will still be possible, but:
- the absence of such information provision will render the research ineligible for the CNIL reference methodologies, and authorization from the CNIL will be necessary;
- the Data Controller must therefore request exemption from the obligation to inform, in the authorization application that must be sent to the CNIL. This request must be solidly documented by justifying the impossibility of informing people, disproportionate efforts, or that informing them renders impossible or seriously compromises the performance of the study.

Reuse and sharing of data between multiple research organizations

The Inserm teams may be required to "share" personal data with other Inserm teams (in which case the Data Controller and DPO remain the same), but also elsewhere, for example with Pasteur (in which case the Data Controller and DPO change).

When setting up a database within the framework of an initial research project, health care, etc., regardless of its regulatory designation, we must consider the question of the reuse of the data contained in several research projects ("warehouse”), inform the data subjects of the secondary reuse of their data as part of scientific research projects, as indicated above, and obtain the explicit consent of those concerned by this reuse.

In the event of explicit consent not being obtained, the processing relating to the establishment of the warehouse must be the subject of a request for “health” authorization (excluding research) and the intended purpose must be of public interest (a CNIL reference framework is in preparation).

In all cases, a data protection impact assessment (DPIA) will have to be carried out.

If it concerns data pertaining to French research in human subjects Category 1 (interventional research) or Category 2 (interventional research with minor obligations and risk), express consent to participate in the research is mandatory and the possibility for reusing the data can be provided through an additional checkbox. In a non-interventional study that requires "only" the manifestation of non-opposition, obtaining explicit consent for the secondary reuse of data for research purposes is more problematic.

The occasional secondary reuse of data constitutes processing of personal data that is distinct from the source processing and, insofar as this new processing is subject to specific formalities and specific

information to the data subjects, Inserm, responsible for the "source" processing, becomes the data supplier for the subsequent scientific project and it is its responsibility to communicate only the data strictly necessary for the project and to ensure that the third party is "authorized". An explicit contract with the partners is required to confirm a change of responsibility (it is the team receiving the data for other processing which is responsible for that processing and therefore for the data it uses for that purpose).

Knowing that if the partner requests data from Elfe, for example, it will be a precisely defined subset, as economical as possible (data minimization), de-identified, and with the approval of Elfe's scientific committee ("raise the awareness of the entry points") regarding its relevance to the research question asked and the conditions of personal data security and protection (no possibility of re-identification due to new possibilities of cross-referencing). In this chain of "successive responsibilities", Inserm must ensure that any transfer is made for the benefit of a recipient authorized to receive them or who undertakes to comply with an RM, before being able to consider that the transferred data are then the responsibility of the recipient. Governance of access to data involving those responsible for the source databases must therefore be put in place to document the project's scientific and regulatory compliance.

But if the partner, for the purposes of its research protocol, needs to administer questionnaires to those having participated in the initial research, it can only do so if it has access to their identity, which is only possible through the intermediary of the team behind the first processing, which alone has the possibility of recontacting people and informing them of the purposes of the new processing. And this secondary research must itself be qualified and follow the ad hoc circuit.

As Dutch researchers Jacobs and Popma (2019) have stated, "*Data governance should not end with sharing*".

I.1.3 *Problems posed by including in the HDH catalog data collected for research purposes*

With the aim of facilitating the sharing of data from varied sources for equally varied purposes, the HDH encourages "data producers" to make the data available on the national HDH. This invitation concerns databases created by researchers and placed under the responsibility of Inserm. While the GDPR appears to provide a satisfactory framework ("processor agreement") for the usual relationships between the "controller" ("the responsible person/organization providing the means and determining

the goals of data processing") and the "data processor" ("processing the data on behalf of the controller"), the majority of research configurations do not fall within this framework and result in responsibility for the data being transferred to the party hosting/processing them. The problem arises from the fact that the consent was given to the party responsible for the collection, whilst the new custodian of the data (the HDH in our case) and the party that will process those data have no personal connection with the subjects having consented to their collection for a specific use or any knowledge of their identity; they have no way of informing them individually with a view to requesting consent to the re-use of those data, generally for a purpose other than that for which they had been collected. This can be a major source of mistrust by research participants, and protocols guaranteeing initial commitments must be envisaged between the parties initially responsible for the collection and those to which they have been transferred (Jacobs and Popma, 2019). However, it is important not to compromise the participation of people in the cohorts, particularly the general population cohorts where participation is on a voluntary basis. And since the lifespan of these collections generally exceeds the time dedicated to the initial project by the principal investigator, the protocol must cover the entire data lifecycle...

According to Jacobs and Popma (2019), the legal basis for the reuse of shared data has four dimensions:

- guarantee that the delegated use remains within the bounds of the commitments made at the time of the initial collection; the user of the transferred data has the real possibility of complying with the obligations of the first "controller";
- reinforce data protection, confidentiality and pseudonymization requirements in accordance with ISO standards;
- protect intellectual property, including that relating to all derived data ("derived from other data or from biosamples");
- allow the initial investigator to verify compliance with these obligations with the possibility to revoke the agreement to use the data in the event of non-compliance.

