OPINION 130

BIG DATA AND HEALTH:

A NEW APPROACH TO THE ETHICAL ISSUES

COMITÉ CONSULTATIF NATIONAL D'ÉTHIQUE
POUR LES SCIENCES DE LA VIE ET DE LA SANTÉ

BIG DATA AND HEALTH:

A NEW APPROACH TO THE ETHICAL ISSUES

Opinion made public on 29 May 2019

TABLE OF CONTENTS

SUMMARY

Continuing previous reflection on the questions posed by the collection of personal data and the new possibilities of their use in the life sciences and health, and following the *États généraux de la bioéthique*[1] (Bioethics Forum), the CCNE in its Opinion 130 considers the ethical issues raised by the exploitation of big data[2]. In the context of accelerating technological and cultural changes linked to the treatment of big data, the CCNE points out how the massive accumulation of data on people and the value created by the increased capacity for processing these data call for debate and ethical reflection.

This opinion:
- sets out the ethical principles *common* to all contexts of the exploitation of big data in health (section 2);
- identifies the ethical challenges *specific* to the use of big data in healthcare, research, management of care, and personal life (section 3);
- proposes 12 recommendations that are essential for the respect of fundamental ethical principles and that enable, but do not hinder, the growth of technologies based on big data.

Changes induced by big data

The CCNE considers that we cannot adopt a position hostile to these new digital technologies just because they carry risks, as it would be unethical not to promote their development if they can benefit the health of everyone and help rationalize costs. But the CCNE reiterates that the main principles that underpin medical ethics – respect for the person (including the respect of his/her independence), justice, relevance, and benevolence (including here the obligation: do no harm) – should not be weakened by the growth of digital technologies.

[1] See CCNE Opinion 129 "Contribution of the Comité consultatif national d'éthique to the revision of the bioethics law", and the report "Digital technologies and health. Which ethical issues for which regulations?" commissioned by the CCNE on artificial intelligence.

[2] Big data means the availability either of a large amount of data or of data of large size that can only be treated effectively by digital tools combining algorithms with great computing power. The *change in scale* is such that only machines, and no longer humans, are able to collect, store, and analyze data, and what is more these data are characterized mainly by three properties: their *permanence* (they can be copied and reused indefinitely); their *dissemination* in time and space, which enables their rapid and borderless sharing; the *generation of secondary data, ie, new information* obtained by the processing and cross-referencing the initial data with other sources, which makes these data usable well beyond the purposes of the initial collection.

The transformations produced by big data modify our way of formulating research hypotheses, acquiring knowledge, and using these to take decisions affecting the individual and the community. They alter our representations and overturn our way of considering measurements of parameters relating to our health and well-being. These personal data tend to become a dynamic relational tool and an instrument of autonomy intended to give us better personal control over our state of health.

The CCNE cautions that people who do not have easy access to digital technologies should also benefit from advances in healthcare and should not be stigmatized or discriminated against in their access to care (***Recommendation 9***).

In view of these transformations, the CCNE makes two observations:

- The first is that any primary data stemming from a human activity – even if apparently not health-related – can, through cross-referencing with other unrelated data, contribute to the creation of new information on an individual's health. The term health data can no longer be limited to personal data collected within the framework of medical care (laboratory tests, genomic characteristics, clinical data, etc.).

- The second is that ethical questions arise against a backdrop that is constantly changing because of the extreme pace of technological innovations and that is complicated by the diversity of situations in which health-related data are collected and processed. This necessitates constant watchfulness and periodic evaluation of the effective implementation of protective measures (***Recommendation 2***). Ethical considerations should take into account the fact that the purpose of some innovations may not be medical care, but rather the exploitation of a market that is presented as concerning well-being.

A new perception of the ethical issues and appropriate responses

The CCNE shows that one of the characteristics of big data in health is the blurring of the distinctions underpinning implementation of the ethical principles that promote the protection of individual rights in health. The separation between *private life and public life* is blurred by the possibility of cross-referencing unrelated data, but also because our perception of "privacy" is changing. The *relation between the individual and the collective* is evolving: the autonomy of the individual is growing, but precise knowledge of individuals and of their state of health creates a risk of profiling, which threatens the protection of private life and may lead to stigmatization of people or groups. Such stigmatization threatens private life, but also the principles of solidarity and equity which are the basis of our health system. *Care and business* are becoming increasingly hard to distinguish, as a result of the transformation of care and of the healthcare market.

The CCNE notes, furthermore, that the very notion of free and informed consent, which is required to protect a person subject to a decision concerning him/her, is called into question by the conditions of exploitation of big data (uncertain aims and data processing that is not understood or even inaccessible). The need for protection of the individual must be reaffirmed and its modalities redefined, to dispel the threat of a society under the surveillance and control of multiple providers acting for various purposes. Although consent remains a cornerstone of the legality of the processing of personal data, the GDPR[3]¶, which means to be firm but practical in terms of its principles, notes that this requirement is not achievable in all cases, notably in the scenario of reuse of data. The GDPR recognizes as legitimate other methods of protection when the aims are for the public good. The CCNE takes note of this evolution from the individual's *a priori* wish for exhaustive control to an *a posteriori* logic of intervention and control based on a quest for intelligibility and empowerment. This logic demands loyal behavior by those in charge of data processing, transparency in their processes, and the ability to check their possibilities of access to data and their ethical approach. The CCNE reaffirms the importance of a governance that can be clearly identified and of engagements that can be verified. Precise and fair information adapted to different contexts of use becomes a major ethical criterion ***(Recommendations 1 to 3)***.

The CCNE considers that a "human guarantee" of the different steps of the process of data analysis is fundamental. These data are in fact the "raw material" needed for the design of algorithms that aid in decision-making, which are playing an ever-greater role in medical practice and in defining health policies. A human guarantee is therefore essential in achieving methodological rigor during the different steps of data collection and processing: (i) the quality and adequacy of the data selected to train the algorithms; (ii) the appropriateness of the choice of algorithmic treatments to the question posed; (iii) the verification, using an independent set of data, of the robustness and accuracy of the result given by the algorithm. Implementation of the human guarantee requires strong actions in three areas: training, qualitative evaluation of websites, applications, and connected objects, and high-level scientific research ***(Recommendations 4, 5, 6, 8)***.

Context-specific ethical questions

To the diversity of sources of primary data corresponds an equally broad diversity of areas for their use in health (care, design of new drugs or medical devices, enhanced understanding, management of clinical trials, better economic performance, improvement in public health, business/commercial objectives). Ethical issues specific to three of these situations are as follows:

- The CCNE reiterates that the care relationship is based on a direct human relation, built on confidence and a set of decisions truly shared between the doctor and the

[3] General Data Protection Regulation, is a regulation in European law that came into effect in the European Union on 25 May 2018.

patient, even if the computerization of health systems is now generalized. Three ethical principles can be weakened by the exploitation of big data: *medical confidentiality*, by the multiplication of information shared and exchanged between various stakeholders, some of whom are not from the medical community; the *responsibility of the medical decision,* by the risk automation creates because of the proliferation of algorithmic software; the *personal relationship* between the doctor and patient, which may be impoverished by innovations foreseen in the treatment of big data, the risk being that the patient is reduced to a data set to be interpreted, seemingly making it unnecessary to listen to what the patient has to say. The CCNE reiterates that the digital technologies must remain decision-making aids and considers that the time thus saved should be devoted to listening to and discussing with the patient, and to the taking into account of the patient's personal circumstances (***Recommendation 7***).

- In the case of research protocols, the CCNE notes that the main ethical concern is to find a good balance between the risk that underexploitation of data will limit research conducted for the public good and the risk of inadequately controlled and unrestricted sharing of data, thus calling into question the basic rights of the individual. The CCNE emphasizes that each stakeholder involved perceives the ethical questions that arise differently. The holder of the data, for example, will have little precise information on the research at the time of collection and will not necessarily benefit personally from the results. The confidence accorded, in this context, is based above all on the process of governance and on the way in which access to data is controlled. The quality of information delivered to the holder of the data is a major criterion of this quality of governance ***(Recommendations 11, 12)***.

  The CCNE analyzes more specifically the case of genomic data : these are distinguished, among other things, by the characteristics of the genome sequence that identify the individual and his/her kin, and allow prediction, and by questions related to the constitution of large banks of genomic data and the sharing of these data in research While their exploitation has enabled spectacular advances in understanding and a notable improvement in the treatment of patients, genomic data also illustrate the risks raised by the exploitation of big data: uncontrolled dissemination, possibility of identification, and loss of confidentiality and hence of security. The CCNE recommends the maintenance of specific regulations and, in particular, the need for express consent and the taking into account of relatives. It highlights the increased risk of incidental discoveries and the accompanying ethical problems and encourages vigilance regarding possible bias in the selection of study populations (***Recommendation 4***).

- ***Social media, applications and connected objects, and internet platforms*** for sharing health information intended for patients have become a very important source of data that is precious notably for medical follow-up, pharmacovigilance, research, and policies for prevention or public health surveillance. However, when data are collected outside the care pathway, the dissemination and use of these real-life data weaken the protection of patients, notably their right to respect, fair information, the limits of their consent, and the dissemination, hosting, and reuse of these potentially health-related data.

Lastly, the CCNE notes that the digital technologies demonstrate that, for a country to keep control of its health policy and its capacity for scientific and medical innovation, there is a need to confront the technological challenges of data storage and security, and to ensure a high scientific and technological level in data exploitation ***(Recommendation 10)***.

INTRODUCTION

On 25 January 2017, the Minister for Social Affairs and Health asked the CCNE for an opinion prompted by the development of precision medicine and the increasingly frequent and common exploitation of big data. Noting that big data offer major opportunities to improve the quality and safety of care, the minister pointed out that they nonetheless raise ethical questions relating, in particular, to how the people concerned are given information and how they give their consent, or exercise their right of opposition, depending on the use to which the data generated are to be put. There is also a need to protect these data. The minister reflected upon the search for a fair balance between, on the one hand, the opportunities for development, for the benefit of the community, patients, and their relatives, and, on the other hand, the need to protect citizens' private lives. This reflection is part of the general debate on the major ethical challenges raised by the collection and treatment of big data in healthcare.

The CCNE was already interested in the questions posed by the collection and possible new uses of personal data in the life sciences and health[4]. CCNE Opinion 124 of 21 January 2016, which relates specifically to genetics, focuses on problems posed by the information contained in the genome sequence, as well as consent procedures, notions that are at the heart of our study[5].

[4] Opinion 46 of 30 October 1995: Genetics and medicine: from prediction to prevention.

Opinion 77 of 20 March 2003: Ethical issues raised by collections of biological material and associated information data: "biobanks", "biolibraries".

Opinion 91 of 16 February 2006: Ethical issues arising out of computerized hospital prescriptions and patient records.

Opinion 98 of 26 April 2007: Biometrics, identifying data and human rights.

Opinion 104 of 29 May 2008: The "Personal Medical Record" computerization of health-related data.

Opinion 116 of 23 February 2012: Ethical issues arising out of functional neuroimaging.

Opinion 124 of 21 January 2016: Ethical reflection on developments in genetic testing in connection with very high throughput human DNA sequencing.

[5] In assessing the challenge of the management of data (pages 19 and 20), Opinion 124 noted that reflection on the management of big data is increasingly prevalent internationally. The Council of Europe, for example, considers ethical questions posed by big data in medicine, a subject that goes far beyond genetic data from high-throughput sequencing of human DNA. One of the characteristics of this trend is that the main providers are large groups (Google, Amazon, Facebook, and Apple, for example), which have no tradition of working with doctors and medical technologists. Another specificity is that the power required for the analysis and storage of these data selects a small number of companies that are the only ones able to provide this power, thereby creating a concentration of power that may appear hegemonic, and a form of appropriation of these data that is, in fact, in contradiction with the foundation and justification of these analyses of large amounts of health data, to wit, the free sharing and open access of information. This question is at the heart of the concerns of the CCNE and was taken up in an independent study by the committee for ethical issues related to the use of health data, beyond the questions of genetics considered in Opinion 124.

The digital revolution extends to all fields and, with economics and the environment, shares the characteristic of being a transdisciplinary expertise which informs the whole field of health. The CCNE has considered the ethical issues of this digital revolution, as witnessed by the digital technology and health report commissioned by the president of the CCNE and published in November 2018[6]. The present opinion continues this work by delving deeper into the specific question of the exploitation of big data in health. These two documents form part of a wider reflection by the CCNE on bioethical questions raised by the transformations of the health system, characterized by "*a tension between great technical know-how and fundamental issues that affect all human beings and in the representation they give for themselves and their species"*[7] *(*see the synthesis report on the Bioethics Forum, and Opinion 129[8]).

But why the particular interest in big data? Big data are certainly an essential component of digital science and technology and, in particular, of machine learning, robotics, and new means of communication. They are also part of the digital revolution that is shaking our society to its very roots. This interest is justified by the specificity of big data, which certainly deserves particular study.

The term big data is now part of everyday language in English, but although often translated in French by "données massives", there is no universally recognized definition. There are numerous meanings, which depend on the area of use. In general, big data designates the availability either of a large number of data, or of large data sets, which only digital tools ranging from algorithms to the calculating power of computers can treat effectively. The information that the data initially contain, whatever their nature (scientific, administrative, personal, or other), is considerably enriched by cross-referencing data. Highly diverse, these data can be related to health or well-being. They may stem from laboratory tests, notably data from genomics, proteomics[9], or metabolomics studies[10], they may be clinical, environmental, or behavioral data from large cohorts collected occasionally or repeatedly from medical records, but also from the individuals themselves, via social media and other means of communication. The originality of the analysis of big data is that it is not necessarily based on pre-existing structures and

[6] "Digital technologies and health. Which ethical issues for which regulations?", a working group report commissioned by the CCNE, in conjunction with CERNA (a commission set up by Allistene, the Digital Sciences and Technologies Alliance, to consider the ethics of research in the digital sciences and technologies).

[7] « Une tension entre une grande technicité et des enjeux fondamentaux qui touchent chaque être humain dans la représentation qu'il a de lui-même et de son espèce ».

[8] The Bioethics Forum (*États généraux*) was held in the first half of 2018 and all the studies, arguments, and opinions were included in a synthesis report published in June 2018. Furthermore, the CCNE published in its Opinion 129 a "Contribution of the Comité consultatif national d'éthique to the revision of the bioethics law 2018-2019".

[9] Study of all the proteins of an organism, or cell.

[10] Analysis of the metabolites present in the organism and resulting from the biological processes occurring during metabolism.

enables the discovery within collected data of correlations, even causalities, using a specific algorithm.

Digital technologies, based on rapidly developing scientific innovations, already govern many aspects of our daily lives (information and documentation, localization, communication, business and financial transactions, management of industrial processes, decision-making aids). They induce profound changes relating essentially to two points:

- the management and communication of data: once collected, data can be reproduced endlessly without loss of quality, stored in repositories[11] or platforms, disclosed and used widely, including for purposes that differ from those for which they were initially provided or captured;

- decision making: the algorithmic treatment extracts from a mass of data information that appears to be of sufficient precision to guide or even replace human reasoning.

Rapid advances in the digital sciences and technologies have already led to important innovations in the care of patients and the organization of the health system. This will be even truer tomorrow, because this evolution is major and irreversible.

But these new, disruptive technologies raise specific ethical questions:

A/ The information given to the person and the person's consent to the collection and use of personal data are complicated once the data can be easily duplicated and reused for purposes not initially defined and their processing reveals new so-called secondary data, which are often more sensitive than the initial data.

B/ The analysis relates to a large amount of data that in general has not been fixed beforehand. Cross-referencing of data often allows precise identification of people and efforts at anonymization of the initial data may therefore no longer be a sufficient guarantee of the protection of human rights[12]. This sheds new light on the ethical issues associated with this protection, but also on the issue of solidarity. This is because precise knowledge of the predispositions of each individual could lead to individualized risk management and threaten the principle of sharing that underpins our system of social protection.

[11] A rich terminology designates data storage devices: warehouses, repositories, libraries, lakes, bases, banks.

[12] Article 29. Data protection working party on the protection of individuals with regard to the processing of personal data. Opinion 05/2014 on anonymization techniques (0829/14/EN WP216). April 2014. https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf

C/ The care relationship and medical care of the person are conventionally based on a human relationship rooted in trust, listening, observation, the use of physiological indicators, and experience. This relationship will be profoundly altered by the role taken in this approach by decision making based on algorithmic exploitation of data.

D/ Beyond the care relationship, the parameters that influence health and well-being become measurable by means of connected devices and applications. Patients are therefore increasingly urged to check their own physical state and to join personalized health networks so as to understand the results of medical research on their disease[13]. As a result, new stakeholders become involved, some of whom can exploit for commercial gain the health and wellness market.

These questions were addressed during the 2018 Bioethics Forum organized by the CCNE prior to revision of the bioethics laws and are dealt with on pages 73 to 84 of the CCNE's synthesis report[14]. Three concerns were expressed repeatedly:

- the need for explanations and information on the digital data path and the use of the data collected;
- the fear that the development of digital tools will diminish the special relationship between patient and doctor and in time lead to the risk that the medical decision is impersonally imposed by the digital tool;
- a shared mistrust regarding the fate of the data and the risk of their malicious use, in particular concerning vulnerable people.

The CCNE's reflection focused on analysis of new ethical issues arising from rapid technological changes in the exploitation of big data, both with regard to individuals and to the values on which our health system is based. The responses proposed to preserve respect of these values should not deprive us of the progress that emanates from the exploitation of big data in healthcare and research. As the CCNE emphasizes in its Opinion 129, insufficient use of digital means in patient care, research, or the development of data-driven guidance leads, on a large scale, to unethical situations in our health system.

1. HEALTH DATA IN THE ERA OF BIG DATA

1.1 Disruption

Since the advent of clinical and laboratory medicine at the end of the Middle Ages, carers have collected data on patients, without their explicit consent. The document used

[13] See notably the websites PatientsLikeMe and Carenity.