1.1.4 Concerns related to GDPR compliance (excluding hosting) in the case of health data warehouses

According to the CNIL[xxx], "Data warehouses are created primarily to collect and have available Big Data (data relating to patient medical care, sociodemographic data, data from previous research, etc.). These data are then reused, mainly for studies, research, and evaluations in the health domain. These

databases are set up for a long period of time (generally more than 10 years), with the objective of obtaining large volumes of data, and can be supplied by multiple sources (health care professionals, patients, pharmacies, health care establishments, etc.). ”

Since the Group began its work, the CNIL has adopted a reference framework relating to the processing of personal data implemented for the purpose of creating health data warehouses. “The reference framework enables organizations wishing to set up a data warehouse in compliance with that reference framework to not request prior authorization from the CNIL: after verifying the compliance of its warehouse project with the reference framework, the organization can declare its compliance. Internally, the organization responsible for this processing is required to document its compliance with GDPR and the reference framework in its record of processing activities. The reference framework only applies to health data warehouses based on the performance of a task conducted in the public interest, within the meaning of GDPR article 6.1.e. ”

In the current context of the operation of healthcare establishments, doubts are expressed as to the reality and effectiveness of the information to those whose data are collected, the nature of the uses that can be made of the data and in particular that of its reuse for purposes for which they had not been collected[xxxi]. This already applies to the health data warehouses independently of the HDH because of the sensitive nature of the data they contain (a single contact with the health system is private information, but it is potentially about data of a more intimate nature (sex life, addictive behavior, mental health, etc.) or that which is of interest to insurers (oncology), especially given the duration of storage of certain data which may have a deferred interest whether for care or research, genetic data, for example.

This concern is reinforced due to the possibilities of matching databases with an increasing number of other databases permitted by the HDH through its catalog. If only because under these conditions anonymous data no longer remain anonymous given the possibility of re-identifying people as a result of the potential cross-referencing of very precise information.

The situation of the health data warehouses sometimes seems opposed to that of research initiatives such as cohorts: the people whose data are recorded in health data warehouses – health service users – are supposed to authorize the recording of their health care data over time on the basis of clear information that also applies to the reuse of these data for research purposes. The information appears minimal, in fact (footnote in small print on "administrative" documents (reports or invitations to various appointments, in particular) and the information on reuse must be the subject of an active search on a website that is more or less explicit and easy to consult. In contrast, in a cohort like Constances [in italics, extracts from the interview report], “*the issue of protecting the data and rights*

of the users is the objective of the 'moral contract' between the volunteers and the cohort, which forms the basis of a relationship of trust (information and consent are also formalized at all stages of participation in the research activities). This trust goes first of all to the cohort coordinators (quality of the relationship established by respected researchers) and to their institutional affiliation (Inserm as a public research organization). It is also based on the establishment of explicit procedures of information and the collection of consent to participate in the research (from the initial acceptance to participate after having been drawn at random from the National Inter-Scheme Health Insurance Register (RNIAM) to the consent requested for each of the research operations: annual and specific questionnaires, inclusion in the biobank [its hosting in Luxembourg also raises some reluctance on the part of some volunteers]. The webinars organized by the cohort coordinators for volunteers, the one on the Secure Data Access Centre (CASD) in particular, are appreciated.[xxxii]*. "*

The cohort managers pay close attention to respecting the contract of trust that binds them to the volunteers.

"While the data are the property of the volunteers, it is clear to everyone that they are collected for the purpose of sharing ('our data belongs to us, we agree to share them, the question is with whom'; 'without researchers, our data are useless' [association]; 'in no way do I consider them to be my data, they are data that have been entrusted to me' [researcher])."

"All research projects are based on the reuse of data that implies the explicit agreement of the volunteers. They are informed by a letter from the cohort, and they can individually oppose the reuse of their data for a specific project. Oppositions are actually rare (a few dozen each time – rather for private-sector projects) but they are enough to show that the information is circulating. "

It seems that the philosophy behind the establishment of health data warehouses may vary. Some health data warehouses have integrated, from their design stage, beyond the respect for the rights of individuals, the necessity to consider them as partners. The Ouest Data Hub, for example, was formed in a "data ecosystem" which considers that it is not data that we share but expertise surrounding data. And that expertise is clinical. But such an ecosystem implies the confidence of the people who entrust their data, and the latter is based on transparency with a constant effort to inform people.

I.2 Data quality and research project quality

I.2.1 Data quality: epidemiological data vs. routine data

Remember that according to the GDPR, the data must be “accurate”. Therein lies the strength of ad hoc data collection systems such as epidemiological surveys and cohorts, organized to collect data of controlled quality (even if some of the researchers we interviewed consider that “very good things can be done with the SNDS”). This quality also lies in the finer granularity than that of routine data. This advantage comes at a price: the cost of operating these infrastructures; the “limited” size (even if a cohort like Constances has around 220,000 volunteers) which prevents the study of rare diseases; absence of pediatric representation (for Constances, although there is Elfe); more generally selection bias linked to the voluntary nature of the participation.

The scientific managers of the Constances cohort give specific examples that illustrate the superiority of cohorts when it comes to producing valid epidemiological knowledge. The benefit of data granularity can be illustrated in terms of diagnostic criteria. Constances immediately distinguishes between type 1 and 2 diabetes, for which knowledge of treatment with insulin – the only information present in the SNDS – is insufficient for distinguishing between the two conditions. It is then possible to calibrate an algorithm to distinguish between the types of diabetes using only SNDS data in relation to this diagnostic reference. Work is underway within the framework of the ReDSiam network to assess the capacity of the SNDS to identify health problems defined on the basis of algorithms using only SNDS data and validating them using additional data from Constances, disease registries, or other sources whose diagnostic information can serve as a reference.

The accuracy and detail of the data also offer possibilities to take confounding factors into account and control bias (in part for classification and selection) that cannot be achieved with routine data alone. Studying the consumption of health care by people with obesity, for example, involves taking into account several confounding factors: isolating the role of obesity alone involves knowing the individuals’ weight (which the SNDS does not contain) and taking into account (for example) the presence of diabetes, how long they have had it and any possible complications, smoking (how much they smoke and how long they have been smoking), … all details that are absent from the SNDS and can only be found in Constances. Without forgetting to reiterate the possibility of taking precise account of people’s social status (see below). Under certain conditions, the partial control of selection biases is possible: Constance has information on the cohort of non-volunteers (n=400,000) which makes it possible to assess the extent of its selection biases and produce corrected epidemiological data.

Another advantage of a cohort such as Constances is that its open infrastructure enables the conduct of research projects using data from the cohort and possibly matched data (over one hundred projects submitted). By comparison, the HDH is a host that is unable to help researchers use the data from the databases in its catalog for at least two reasons:

- Some research projects involve the collection of specific data that requires being able to enter into contact with volunteers whose profiles are of interest; only an initiative like Constances has this possibility for nominative identification that makes it possible to inform people and collect their consent;

- The research projects require in-depth knowledge of the highly detailed data collected; only the researchers having generated them have this knowledge and are therefore in a position to guide other researchers in choosing the relevant data to answer a specific research question. For example: many variables can be used to characterize social status in Constances, which should be used? Constances permanently employs four epidemiologists tasked with guiding researchers in choosing and processing the relevant variables in relation to a specific problem.

I.2.2 Research project quality

Doubts about the quality of the projects arise from the fact that they will be conducted in an environment (HDH) that does not have research as its primary purpose, but rather goals to optimize health system functioning by taking advantage of the combination of immense possibilities for data storage and the development of innovative processing methods thanks to the mobilization of Artificial Intelligence. The purpose of optimization is hardly debatable in itself (and to not take advantage of the possibilities would be unethical); the perspective can be (depending on one's appraisal of the various concepts of "optimization"; will cost-effectiveness issues, for example, be compared with equity issues?), regardless of the fact that we are probably deluding ourselves about the scope of the expected innovations. One question may arise as to the scope of what is considered to come under research. An entire field of research on health services, programs and policies is based on the mobilization of "medical-administrative databases" with full awareness of the limits of validity of the data they contain (supposedly offset by their huge volume and the power of the processing algorithms). Another area is the development of Artificial Intelligence methods for data analysis and decision support. Artificial Intelligence will never compensate for the difference in validity between routine data (those of the Medical-Administrative Databases [BDMAs] are not collected for research) and data generated by researchers for epidemiological research purposes (whether descriptive, analytical, or evaluative). The constitution (under-representation of researchers) and the operating methods (workload) of the Ethics and Scientific Committee for Research, Studies and Evaluation in Health (CESREES) raise questions in this regard in comparison with the functioning of the committees that came before it (Consultative Committee on Health Research Information [CCTIRS] followed by Expert Committee for Research, Studies and Evaluation in Health [CEREES]).

II Avenues for reflection and action regarding the ethical and responsible use of Health Big Data

The following section provides starting points for further reflection. These are inspired by the content of the initial interviews conducted to start defining the scope of the subject. Therefore, they do not represent a structured program but rather initial elements to be enriched with other interpretations and interviews, with the HDH managers, particularly for the European dimension, and different research teams active in the field (particularly the fields of Artificial Intelligence and the Internet of Things).

Guaranteeing the security of data storage

The reflection by the CEI began out of a deep concern regarding the decision to entrust the hosting of the SNDS to a foreign private company, which in addition is governed under US law. Given that the reflection by the public authorities is now oriented towards reconquering a certain level of sovereignty, it appears all the more important to study hosting possibilities at national (or European) level. Alongside private companies such as OVH (especially given the major damage to its credibility by the fire in some of its data centers in Strasbourg) and Thalès, the existence of a public solution such as the Secure Data Access Centre (CASD) appears to be worth reiterating. Especially since another aspect of security deserves consideration: the issue of centralization versus the possibility of processing by distributed databases (instead of moving the data en masse to centers where storage and processing are carried out together, it is the processing software that moves to the data storage locations, thereby reducing both the size of the “honeypot” – there are several of them – and the risk of leaks). One advantage of the CASD is that it already hosts extremely sensitive data from public operators (the French inland revenue service having particularly high security requirements), data which also may be of scientific relevance to health research (on social inequalities in particular). But the CASD has also developed a comprehensive security procedure (physical and logical) that appears particularly suited to matching research data with routine data: the creation of secure bubbles containing only the data necessary for a project and the relevant processing software (including Artificial Intelligence), hermetically sealed and accessible only from secure, locked, and sealed SD-Boxes, with smart card and biometric authentication. Some Inserm teams, such as Constances and the

Pierre Louis Center, use it. Other advantages of the CASD include its participation in the International Data Access Network project and the development of an original procedure for certifying the quality of publications by scientific journals, in partnership with the CASCAD certification agency (dedicated secure bubbles for access by the auditor, by CASCAD, to the source data of the scientific publications).