[14] Synthesis report of the *États généraux de la bioéthique*. June 2018. https://www.ccne-ethique.fr/fr/publications/rapport-des-etats-generaux-de-la-bioethique-2018).

to record findings was an "observation" and formed the basis of the medical record. This observation, the property of the institution (and not of the person), was the basic tool used to make progress in the identification of morbid situations through comparisons with similar observations.

Since the second half of the 19th century, practitioners have sought to use clinical signs and symptoms and laboratory parameters to predict the diagnosis, and even to predict the success of a prescribed treatment.

Over time, these clinical signs benefited from Bayesian statistics (long unrecognized because it dates from the mid-17th century) in terms of adding predictive probabilities. Bayesian approaches in clinical medicine are recent[15] and have considerably improved diagnosis and even therapeutic precision. They long preceded the massive data collections that human intelligence can no longer process directly.

A major innovation stems from the fact that in France, for fifteen years now[16], medical records have been the patient's property. The use of medical records for collective purposes (medical research or medical decision making) has therefore become dependent on the explicit and informed consent of the holders of the medical records, which introduces an important change in logic in the use of the data collected in the framework of healthcare.

Moreover, the digitization of clinical and paraclinical information collected in healthcare establishments (observational data and medical records including the results of clinical exams, lab tests, static and dynamic imaging, and administrative medical information) has enabled the constitution of huge collections of data. These large datasets have appeared in public institutions (large university hospitals and even private hospitals in the English-speaking world, sickness insurance funds, national medical databases [SNIIRAM], in France) and also in private institutions that deliver care to the public sector (notably, cancer treatment centers).

Big data are disruptive by virtue of their four main characteristics:

[15] Bayesian theory provides a mathematical model of the optimal way of plausible reasoning in the presence of uncertainty. Bayes' theorem indicates how to combine optimally the *a priori* information stemming from our evolution or from our memory with data from the outside world (source: Stanislas Dehaene, courses at the Collège de France). Bayesian inference is an inductive approach to calculating the probability of a hypothesis from *a priori* knowledge in the form of probability measures. In Bayesian reasoning, in contrast to classical statistics, the parameters are considered as random variables to which a probability density can be ascribed. The probability is a measure of the degree of belief (or of confidence) in the occurrence of an event or in the truthfulness of a proposition. It is the digital expression of knowledge; it measures a degree of certainty in the truth of a hypothesis.

[16] This innovation dates from law No. 2002-303 of 4 March 2002, also known as the Kouchner law, pertaining to the rights of patients and to the quality of the health system.

- a change of scale, because of the considerable increase in the data available[17] and in our capacity to analyze them;
- their sustainability: data are not destroyed by use and can therefore be reused;
- their rapid dissemination, which allows their sharing and their use beyond medical teams and national borders;
- their capacity to generate new information (secondary data) and new hypotheses, by means of their processing.

This technological disruption greatly modifies the way questions of health are addressed and forces us to re-examine how we consider the ethical concerns raised over the last thirty years by the introduction and growth of digital techniques in healthcare, for the following reasons:

- All primary information taken in isolation and apparently inconsequential must now be considered as being able to contribute to information on health. There is the possibility, which is fundamentally new, of using and cross-referencing data in a different context with other unrelated data. From this emerges new information (secondary or deduced data), which may be personal data that is sensitive and identifying. These data are saved and can be used unbeknownst to the initial holders, while they can be used for the benefit of the holders, but also possibly for harm. This impossibility of defining *a priori* health data is worsened by the massive and unselected nature of data collection, which leads to the storage of vast datasets which give rise to dynamic and constantly updated analyses that can be consulted by researchers, physicians, or even other stakeholders.

- Digital technology abolishes the particularity of the data collected and allows their connection, whatever the carrier (writing, sound, image), area (health or other), or source (social media, mobile apps, computer files). It therefore becomes harder to characterize a finding as sensitive and so to provide specific protection.

- The very notion of health today is defined broadly and extends to well-being, as recommended by the WHO (World Health Organization), but also takes into account environmental and social factors, as well as lifestyle data, which can be collected in real time using connected objects.

- Primary data are collected in multiple situations: they are provided by people themselves (social media, mobile applications, geolocalization, connected objects, digital payments), whether or not they are aware of this; data collected in

[17] These data today represent terabytes (10^{12}) accumulated each day, and therefore petabytes (10^{15}) if we consider the sum of the data. The total quantity of data collected in the world doubles every 2 or 3 years.

the framework of healthcare (hospital databases, shared medical records, the future "personal digital health space"[18]), or by researchers to enrich databases, registries, and cohorts; to which should be added data collected by the medical administration health bodies that constitute the national health data system (SNDS)[19].

All of the above shows that the primary data form a "raw material" that providers of extremely varied origins, training, and motivations seek to use. The providers may be pursuing a health objective (health professionals), the design of new drugs or devices (pharmaceutical firms, start-ups), improvement of knowledge (researchers), business (internet start-ups and multinationals), improved economic performance or public health (public institutions). These highly diverse stakeholders do not share the same professional culture or the same values, notably the protection of private life, and this compounds the problem.

To the disruptions generated by the generalization of the collection, storage, and treatment of big data is added an environmental challenge that results from these operations and concerns several issues:

- the high energy consumption of computers, data centers, and networks, which together account for 10% of the world's electricity consumption, a figure that is rising steadily;
- the large carbon footprint of functioning infrastructures;
- the consumption of rare metals needed to make computers and smartphones; they are extracted using techniques that are destructive and use environmentally harmful products and, after use, produce waste a large part of which is found in uncontrolled landfills in Asia or Africa.

This is an important subject, which will not be developed further in the present opinion because it is not specifically health-related, even if the consumption related to health is certainly great. But it is certain that, in this field where the availability and continuity of care are of primordial importance, we cannot lastingly trust these new technologies if we are unsure of being able to control the conditions of their functioning. This control requires sufficient energy resources (whether by marked reduction in consumption or

[18] “espace numérique de santé”. See the final report (October 2018) on the strategy for transformation of the French health system. "Accelerate the digital revolution" (in French). The aim is to create from birth for each user a secure, personalized digital health space enabling the user to access all his/her health data and services throughout life. https://solidarites-sante.gouv.fr/IMG/pdf/masante2022_rapport_virage_numerique-2.pdf

[19] The national health data system comprises: national medical databases (SNIIRAM); data from hospitals and other health establishments (French medical information system - PMSI); statistical data on the causes of death (BCMD). Data on disabilities should complete the health data system.

by the development of reliable, environmentally friendly sources of energy) and the availability of components needed for the manufacture of digital tools that can be extracted, used, and reprocessed without harming the environment[20]. This underscores the importance of fundamental research on digital technologies, with a view to improving their cost effectiveness.

1.2 A technological mutation that induces behavioral changes

Big data, by their very nature, call into question distinctions about which we are accustomed to think and which conventionally serve as a support for ethical reflection: public *vs* private life, the care relationship as opposed to the economic market, individualism *vs* solidarity.

- *Public and private life*

The prospect of cross-referencing data of different types and from different sources, collected by public and private health providers in the framework of new modes of living and social relations, erases the traditional distinction between public life and private life, since seemingly harmless data can, once correlated with other data, yield information on health. This reconsideration is accentuated by the impossibility of guaranteeing the long-term efficiency of anonymization of sensitive data. Loss of the confidentiality of private life – resulting sometimes from disclosure by the people themselves – is a major feature of our times. It opens up to exploitation of big data a potentially vast field of investigation. The notion of health data can no longer be limited to personal data collected in the framework of healthcare.

- *Transformation of care and of the health market by the intervention of new stakeholders*

To the traditional public and private health stakeholders we need to add new stakeholders in the care relationship and in the health and wellness market. First there are data scientists[21] who are involved whatever the area of application. They occupy a central place because they are responsible for the management and exploitation of data with a view to producing new information.

[20] The reader will find on the website ecoinfo.cnrs.fr the article "*Impacts environnementaux du numérique, de quoi parle-t-on ?*" by Françoise Berthoud, in the blog *Binaire* (l'informatique : science et technique au cœur du numérique) / Le monde, 29 January 2019 (http://binaire.blog.lemonde.fr/2019/01/29/impacts-environnementaux-du-numerique-de-quoi-parle-t-on/), and the Unesco ethics committee report.

[21] This term designates experts in the management and analysis of big data, who design models and algorithms for the collection, treatment, and restitution of data. Also included should be the curators who clean the data.

Many initiatives come from patients themselves who search for and share medical information, as with the personalized health network PatientsLikeMe[22]. But it is mainly private companies that drive innovation in the digital technologies, including healthcare. Some reach out to people and patients as consumers. Profiting from the multiplication of connected objects[23] and of the widespread use of social media, they seek to attract their customers by inciting individuals to be stakeholders in their own health. They do this by an individualized approach that uses the notion of performance.

Beyond the hard-to-control risks of disclosure of sensitive personal data, this trend marks the increasing importance in healthcare of stakeholders whose purpose is not healthcare but the exploitation of a market. These stakeholders do not necessarily feel bound by the ethical obligations or the confidentiality rules that specifically apply to health professionals, even if they do often seek the assistance of these professionals.

This does not mean that we condemn these new practices. They can have the real virtue of encouraging our contemporaries to take greater care of their own health. And they allow measurement of health parameters and take into account, by real-time observation, factors corresponding to the environment and lifestyle, data that are determinant for patient care.

- *Precision medicine and solidarity*

The processing of large amounts of health data relating to treatment itself, but also to a person's genetics, lifestyle, and environment, allows an individualized approach to the prediction of a disease, its prevention, or, if it occurs, its treatment. With so-called precision medicine, each patient in a way becomes a unique case. All the information enables increased and better quality diagnostic and therapeutic precision, thus improving the prognosis. Such precision favors the quality of care, but is not without consequences. Notably, it allows the identification of possible risk factors shared by those who will later develop the disease. Precision medicine could therefore enable profiling that could dispense with the need for the notion of risk sharing within a community. Yet, risk sharing is a principle that underpins national solidarity, an essential value of our health policy.[24]

[22] https://patientslikeme.com

[23] In this digital era, each of us is invited to be a stakeholder in our own health. Many apps measure parameters such as weight, heart rate, blood pressure, blood glucose, blood cholesterol, and from these parameters derive information on lifestyle and the search for well-being (the notion of the "quantified self"). The rapid dissemination of these practices induces behaviors enabling everyone to follow the advice of a coach and to monitor one's progress, possibly through comparisons with others.

[24] This question is dealt with in greater detail in section 2.2- Justice: solidarity and the challenge of individualization

The distinctions between public and private life, business and care, and individualism and solidarity have become blurred, showing that we are confronted by a new situation marked by great uncertainty and risks of unethical behavior. This should encourage us to be vigilant, particularly in how we address the essential notions of information, consent, and control in health.

1.3 The notion of health data

The definition of health data is complex and changing. It has broadened over time. This notion is not limited to health data by nature (ie, data collected in the framework of medical care), as is apparent from section 1.1 concerning the expansion of the field of health and the possibility of obtaining sensitive secondary health data from primary data that in theory have no direct relation to health. The definition necessarily also includes data that – without in themselves being qualified as health data – become health data, either by cross-referencing with other data, thus allowing conclusions to be drawn on the state of health or the risk for an individual's health, or by their intended use (eg, in a care pathway)[25]. In these two cases, it is the purpose of the treatment that qualifies data that in theory are not health data as health data. To refer to these data altogether, we propose to use the expression "health-related data" in this opinion.

It is therefore not surprising that the General Data Protection Regulation (hereafter GDPR or the European regulation; see next section) concerning the protection of natural persons regarding the processing of personal data and the free movement of these data gives an extensive definition of 'data concerning health' in its article 4: "*For the purposes of this Regulation*, '*data concerning health' means 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*."[26]

[25] See the site of the French Data Protection Authority (CNIL) under "Qu'est-ce qu'une donnée de santé?" (cnil.fr).

[26] No. 35 of the GDPR states this as follows: "Personal data concerning health should include all data pertaining to the health status of a data subject which reveal information relating to the past, current or future physical or mental health status of the data subject. This includes information about the natural person collected in the course of the registration for, or the provision of, health care services as referred to in Directive 2011/24/EU of the European Parliament and of the Council (1) to that natural person; a number, symbol or particular assigned to a natural person to uniquely identify the natural person for health purposes; information derived from the testing or examination of a body part or bodily substance, including from genetic data and biological samples; and 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, for example from a physician or other health professional, a hospital, a medical device or an in vitro diagnostic test."

This defines data concerning health as a subset of personal data, the definition of which it is worth quoting from the same article 4: "*For the purposes of this Regulation, 'personal data' means any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an to identify such as a name, an identification number, location data, an online to identify or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person.*"
The updated data protection law[27] includes a part devoted to the processing of personal data in health (section II, part 3, art. 64-73).

The variety of health-related data – as in our definition – does not allow an exhaustive enumeration. The essential categories are historical data (family history, personal history), clinical data, laboratory data, data from analyses and investigations (including imaging data), data on treatments, prescriptions, and medicinal drug use, as well as environmental, socioeconomic, demographic, and behavioral data (which are informative regarding lifestyle and habits).

In the rest of this opinion, we will call health data any information on physical or mental health and well-being, and personal health data any health data concerning an identified or identifiable natural person.

1.4 Protection of health data

In the framework of medical care, health data are subject to several protective regulations, essentially the GDPR (art. 9.1), the data protection act (LIL), and the public health code. The modified LIL (art. 6.I), like the GDPR (art. 9.1), provides for the principle of a ban on the processing of health data, tempered nonetheless by a series of exceptions (see section 2.1.3).

The public health code regulates medical confidentiality (art. L. 1110-4), the hosting of health data[28] (art. L. 1111-8 and R. 1111-8-8 and s.), their availability (art. L. 1460-1

[27] This modification of the data protection and civil liberties law No. 78-17 of 6 January 1978 resulted in law No. 2018-493 of 20 June 2018, which enshrines the GDPR in French law. A ruling of December 2018 clarifies the national legal framework.

[28] Art. 1111-8 of the public health code: "Any person who hosts personal health data (...) does so under the conditions stipulated in the present article.

The hosting, whatever the carrier, paper or electronic, is done after the person in charge has been duly informed, and unless there is opposition for a legitimate reason.

(...) The host of the data mentioned in the first paragraph (...) is the holderof a certificate of conformity."

and s.), the conformity of information systems (art. L1110-4-1), data sharing (law on the modernization of our health system, art. R1110-1 and 1110-3), and the ban on the sale or commercial exploitation of health data (art. L. 1111-8, art. L 4113-7 of the public health code).

The adoption of particular security measures has the obvious advantage of improving confidentiality. These measures can, however, have the drawback – in particular for hosting – of compartmentalizing databases (clinical, medical-administrative, genomic), which may limit the possibilities of communication between them. Now, cross-referencing of these databases is of major interest for medical research, the effectiveness of which is impaired as it cannot derive the benefit full of the rich data collected in our country[29]. To allow cross-referencing of data, European projects (including those funded in the framework of Horizon 2020) incite those responsible for governance of data repositories to follow the FAIR ("findable, accessible, interoperable, reusable") guiding principles and criteria[30]. The FAIR principles enable the construction, storage, presentation, and publication of data, so as, among other things, to ensure facilitated and regulated sharing. This is the meaning of the French open science policy laid down by the minister of higher education and research on 4 July 2018[31].

This specific protection applies only to data "*collected on the occasion of activities of prevention, diagnosis, care, or medical-social or social follow-up*" and so does not cover the broad notion of personal health data that we used in section 1.3.

More generally, our country is distinguished by its longstanding and strong protective legislation. The 1978 data protection law (LIL; Loi informatique et libertés) stipulates that, for the processing of personal data, prior registration is mandatory, as is also sometimes an opinion or authorization from the CNIL (French Data Protection Authority). In general, this law, like the GDPR, states that the processing of personal data must observe certain key principles: a determined, explicit, and legitimate purpose, minimization of data collection, limited duration of data storage, mandatory security, and respect of human rights. The CNIL has been charged with enforcing observance of these principles,

Art. L 1110-4-1 of the public health code: "To guarantee the quality and confidentiality of personal health data and their protection, health professionals, health establishments and services, and any other body participating in prevention, care, or medical and social follow-up, the conditions and activities of which are governed by the present code, use, for the treatment, storage on data carriers, and electronic transmission of these data, information systems that comply with interoperability and security benchmarks drawn up by the public interest group mentioned in article L. 1111-24. These benchmarks are approved by decree of the minister of health, issued after seeking an opinion from the French Data Protection Authority."

[29] This question is considered in greater detail in section 3.2.1.

[30] See the site https://www.force11.org/fairprinciples; and Wilkinson M, et al. The FAIR guiding principles for scientific data management and stewardship. *Scientific Data* 3, *Article number*: 160018 (2016).

[31] The national plan for open science announced by Frédérique Vidal, on 4 July 2018. http://www.enseignementsup-recherche.gouv.fr/cid132529/le-plan-national-pour-la-science-ouverte-les-resultats-de-la-recherche-scientifique-ouverts-a-tous-sans-entrave-sans-delai-sans-paiement.html.

which, for the large part, have since been reaffirmed and completed. A few adjustments have been made to take into account technological advances stemming from the generalization of digitization and the exploitation of big data, notably through law No. 94-548 of 1 July 1994 concerning the processing of named data for research in healthcare and law No. 2016 of 7 October 2016, the "Law for a Digital Republic", which covers data and other issues in the digital age. Its principles include the neutrality of the internet, the right to data portability and recovery, platform fairness, and the provision of fair, clear, and transparent information to people whose data are used. Notably, the right to protection of personal data is assured by acknowledgement of the right to decide and to control the uses made of the data and the right to delete these data. Specific provisions are planned for when the holder of the data was a minor at the time of data collection.

But the above considerations regarding the new problems raised by the exploitation of big data show that the general requirement for prior authorization or registration became unrealistic, and this requirement was dropped by the European regulation of 27 April 2016, which came into force in the member states on 25 May 2018.

This regulation adopted an ambitious legislation that reaffirms the protective principles and provides for several devices to ensure their implementation[32]. This text does not call into question the special nature of consent, the purpose of personal data processing, or limiting the amount of data collected to what is really useful. The sanctions it provides for in the case of infringement serve as a deterrent.

To avoid the circumvention that could result from the transfer of personal data in a non-member country of the European Union that has no law on the protection of personal data equivalent to the GDPR and where providers could contest application of the European regulation, article 44 states that such a transfer can only be made if the controller and the subcontractors, including in cases of later transfers, respect the protective provisions of the GDPR. Companies that have their head office outside Europe, but which offer goods or services in Europe, must also respect the GDPR (article 3). The responsibility of the person in charge of data processing and the subcontractors is thoroughly regulated and sanctionable. Each member state, furthermore, must appoint an independent supervisory authority (articles 50 to 59).

The GDPR emphasizes the self-regulation and increased empowerment of stakeholders. Its essential contribution is recognition of the principle of accountability, that is, the person in charge of data processing has to take measures to guarantee and prove that data

[32] See the GDPR text (in French) on the CNIL website: https://www.cnil.fr/fr/reglement-europeen-protection-donnees.

protection has been a constant concern and that the rules guaranteeing it have been respected.

The GDPR comprises provisions applicable to the processing of personal data that are intended to facilitate research, in its broad sense, including the development and demonstration of technologies, basic and applied research, and research funded by the private sector (see notably article 89). Law No. 2018-493 of 20 June 2018, which covers the adaptation of this regulation to France, establishes the CNIL as a national supervisory authority. The role of the CNIL is that of a genuine regulator. In health research, the CNIL has adopted five new reference methodologies adapted to the legal framework in terms of health data, and a reference for accessing certain data of the SNIIRAM[33]. These methodologies streamline the formalities linked to the processing of these data for the needs of health research, since a request for authorization is not necessary in the case of declaration of conformity[34]. An audit committee has selected service providers to conduct audits of all systems that pool, organize, or make available part or all of the data of the national health data system.

The committees for the protection of human subjects also participate in data protection[35]. They are charged with giving a prior opinion on the conditions for the validity of all research involving human subjects, with regard to the criteria defined by article L 1123-7 of the public health code. Their role is specified by articles L 1121-1 to L 1126-11 of the same code. The committees are, in particular, consulted for any project involving the reuse of health data and regarding any need to recontact the holder of the data. The CEREES[36] (expert committee on research, studies, and evaluations in health) is called upon for requests to authorize studies, evaluations, or research not involving human subjects.

2. ETHICS CHALLENGED BY BIG DATA IN HEALTH

[33] https://www.cnil.fr/fr/quelles-formalites-pour-les-traitements-de-donnees-de-sante-caractere-personnel. SNIIRAM: Système national d'information inter-régimes de l'Assurance maladie (health insurance database).

[34] These reference methodologies concern research involving human subjects, studies or evaluations not involving human subjects, access to PMSI data (French medical information system) by health establishments, federations, and industrialists of the health sector, with a view to conducting studies in strict conditions of confidentiality and security.

[35] LIL, ruling December 2018, art.76.

[36] Comité d'expertise pour les recherches, les études et les évaluations dans le domaine de la santé (CEREES).

Big data undermine the benchmarks that conventionally support ethical reflection (the distinctions between public and private life, individualism and solidarity, economic concerns and the purpose of treatment[37]). Apart from the resulting uncertainty, this heightens the tensions between the individual and the community, and between economic logic and ethical values that require the protection of private life and of individual autonomy.

But it would be a caricature to contrast individualism and economic liberalism, on the one hand, and ethical principles, on the other.

Ethical values, such as autonomy, respect of private life and dignity, equality, solidarity/fraternity/justice, cannot be taken in isolation, as if each of them constituted an absolute. We cannot establish a hierarchy between them, and these ethical values are themselves subject to the tensions we have just considered. So, the principles of autonomy and justice can encourage the favoring of an individual's capacity for action and can conflict with the values of equality, solidarity, and fraternity. The ethical issues stemming from the interaction between these values should be viewed from a dynamic rather than a static perspective. What is at stake changes as a function of the constant and rapid changes in our society and of scientific and technological advances. A point of equilibrium should therefore always be sought and should be adapted to the specificities of each context. This more specific and concrete dimension will be dealt with in part 3.

But before this, it is worth considering how big data force us to renew our ethical reflections in healthcare. Such reflections relate to the main ethical principles that underpin medical ethics: respect of the person (including the respect of autonomy), justice, relevance, and benevolence (including here the obligation to do no harm).

2.1 Respect of the person

2.1.1 A value faced with a new situation

In its contribution to the revision of the 2018-2019 bioethics law (Opinion 129), the CCNE pointed to recent upheavals in representation of the human subject, and mentioned the genome and data among "*the qualitative changes occurring in the representation of the body and of its relation to the human subject, be it progress or fragility*". The CCNE continued by saying that "*all of this affects the very image of the species and of the human being, as much as the place of the individual, of the patient, and of the*

[37] See §1.2.

citizen"[38]. This question was addressed during the Bioethics Forum through consideration of a *"a new objectification of the human body to which, the genome and health data are added to traditional corporeal characteristics."* [39]

This raises the question of the link between data and the person. Should data be viewed as an "object detachable from the person," which results in the person being ignored and limits the protection of the person due to the use that can be made of the data? Or, should we on the contrary consider that data give a representation, an image, of the person? We then speak of identity or of digital profile, what Antoinette Rouvroy[40] summarizes as follows: "*subjects only exist at an infra-individual level (fragmented in various data banks) or a supra-individual level ("profiles" that only ever address sets of individuals, or, more accurately, sets of behaviors*)." [41]

Once digital technology allows infinite reproduction of data without dispossessing their holder, this raises the question of the relevance of the application of the notion of property. The question of ownership of personal data is at the heart of legal debates between the personalist conception and the proprietary conception. Personal data are at one and the same time extensions of personality of individuals (revealing the identity, the private life, personal choices, etc.) and information that can circulate, be transferred, and reproduced. From a legal point of view, "*data are at the border between the person and the thing, and can reveal equally well the one and the other*".[42] French law has opted for a non-proprietary qualification of data (notably, the law of 7 October 2016, the "Law for a Digital Republic", which retains the "right to informational self-determination"). Even though voices have been raised in favor of property rights, using the argument of the economic value of data[43], it seems that the non-proprietary qualification should be favored. An astute author objects to the hypothesis of the qualification of personal data

[38] *Le génome et les données sont « parmi les sauts qualitatifs en cours dans la représentation du corps et de son rapport à la personne, qu'il s'agisse de progrès ou de fragilité ». Tout ceci « affecte l'image même de l'espèce et de l'homme autant que la place de l'individu, du patient et du citoyen ».*

[39] "*Une nouvelle objectivation du corps humain où le génome et les données de santé s'ajoutent aux caractéristiques corporelles traditionnelles*". See on the CCNE website the 2018 *États généraux de la bioéthique* (Bioethics Forum) report, made public on 2 July 2018.

[40] FNRS researcher at the Research Centre in Information, Law and Society (CRIDS), University of Namur, Belgium.

[41] *« Les sujets (qui) n'existent que de manière infra-individuelle (fragmentés dans diverses bases de données) ou supra-individuelle (les profils ne s'adressant jamais qu'à des ensembles d'individus, ou plus exactement, à des ensembles de comportements) ».* Antoinette Rouvroy. Pour une défense de l'éprouvante inopérationnalité du droit face à l'opérationnalité sans épreuve du behavioralisme numérique. *Dissensus*. Revue de philosophie politique de l'université de Liège, April 2011, pp. 127-149.

[42] Philippe Mouron. Pour ou contre la patrimonialité des données personnelles. *La revue européenne des médias et du numérique* 2018, n°46-47, pp. 91.

[43] Le Monde of 12 January 2019. Inventer un droit patrimonial sur les données de santé, p.7. Rapport Mes data sont à moi - Pour une patrimonialité des données personnelles. *Génération libre*, January 2018 (https://www.generationlibre.eu/wp-content/uploads/2018/01/2018-01-generationlibre-patrimonialite-des-donnees.pdf).

as property and even considers that the texts do not go far enough in the adoption of the personalist approach.[44] This is also the position of the Conseil d'État, which considers that "*while we should strengthen the individual stakeholder's right to data protection, this should be done by envisaging this as a right to self-determination rather than as a right of ownership*".[45]

But as digital data can be transferred without any obvious inconvenience to the holder, one may wonder why we should seek to protect the data if their use by third parties in no way hampers the holder's ability to act.

The need for this protection cannot be overemphasized. We have seen that the logic of digital treatment is based on correlations established between a multitude of data which, taken separately, can be insignificant, but which, by cross-referencing, give precise indications on the most sensitive individual rights (philosophical and political opinions, religious beliefs, state of health, sexual orientation, lifestyle, people seen and places visited). Unlimited appropriation of personal data would distance us from a democratic society. It would, in fact, be tantamount to admitting the threat of a surveillance society in which individuals are monitored by numerous public or private providers acting for various purposes, be they commercial, political, or security.

But while consent is consistent with the vision of a "right to informational self-determination," it is only fully effective if it is free and informed, that is, if the holder of the data has knowledge of those (health stakeholders) who may use the data, of the immediate use to which they intend to put these data, and potential future uses. These big data may assume a "supraliminal" dimension, to borrow the term proposed by Gunther Anders[46], revealing a dimension that is too large to be apprehended by the individual, and thus questioning the very notion of consent.

Some features of big data make this objective of free, informed consent extremely difficult, even impossible, to achieve in algorithmic exploitation of big data:

- Digital technology allows the reuse of data, for purposes that may differ significantly from the purposes for which they were collected initially. The result is that

[44] Judith Rochfeld. Contre l'hypothèse de la qualification des données personnelles comme des biens. In: Les biens numériques, éd CEPRISCA, 2015, pp.221-236.

[45] *« S'il convient de renforcer la dimension de l'individu acteur dans le droit à la protection des données, c'est en envisageant celui-ci comme un droit à l'auto-détermination plutôt que comme un droit de propriété ».* 2014 Conseil d'État study. Le numérique et les droits fondamentaux, pp.264. http://www.conseil-etat.fr/Decisions-Avis-Publications/Etudes-Publications/Rapports-Etudes/Etude-annuelle-2014-Le-numerique-et-les-droits-fondamentaux.

[46] Günther Anders. Et si je suis désespéré, que voulez-vous que j'y fasse ? (1977). Allia, collection, Petite collection.

it is difficult to predefine the purposes of the exploitation of big data, which explains the notion of purposes that are not incompatible with the initial purposes, as developed by the GDPR (art. 5.1b and art. 6.4).

- Significant correlations not corresponding to a predefined hypothesis can be made by reuse of data that does not necessarily require a prior and precise request for the holder's consent to this new use. The absence of predetermination concerns as much the data used (it is difficult to know beforehand which data will be concerned by the comparisons made by the algorithmic treatment) as the results of this use.

- The holder of the data that a provider wishes to access is at a disadvantage. This results as much from the asymmetry of knowledge of the technologies used or the value of the data as from the resulting pressure – in particular in the case of access to the websites of the private internet providers who offer a service – from the subordination of this service to the provision of data.

- The complexity of the processes implemented in the exploitation of the data is such that even if fair information on the algorithms or the programs that result therefrom is given to the people concerned, the latter are unlikely to be able to extract information enabling them to make a choice concerning the fate and use of the data. This difficulty is heightened when use is made of deep learning techniques. The results obtained using these algorithms are currently hard to explain for programmers themselves, which makes this an important research question.

We shall consider below (§ 3.3) the possible answers to this question of consent and what possible alternatives there may be to individual consent as conventionally defined.

2.1.2 A new situation that mandates a new perception of what is at stake

European regulations reaffirm – while adapting – the principles of the special nature of consent[47], of determined, explicit, and legitimate purposes for data processing, and of the minimization of data collection and processing. This was ethically necessary. In its 2014 study report on digital technologies and fundamental rights, the *Conseil d'Etat* justified the need to specify the purpose of data processing by noting that "*the principle of specified purposes is at the heart of the trust that people have in the services of the digital society. When people have recourse to such services and when data concerning them are collected in this context, they must have the assurance that these data will not be used for purposes other than those of the service, unless they are informed*

[47] When consent constitutes the basis of legality, which is not always the case: it is not required for the execution of a legitimate interest contract or of a public interest mission.

thereof or the law stipulates this. The principle of specified purposes holds that personal data are not goods or, at least, that they are not like other goods"[48].

The data subject shall have the right of access to and rectification of personal data (articles 16 and 17 of the GDPR).

Prior authorization by the CNIL is not required if use of the data complies with a reference methodology[49], so what is at stake is data control, which must be effective to preserve users' trust, but which should not be crippling. It should be emphasized that it would be ethically unacceptable to hinder the major advances expected from the exploitation of big data in medical research and care.

2.1.3 How to achieve control?

As it is impossible to control every use of big data, alternatives are sought. Codes of good practice are therefore encouraged and submitted to the supervisory authority, as are the mechanisms of certifications and endorsement, thus providing proof of compliance with legal requirements. For each treatment, a controller is appointed to determine the purposes and the means and to take the requisite measures to supply the person concerned with all useful information (articles 12 to 15 of the GDPR). The controller is assisted, if need be, by a data protection officer (mandatory for the public sector [articles 37 to 39]), who functions independently (article 38).

A European data protection committee[50] is tasked with ensuring the coherence of the application of GDPR.

[48] *« Le principe de finalités déterminées est au cœur de la confiance que les personnes peuvent avoir dans les services de la société numérique. Lorsqu'elles recourent à de tels services et que les données les concernant sont collectées dans ce cadre, elles doivent avoir l'assurance que ces données ne seront pas utilisées pour d'autres finalités que celles du service, sauf à ce qu'elles en aient été informées ou que la loi le prévoie. Le principe de finalités déterminées est ce qui fait que les données personnelles ne sont pas des marchandises ou, du moins, qu'elles ne sont pas des marchandises comme les autres. »*

[49] The creation and updating of reference methodologies has been made necessary by changes in national legislative texts, by the coming into force of the General Data Protection Regulation (GDPR), and by feedback on existing frameworks from stakeholders in the field. The adoption by the CNIL of these methodologies is designed to create a protective framework for the people concerned that is favorable to research, innovation, and competitiveness (https://www.cnil.fr/fr/recherches-dans-le-domaine-de-la-sante-la-cnil-adopte-de-nouvelles-measures-de-simplification).

[50] This committee followed the Article 29 Working Party (G29), which was the forum for exchanges and for drawing up the doctrine instituted by directive 95/46. Apart from the formal opinions that it will be called upon to express on article 64, and the binding decisions that it will take on article 65 in the event of disputes between authorities, the European Data Protection Board will continue the work of drawing up the joint doctrine of the data protection authorities of the European Union through guidelines, opinions, etc. (see CNIL website).

The GDPR is designed to be firm regarding principles, but realistic. Although consent remains a major legal cornerstone of the processing of personal data, the GDPR, taking note of the fact that this requirement is not achievable in all cases, notably when data are reused, makes the processing of personal data lawful for certain purposes (see article 6.1 of the GDPR). In the case of particular categories of personal data, including health data (article 9 of the GDPR, article 8 of the data protection act in its version derived from the law of 20 June 2018), the ban on processing is the rule, but this ban can be lifted insofar as the purpose of the data treatment requires it, for example, for "*treatments to which the person concerned has given express consent*", "*treatments comprising health-related data justified by the public interest*", "*treatments relating to personal data made public by the data subject*", or "*treatments necessary to public research*"[51]. Either way, if consent is not required, the information must be delivered in a "*concise, transparent, intelligible and easily accessible form, using clear and plain language*," and the controller must take appropriate measures to achieve this. "*The information shall be provided in writing, or by other means, including, where appropriate, by electronic means*" (GDPR art. 12).

Any member state can, furthermore, introduce additional conditions to take account of its specificities, provided that this does not impair the effectiveness or alter the spirit of the European regulations. This is the very meaning of the change, dated 21 June 2018, to law 78-17 of 6 January 1978. Ruling No. 2018-1125 of 12 December 2018 legislatively finalizes the compliance of national rights with the general regulations on data protection (GDPR).

So, there is a progressive move from a wish for *a priori* exhaustive control to a logic of *a posteriori* intervention and control based on a quest for intelligibility and accountability which demands trustworthy behavior from those performing data processing.

In ethical terms, the CCNE endorses this approach, which must be based on a relation of trust between the holder of the data and those who collect, have access to, and treat the data. Transparency is a fundamental value established as a principle by article 5 of the European regulations, but can sometimes appear hard to reconcile with the working principle underpinning the exploitation of big data. However, it is necessary to seek sufficient intelligibility of the process employed to exploit the data and of its possible consequences. The use of such an approach may sometimes be doubted given the reality of the policies of certain providers. In addition, with the use of programming techniques

[51] *See article 8* modified by law No. 2018-493 of 20 June 2018. *Insofar as the purpose of the treatment requires it for certain categories of data, the ban provided for in I does not apply to 11 categories of treatment, including treatments necessary for public research, as defined in article L. 112-1 of the research code, used in the conditions provided for in 2 of article 9 of the aforementioned (EU) ruling 2016/679 of the European Parliament and the Council of 27 April 2016, after the considered opinion published by the French Data Protection Authority in accordance with article 28 of the present law.*

based on deep learning, it becomes impossible to follow step by step the path taken by the machine to answer the question asked. One has the right to expect that the person in charge of treatment masters the logic followed and the parameters taken into account[52]. Renunciation of this requirement would amount to renouncing all possible control over the bias that can affect the algorithmic treatment and all possible responsibility of the stakeholders. Given that the complexity of the process, exceptions apart, does not allow an individual to exercise this type of control, the individual should only give his/her trust and consent to the use of personal data if this use is made within the framework of a governance identified by a designated person who has made clear commitments. These commitments must, of course, be verifiable by a supervisory authority that can call upon assistance from experts[53].

It is essential to have these means of control and to ensure that our fellow citizens are fully informed regarding:

- the details they should receive when their personal data is used or reused,
- the commitments made by whoever requests use of their personal data,
- the checks that can be implemented to verify the seriousness of these commitments and their long-term maintenance.