The economic and industrial question also arises at local level. While some establishments, such as the Paris Public Hospitals Group (AP-HP), choose free software, it is rarely comprehensive and an initial constraint for the development of health data warehouses is the control of industrial partnership. According to one of our key sources, “there is an economy surrounding platforms” and “so much the better if it is French companies that are helping to develop Research & Development.” The challenge according to him – and it is national – is not to miss the convergence with a heavy financial dimension. The human dimension is also major; it is difficult to recruit at market price and the question arises of knowing how to expand the teams (of the Clinical Data Centers in the Grand-Ouest organization) and recruit high-level data scientists in the public sector under public sector remuneration conditions.

Respecting people’s right to honest and accessible information

At a time when France’s Data Protection Act is celebrating its 40th anniversary and whose principles, which have remained essentially unchanged, are known and respected in the conduct of research, the introduction of GDPR has profoundly modified the approach to data protection by inviting itself into the governance and organization of the stakeholders who must document compliance and the respect of it; respect that can still be greatly improved upon by health professionals across the board. Much remains to be done to ensure that data is always “processed lawfully, fairly and in a transparent manner in relation to the data subject”, “adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed”, “accurate” or “processed in a manner that ensures appropriate security of the personal data, including protection against unauthorised or unlawful processing and against accidental loss, destruction or damage, using appropriate technical or organisational measures” and so that the organization and the procedures attesting to it are in place.

Our interviews show that there is still a long way to go in order to guarantee fair information to people that is a guarantee of informed consent. This arises within the primary framework of research protocols, especially regarding the reuse of data, whether they are collected within the secure framework of research or more particularly in the context of health care. In the latter case, the data collected as part of the doctor-patient relationship are often highly sensitive; their potential re-

use in research is neither systematically envisaged nor information on this subject given. In the knowledge that any reuse is supposed to be the subject of a new information procedure specific to the intended scientific purpose. Procedures for systematic information and effective collection of specific consent remain to be put in place, particularly regarding health data warehouses. Very precise indications are provided by MR004. The modes of providing general information include display on the premises, Patient Charter, etc., and these must be supplemented by specific individualized methods. The time constraint cannot be minimized, and innovative procedures are emerging, whose possibilities for widespread application remain to be evaluated before such application. It is a substantial undertaking but one that determines the public's trust in the conduct of research that respects people's rights and their ongoing participation.

Recognizing a partnership between subjects and researchers in the production of scientific knowledge

Like others, the HDH managers told us that they want to promote a "data culture". This seems essential and it appears to us that it must take into account the specific dimension of research data. Unlike "routine data", research data are the result of work done on two levels: that of people who commit and lend themselves to research, to contribute to the advancement of knowledge, on a voluntary basis (the time they devote to this cannot be overlooked); and that of researchers who have designed and implemented systems for the collection (mainly by questionnaire as part of cohort studies) and processing of data with an ongoing concern for validity at all stages of the data life cycle. The whole is based on a contract of reciprocal trust that is formed as part of a demanding long-term partnership, trust that is extremely fragile however and potentially threatened by the introduction of third parties that have neither the history nor the culture of this relationship, as could happen were research mechanisms to be listed without precautions in the HDH catalog. This work also calls for mechanisms of value creation that break its invisibility and recognize it at its true value when these data are poured into an undifferentiated shared pot where they are aggregated with routine data. This need for recognition is plural, at the very least being symbolic (recognition of the effort to participate in the production of a common good) and financial (no good is without cost and data has a cost that must be compensated in order to maintain high levels of quality).

Developing the participatory aspect of epidemiological research

The scientific leaders of the Constances cohort, just like the members of the association of the same name, give an example of the contractual relationship between volunteers and researchers. The researchers are aware that they work with "entrusted data"; the volunteers are willing to be able to

decide whether or not to participate in various projects depending on how relevant they understand them to be. This goes beyond the respect of the right to informed consent to represent a veritable research partnership in a public framework protected from private interests (this is what the volunteers say). This implies the implementation, over the long-term, of sophisticated information-sharing and consultation mechanisms in which digital technology has a central place.

The manager of the Ouest Data Hub explains the importance of this participatory aspect in the constitution and routine operation of health data warehouses as well. It is not data that is produced but expertise, and without that expertise and the transparency of that expertise there can be no trust.

Guaranteeing the scientific validity of research projects

A prominent observer of the functioning of expert committees has drawn attention to the temporal and technical constraints that weigh on the evaluation of research projects. A delicate balance must be struck between the “public interest” and scientific rigor (which should go hand in hand). This is based on adversarial peer review and the building of a shared culture of the examination of cases that facilitates their collective expert appraisal (anticipation also facilitating the management of urgent situations). This implies at the very least the sufficient presence of researchers on the dedicated committees (number, time spent, and deadlines for the expert appraisal compatible with rigorous examination).