This process could result in extra work that would be hard for the health system to take on, a difficulty that should be anticipated, taken into account, and overcome.

The information should be adapted to the different contexts of use (healthcare, occasional research, participation in an international database, etc.), so that the holder of the data has a sufficient grasp of the technologies used and of the terms employed, should he or she be in a vulnerable situation. This information should enable the holder to choose freely when consent is required. The complexity of usages and contexts is such that it is necessary to reflect upon the notion of consent and on new ways of collecting it that guarantee the respect of ethical principles and human rights (see section 3.4.1) (RECOMMENDATION No. 1).

These are the conditions of a trust that is necessary to meet the ethical requirement for respect of the dignity of the data subject. These conditions echo the need for fairness and vigilance highlighted by the CNIL in its aforementioned report filed in December

[52] Decision No. 2018-765 DC of 12 June 2018 by the Constitutional Council stipulates that the person responsible for data treatment must control the algorithmic treatment and how it evolves, so as to explain in detail and intelligibly to the person concerned the way the treatment has been implemented. As a result, it is not possible to use, as an exclusive basis for an individual administrative decision, algorithms likely to revise the rules that they apply, without the control and approval of the person in charge of data treatment.

[53] See the CNIL report: Comment permettre à l'homme de garder la main ? Rapport sur les enjeux éthiques des algorithmes et de l' intelligence artificielle, 15 December 2017, https://www.cnil.fr/fr

2017 following the public debate it oversaw on the ethical issues associated with algorithms and artificial intelligence.

Even though this approach is not specific to health, it is of great interest therein, with the caveats that we examine in section 3, notably in terms of the particularities of research and healthcare.

Trust should be mutual. This would not be the case if the holder of the data, after having consented to their use in areas affecting the community, notably medical research, could make improper use of the right of erasure, at the risk of weakening or even calling into question a research project. The trade-off for the obligation to continue providing information on the results of the data treatment should be that the right of erasure is contingent upon a legitimate reason, such as proven bad use.[54]

While the European regulations and the resulting legislation are intended to establish an ambitious system of protection, which is tending to become a reference even outside the European Union, they are not yet fully implemented and will only achieve their aim if the protective principles they provide for are effective. This is the major issue from an ethical viewpoint.

Given the particularly fast rate of scientific and technological innovations and the resulting changes in the collection and use of health-related data, the CCNE considers that it is necessary periodically to check that the legal provisions over time effectively maintain the system of protection of personal data (RECOMMENDATION No. 2).

Two points in this regard are of particular note:

- The European regulation only applies to so-called personal data and does not concern anonymous information, defined as information that does not concern an identified or identifiable natural person. To determine whether a natural person is identifiable (see note 9), account should be taken of all the means reasonably likely to be used, such as singling out, either by the controller or by another person to identify the natural person directly or indirectly (recital 26). This notion of all means reasonably likely to be used could be an unstable criterion in a world marked by a powerful continuous flow of technological innovations[55].

[54] It is to deal with this concern that the GDPR excludes the right of erasure when the treatment is necessary for archival purposes in the public interest, for the purposes of scientific or historical research, or for statistical purposes (article 17-3-d). Under these circumstances, appropriate guarantees are needed for the rights and liberties of the person concerned, such as the minimization of data and their pseudonymization (article 89-1).

[55] It is without doubt conceivable that data at one time anonymized could become personal again because of progress in techniques of identification, and the data would again come within the scope of the GDPR.

- The regulation does not apply to treatments of personal data performed by a natural person during strictly personal or domestic activities, unrelated to professional or commercial activity[56]. However, the regulation applies to those performing the data processing or to subcontractors who provide the means of treating the personal data for such personal or domestic activities (recital 18).

2.2 Justice: solidarity and the challenge of individualization

In France, health is funded largely by the welfare system, and particularly by the sickness insurance fund, and is rooted in the sharing of health risks. This sharing is one of the fundamental manifestations of solidarity, an essential ethical value of our health system since it contributes to the implementation of our founding principles of equality and fraternity.

The organization of healthcare is highly regulated and is based on two types of coverage: mandatory social security contributions and private supplementary health insurance. The former is independent of income and age, and considers only the cost of treatment, which should not be an impediment to care; it is completed by private insurance (a little under 14% of the care), which is divided into categories (income, age) and which introduces sub-groups. However, the taking into account of these costs hides disparities between types of treatment. For certain treatments (optical, dental protheses) coverage by the private health insurers is predominant and may leave the patient to make a substantial payment.

For stakeholders who must bear ever increasing health costs, there is a strong temptation to personalize the risk so as to achieve better economic management. Risk sharing underpins the insurance contract (insurers are greatly involved in health matters) and is also based on uncertainty. Each person contributes a relatively modest amount and as there are a large number of contributors it is possible to cover the sometimes-high costs that accrue when the insured risk becomes a reality for a much smaller number. As CCNE Opinion 124 puts it, the French people's strong attachment to the principles of non-discrimination and risk sharing could be weakened by the emergence of predictive medicine. Detailed knowledge of individual risks, if this proved possible, could thus lead to a breakdown or at least a weakening of solidarity and of shared coverage of health risks.

[56] These activities could include the exchange of correspondence and the keeping of an addresses book, or the use of social media and online activities within the framework of these activities.

The health-related question posed by the difficulty of reconciling autonomy and individual liberty and by equal access to care and national solidarity is not new[57]. But big data and their algorithmic treatment, in particular using the techniques of artificial intelligence, shed new light on this question.

2.2.1 A value faced with a new situation: national solidarity and the individualization of medical risk

The intrinsic ambivalence of digital technologies means that an overly strict intervention designed to prevent their negative effects may restrict their positive effects. Likewise, the exploitation of big data in healthcare reveals a conflict of values and of interest between the benefits hoped for by the individual and those that could result, for the community, from an intrusive approach intended to favor a better economic rationale.

The availability and cross-referencing of databases – clinical, biochemical, genomic, imaging, environmental – enable analysis of people's state of health with a precision that leads to individualization: each individual is a unique patient because of the particularities of his/her disease or risk. This is so not only for overt disease whose diagnosis can be refined, thereby allowing precise adaptation of treatment and improved care (personalized medicine), but also – when disease is not present – for risk analysis. The way is therefore open to prediction and prevention, with a view to preventing the emergence of the disease.

How can we prevent this more precise knowledge of risk and treatment personalization from leading, for economic reasons, to discrimination and reconsideration of national solidarity, which is required for living together and espouses the idea of risk sharing? Risk sharing may be compromised if an action to prevent risk is known and can be imposed, or if prediction of an unfavorable prognosis prompts minimization of care for economic reasons.

[57] Paul Ricoeur considers that, in the last analysis, this conflict on the public health front is in no way surprising. The medical contract could be rewritten in terms of a series of paradoxes. First paradox: the human subject is not a thing, and yet the human body is a part of observable physical nature. Second paradox: the person is not merchandise, nor is medicine a business, but medicine does come at a price and a cost to society. The last paradox covers the first two: suffering is private, but health is public. One should therefore not be surprised if this public health conflict worsens, given the increasing costs of research in medical biology, of exploring the human body, and of highly sophisticated surgical interventions, all of which is compounded by increasing life expectancy, not to mention the unreasonable expectations of a public that demands too much from medical professionals, whose possible misuse of power also creates fear. In short, the gap will only widen between demands for unlimited individual liberty and the preservation of equal public distribution of healthcare under the rule of solidarity (Les trois niveaux du jugement médical – le contrat médical *Esprit* No. 227 - December 1996, pp. 21-33).

As we shall see in greater detail in part 3, progress in genetic testing is emblematic of this threat, as it uses powerful predictive markers. This is why the CCNE noted in its Opinion 46 of 30 October 1995 "Genetics and medicine: from prediction to prevention", that *" Genetic tests give information on the identity of persons and emphasise their diversity which contributes to the rich nature of humankind. To use such information for the purpose of selection or of discrimination in social or economic terms, be that in the realm of public health policies, employment, or insurance systems, would be crossing a boundary of the most extreme gravity and would question those principles of equality of rights, dignity and solidarity for all human beings upon which society as we know it is based. The CCNE insists on the necessity of observing those fundamental principles whatever aims may be pursued by genetic testing. Human Rights are at stake"*[58]".

2.2.2 A new situation that demands another conception of what is at stake

Profiling, inasmuch as it provides greater knowledge about an individual, also offers useful ways of helping the State to improve its management of health economics. By its very nature, however, the individualization of risk can impair sharing. It provides a justification for offering price benefits to those who are not at high risk, to the detriment of others. Furthermore, it encourages the persuasion of individuals to adopt behaviors likely to prevent or to limit the extent of the diseases that threaten them. Yet, it is one thing to contribute to the general education of the population by campaigns promoting a healthier lifestyle, and quite another to target people or groups determined on the basis of the risk they present. The nature of prevention would then change, and it would lose its educational character and become intrusive if accompanied by deliberate monitoring, by means of connected objects, for example. Prevention would lose its favorable character if it were to involve a regime of sanctions or rewards for the results obtained[59]. Such a deviation would threaten not only the principle of solidarity, but also the protection and care that a health policy should provide to the most vulnerable. It would also jeopardize individual liberties by bypassing people's free will while imposing on them an inappropriate restriction. The differentiation between people with "good" and "bad" risks could only be justified by advancing the merits of those who make an effort to have a

[58] *« Les tests génétiques apportent des informations sur l'identité des personnes et soulignent leur diversité qui contribue à la richesse de l'humanité. L'utilisation de ces informations à des fins de sélection ou de discrimination dans la vie sociale et économique, que ce soit dans le domaine des politiques de santé, de l'emploi ou des systèmes d'assurance, conduirait à franchir une étape d'une extrême gravité vers la mise en cause des principes d'égalité en droits et en dignité, et de solidarité entre tous les êtres humains, sur lesquels repose notre société. Le CCNE insiste sur la nécessité de respecter ces droits fondamentaux, quelle que soit la finalité de l'utilisation des tests génétiques. Il y va des droits de l'homme ».*

[59] This is already so for certain private supplementary health insurance contracts (see note 56).

healthy lifestyle and by stigmatizing those whose behavior is deemed irresponsible. Beyond the inadmissible nature of such a classification, the restriction that it would impose would be unfair because it would overlook the fact that risk depends not only on individual behavior, but also on factors we do not control and which are linked to heredity and misfortune. One cannot ignore the impact of the social environment, all studies of which show that it is determinant for people's state of health and life expectancy[60].

The individualization of health risks tends to establish correlations between people as a function of behavioral criteria, as well as criteria concerning choice and lifestyle. The worldwide communication allowed by the internet and social media favors the development of affinities and of a sense of solidarity between those who see themselves as similar, whatever their geographical location, rather than between fellow citizens. This can make it harder to accept national policy as a reference of solidarity and public health policies. Yet it is within this national framework that health systems are funded and that solidarity is mainly practiced in the face of disease and life's misfortunes.

2.2.3 How can control be preserved through national solidarity?

Today in France, the ethical imperative of equal access to care prohibits public health insurance bodies, set up because of the principle of mandatory public health insurance, from operating risk selection in health. But, while the cost of serious acute diseases is fully covered, discriminatory practices may be sought for minor diseases, or for diseases that become chronic and which account for a large part of public health expenses[61]. "There could be a blurring of the traditional distinction between a risk inherited genetically and as yet uncontrollable, and a risk chosen by the adoption of a lifestyle." (CCNE Opinion 124).

The risk of establishing an individual health profile would stem less from potentially unequal access to the healthcare system than from being forced to accept a treatment for the sake of greater efficiency of that system. The benefit expected by the community would be a reduction in the expenses occasioned by delayed diagnosis and treatment.

[60] Notably: INSEE première, No. 1372, 5 October 2011: L'espérance de vie s'accroît, les inégalités sociales face à la mort demeurent.

[61] The idea that behavior retroactively acts on healthcare is not excluded. For example, home-based management of sleep apnea is subject to the patient's adherence to treatment, as assessed by remote monitoring.

A temptation to discriminate in insurance costs is a real problem because it is based on a powerful economic logic. The ban on selection can be circumvented by private insurers, notably because big data multiply profiling possibilities, even without the use of medical questionnaires or health data. The availability of information on the daily behavior of those insured already prompts private insurers to establish partnerships with companies and thereby modulate their premiums by awarding "rewards" to those whose behavior is deemed responsible[62].

The conflict between economic logic and individual interest, on the one hand, and the requirements of living together and solidarity, on the other, can only be arbitrated by political action. Only the law can set limits on the individualization of risks and enact the rules needed to preserve national solidarity.

It would be particularly useful if the same approach were pursued on a supranational scale, and notably within the European Union.

The CCNE considers that there is a need for conditions of vigilance implemented collectively by all health stakeholders, to ensure that the logic of personalization, which can be beneficial, does not transgress the values of equity and solidarity by progressing towards profiling that is discriminatory, notably economically (see **RECOMMENDATION No. 8**).

2.3 Non-harm and benevolence: big data, a factor of innovation in health but a risk of harm if data quality is not ensured

2.3.1 Quality of care and access to innovation confronted with a new situation

It would be unethical not to take advantage of the collection and analysis of the data of patients (or of healthy individuals participating in research) with current computer technologies for the health benefit of patient and community alike. It would be equally unacceptable ethically to ignore the risks of harm that can result from this approach and not to seek to reduce them[63].

[62] In 2014, Axa, together with the startup Withings, since taken over by Nokia Health, provided its insurees with a connected bracelet by offering gift certificates if the insuree takes 7000 steps a day. Audiens (insurance company for professionals working in the cultural field) and intelligent watch and bracelet producers Garmin made an offer to very small companies for the generalization of a complementary health plan.

[63] “The obligations of beneficence affect both individual investigators and society at large, because they extend both to particular research projects and to the entire enterprise of research. In the case of particular projects, investigators and members of their institutions are obliged to give forethought to the maximization of benefits and the reduction of risk that might occur from the research investigation. In the case of scientific research in general, members of the larger society are obliged to recognize the longer term benefits and risks that may result from the improvement of knowledge and from the development

The health-related innovations stemming from data processing will probably be numerous, even if they come up against important technical obstacles (notably for the techniques that use machine learning) in everyday practice[64]. Among expected innovations are the prediction of risk using genomic data, early detection of warning signs, therapeutic monitoring, more precise classification of diseases[65], diagnostic aids[66], and the management of patient flow. The perspectives in public healthcare and organization of the health system are equally important and can help limit the current rate of increase of healthcare expenses, which will be hard to bear in a context marked by low economic growth and aging of the population. Assuming we are able to cross-reference these different data and extract from them health indicators, epidemiology, which uses these indicators to study risk factors and diseases in the population, would constitute a powerful tool for a public policy of disease prevention and improvement of therapeutic decisions. The development of connected objects can be useful in this regard. French interest in the quantified self, which is enabled by digital technologies, by empowering people to measure their actions, study the consequences thereof, and note their day-by-day progress, constitutes a genuine opportunity to improve health-related behavior, if the state gives itself the means to invest in a policy of prevention using digital technologies. These applications will be illustrated in part 3.

However, two factors are currently holding back this development:

- the collection and exchange of data are not sufficiently developed in France because of the dispersion of data warehouses, registries, and cohorts of patients, gaps in the interoperability of information systems, and inadequate sharing of sources. To overcome this difficulty, there is a plan for the setting up in the near future of two shared platforms to pool on a national scale big data in the fields of genomics (France Médecine Génomique 2025[67]) and health data collected in the framework of healthcare or its reimbursement (Health Data Hub[68]);

of novel medical, psychotherapeutic, and social procedures." Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research, Report of the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. 1974).

[64] Yu Kh, Kohane IS. Framing the challenges of artificial intelligence in medicine. *BMJ Quality & Safety* 2019, 28: 238-41. Weintraub WS, et al. Translational medicine in the era of Big Data and machine learning. *Circulation Research* 2018; 123: 1202-4.

[65] Rumsfeld JS, et al. Big data analytics to improve cardiovascular care: promise and challenges. *Nature Reviews Cardiology* 2016; 13: 350-9.

[66] For instance, artificial intelligence algorithms for the diagnosis of cancer and depression, the management of chronic pain, the prediction of suicide, or help with dietary prescriptions in diabetic patients. *Nature Medicine* 2018, 24, 1304-5.

[67] https://www.gouvernement.fr/sites/default/files/document/document/2016/06/22.06.2016_remise_du_rapport_dyves_levy_-_france_medecine_genomique_2025.pdf

[68] The Health Data Hub, a platform for the exploitation of health data, could be the instrument whereby the state achieves an ambition: put the wealth of health data financed by national solidarity at the service of the patient and of the health system while respecting the ethics of the basic rights of our fellow citizens.

- the difficulties of access to communication technologies experienced by some people because of poor geographic coverage or unfavorable socio-economic conditions create what has been called a digital divide ("*73% of the population owns a smartphone [99% of 18- to 24-year-old], and 94% a mobile phone; half of the 6% who do not own one are aged 70 or over, and one-third of them have a low income; 10% of the population aged 12 or over have no computer, no smartphone, and no tablet*"[69]). This difficulty of access is all the more significant as it affects populations that are vulnerable and most in need of regular support. The *Observatoire des Inégalités* noted in 2016 that "*15% of the population does not use the internet and half are not on social media. This proportion reaches 25% among the most disadvantaged (households with median revenue of about 1200 euros per month) and 43% for those without diplomas. [...] The divide that remains is above all generational: the oldest people are outside a world whose usefulness they do not really grasp and do not always understand. Yet, constant reference to the internet and to social media, in particular, as if it is self-evident that everyone uses them, constitutes a symbolic violence for those who do not have the means to participate*".[70]

The CCNE emphasizes the need to make sure that people who have no access to digital technologies, for economic reasons or because they have difficulty understanding how they work, benefit, like others, from the advances in healthcare and are neither penalized nor discriminated against in their access to medical resources (**RECOMMENDATION No. 9**).