The scientific guarantee of the research projects also implies detailed knowledge of the data (relevance in relation to the knowledge production objective, limits of validity) and therefore the close association of the researchers with the research that mobilizes the data in which they have expertise. The quality of this association particularly depends on the consideration given to confounding factors and the control of the multiple biases against which all epidemiological research must guard as much as possible. Those in charge of the Constances cohort are particularly insistent on this, just like the Ouest Data Hub partners ("never alone in the face of data!”).

According to the Rennes experience, the dematerialization of the patient record at the origin of the construction of a health data warehouse succeeds if it is driven by the central idea that the data collected for health care purposes may be used again, particularly for research purposes. In this approach, the technical tool is certainly important, but it is around the medical expertise of clinical data processing that it must be developed (the organization around clinical data centers in each establishment of a network as an example). Research consists rather of studying the questions that arise and examining whether the data to answer them are available. This model is the reverse of that of the “data brokers” who deploy tools to capture data first and “provide” it (for a fee) afterwards.

Reinforcing researcher acculturation and support

Here we come back to the imperative of respecting the rights and freedoms of people who participate in research and its conditions of effectiveness.

It is not easy for researchers, especially when they are health care professionals, to be aware of and continuously respect the requirements of managing clinical information on two levels, that of the clinical relationship in which the collection of nominative individual data goes without saying – these data being excluded from the regulatory framework of personal processing in the health domain, and that of research, where the protection of anonymity must be strict, the information to people receiving care must be permanent and accessible in order to guarantee in both cases informed consent or the manifestation of non-opposition to the collection and reuse of data. A general educational effort appears necessary, and a suitable training system remains to be established.

But the requirements for compliance with GDPR and French laws in research are complex. While the drafting of procedures is sometimes simple, it most often requires subtle appraisals of the situation in relation to CNIL requirements and arbitrations that must be manageable between legal or regulatory constraints and their possibilities of compliance (information to people lost to follow-up, retention period and procedures for periodic updating). The CEEI-IRB does its part in relation to its field of competence; the DPO likewise by innovating strongly during the COVID period to accelerate the examination of authorization requests with the CNIL. A significant need to strengthen their resources is clearly apparent, as well as that of local research support bodies, these local bodies being increasingly required to expand their field of competence beyond that of clinical research for which they were originally established. This represents a major challenge when it comes to realizing the promises of ethical and responsible research at Inserm and among its partners.

Acknowledgements

The CEI Working Group leaders would like to thank all the group members in addition to all the CEI members, first and foremost its Chair, Hervé Chneiweiss. The contribution of the Inserm Ethics Evaluation Committee (CEEI-IRB) President, Christine Dosquet, was also particularly important. The group also received the support of Inserm's former Data Protection Officer (DPO), Frédérique Lesaulnier. Finally, Catherine Bourgain provided the group with the benefit of the reflection by a circle of social science researchers from the School of Advanced Studies in the Social Sciences (EHESS).

Our thanks first and foremost to those who agreed to be interviewed in order to deepen the group's understanding. This Memo attempts to faithfully translate their remarks but does not commit them in any way. The responsibility for its content lies exclusively with its authors.

The following contributed successively up until the writing of this Interim Memo:

- Prof. Catherine Quantin, University Professor and Hospital Practitioner (PU-PH) in biostatistics and medical informatics, specialist in medical information at Dijon University Hospital
- Dr. Grégoire Rey, Director of the Inserm Epidemiology Center on the Medical Causes of Death (CépiDC)
- Ms. Frédérique Lesaulnier, Inserm Data Protection Officer until autumn 2021
- Prof. Jean-Louis Serre, President of the Consultative Committee on Health Research Information (CCTIRS) followed by the Expert Committee for Research, Studies and Evaluation in Health (CEREES)
- Mr. Kamel Gadouche, Director of the Secure Data Access Centre (CASD)
- Dr. Marie Zins and Prof. Marcel Goldberg, Scientific Managers of the Constances cohort
- Ms. Florence Ghioldi, President, Frédérique Anne, Vice-President, and Martine Dréneau-Fénerol, General Secretary of the Constances association
- Mr. Gérard Raymond, Vice-President of the Health Data Hub, Ms. Caroline Guillot, Deputy Director of Association and Citizen Relations, Mr. Alexandre Romainville, Department of Association and Citizen Relations
- Ms. Laure Maillant, Director of Innovation and Data of the Paris Public Hospitals Group (AP-HP) Information Systems Division
- Prof. Marc Cuggia, University Professor and Hospital Practitioner (PU-PH) in biostatistics and medical informatics in Rennes. Member of the Inserm/University of Rennes 1 Joint Research Unit

(UMR) 1099 Signal and Image Processing Laboratory), Leader of the Health Big Data Team. Coordinator of LabCom LITIS and of the Grand Ouest Interregional Network of Clinical Data Centers (GCS HUGO).

[i] The description continues: "It designates the Data Controllers, defines their role and missions. It also modifies the composition of the list of organizations, establishments and services having permanent access to National Health Data System data due to their public service missions. It specifies the rules applicable to this permanent access. It envisages the procedures for exercising the rights of data subjects and in particular the conditions for informing those to whom the data relate."

[ii] Zins M, Cuggia M, Goldberg M. Les données de santé en France : abondantes mais complexes. Méd Sci (Paris) 2021 ; 37 : 179-84.

[iii] Chevreul K, Delpierre C, Dourgnon P, et al. Les données de santé, un patrimoine commun qui doit servir à améliorer le bien-être de tous. Le Monde Idées, October 30, 2020, p27.

[iv] Chneiweiss H. Big data et santé : questions éthiques p200-201 in Big Data à l'échelle de la société. Rémi Mosseri ed. Presses du CNRS 2016.

[v] Deliberation no. 2021-067 of June 7, 2021 giving an opinion on the draft decree implementing section II of article 1 of law no. 2021-689 of May 31, 2021 on the management of the exit from the health crisis (opinion request no. 21010600).

[vi] Decree no. 2021-848 of June 29, 2021 on personal data processing through the National Health Data System. Official Journal of the French Republic (JORF) no. 0150 of June 30, 2021.

"Concerning: people whose data are collected as part of prevention, diagnosis, medical care or medical-social follow-up activities or from health surveys and which are added to the National Health Data System; public and private organizations tasked with conducting projects for the purposes of research, study and evaluation in the health domain.

Subject: modes of National Health Data System implementation.

Entry into force: the text enters into force on the day after its publication.

Description: the decree envisages the modes of governance and operation of the National Health Data System, the scope of which is extended to new databases. It designates the Data Controllers, defines their role and missions. It also modifies the composition of the list of organizations, establishments and services having permanent access to National Health Data System data due to their public service missions. It specifies the rules applicable to this permanent access. It envisages the procedures for exercising the rights of data subjects and in particular the conditions for informing those to whom the data relate.

References: the provisions of the decree are made pursuant to law no. 2019-774 of July 24, 2019 relating to the organization and transformation of the health system

[vii] "In addition, it is intended to integrate the majority of these data into a 'centralized SNDS' composed of a primary database (currently including the 'historical SNDS' and which may be enriched in the future) and a database catalog including other databases considered relevant for research players. This integration will involve the migration of data. Initially envisaged as a decentralized system, the Commission notes that the choice of the ministry is finally oriented towards centralization of the SNDS data. The Commission takes cognizance that this draft decree aims to initiate this centralization with the French National Health Insurance Fund (CNAM) and the HDH and to govern only the implementation of this 'centralized SNDS.'"

[viii] Even the existence abroad of contracts for the provision of public health service data to private companies. Lemke C. Ma santé, mes données. Premier parallèle, 2021, 171p.

[ix] "Nvidia is commissioning its new supercomputer in the UK, the most powerful in the country, with five health sector partners: AstraZeneca and GSK, Guy's and St Thomas' NHS Foundation Trust, King's College London, and Oxford Nanopore. With a computing power of 8 petaflops, it will be used for various research projects, including the development of a deeper understanding of brain diseases, such as dementia. But also to reinforce the use of AI to design new drugs and improve the precision of the search for genetic variations causing diseases in humans. 'Cambridge-1 will empower researchers (…) with the ability to perform their life's work– unlocking clues to disease and treatments– at a scale and speed that was previously impossible,' declared Jensen Huang, Founder and CEO of Nvidia. AstraZeneca, for its part, wants to accelerate its work on the use of AI in digital pathology [?], which involves the capture, management, analysis and interpretation of digital information. Whereas GSK hopes to better predict human health and develop better drugs, more apt to show positive results in clinical

trials." ("Nvidia démarre son supercalculateur au Royaume-Uni, avec des projets en santé." L'Usine Nouvelle - July 7, 2021, reported in Pharmaceutiques, July 8, 2021)

[x] APMNews July 19, 2021: "Creation of the French Alliance for Real-World Data to build a bridge between industry and the Health Data Hub". "Launched on Tuesday on the initiative of consulting firm Kynapse and the 'AI for Health' think tank, the French Alliance for Real-World Data will welcome '5 or 6 pharmas' from September to 'facilitate and accelerate real-world data research projects' and promote collaborations between the Health Data Hub (HDH) and industry, explained Kynapse CEO Stéphane Messika Thursday afternoon to APMnews. "

[xi] "AstraZeneca, Docaposte – the digital subsidiary of La Poste, and Impact Healthcare – a company specializing in innovation consulting, are launching Agoria Santé: a platform to collect and analyze health data for better patient care. 'The objective is to provide a legal, ethical and secure framework for health care players, enabling them to accelerate research,' explains Olivier Nataf, head of AstraZeneca's French subsidiary, to Figaro. The catalog of data and services offered should be enriched over time, and partnerships established around shared themes. Five pharmaceutical companies and some ten other players in the sector (hospitals, universities, etc.) are already in discussions to join the initiative. Agoria will also operate as a trading platform, offering paying services to hospitals, universities, and pharmaceutical companies. Its founders, who decline to communicate the amount of their investments or the prices of the services, hope in this way to make the platform profitable. The new members of the consortium will have to pay to join. And the 'users' will pay according to the services used. According to Docaposte CEO Olivier Vallet, the aim is also 'to accelerate the use of digital technology and Artificial Intelligence to make France a leader. The health crisis has only accelerated awareness.'" (AstraZeneca, Docaposte et Impact Healthcare s'unissent pour l'accès aux données de santé. Le Figaro - June 18, 2021. Cited by Pharmaceutiques on July 6, 2021).