2.3.2 A new situation that requires another view of what is at stake: the preservation of human control to ensure the reliability of data and of decisions generated by exploitation of big data

Conventional medicine is based on listening to the patient, the analysis of clinical, laboratory, and imaging data, combined with the doctor's education and experience. The patient-doctor relation is based on a contract of trust established using honest and personalized information, leading to decisions taken jointly and genuinely shared[71]. A more innovative practice of medicine can also benefit from the use of decision-making algorithms based on a learning process built on the treatment of big data.

It would be a single window that would facilitate access to data. It would guarantee the quality of the data shared and would represent a trusted third party for the sharing of data while respecting patients' rights (see report https://solidarites-sante.gouv.fr/ministere/documentation-et-publications-officielles/rapports/sante/article/rapport-health-data-hub-mission-de-prefiguration). 12 October 2018.

[69] All 18- to 24-year-olds and 98% of those aged 25-39 report owning a mobile phone. This proportion drops to 76% among those aged 70 or over. Today, 6% of the French population aged 12 or over does not have a personal mobile phone. This lack is more common among older people (24% of the over-70s have no cell phone), those with no diplomas (22%), retirees (16%), and people living alone (12%) (Baromètre du numérique, Credoc 2017).

[70] Observatoire des inégalités. Qui a eu son iPhone 7 à Noël ? December 2016.

[71] CCNE Opinion 58: Informed consent of and information to persons accepting care or research procedures (12 June 1988).

Big data are conventionally defined by the 3 V's of volume, velocity, and variety, to which can be added veracity and value. As we have already discussed when we examined the ethical issues of the protection of the person and of the respect of private life, the challenge is intelligibility and the control of both the quality of the data themselves and the quality of the conclusions drawn by the operating system.

How then can the person (or the community) be assured of the reliability of the results obtained regarding known scientific knowledge? The main risk is that of bias or error (involuntarily or willfully) in the collection, annotation, or processing of data upstream[72] (see part 3, section 3.3 on genomic data). This would lead to inexact information downstream, resulting in an erroneous decision or discrimination (flawed diagnosis, unsuitable behavior or care protocol, research or administrative body led to draw inappropriate conclusions).

To prevent this risk, it is essential – particularly in healthcare – to have a "human guarantee"[73,74] in order to comply with the methodological rigor of the quality of the data, the appropriateness of the algorithmic treatments to the question posed, and the verification using an independent data set of the robustness and accuracy of the result given by the algorithm[75]. This guarantee is all the more important as the infallibility and objectivity generally accorded to analyses based on computer and mathematical models result in excessive trust, which can increase the risk of error.

This human guarantee must be given repeatedly because, apart from the biases mentioned, there is the question of the validity and sustainability of decisions deduced from algorithms trained with data that are retrospective and often purposely selected. The results obtained could be weakened by the constant flow of new data integrated automatically in the learning processes[76].

[72] The future of biocuration. *Nature* 2008; 455: 47.

[73] The principle of a "human guarantee" was proposed by the working group that drew up the report entitled Digital technology and healthcare: which ethical issues for which regulations? commissioned by the CCNE and made public in November 2018. The term was used in the CCNE Opinion 129, in the section on digital technologies and health, pp. 94-106.

[74] In its decision No. 2018-765 DC of 12 June 2018, the Constitutional Council stated that the exclusive use of an algorithm to process health data is excluded if this treatment relates to sensitive data as mentioned in section I of article 8 of the law of 6 January 1978, ie, personal data, which reveals the purported racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership of a natural person, genetic data, biometric data, health data, or data on the sex life or sexual orientation of a natural person.

[75] Beam AL, Kohane IS. Translating artificial intelligence into clinical care. *JAMA* 2016; 316: 2368-9; Big Data and machine learning in health care. *JAMA* 2018; 319: 1317-8.

[76]KH Yu, IS Kohane. Framing the challenges of artificial intelligence in medicine. *BMJ Quality & Safety* 2019; 28: 238-41.

A flaw in the human guarantee of the process could have two major consequences:

- the risk that a bias or methodological error, based on the processing of personal data, results in a conclusion or an erroneous decision that has direct consequences for the person;
- the risk of an inconsistency and of an error of judgment if it is not possible to verify the result obtained or to check the path that leads to it, when personal data are involved. This risk would be all the more unacceptable as it would concern decisions for which the patients must benefit from the right to information, to which they must consent or be able to refuse, and about which they must have all the facts (see RECOMMENDATION No. 4).

2.3.3 How can control be ensured?

These risks can be more effectively controlled in terms of clinical research and care (or of connected objects recognized as “medical devices”), despite the heterogeneity of the situations, because those who apply the information deduced from data processing are health stakeholders (doctors, nursing staff, researchers) who act within a professional framework. They are constrained by strict ethical rules, such as medical confidentiality, the validation of treatment protocols, and medical responsibility. The data themselves are health data in the sense of article L 1111-8 of the public health code, and we saw in part 1 that they benefit from particular protection. However, an erroneous conclusion drawn from data analysis cannot be excluded, for example, if the data do not include data from certain minority populations (see section 3.3 on genomic data). There may also be the temptation to bias algorithmic learning so as to favorably evaluate certain healthcare establishments.

This question of the veracity of measurements and data is even more acute in the case of new applications offered to consumers by private providers. The development of these apps is unsupervised and their efficacy is not scientifically evaluated, either for the algorithms used or for the conclusions drawn and delivered remotely, which can be partisan or sponsored.

Given the risks thus revealed, there is a danger of refusing any protocol not based on proven techniques, which in most cases would exclude the exploitation of big data. On the contrary, it is at the cost of an opening and of sharing of knowledge – a condition of a genuinely fruitful scientific approach (see below: section 3.2) – that one can hope to advance medicine and multiply prospects. It is to meet this requirement that the Villani report[77] recommended the creation of a platform for access to and sharing of data corresponding to health research and innovation (Health Data Hub[78]). This should be implemented in the first half of 2019[79].

The control of the quality of algorithmic treatment of digital data imposes actions in several areas:

- *An ambitious training program*. The CCNE considers that health professionals should, during their initial training and throughout their career, undergo suitable training on digital technologies, the ethical and legal principles that govern data collection and treatment, the means, notably technical, that should be put in place to respect these principles, and the risks of bias and of infringement of confidentiality and respect of human rights that would result should these principles be overlooked.

  Experts in the management and analysis of big data (data scientists), as well as researchers, must be aware of ethical questions raised by these technologies in order to protect fundamental rights and individual liberties (RECOMMENDATION No. 5).

- *A qualitative evaluation*. The proliferation of websites and applications that give advice on improving lifestyle and well-being, outside the care pathway, pose the question of the rigor with which they collect, interpret, and treat health-related data. The CCNE considers that these websites and apps and also the quality of the information given to users should be evaluable, so as to avoid insufficiently rigorous approaches having harmful effects on people's behavior and health (RECOMMENDATION No. 6).

  In this regard, the French National Authority for Health (HAS) has recently published a reference document designed to define good practices concerning cell phone apps[80]. As a national supervisory authority, the CNIL should play a determinant coordinating role in this field.

- *High-level scientific research* in informatics and basic mathematics so as to meet the challenge of "explicability". There is a need to enable human subjects to understand the basic steps of algorithms based, notably, on algorithmic learning, and hence to understand the performance of the machine in reaching a result.

[77] Donner un sens à l'intelligence artificielle : pour une stratégie nationale et européenne . Report of 28 March 2018. https://www.ladocumentationfrancaise.fr/rapports-publics/184000159/index.shtml

[78] Report (in French) downloadable from https://solidarites-sante.gouv.fr/IMG/pdf/181012_-_rapport_health_data_hub.pdf

[79] It is inscribed in the bill on the organization and transformation of the health system, title III: Développer l'ambition numérique en santé (adopted at the first reading at the Assemblée nationale on 26 March 2019, and the subject of parliamentary debate, May 2019).

[80] Good practices concerning apps and connected objects in health (in French). Haute autorité de santé. October 2016. https://www.has-sante.fr/portail/jcms/c_2681915/fr/referentiel-de-bonnes-pratiques-sur-les-applications-et-les-objets-connectes-en-sante-mobile-health-ou-mhealth

This process can still today be likened to a black box, which hinders applications in healthcare.

3. WHICH PRINCIPLES OF ACTION FOR DIFFERENT CONTEXTS?

3.1 Exploitation of big data for healthcare innovation

Whether it concerns care, research, prevention, or public health policies, the aim of medicine is to relieve suffering and to do everything to promote the health and well-being of the population. This care relationship, based on a direct human relation rooted in trust and in genuinely shared decisions, is subject to strict ethical conditions, including medical confidentiality. It is controlled by professional bodies.

Digital technologies have long since changed the carer-patient relationship. Thus, the computerization of healthcare establishments has led to the creation of databases for daily practice. Doctors have online access to constantly updated knowledge, to expert systems, and to computer-assisted prescription, which are certified by the reference system of the French National Authority for Health (HAS). The law of 16 January 2016 on the modernization of our health system introduced the principle of the sharing of health information between professionals involved at different stages of the care pathway, under the patient’s control, increasingly often through shared medical records and tomorrow through the patient’s digital health space[81]. Personal data collected in the framework of this care relationship, or via medical-administrative data, are considered highly sensitive and as such benefit from particularly strong protection[82] (see section 1.3.2).

3.1.1 What changes with big data?

The capacity to treat data and to extract new information on a scale that is only possible with a machine – one of the elements of the disruption induced by big data that we examined in section 1.1 – leads to a new approach to the acquisition of scientific knowledge. This capacity is not based solely on testing theories and models that need to be verified, but on the production of new hypotheses and knowledge from data accumulated without having been selected to answer an explicitly stated (inductive ap-

[81] The creation of the digital health space is a provision of the bill on the organization and transformation of the health system, title III: Développer l'ambition numérique en santé (subject of parliamentary debate, May 2019).

[82] The conditions of sharing health data, of the right to oppose, of information for the patient and the collection of consent vary depending on whether the professionals who share these data are part of the same care team or not, and are specified by the rules of exchange and sharing fixed in articles L.1110-4 and R1110-3 of the public health code.

proach) question (or intention). This has already been done in studies like the Framingham study[83], but at that time this involved human learning exclusively and not machine learning, in which the human role is more or less reduced and the path of which is currently opaque. This shift from human- to machine-based treatment of data makes inevitable the intervention of private internet providers for the collection, storage, and use of health data.

From this disruption, one can induce several likely innovations even if the daily practice of health professionals has not yet really changed and even though some people temper the promise of rapid and major benefits in health stemming from these technologies[84]. Innovations are expected essentially in five areas[85]:

- prediction of the risks of death or of accidents thanks to improved knowledge of risk factors and to early detection of warning signs[86];
- improvement in pharmacovigilance thanks to real-life screening for the side effects of drugs[87];
- more precise classification of diseases[88];
- refinement of treatments;
- automation and refinement of interpretation, notably in imaging and radiology.

Many of these innovations depend upon the combination of multiple sources of data: data collected by health professionals during the care pathway, notably genomic data (see section 3.3), medical-administrative data, data collected from the internet by smartphones or computers using applications that become accessories for health monitoring and research tools. While some of these devices are used in the care pathway

[83] *The main aim of the framingham study, which started in 1949, was to shed light on and observe the development of cardiovascular diseases. The cohort – an unselected geographic population – initially comprised one-third of the inhabitants – aged between 30 and 60 – of the town of framingham in the united states. It currently includes and is studying the third generation. Mahmood ss, et al.* Lancet *2014; 383: 999-1008.*

[84] *Chen jh, et al. Machine learning and prediction in medicine — beyond the peak of inflated expectations.* New england journal of medicine *2017; 376: 2507-9.*

[85] Rajkomar A, et al. Machine learning in medicine. *New England Journal of Medicine* 2019; 380: 1347-58.

[86] Torous J, et al. Smartphones, sensors and machines learning to advance real-time prediction and interventions for suicide prevention: a review of current progress and next steps. *Current Psychiatry Reports* 2018; 20 : 51; De Fauw J, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature Medicine* 2018; 24: 1342-50.

[87] Les données de vie réelle, un enjeu majeur pour la qualité des soins et la régulation du système de santé. L'exemple du médicament. Bernard Bégaud, D. Polton, F. von Lennep. Report commissioned by the minister of health. May 2017.

[88] Rumsfeld JS, et al. Big data analytics to improve cardiovascular care: promise and challenges. *Nature Reviews Cardiology* 2016; 13: 350-9.

and are supervised by health professionals and are subject to specific French or European regulations, this is not so for devices or applications controlled by commercial providers outside France.

Two examples illustrate these innovations:

- The progression towards precision medicine and preventive medicine could be a major consequence of the exploitation of big data.
  In their daily practice, doctors are generally consulted by someone who expresses a request or reports suffering caused by a disease. Now, this disease is the outcome of a series of bodily dysfunctions that have accumulated and which could have been detected earlier, if monitoring of certain parameters associated with personal risk factors had been instituted[89].
  The exploitation of big data from multiple sources and measured continuously independently of any medical care could lead to improved prediction of risk[90]. This shift from symptomatic diagnostic medicine to asymptomatic predictive medicine (CCNE Opinion 77) is based not on management of a reported symptom, unlike current medical practice, but on the anticipated emergence of a dysfunction. Because this dysfunction occurs too late with respect to later or insidious manifestations of the disease, classic clinical medicine is often not adapted to combat the disease in a timely fashion. Things would be improved significantly by setting up real-time monitoring of various laboratory, clinical, and environmental parameters. This monitoring establishes for each person an individual map of a "baseline state", deviation from which during continuous and situational measurement should serve as an alert and enable adaptation of a strategy or implementation of therapeutic measures. This preventive approach is then based on the measurement of parameters and not on someone's formulation of a request. The other innovation is that the disease state is no longer defined as deviation from a zone of normality established in a large control population including highly diverse phenotypes[91], but rather as a deviation from the person's own baseline state. Everybody becomes his/her own statistical reference. The first studies published recently following exploitation of UK Biobank (500 000 participants)

[89] Atul J Butte. Big data opens a window onto wellness. *Nature Biotechnology* 2017; 35: 702. Price ND. et al. A wellness study of 108 individuals using personal, dense, dynamic data clouds. *Nature Biotechnology* 2017; 35: 747-56.

[90] The Hu-PreciMED (Human Precision MEDicine) project, an industrial initiative, brings together all public and private stakeholders working in precision medicine and focuses on the health data of patients of the Health Data Hub to valorize clinical data with a view to improving treatments and diagnostic tools, and to develop new approaches in predictive and preventive medicine, while using the latest advances in big data and artificial intelligence. Forty-five public and private organizations have joined the project.

[91] All the observable, apparent characteristics of an individual. The phenotype is the resultant of genetic (genotype), behavioral, and environmental factors.

clinical, genomic, and brain imaging data are convincing in this regard[92]. In an extreme view, one could envision building a "digital twin" of the person, based on the model of what is already happening in robotics[93].

- Other innovations are foreseen notably in the interpretation of radiological images and imaging data[94], as well as for the management of patient care. Regarding the latter, we can cite prioritization in an emergency department[95] and improved response to stroke or septic shock[96], two situations in which the speed of treatment implementation has a major impact on prognosis. One can also expect improvement in the prediction of hospital readmissions[97].

3.1.2 Respect of ethical principles in exploitation of big data in the care relationship today

Schematically, in the framework of healthcare, data collection and exploitation are done with a defined aim, concerning an identified person – the patient – who expects a benefit, during an exchange with a health professional who cares for and takes decisions with the patient. In the new context that we are examining, while the medical profession today relies on the treatment of big data to improve patient care, there is involvement of other stakeholders, who often have no relation with health professionals and treat health not in the framework of care and medical ethics, but as a market. Through applications or connected objects that they alone produce, and social media that they control, these providers can credibly compete with care and research. But the storage of

[92] Since 2006, the UK Biobank in the United Kingdom has collected clinical, genomic, and behavioral data from 500 000 people. All the genomic data were published in 2017 and access to them was requested by 7000 researchers. See: UK Biobank debuts as a powerful resource for genomic research. *Nature Medicine* 2018; 24: 1792-4.

[93] michael grieves introduced this notion. The digital twin is the digital expression of the information of a physical system. This digital information is linked with this physical system throughout the system's entire lifecycle. The aim is to be able to access all information concerning a product, its requirements, behavior, etc., without possessing it physically. So, the concept of digital twin is divided into two spaces, one physical, the other virtual. Data flows across the physical space to the virtual space, and information flows from the virtual space to the real space and virtual subspaces (le monde 26 february 2018 - https://abonnes.lemonde.fr/les-cles-de-demain/article/2018/02/26/le-jumeau-numerique-est-un-interessant-moteur-de-l-innovation_5262662_4758288.html).

[94] launch of a project for a french artificial intelligence ecosystem dedicated to medical imaging. This system will be independent, notably of the american gafam (google, apple, facebook, amazon, and microsoft) and the chinese batx (baidu, alibaba, tencent, and xiaomi). Hosny a, et al. Artificial intelligence in radiology. *Nature reviews cancer* 2018; 18: 500-10.

[95] *hong ws, et al. Predicting hospital admission at emergency department triage using machine learning.* Plos one *2018; 13: e0201016.*

[96] liu vx, walkey aj. Machine learning and sepsis: on the road to revolution. *Critical Care Medicine* 2017; 45: 1946-7.

[97] Lynch CJ, Liston C. New machine-learning technologies for computer-aided diagnosis. *Nature Medicine* 2018; 24: 1304-5.

health-related data and these providers' use thereof poses a problem[98]. There is an ethical imperative to clarify roles.

3.1.2.1 from the patient's point of view: information and collection of consent

Consent is an essential feature of the relation between doctor and patient[99]. The patient must be informed of the approach used by the doctor to establish a diagnosis, determine the treatment, or ensure follow-up, and the patient must consent to this. The information on the approach used implies that the patient is informed of the use of treatment of big data, either to interpret radiological or imaging examinations or as a diagnostic aid. But the patient does not have to give specific consent to the use of personal data collected in the framework of this medical care, the doctor being subject to medical confidentiality and so to data confidentiality. In contrast, the question arises regarding subsequent use of these data to constitute a data collection or warehouse for clinical research, a frequent situation following hospital treatment. If consent is not required, the patient must be informed that his/her data may be used for research purposes. The patient can refuse this use and demand the return of the data. It is for data collected and stored outside care by health professionals that the questions become more acute regarding information and consent, the intelligibility of the treatment, and control over the fate of personal data, given that the use of connected objects can be very profitable in a care pathway. The question of technical reliability and of the protection of personal data in this context will be examined in section 3.4.3, which covers the exploitation of big data outside the care relationship.