[xii] Interview with Frédéric Dufaux, Deputy CEO of Docaposte in charge of health (TechmedInfo, July 5, 2021).
Docaposte is strengthening its digital health footprint with the launch of Agoria Santé, in partnership with AstraZeneca and Impact Healthcare. A new platform dedicated to companies in order to carry out and safeguard their research projects on health data. What is Docaposte's current role in the digital transition in health?
Docaposte is a Health Data Host (HDH), with the particularity of being certified in the six domains of activity, making it possible to include an application dimension in our offering. Our strategy centers around two major orientations: the structuring, collection, and analysis of data at the service of industry, including the medtechs and pharmaceutical companies, whilst securing processes, including the use of algorithms. It is not about judging their scientific relevance, but about ensuring that the data processing meets regulatory requirements in terms of its form, and that it will be auditable. This is the added value we represent in relation to a more conventional host. Our second orientation is to act as a health data operator to facilitate exchanges and the interconnectivity of systems between operators. A typical example is our work on the Pharmaceutical Record (DP) for the French National Chamber of Pharmacists, or with the Elsan private hospitals group, which we support on its virtual assistant, Adel.

[xiii] A leader in France in pharmacy management software, the international company Iqvia has attracted the attention of both the media and the CNIL: "05/17/2021 – The operation of the data warehouse of the company IQVIA authorized in 2018 was called into question in Cash Investigation, a TV program that will be broadcast on May 20. The CNIL specifies that to date it has not received any complaints relating to the operation of this warehouse but announces, in light of the elements brought to the attention of the public, that it will carry out checks. "

[xiv] We must therefore be wary of a misleading resemblance to the inductive reasoning implemented in the social sciences. If the sociologist "constructs his or her object" from the observation, 1) this observation is conducted methodically, with the sociologist having to "maintain control of his or her questions by keeping a distance from the 'prenotions' of common sense or current controversies"; 2) ("based on the Weberian principle that individuals are the 'elementary atoms' of a society, one of the goals of sociology is to analyze the relationships between them" (…) by taking into account "the various criteria that play a role in the activities and bind the individuals to each other". "There is no universal theory that would explain how [the] variables are articulated ... we cannot just align the variables assuming that they have the same weight, a common mistake among statisticians." Extract from Race et sciences sociales, Essai sur les usages publics d'une catégorie, Beaud S and Noiriel G, Agone Paris 2021 [pp 188-9].

[xv] Especially at a time when the Health Data Hub is launching a call for expressions of interest centered around targeting algorithms, as part of the Open Library of Health Algorithms (BOAS) project.

[xvi] Eric Sadin. L'intelligence artificielle ou l'enjeu du siècle. Ed L'échappée, Paris, 2021, 298p.

[xvii] Interview by Laure Belot as part of an article published in the Sciences et Médecine section of Le Monde, October 28, 2020.

[xviii] The CNIL deliberation dated June 30, 2021 emphasizes this: "The Commission notes, however, that an organization responsible for processing a source database that supplies the principal database or database catalog may continue to make the data from the source database available to other Data Controllers (for example, the Technical Agency for Information on Hospitalization (ATIH) makes PMSI data available to other Data Controllers or a university hospital makes hospital warehouse data available to a company specializing in Artificial Intelligence). It notes that this provision will be governed by the provisions of the French Public Health Code (CSP) (prohibition of the pursuit of forbidden purposes, compliance with the SNDS security reference framework, etc.). It takes cognizance of the clarifications provided by the ministry according to which the organization is responsible for the processing of its source database as long as it processes the data and until the data are processed to supply the 'centralized SNDS'. "

[xix] But the field of reflection stops there, the group does not address the issue of the security of mobile applications, despite this being problematic. See for example Data sharing practices of medicines related apps and the mobile ecosystem: traffic, content, and network analysis. Grundy Q, Chiu K, Held F, Continella A, Bero L, Holz R. BMJ 2019;364:l920

[xx] Emphasized in the CNIL deliberation of June 30, 2021: "The ministry has confirmed that the HDH will have a copy of the principal database, for the efficient response to requests and in particular to make ad hoc matches between the principal database and the catalog. Without calling into question this operational necessity, the Commission is concerned about the duplication of a database that by nature contains sensitive data covering the entire population. Duplication which involves the regular transfer of a large volume of data between the French National Health Insurance Fund (CNAM) and the HDH, as well as the sharing of pseudonymized identifiers; in addition, the Commission reiterates that the HDH does not have - unlike the CNAM - its own data centers and uses a service provider in a data center that is shared with several customers. It recalls that these different operations automatically increase the surface of attack and the risk of violation of these data. "

[xxi] Goldberg M, Zins M. La plateforme « Health Data Hub » pose des questions de sécurité majeures. Le Monde Idées, October 30, 2020, p27.

[xxii] Banck A. RGPD : la protection des données à caractère personnel. 19 fiches pour réussir et maintenir votre conformité. 3rd edition, Gualino, Paris 2020, 79p.

[xxiii] According to the CNIL, "The processing of personal data within the scope of research is implemented under the responsibility of the Data Controller, and/or with third parties acting on his or her behalf. The Data Controller must conduct a data protection impact assessment, which must particularly cover the risks to the data subjects' rights and freedoms. He or she implements the appropriate technical and organizational measures to guarantee a level of security in line with the risks identified. The one same analysis can cover a set of similar processing operations. The Data Controller must implement and monitor the application of a policy of security and confidentiality pursuant to the reference methodology.

[xxiv] Article 4 of the GDPR defines them as follows: "'Personal data' means any information relating to an identified or identifiable natural person [...], directly or indirectly, [...] by reference to [...] an identification number [...] or to one or more factors specific to [...] that natural person" (in practice, identification that is direct, indirect, or by cross-checking).

[xxv] The same article 4 of the GDPR specifies: "'processing' means any operation [...] which is performed on personal data [...], such as collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction".

[xxvi] "Digital data are taking on an ever more central role in biomedicine today. But these data are increasingly generated outside the traditional spaces of the medical system, as individuals go about their daily lives interacting with consumer devices. Moreover, the technological tools needed to produce, store and analyze these data increasingly lie beyond the remit of traditional medical scientists. In other words, the health data ecosystem is expanding, to include new types of data, new methods for capturing and analyzing them, and new stakeholders. Pressing questions emerge concerning privacy, informed consent, the commodification of personal health data, and the drawing of new power asymmetries between data subjects and Data Controllers, the public and the private sector. At the same time, concepts and values that previously acted as normative anchor points, such as "solidarity", the "public" or the "common good", are destabilized, re-conceptualized, and mobilized in new ways. This special theme addresses the question of how the expansion and decentralization of the health data ecosystem disrupts existing norms and frameworks of data ethics and data governance, and what kinds of re-thinking of ethics and governance this solicits. The collection of articles and commentaries provides a combination of conceptual and practice-based reflection." Presentation of the Health Data Ecosystem section of the journal Big Data and Society by its editors, Tamar Sharon, Associate Professor, Radboud University and Federica Lucivero, Senior Researcher, University of Oxford (accessed online on June 18, 2021).

[xxvii] "Digital data are taking on an ever more central role in biomedicine today. But these data are increasingly generated outside the traditional spaces of the medical system, as individuals go about their daily lives interacting with consumer devices. Moreover, the technological tools needed to produce, store and analyze these data increasingly lie beyond the remit of traditional medical scientists. In other words, the health data ecosystem is expanding, to include new types of data, new methods for capturing and analyzing them, and new stakeholders. Pressing questions emerge concerning privacy, informed consent, the commodification of personal health data, and the drawing of new power asymmetries between data subjects and Data Controllers, the public and the private sector. At the same time, concepts and values that previously acted as normative anchor points, such as "solidarity", the "public" or the "common good", are destabilized, re-conceptualized, and mobilized in new ways. This special theme addresses the question of how the expansion and decentralization of the health data ecosystem disrupts existing norms and frameworks of data ethics and data governance, and what kinds of re-thinking of ethics and governance this solicits. The collection of articles and commentaries provides a combination of conceptual and practice-based reflection." Presentation of the Health Data Ecosystem section of the journal Big Data and Society by its editors, Tamar Sharon, Associate Professor, Radboud University and Federica Lucivero, Senior Researcher, University of Oxford (accessed online on June 18, 2021).

[xxviii] Amiel P, Dosquet C, Inserm Ethics Evaluation Committee (CEEI). Guide de qualification des recherches en santé. Inserm, 2021.

[xxix] See also Astruc A, Jouannin A, Lootvoet E, Bonnet T, Chevallier F. Les données à caractère personnel : quelles formalités réglementaires pour les travaux de recherche en médecine générale ? Exercer 2021 ; no. 172 : 178 – 184.

[xxx] Traitements de données de santé : comment faire la distinction entre un entrepôt et une recherche et quelles conséquences ? November 28, 2019 Consulted on the CNIL website on October 2, 2021.

[xxxi] CNIL deliberation of June 30, 2021: "The Commission notes that, despite the scope of the processing, both in terms of data sensitivity and volume, the draft decree does not envisage the individual information of the data subjects. Moreover, taking cognizance that the information will almost exclusively be provided in a dematerialized way (websites, French National Health Insurance personal 'Ameli' account), the Commission asks the Ministry to consider additional alternative methods (poster or media information campaigns, provision of information notices in National Health Insurance offices, transmission of a comprehensive information notice if requested by data subjects, etc.). As for the 30% of beneficiaries who do not have an 'Ameli' account, the Commission requests that comprehensive individual information be mailed to them, for example, when sending the statements of reimbursement. "

[xxxii] Extract from the report on the interviews with the Constances association representatives and the Constance cohort scientific managers held on April 12, 2021.