Finally, in this context where everything is known, is predicted, and can and should be measured and announced, one may wonder about the role to be accorded to the right to not know.

[98] An emblematic example is provided by the breaches of ethical requirements during the collaboration between DeepMind and the British National Health Service (NHS). In 2016, DeepMind, the artificial intelligence system of Alphabet Inc. (Google) – which beat the best Go player – announced a collaborative project with the Royal Free London NHS Foundation Trust in designing an algorithm and a mobile application for the care of patients with acute kidney injury. But the data of nearly 1.5 million patients – a number disproportionate to the stated purpose – were transferred to DeepMind without information and without the consent of the patients concerned, testifying to a clear breach of ethical principles, revealed by journalists. Two other projects between DeepMind and the NHS are ongoing
Powles j, et al. Google deepmind and healthcare in an age of algorithms. Health technology 2017; 7: 351–67; hodson h. Google's new nhs deal is start of machine learning market place. New scientist 6 jul 2016; le partenariat entre google deepmind et les hopitaux londoniens juge non conforme a la loi. Le monde of 4 july 2017 (pixels).

[99] everyone, together with the health professional and in light of the information and recommendations the health professional provides, takes decisions concerning their health (article l.1111-4 of the public health code).

3.1.2.2 From the health professional's point of view

Respect of medical confidentiality. Medical confidentiality ensures the non-identification of the patient by third parties. Conventionally, the personal, direct, and exclusive relation set up between the patient and the medical profession allows the latter to ensure the protection of this confidentiality, provided only that the medical profession respects its own ethical principles. But in this era of the informatization of medical practices and of a necessary transmission of medical information, confidentiality is already weakened by the notion of "information shared" between the carers (via shared medical records, pharmaceutical records[100], or, soon, the patient's digital health space[101])[102] . But confidentiality is still further weakened by use of providers outside the medical community, by the disclosure on the internet of data that can secondarily become health data, and by the practice of health professionals using consumer digital services.

The responsibility of decisions concerning the patient. Whether it relates to the diagnosis or the treatment of the patient, the use of algorithms should be considered as an aid to human decision making, excluding any automation of the medical decision. But this new type of intervention first poses the question of the verification of the quality of the services offered by companies that operate in this market. The results provide by their software programs must prove useful and reliable, which brings us to the quality control referred to above. The American FDA (Food and Drug Administration) recently approved and authorized two artificial intelligence software programs that assist in the diagnosis of diabetic retinopathy and the interpretation of mammograms[103]. But these results must remain an aid that consolidates the diagnosis made by a health professional. Telling the patient of the diagnosis and the resulting decisions is the sole responsibility of the doctor (see RECOMMENDATION No. 4).

[100] The pharmaceutical file is a computer file created and consulted by your pharmacist, with your agreement. It lists the drugs delivered to you over the last 4 months, as well as ongoing treatments and dosages. The drugs figuring in the file may have been prescribed by a doctor or bought over the counter. https://www.service-public.fr/particuliers/vos rights/F16033.

See in particular the deliberation of the CNIL (No. 2017-285 of 26 October 2017) authorizing the Open-Health Company to implement treatment of personal data with a view to constituting from the data of retail pharmacies a warehouse of personal data for the purposes of research, study, or evaluation in healthcare.

[101] See note 75 and the final report (in French) on the digital revolution in health. https://solidarites sante.gouv.fr/IMG/pdf/masante2022_rapport_virage_numerique.pdf

[102] Law on the modernization of the French health system, enacted on 26 January 2016, art. L1110-4 and R1110-3.

[103] https://www.fda.gov/newsevents/newsroom/pressannouncements/ucm604357.htm. The approval will enable a family physician with an appropriate camera to make this diagnosis, facilitating the monitoring of diabetic patients, and early diagnosis. The FDA has also approved an artificial intelligence system for interpreting mammograms (https://www.fdanews.com/articles/189314-fda-clears-screenpoint-medicals-ai-system-for-reading-mammograms). Others will likely follow (eg, for the diagnosis of skin lesions).

The preserved care relationship. There must be a direct personal relation between health professionals and patients[104]. This is essential to the trust that is at the very heart of the care relationship. As they are in direct contact with patients, medical personnel can acquire a sufficiently accurate idea of each patient's personality and thereby understand what he or she needs in order to have a genuine understanding of the relevant health-related questions. It is worth recalling how the world of medicine and the world of new technologies differ in their vision of the human body. Traditionally, medical personnel seek to make a diagnosis through history taking, observation, auscultation, and palpation, and this is how the carer-patient relationship is established. What makes digital technologies possible is the power to access all this information when the patient is not present. The sustainability of the know-how of classic medical practice could be called into question if the decision-making assistance provided by the new technologies were to make less necessary, even unnecessary, the teaching and practice of observation of the human body. Jean-Michel Besnier has expressed this by saying that the health of the future is the "disenchantment" of bodies[105].
The risk would be that the treatment of big data determines a result by digging into the patient's data and comparing them with other comparable data and with scientific facts, without according any place to the patient him- or herself. That would also involve dependence on the machine, thereby appreciably reducing the substance of the human guarantee referred to in recommendation No. 4 above. Now, the patient cannot be reduced to a set of data to be interpreted, thus rendering it unnecessary to listen to and note the patient's experience. However useful data may be as an aid to diagnosis and in guiding treatment, they cannot replace dialogue. On the contrary, the use by health professionals of recent technologies must also aim to free up time for listening to the patient and for physician-patient exchanges, by simplifying the collection of relevant information. New technologies should empower patients to become greater stakeholders in their care pathway by allowing them ownership of their data, which is a condition of a fully responsible attitude (see **RECOMMENDATION No. 7**).

3.1.2.3 From the public authority's and the community's (or the health system's) point of view

We have seen that the development of precision medicine born from the exploitation of data could harm risk sharing. Such an evolution towards precision medicine is still only

[104] The importance of this relation is reiterated in the Livre blanc on physicians and patients in the world of data, algorithms and artificial intelligence, published in January 2018 by the Conseil national de l'ordre des médecins (CNOM; French Medical Association). In its recommendation No. 10, the CNOM recommends that the development of technical devices using artificial intelligence should tend towards an industrial market of decision-making aids and not a market that would impose on physician and patient alike a decision made by an algorithm that cannot be criticized or overruled.
[105] J.M. Besnier interviewed by Hugo Jalinière. *Science et Avenir*, 6 September 2015.

hypothetical, because it requires organization and infrastructure that is as yet not operational. But if this does happen, it would raise formidable ethical concerns.

The main concern, as we have just seen, would be that the citizen is reduced to a set of data which supposedly depicts the citizen holistically, as the CCNE has already emphasized in its Opinion 98 on biometrics[106].

But such an evolution would, furthermore, heighten the underlying tension between the expectations of the individual and those, guided by economic concerns, of society. If the analysis and systematic monitoring of the health, environmental, and behavioral data of everyone, ill or healthy, became the rule, two ethical major principles would be threatened: on the one hand, individuals' freedom in making life choices and their right to know or not to know, notably when there is no effective curative or preventive strategy to fight against a detected vulnerability; and, on the other hand, the risk of a loss of sharing and solidarity in the treatment of disease for people who do not implement the measures recommended to them. The prediction of medical risk would end up determining health policies and imposing their application.

Another risk, common to all contexts of the exploitation of big data in health, is that of a loss of autonomy and sovereignty for our country (see also the end of 2.2.2, and 3.4.2). This loss arises because technologically we are lagging behind in the fields of data hosting and the processing of data, the volume of which is skyrocketing with the new techniques of genomic analysis and with the policy of systematic digitization of radiological and imaging examinations practiced by healthcare establishments[107]. The current places for storage of health data are subject to law 2016-41 of 26 January 2016 on the modernization of our health system. This law institutes a state-supervised certification procedure entrusted to public or private bodies. The trend then is towards the sharing and migration of data to platforms managed by private or public service providers. The American giants (Amazon, Microsoft) have set up in France and in November 2018 Microsoft obtained certification as a web hosting service for its four centers in France.

Faced with the technological challenge to national and European sovereignty posed by the storage, sharing, and treatment of big data in healthcare, the CCNE recommends the development of shared and interconnected national platforms. Such platforms, open to public and private stakeholders according to modalities yet to be defined,

[106] "The combination of all this information provides an almost infallible result and encloses each of us within a well-defined framework. Society seems to be content with characterisation of a person by a set of data assembled in this manner." Biometric data "may reduce human beings to an accumulation of data and cartographic criteria, paradoxically at a time when biology is moving away to some degree from the reductionist and analytic approach and is seeking to apprehend systems holistically through an integration of all the properties of an organism or of a life form (integrative biology)." CCNE Opinion 98: Biometrics, identifying data and human rights.

[107] for example, the volume of medical imaging is increasing by 20% to 40% year on year.

should enable our country and Europe to preserve strategic autonomy and to avoid losing control of the riches constituted by the data, while promoting controlled sharing, which is indispensable to the efficacy of healthcare and of medical research **(RECOMMENDATION No. 10)**.

This is the rationale for the health data platform that should be created in the framework of the bill on the organization and transformation of the health system[108], following the recommendations of the Villani report[109].

3.2 Exploitation of health data in the framework research protocols

The accumulation of data collected in highly varied contexts, sometimes in real time, and which can be reused, weakens the frontier between care and research. This data accumulation has a specificity in line with its aim: advances in medical understanding, for the public good, and not preventive or curative treatment necessitated by the state of a given patient.

The exploitation of big data (pooled in immense local, national, or international databases) for research helps identify groups of people who share the same characteristics derived from cross-referencing of genomic, clinical, and imaging data, etc. These correlations – obtained in certain cases without a prior hypothesis – yield new lines of research on disease mechanisms (research study of causality) and the definition of early diagnostic markers, prognostic indicators, or treatments.

In these databases, the individual is usually not identified by name, and in a way is "erased." While there is always a person who is the source of primary data, that person has no link with the people who exploit the data and generally remains anonymous. Unlike the clinician who is face to face with the patient, the researcher using a large database, or the data curator or scientist who manages the data, do not know who owns them. Two risks result from this: not seeing the data holder as a human being and not ensuring genuine anonymization (or pseudonymization) of the data.

The aim sought is the public good, the point being to advance understanding or the establishment of public health measures from which the individual could benefit, often indirectly and uncertainly. How then does this relate to the ethical principles enunciated in part 2, from the point of view of the data subject, the data scientist, the researcher,

[108] See title III of the bill "Développer l'ambition numérique en santé " introduced in February 2019. https://solidarites-sante.gouv.fr/actualites/actualites-du-ministere/article/presentation-du-projet-de-loi-relatif-a-l-organisation-et-a-la-transformation. See note 73.

[109] See Health Data Hub report, preparatory mission. 12 October 2018. https://solidarites-sante.gouv.fr/ministere/documentation-et-publications-officielles/rapports/sante/article/rapport-health-data-hub-mission-de-prefiguration.

and the community that will draw conclusions therefrom? It is in this context that several lines of thought are at odds and there is discussion of new forms of consent and access to data.

3.2.1 The ethical issue of data sharing

The sharing of (or access to) data is the very foundation of the exploitation of big data. These data are only of interest if they are accessible to the greatest possible number of researchers or clinicians and if they can be cross-referenced with clinical and environmental data, often stored in the same cloud or repository or warehouse.

This sharing results in inevitable tension between the risk of under-exploitation of the data that may imperil research conducted for the public good and that of sharing that is too extensive for the basic rights of the person to be respected. From this difficulty of ensuring both broad access to the data and effective protection of individual rights[110] are born reflections on collective rights concerning health data[111] or the "right to science,"[112] which we shall consider in section 3.4. This difficulty is heightened by the absence of international standards of good practice and governance[113], even though the European regulations constitute an important advance[114].

One should consider the ethical principles as stated by the three bodies involved in the research process: the person who owns the data and is concerned by their treatment (the question of consent and of data security), the governance of the databases (who controls security and access?) and the researcher, who must be vigilant in the use of data for research.

3.2.1.1 The person's point of view: what role for consent/new forms of consent?

The very characteristics of the data described in section 1.3 create two difficulties in the context of research. The first difficulty stems from the absence of a clear definition of this activity of "research," a term which is not defined by the GDPR, whereas the objective of scientific research can be a legal criterion in the processing of health data or in the reuse of data, independently of consent. The second difficulty arises because of the overlap between care and research. A frequent situation (recognized by the GDPR) is

[110] Joly Y, et al. Are data sharing and privacy protection mutually exclusive? *Cell* 2016; 167: 1150.

[111] Bourcier D, Filippi P. Vers un droit collectif sur les données de santé. *Revue de droit sanitaire et social* (Dalloz revues) 2018: pp.444-50.

[112] Knoppers BM, Thorogood AM. Ethics and Big Data in health. *Current Opinion in Systems Biology* 2017, 4 : 53-7.

[113] In this regard it is interesting to note the recent condemnation in China of companies that break the law on the sharing of data (*Nature* 15 November 2018, page 301).

[114] Stein L, et al. Data analysis: Create a cloud commons*. Nature* 2015; 523: 149.

imprecision in the purpose of data processing when explained to the person being informed and – if required – in the person's initial consent to data collection and treatment[115]. It is in this context of imprecision in consent that we see the importance of the quality of the information provided to the person and hence of the relation of trust with the contact person, and that there must be guarantees of transparency, respect of ethical requirements applicable to scientific, in particular medical, research, and commitment to regular communication regarding use of the data.

It is difficult to transpose the traditional model of consent, which was designed within the framework of a relation between an individual (sometimes a patient) and a stakeholder (researcher/doctor) working for a specific project (purpose), limited in time, collecting specific data consistent with the project, governed by stable regulations. This model is no longer appropriate for such a data flow, nor for generalized sharing, while the possibilities of exploitation are myriad and no precise information on the fate of these data is available at the time they are collected. It is not always realistic to envisage recontacting the initial data subjects to request consent for the use of the data in another project. The GDPR dispenses with this obligation if the purpose of the other project is not incompatible with that of the initial project (art. 6.4 of the GDPR).

Henceforth, other modes of consent are discussed in the context of research:

- broad consent, which does not specify a precise purpose, but just a field of applications, which may be varied. The participants can consent to the use of their data by a biobank, which in turn guarantees security and controlled access via a committee of governance (in this regard, see the organization of the type of consent of the UK Biobank[116]);
- consent with options, in which the data subject selects the areas of research for which he/she authorizes use of the data (for example, cardiovascular research, but not cancer research, or consent for the use of genetic data distinct from that accepted for conventional laboratory data);
- dynamic consent (often linked to broad consent), which makes use of digital technologies; the holders of the data are considered as participants in the research. They alter or update their consent as a function of the new purposes, about which they are informed via a dedicated website used for information exchange;
- opt-out consent, in which the data can be used except in the event of refusal, and it is this refusal (and not approval) that is indicated by the holder of the data.

If the data are completely anonymized, they cannot be qualified as personal data and

[115] Working group, article 29: guidelines on consent under ruling 2016-679, final version of 10 April 2018 (in French). ttps://www.cnil.fr/sites/default/files/atoms/files/ldconsentement_wp259_rev_0.1_fr.pdf.

[116] https://www.ukbiobank.ac.uk/the-ethics-and-governance-council/. See also the annual report of the Governance Council. https://egcukbiobank.org.uk/sites/default/files/UKBEGC_Review2016_2017.pdf

are not covered by the protection of the GDPR or by that of the current version of the data protection act. Access to these aggregated clinical data can thus be completely free. But then the question arises regarding the irreversible nature of the anonymization of the data. Not only can later re-identification no longer be excluded now, but anonymization strips the data of a great part of their usefulness, because it obligates the deletion or scrambling of part of the useful information[117]. Can consent then be ethically dispensed with?

When consent is required, it cannot be unique and standardized, for whatever research context in which it is required. The ethical imperative is that consent be adapted to each specific situation, that the information given in return on the research be available, constantly updated, and clear and fair, so as to establish and justify a relation of trust between the holders of the data and those (curators and researchers) who have access to and treat these data. This information should include both open access scientific publications and letters of information written in clear and plain language for nonexperts.

Remember that consent is not the only basis for the legality of the processing of personal health-related data and that, in the many situations where consent is not required, concise, transparent, intelligible, an easily accessible information is, in contrast, required. The question arises as to the management of this complex information by the person in charge of data processing: prioritization of information, channels to be used for delivery – human interactions or digital tools –, possibility for the data subject to determine in advance the significance and consequences of data treatment, particularly as the he/she may be in a vulnerable situation. While the characteristics of this information have been detailed by the Article 29 Working Party[118], in what measure will they really be monitored and how can we be sure that they have been delivered to and understood by the person concerned?

Another difficulty, already referred to above (section 2.1.3), is if the data subject withdraws the data for no legitimate reason. Apart from the fact that a discretionary withdrawal may not be technically possible, it could also be unethical. The quality of the

[117] De-identification and pseudonymization are rather complicated words that are virtually synonymous in the sense that in both cases the true identity of the person (family name, first name, social security number…) is absent or masked. The use of a pseudonym means furthermore that the true identity has been replaced by a conventional identifier (often a number), which in a given context always designates the same person, so as to allow longitudinal follow-up (tracking). The allocation of a pseudonym by a process that prevents the data manager from obtaining the name of the person concerned (irreversible encryption, for example) was often called anonymization and still is sometimes, but we know now that the dataset thus modified is not necessarily anonymous. This is why it is better in this case to speak of pseudonymization (André Loth, DRESS, Solidarité Santé, July 2015).

[118] Article 29 Working Party. Guidelines on transparency under Regulation 2016/679 – version of 11 April 2018 (in French). https://www.cnil.fr/sites/default/files/atoms/files/wp260_guidelines-transparence-fr.pdf.

research would be impaired by a risk of bias and of flawed conclusions if certain categories of data were removed from the research project.

Given the diversity of uses and of contexts regarding big data, notably in research, it is necessary to reflect upon the notion of consent to the treatment of big data. This reflection should relate to consent and how it is collected, so as to ensure a lasting balance between the respect of human rights and the dynamics of uses. This reflection should also foster the public debate on ethical recommendations and will allow periodic updating of the law
(RECOMMENDATION No. 3).

The modalities sought assume a relation of trust between the holders of the data and the technicians, engineers, and researchers who have access to and treat these data. It is essential that the holder of the data be informed of the modalities whereby the supervisory authority fulfills its function of trusted third party. In this way, a threefold ethical requirement is ensured:
- rigorous and transparent evaluation of the value of the research, which must help to increase knowledge in healthcare to the benefit of everybody (notion of public good);
- sharing of information on the progress of research with the participants, according to diverse modalities;
- assurance of the security and traceability of the data, and of the absence of malicious use
(RECOMMENDATION No. 11).

3.2.1.2 Governance: the importance of access to data guaranteed by the institution

The trust accorded by the holder of the data is also based on how the control of access to databases is safeguarded. It is ensured by a third-party committee, often a committee of governance, which checks compliance with ethical principles and regulation by the public or private institution, which is the data repository. This limits the risks of misuse. Access to data, for research or public health projects, is more often called controlled or restricted. A committee evaluates the relevance of the research project, its adaptation to the consent of the participants, compliance with ethical and security rules, and commitment to divulge the results to the community. An intermediate procedure is that of declared access, where the researcher or the clinician gives only their identity and agrees to conform to the terms of use of the data (for example, conformity with the reference methodologies published by the CNIL for access to the INDS [Institut national des données de santé] in France[119]). These different modes of control are more or less

[119] The CNIL published, on 13 July 2018, three new reference methodologies that regulate the treatment of personal data at the end of studies, evaluations, or research not involving human subjects. These are simplified procedures for access to personal health data that avoid a request for authorization (INDS-

adapted depending on the use of the database (basic or clinical research), the type of data, and the status of the applicant. Today, the model of controlled access is the surest. The respect of ethical principles requires that governance of sensitive data be free of ambiguity and rigorous insofar as certain data of international bases, notably genomic databases, can sometimes be hosted and hence accessible via a cloud managed by a third party (Amazon, for example). The development of privacy enhancing techniques could also provide a response[120].

In France, examples of controlled access are those of access to the INDS, to hospital data warehouses, and to genomic databases (see France Médecine Génomique). The UK Biobank mentioned above is another example. There is also the creation in France, as inscribed in the February 2019 bill on the transformation of the health system, of a health data platform (Health Data Hub), a project undertaken by the Villani mission. It will bring together major French research bodies□ Inserm, CNRS, and Inria□and numerous public and private partners with a view to providing within a highly secure framework a working space for algorithmic learning. The project proposes governance of the platform, principles for intervention, and legal and operational modalities to manage the sharing of data.

3.2.1.3 The transdisciplinary chain of participants-researchers: ethical reflections

The fate of data escapes the initial holder's control because of the intervention of multiple processes and stakeholders and of a technology that the holder most often cannot understand. So, the holder's consent will rest on a relation of trust and reciprocity and not on a restrictive definition of purposes, established *a priori*. Researchers must deserve this trust and one should not underestimate the public's resistance to certain aspects of the sharing of data in terms of research[121].

- *A person is the source of the data.* Unlike the doctor who is face to face with the patient, the researcher using big data has no contact with the holders of the data and knows nothing of them. The danger is therefore that the researcher considers the data as a simple research instrument, forgetting that they are linked to people who are exposed to a risk of harm. This renders all the more important the researcher's knowledge of the quality of the information given to the holders of the data and the terms of the consent that the holders have given. When the consent given is broad, even more attention should be paid to the risk of lack of respect of private life, to the security of the data, and to flaws enabling re-identification. The researchers must also ensure that the data they are using have not

CEREES-CNIL) without, however, avoiding the need for the study, evaluation, or research to be in the public interest. As a trade-off for this simplification of formalities, the person responsible for data treatment undertakes to respect several obligations.

[120] Erlich Y, Narayanan A. Routes for breaching and protecting genetic privacy. *Nature Reviews Genetics* 2014; 15: 409-21.

[121] Majumder MA, et al. Beyond our borders? Public resistance to global genomic data sharing. *PLoS Biology* 2016; 14: e2000206.

been obtained unethically. An example in this regard is the company 23andMe, which sold the genetic data of its five million clients to the British pharma giant GlaxoSmithKline. 23andMe asserted that 80% of its clients had agreed that their data could be used for medical research purposes[122].

- As a "*producer*" of new data, which could be used to make healthcare decisions, the researcher is responsible for the quality of the sequence of steps needed to obtain the data used, for the absence of bias that may affect data collection, for the questions on which treatment is based, and for the evaluation of the algorithm. Big data bring a new transdisciplinarity, so the researcher in biology who wants to use big data must be trained in algorithmic analyses or collaborate with data scientists, in the established framework of the rules that govern these databases.

- *Mathematicians/computer scientists* constitute a third party between the person who provides the data and the researcher or doctor who uses them. Like the researcher, they must be familiar with the ethical issues related to the exploitation of big data and be trained in the protection of people's private lives. They must be informed of the consequences of mistakes and imprecisions in data collection, in the taking into account the number of data, their heterogeneity, and the diversity of contexts of data collection (patient, researcher, etc.). The consequences are not negligible and can have harmful and even deleterious effects on machine learning systems (see RECOMMENDATION No. 5).

- *Organization and governance of data, responsibility of stakeholders.* Transdisciplinarity increases the importance of the organization and governance of data (collection, annotation, hosting) so that these data promote medical discoveries and can be reused usefully. This governance must be ensured under the responsibility (accountability) of the stakeholders, who must put in place mechanisms and internal procedures that show that rules relating to data protection are followed. In a way, this is a contract of trust between the people who agree to provide their data and the organization that oversees access to the data and the fate of the data. One possible answer is the concept of privacy by design, or protection of private life from the conception of the project, a concept created in the United States in the 1990s and included in the GDPR[123]. Each program treating personal data must guarantee from its conception and at each use, even if not planned for at the outset, the highest possible level of data protection.

[122] 23andme's pharma deals have been the plan all along. Wired, 8 march 2018 (https://www.wired.com/story/23andme-glaxosmithkline-pharma-deal/).

[123] Article 25 – Data protection by design and by default.

3.2.2 Sharing of data in research with private drug companies and internet providers

- The pharmaceutical industry considers that the exploitation of big data can substantially expedite research and development and support a model that counters rising costs and the absence of blockbuster drugs[124]. This aid can relate to models that predict the action of new drugs, to the use of clinical trial data, to more adapted and faster selection of patients participating in these trials, and to faster and precise real-life detection of the side effects of new drugs[125]. The real-life aspect is important because clinical trials are generally conducted in a restricted and highly targeted population, which does not allow anticipation of unexpected harmful effects that can result from the wide distribution of a drug product. When private providers are associated with the research, to access health data and personal data collected and stored by the public authority they must make a commitment not to divulge the data and not to use them for other purposes. A control is needed to check that the commercial project or corporate projects (big pharma) fulfill these requirements. This contractual basis is all the more important as research today is internationalized and so brings together stakeholders subject to unequally restrictive legislation.

- *Data from internet platforms in research:* Social media (Facebook, Twitter) and internet platforms that share health information intended for patients (eg, PatientsLikeMe created in 2004 in the United States, or Carenity in France[126], built on the same model) have become an important source of health information: for care and alerts regarding adverse drug reactions, but also for clinical research (targeting for the purposes of recruitment of patients), or surveys in the framework of policies of prevention or health monitoring. The value of these data comes in part from the fact that they relate to "real life" – because they are collected outside the usual treatment modalities. The importance of these real-life data is now recognized[127]. More and more research programs directly ask the owners of social media accounts for their participation – with their consent – in clinical research studies. Examples include the World Diabetes Distress Study[128],

[124] AI-powered drug discovery pharma invest. *Nature Biotechnology* 2017; 35: 605.

[125] Les données de vie réelle, un enjeu majeur pour la qualité des soins et la régulation du système de santé. L'exemple du médicament. Bernard Bégaud, Dominique Polton, Franck von Lennep. Report commissioned by the minister of health. May 2017.

[126] Carenity (created in 2011, over 15 000 users) facilitates contacts between men and women concerned by the same diseases by making available social media free of charge: groups of friends, news feeds, discussion forums, private messaging, etc. The data can be used in the framework of public or private research programs. All the data are aggregated. https://www.carenity.com/

[127] Bégaud B, Polton D, von Lennep F. Les données de vie réelle, un enjeu majeur pour la qualité des soins et la régulation du système de santé. Report commissioned by the minister of health. May 2017. https://solidarites-sante.gouv.fr/IMG/ pdf/rapport_donnees_de_vie_reelle_medicaments_mai_2017vf.pdf

[128] Fagherazzi G, et al. Étude mondiale de la détresse liée au diabetes : le potentiel du network social Twitter pour la recherche médicale - *Revue d'épidémiologie et de santé publique*. Doi : 10.1016/j.respe.2018.04.002. The World Diabetes Distress Study (WDDS) is an international research project (in which Inserm participates) on diabetes (type 1 and type 2) designed to identify distress markers linked to diabetes, poor quality of life, and the risk of complications. Distress linked to diabetes is defined by the

the recognition of depressive episodes via exchanges on Facebook[129], the monitoring of mental diseases on Twitter[130], the therapeutic follow-up of Parkinson's disease[131], and the prevention of suicide[132].

Several studies argue for greater opening of these data. The recent CERNA report on digital sovereignty recommends that platforms collecting big data (eg, GAFAM [Google, Amazon, Facebook, Apple, Microsoft] and BATX [Baidu, Alibaba, Tencent and Xiaomi]) be obliged to open these data to scientific purposes under strict conditions of ethics, integrity, and scientific ethics. This is also the conclusion of a white paper by the Healthcare Data Institute: give stakeholders of public research simplified and free access to databases containing data made public on social media by their users[133]. These media induce a dialogue between researchers and patients, enabling the latter to raise health research topics.

However, the dissemination and use of data outside a secure institutional framework poses important ethical questions for the protection of patients, notably the respect of the limits of their consent to the dissemination, hosting, and reuse of these data. One may also wonder about a possible questioning of their free will when taking medical decisions concerning them. It is possible, by analyzing behavior on social media, to predict certain social characteristics[134], which could lead to unacceptable stigmatization.

The CCNE considers that it is necessary to facilitate the sharing of health data for the needs of research. It is notably in favor of allowing, for research protocols with strictly defined purposes and while respecting the rights of the people who consent to provide data, researchers to access data collected on the internet or from social media by platforms governance of which is controlled (**RECOMMENDATION No. 12**).

burden of stress, fears, or emotions related to the everyday management of diabetes. It is considered as the most important psychosocial health factor in the management of diabetes.

[129] Eichstaedt JC, et al. Facebook language predicts depression in medical records. *Proceedings of the National Academy of Sciences of the USA* 2018: 115: 11203-8.

[130] Reece AG, et al. Forecasting the onset and course of mental illness with Twitter data. *Sci Rep* 2017; 7: 13006; Reece AG, Danforth CM. Instagram photos reveal predictive markers of depression. *EPJ Data Science* 2017; 6: 15.

[131] *Gravitz I. Technology: monitoring gets personal*. Nature *2016; 538: s8–s10.*

[132] Rous J, et al. Smartphones, sensors, and machine learning to advance real-time prediction and interventions for suicide prevention: a review of current progress and next steps. *Current Psychiatry Reports* 2018; 20: 51.

[133] Livre blanc : les réseaux sociaux et la santé : un enjeu pour le suivi des patients et la recherche scientifique.
https://healthcaredatainstitute.com/wp-content/uploads/2015/02/livre-blanc-hdi-2018-print_bd-bd.pdf

[134] *Kosinski m, et al. Private traits and attributes are predictable from digital records of human behavior.* Proceedings of the national academy of sciences of the usa *2013; 110: 5802-5.*

3.3 An emblematic example at the line between care and research: genomic data

Among health data, "personal data relating to inherited or acquired genetic characteristics", according to the definition of the GDPR[135], are characteristic in several regards: they pose all the questions referred to above: volume and storage of data, risk of uncontrolled dissemination, possibility of identification, loss of confidentiality and hence of security. They also illustrate the questions raised by sharing of data for the benefit of health and their cross-referencing, often transnational, which alone allows the acquisition of meaning, but which raises legal and ethical questions that are not only technical but also difficult. Genomic data reveal the tendency, which is growing above all in the United States and Asia, to want genetics to no longer be exclusive to the medical world[136] and to academic circles. The data, above all in North America, acquire a market value and are of interest to the economic sector: they are essentially exploited by biotech companies, for the purposes of family genealogy, but also for health and research.

From a legal point of view, several texts regulate genetic data: the civil code, bioethics laws (specifically concerning the conditions of genetic testing), the data protection act, notably the version of the law of 20 June 2018[137]. This law identifies genetic data as sensitive (which is the general case of data concerning health), which leads to a ban on treating them (article 8 of the updated LIL [data protection law]). There are, however, some exceptions[138], notably concerning research and health, if the person has given

[135] "... 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" (GDPR, article 4) "in particular chromosomal, deoxyribonucleic acid (DNA) or ribonucleic acid (RNA) analysis, or from the analysis of another element enabling equivalent information to be obtained" (GDPR, recital 34).

[136] Sharon T. The googlization of health research: from disruptive innovation to disruptive ethics. *Personalized Medicine* 2016; 13: 563-74.

[137] See notably article 75 of ruling No. 2018-1125 of 12 December 2018 in application of article 32 of law No. 2018-493 of 20 June 2018 relating to protection of personal data.

[138] Extract from article 8 of the data protection action (LIL) modified to comply with the GDPR (June 2018).
I- It is prohibited to treat personal data that reveal the purported racial or ethnic origin, the political opinions, the religious or philosophical beliefs or the trade union membership of a natural person or to treat genetic data, biometric data with a view to uniquely identifying a natural person, or data concerning health or data concerning the sex life or sexual orientation of a natural person.
II. - Insofar as the purpose of the treatment requires it for certain categories of data, the following are not subject to the ban [] provided for in I:
4. Treatments of personal data made public by the person concerned;
6° Treatments necessary for the purposes of preventive medicine, medical diagnoses, the administration of care or treatments, or the management of health services implemented by a health professional or by another person whose profession requires medical confidentiality.
11° Treatments necessary for public research in the sense of article L. 112-1 of the research code, implemented under the conditions provided for in point 2 of article 9 of the ruling (EU) 2016/679 of the

free, informed consent.

3.3.1 From genotype to phenotype[139]: a change of logic due to advances in genome analysis technologies

The sequencing of the human genome in 2001 played a major role in our understanding of the genetic determinants of diseases. It underpins a strategy which correlates genetic variants and vulnerability to certain diseases, a strategy involving the analysis of some families followed by studies of genetic association in thousands of participants (patients, relations, and controls). These so-called genome-wide association studies (GWAS)[140] require the pooling of data via consortiums and databases[141]. Even before the terminology designated them as big data, these studies already related to cohorts of thousands of people. The aim was to determine the regions of the genome (genes or regulatory sequences) influencing people's vulnerability to a disease. The technological revolution of new-generation sequencing has made almost routine the analysis of the whole genome, or more frequently of the whole exome (the sequencing of only those parts coding for genes), which until recently seemed to be a chimera. Large-scale sequencing is made possible by its decreasing cost (less than 800 euros, early 2019) and by the considerable increase in data storage and treatment capacities[142]. The result is today's expanded use in oncology, for the identification of tumor mutations that could be therapeutic targets in a given patient. Another field of intervention concerns the clinical exploration of new genes, the sites of rare mutations, in the perinatal period, so as to limit misdiagnosis. Finally, the exploitation of big data has recently shown that as a complement to the analysis of rare variants, the risk resulting from the accumulation of frequent variants in the general population could be estimated by a "polygenic score"[143]. This score can represent a major clinical marker and so it is highly likely that genome

European Parliament and Council of 27 April 2016 mentioned above, after a considered opinion published by the French Data Protection Authority, rendered according to modalities provided for in article 28 of the present law.

[139] The genotype designates an individual's genetic characteristics, determined by the analysis of his/her genome. The phenotype designates an individual's physiological or pathological state, such as the expression of a disease. For example, a mutation in the DNA (genotype) will result in the development of a disease, which represents the phenotype.

[140] They use the DNA microarray technique and very large-scale analysis of identified human genetic polymorphisms - the SNPs, or single nucleotide polymorphisms.

[141] A database is a structured and organized collection of data that allows the storage of large amounts of information so as to facilitate its use. There are many databases, depending on the nature of the data (raw or annotated) and on whether or not they directly supply the data and so are responsible for their quality.

[142] The data produced by high-throughput sequencing are more extensive than any data ever produced in the past. For example, the plan France Médecine Génomique 2025 provides for the production of several dozen petaoctets (Po) of data per year within 5 years. Note that all the files for a human genome (3 billion letters A, T, G, C) represent 300 gigaoctets (and 20 gigaoctets for an exome).

[143] The polygenic score represents the set of thousands of variants detected in certain regions of a genome and whose combination can give a reliable indication of the risk of developing a disease.

analysis, in a few years, will be part of the care pathway of all diseases.

We are witnessing above all a change in the logic of the approach, which henceforth goes from a genotype to a phenotype, from a genotype to the prediction of a disease, and no longer only from the disease to the genotype. This highlighting of risk evaluation is what underpins the constitution of large genomic databases, often supported by states and public institutions[144].

A major fact is that genomic data are no longer limited to the conventional medical sphere, governed by medical ethics or by research bodies. Genomic data now circulate on the internet and social media with little protection since the advent of direct to consumer tests. For ten or more years now, companies have offered clients full DNA analysis supposed to reveal all genetic vulnerabilities. They deal directly with clients, without the intervention of doctors or medical institutions. In 2018, one company offered genome sequencing for 199 US dollars.

The CCNE, in its Opinion 124 and, more recently, its Opinion 129, has addressed the ethical questions posed by the expansion of genetic testing in the medical field. Specifically, the CCNE considered ethical questions linked to the collection, storage, and processing of genomic data. For some years now, genomic data have been associated with other health data in large projects, essentially in North America, United Kingdom, and, more recently, Asia[145]. In France, where the cohorts were smaller, Aviesan[146] unveiled in 2017 the plan France Médecine Génomique 2025, the aim of which is to provide France with a medical and industrial sector with a view to introducing and developing precision medicine in the care pathway.

3.3.2 Big data in genomics: a challenge for public health and research

[144] Examples include the UK Biobank (500 000 people), the 100K Wellness Project and the All of Us Research Program (one million people) in the United States, in Asia the China Kadoorie Biobank funded by the Wellcome Trust (515 000 Chinese), and others. The NHS (National Health Service of the United Kingdom) and the Ministry of Health also encourage people to pay for analysis of their genome (Genomics England), with guarantees of reliability. Their data could be made available to researchers (Life Sciences Sector Deal 2, December 2018, UK government).

[145] Examples include the Precision Medicine Catapult Centre (1 billion £) in the United Kingdom and the Precision Medicine Initiative (250 million $) in the United States.

[146] Created in April 2009, Aviesan (national alliance for the life and health sciences) comprises the large stakeholders in the life and health sciences in France (Inserm, CNRS, Inra, Inria, IRD, Institut Pasteur, Conference of University Presidents).

It is probably in genomics that the use of data has enabled the most spectacular advances in understanding and the most notable improvement of patient care. The identification of the causes of diseases, the analysis of the genetic diversity of a category of patients for the establishment of new sub-classifications, and the exploration of mechanisms are determinant stages in discovering new effective treatments. Oncology was the pioneer in this regard, but this approach is now applied to other diseases or deficiencies.

The exploitation of big data in genomics has major consequences both immediate and future. Genomic data extend to public health, patient care, and research, illustrating the close link between basic research intended to shed light on nature and applied research designed to improve public health. As Louis Pasteur said, "There are no such things as applied sciences, only applications of science". But the exploration of these genomic data emphasizes also the alignment of the public and commercial sectors, because its development in healthcare depends on sequencing tools, generally provided by private for-profit companies, and the potentially valuable results represent an important part of the health market.

Databases: indispensable cross-referencing

Analysis of the whole genome or exome identifies several thousand or million variants. A small number – mutations that cause monogenic diseases – have a known and validated significance. The contribution of the other variants is currently unknown; it may be nil, small, or associated with a risk. Certain associations are only detected by the analysis of a large population of patients and controls. There are, nonetheless, statistical methods that aggregate the cumulated risk of the presence of these variants and give a "polygenic risk score" for a given disease and for a given individual. These methods are expanding greatly. For certain diseases, a high polygenic risk constitutes a risk as significant as the presence of a rare mutation[147].

While these genetic methods are increasingly efficient, we should remember that the information yielded by analysis of the phenotype is at least as important as the genomic information and that it is the cross-referencing of the two types of data – genomic and phenotypic – which best characterizes the physiological state of an "individual" (CCNE Opinion 124).

Cross-analysis of at least three types of data□clinical, genomic, and environmental□is the most relevant in providing the best diagnosis and in calculating the patient's clinical trajectory.

[147] Khera AV, et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nature Genetics* 2018; 50: 1219-24.

These data can be collected in a large number of individuals and stored in huge databases containing the data of several hundred thousand people. This allows identification of correlations between the variants and clinical or biochemical characteristics. From these correlations are deduced biomarkers useful for the diagnostic management of groups of people/patients, indications concerning the causes of diseases, and targeted therapeutic perspectives.

This is what is done in the United States and in China with millions of participants and in the United Kingdom with hundreds of thousands of participants, and what is proposed on a smaller scale by the France Médecine Génomique plan.

There are many national and international genomic databases[148]: hospital establishments (eg, Institut Curie, Institut Gustave Roussy), national data platforms (UK Biobank, Genomics England and the 1000 genomes emanating from the public sector National Health Service [NHS], the All of Us project, China Kadoorie Biobank, and, in France, the database planned for the France Médecine Génomique 2025 plan); or internationally (eg, the GENIE initiative[149] [under the aegis of the American Association for Cancer Research], the Global Alliance for Genomics and Health[150], ICGC [International Cancer Genome Consortium], and the databanks of biotechnology start-ups[151] or of private firms (eg, 23andMe, Helix, etc.). Much debate surrounds the legal and ethical questions posed by the organization of international databases and the sharing of data. Another challenge is posed by the heterogeneity of the different stakeholders who sometimes share these data, notably those from biotech companies, as this blurs the boundaries between the public and business sectors.

These approaches require considerable financial, human, and computing resources and highlight economic and sovereignty issues. Sharing is essential given the cost of acquisition.

The aim of the France Médecine Génomique plan is to put in place by 2025 the coverage by genomic medicine of all patients (with cancer) in France. This implies taking charge by 2020 of approximately 235 000 genome sequences a year.

[148] See also note 140. Cook-Deegan R, et al. Sharing data to build a medical information commons: from Bermuda to the Global alliance. *Annual Review Genomics Human Genetics* 2017; 18: 389-415.

[149] GENIE is developing a registry that allows the linking of genomic data (deidentified) in oncology with the clinical data of 50 000 patients treated in France (IGR), but also in the establishments of other countries.

[150] The Global Alliance comprises 800 people belonging to 400 organizations (including Inserm) in 70 countries, which gives an idea of the challenge of governance acceptable to all parties.

[151] Including, Arivale, Human Longevity, Verily, Amgen's deCODE Genetics, Regeneron, and iCarbonX.

3.3.3 Are genetic data like other health data?

The great decrease in the time needed for analysis and in the cost of sequencing (a human genome can be sequenced in 40 hours for under 800 US dollars) and the increasing power of calculation and the interpretation of sequences mean that today the genomic sequence (relating to the whole genome or exome or panels of genes) is commonplace information as easy to obtain as a laboratory finding[152], such as cholesterol level or ultrasound imaging. In a few years, the genomic sequence will probably be included in shared medical records (henceforth "shared" and no longer "personal" medical record) and soon in the digital health space, a veritable digital health record and instrument for the coordination of care.

High-throughput sequencing data will be the largest volume of data produced in healthcare.

The question relates to what has been called genetic exceptionalism, in other words what distinguishes genomic data from other health data and could in this way justify a specific treatment.

- First and foremost, there is the *unique character* of the genomic sequence: each of the roughly seven billion people on Earth has a genome that is not only unique, but invariant. The specificity relates to the existence of multiple genetic variants, comparison of which is specific to a given person (approximately 3 million variants distinguish two individuals). Which is not the case for other laboratory data (several individuals can have the same red blood cell count, or the same cholesterol level), or for imaging results. This DNA sequence is therefore "identifying", in the same way as fingerprints, and it remains so throughout life since it is invariant. It is this characteristic that explains the legal interest (notably via genetic fingerprints[153]). Lastly, insofar as the genome is passed on to offspring, all information deduced from the sequence concerns not only the person, but also his/her family. The question then arises – more so than for other data – of informing the person of the results obtained. It is this identifying characteristic which creates the greatest risks in the exploitation of genomic data, above all the exploitation of big data: the sequence alone cannot make the link with a given person, but cross-referencing with other data greatly facilitates this link, which explains the compelling need for strict respect of confidentiality. Then there is the question of the legitimacy of a particular regulation. Ruling No. 2018-1125 of 12 December 2018 added article 75 to the data protection act of 6 January 1978. This article, which came into effect on 1 June 2019, reiterates the need

[152] Over 20% of the data from sequencing of the genome and the exome will be used in the context of genomic medicine. In 2030, the number of sequenced genomes of rare diseases will reach 83 million, and 250 million in the framework of diagnostic approaches to cancer.

[153] These fingerprints are based on measurement of the length of a number of repeat sequences of DNA, the number of repeats being specific to an individual.

for the express consent of the person concerned for the examination of his/her genetic characteristics[154].

- A second characteristic concerns the *stages necessary to obtain the DNA sequence of an individual*. Unlike other health data, the production of medical information from the genomic sequence remains a challenge and requires numerous stages between the sequencing of DNA and the interpretation. It is necessary in the first place to sequence the DNA (determination of the succession of nucleotides), and then to assemble the genome (align the sequence of fragments) and annotate all the genetic variations identified. These technical stages are done using standardized algorithms established by bioinformaticians. As the methodologies differ and are constantly evolving, it is important that these different stages are transparent, because it is from these data that the geneticist will make deductions about the health of the individual.

- A third characteristic concerns the stages *necessary to deduce from the DNA sequence information significant for the health of the individual.* Currently, the geneticist can compare the sequencing data with the literature data so as to identify the mutation responsible for a disease in a patient. This can be done using a simple algorithm that filters genetic variations so as to identify the most significant. With big data and algorithmic analysis, it will probably be possible to specify the risk of a disease by taking into account not only genes already characterized as responsible for the disease, but also the whole genome, which contains variants frequent in the population. Each of these variants in isolation has a small effect, but together they can strongly influence risk[155]. In hypercholesterolemia, for instance, some people who accumulate numerous frequent variants have a risk as high as bearers of a rare monogenic mutation. Whereas a causal mutation can be identified by means of a panel of genes and a simple comparison with the literature data, the algorithmic analysis of the whole genome will allow a more precise diagnosis, albeit using such a quantity of information that it will be difficult for the geneticist to understand how the prediction has been made. It is also important to remember that the prediction scores can differ from one population to another. At present, the vast majority of the data on which the algorithms are based come from European populations. The prediction scores are much less informative about non-European populations[156].

[154] In the case where the research requires the examination of genetic characteristics, the informed and express consent of the people concerned must be obtained before implementation of data treatment. The present article is not applicable to research done in application of article L. 1131-1-1 of the public health code.

[155] This is called the polygenic risk score: the analysis of millions of variants at multiple parts of the genome (polygenic) will give information that is very robust and precise in terms of the prediction of a common disease. See also M. Warren. The approach to predictive medicine that is taking genomics research by storm. *Nature* 2018; 562: 181.

[156] *Martin ar, et al. Hidden 'risk' in polygenic scores: clinical use today could exacerbate health disparities. Biorxiv. Https://doi.org/10.1101/441261. 11 october 2018.*

- Another important point is the *possibility of identifying information that was not initially requested*. For example, someone consults a doctor regarding diabetes, and, thanks to the genetic analysis, the geneticist finds that the person also carries a breast cancer susceptibility gene. We then speak of *incident* data, and their discovery will become increasingly common with whole genome sequencing. This information can lead to preventive measures designed to stop the onset of a disease, but it may also needlessly worry the person if no treatment is proposed. This is why the American College of Medical Genetics and Genomics has drawn up a short list of medically "actionable" genes[157] for which suitable treatment is available (in January 2019, there were 66 actionable genes[158]). With the generalization of genetic analyses and the informing of patients of the results, the management of incident data will probably become an inevitable stage in all care pathways.

- The statistical power needed to analyze certain genetic questions requires a number of participants that often exceeds the data available in a single country. The exchange of international data is made difficult by the heterogeneity of legislations and regulations, but also by the diversity of modes of access to data repositories.

3.3.4 Ethical principles and the challenge posed by advances in the exploitation of genomic data

Perhaps more than in other areas, the use of genetic data generates *strong tension* between two ethical positions, one based on the principles of benevolence and non-harm, and the other based on the principles of autonomy and the respect of people. The perspectives of medical benefits, which are real and major, require increases in the exploitation of the data and in extended sharing. But such dissemination of genomic data generates fear of a threat to the respect of the basic rights of the person. There is then a real risk that the obstacles and challenges created by access to genetic information will hamper research, thus delaying the acquisition of a medical benefit.

The question then is how to increase access to genetic data without compromising the ethical principles respect of which is indispensable for trust. We shall examine three of these ethical principles: consent and transparency in the use of the data, the response to risks of discrimination, and informing the person and management of incident data, while emphasizing what is specific to genomic data.

3.3.4.1 Protection of the person and respect of identity

[157] The term "actionable" is defined here as meaning that there are preventive or therapeutic measures that can avoid the occurrence of harmful consequences of this variant or can alter the natural progression by one means or another.

[158] https://www.ncbi.nlm.nih.gov/clinvar/docs/acmg/

Once the genome can "identify" not only the person, but also his/her ancestors and descendants and affects privacy, as it says something about confirmed or possible diseases, it is particularly important to respect the protection of the person and his/her identity. This is based on:

- *Respect of the person and collection of consent*. In accordance with the provisions of the civil code, the public health code, the GDPR, and the data protection act (see sections 1.4 and 2.1 above[159]), *written consent* of the people concerned or of their legal representatives must be collected if an examination of hereditary or acquired genetic characteristics is envisaged. The list of information to give to the person is defined by the GDPR (art. 13)[160]. Furthermore, analysis of the impact on data protection is required for the implementation of research relating to patients and their genetic data. As for any genetic analysis independent of big data, *the person* must make a double choice: on the one hand, accept or not the study (for the patient's benefit, or for the community's if the study involves participation in a cohort or database) and, on the other hand, to know the results or not concerning the disease or incident data (see below).

  But genomic big data raise several questions: consent becomes ill-adapted as the research progresses, the difficulty of defining the limits of the "purpose" of research in this context marked by substantial changes in and sharing of the data, notably if sharing with commercial companies is envisaged[161], the decision to recontact research participants, the fictional nature of the anonymization when one seeks to correlate genomic, clinical, environmental, and behavioral data for precision medicine. The GDPR takes into account all these questions (art. 4(11) and recital 33) and offers some scope for interpretation (art. 5(1)(b) and 6(4), recital 50). One may then wonder about the efficacy of consent as a method of protection of the person.

[159] CNIL: deliberation No. 2018-153 of 3 May 2018 on certification of a reference methodology relating to treatment of personal data used in the framework of research in healthcare with collection of the consent of the person concerned (MR-001).

[160] The identity and contact details of the data processor and of the person who protects the data processor's data; the purpose of the treatment of data (presentation of the research project); the legal basis of the treatment (article 6 of the GDPR); the nature of the information that will be used in the research; the recipients or categories of recipients of the data; the right to access, rectify, oppose, erase the data, and to limit their treatment; the modalities of the exercise of these rights; the optional nature of participation; if need be, the transfer of personal data outside the European Union and reference to appropriate guarantees; the length of data storage; the information provided for by article L.1122-1 of the public health code; the inscription of the patient in the national register of people who take part in research.

[161] The 2016/2017 activity report of the UK Biobank Ethics and Governance Council mentions, for example, the authorization given to two pharmaceutical firms to sequence the exome of UK Biobank participants, ie, 500 000 people. The agreement stipulates that the results of the sequencing be included in the UK Biobank and made available to researchers. Exclusive use of the results for 9 months was accorded to the firms. The governance of the Biobank considered that it was not necessary to recontact the participants because they had agreed to the use of samples and of data by commercial companies in their initial consent. Information was nonetheless sent to all participants.

- We have seen the *change in logic*, which is no longer based on a guarantee of security, but on fair information about the fate of the data, on broader consent, and the assurance that the risks will be minimized by means of attentive governance, including progress in bioinformatics[162], the use of pseudonymization[163], and sanctions against breaches of good practice.

- This fair information depends on the establishment of a relation of trust between the data subject and the person in charge of processing of the data, instituted by the GDPR. The quality of the *information delivered to the person* in terms of the flow and use of the person's genomic data (purpose, type of processing, storage, no commercial exploitation) and the quality of the means deployed to ensure confidentiality are fundamental. With genomic data, there is also the question of telling the person about *the information yielded by the processing of his/her data.* This is particularly important if there is a lack of precision regarding the initial purpose of the research or if the research evolves.

It is necessary to distinguish two types of information that the person should be able to choose to know or not to know:

- regular information given to the participants on the progress of the clinical or basic research using the genetic data. This is general information that does not directly or personally concern the participants;
- personal information deduced from the interpretation of variations identified in the genome of a given person, notably information on other health problems independent of the original disease. CCNE Opinion 124 deals with this question in the framework of care, and in section 3.3.4.3 we consider the management of incident data. In research, the handing over of this information is a highly debated subject: validity, clinical significance, and actionability are in general required to decide on what information to give to the person, and nowadays only a geneticist is able to explain this information.

- These information modalities are those applied by healthcare establishments or research institutions, in a secure context. However, one should not underestimate the disclosure, by the people themselves, of their genetic data, in response to commercial offers concerning genealogy, or to create links between people, even when the aim is to expedite research[164].

3.3.4.2 The risks of discrimination.

The interpretation of genomic data raises several risks.

[162] It is now possible to query the databases of different international sites without having to move the data.

[163] Treatment done such that the data can no longer be attributed to a specific person without complementary information, as long as these sets of information are stored separately and are subject to technical and operational measures to prevent their attribution to an identified person.

[164] For example, Helix, deCODE Genetics, personal genome project, 23andMe.

- *Stigmatization of at-risk groups.* The use of genomic databases and genetic population studies are designed to find correlations between sequence variants and clinical characteristics or laboratory data. These correlations identify groups of individuals or patients with certain characteristics – disease risk factors, geographic origin – who can be targeted because of these characteristics.

- *Unequal genetic prediction because of bias in the constitution of genomic databases.* Interpretation of sequencing data that is not rigorous or is based on a biased dataset could result in a missed medical opportunity. There is a bias in the constitution of databases used to train algorithms, because certain populations – in particular non-European – are insufficiently represented[165]. So, while the risk scores are reliable and of good quality for European populations, this is not the case for other populations (Asian, African), so there is a possible missed medical opportunity[166]. Note should be taken of this inequality in access to genetic tests and, more generally, to the technological advances of the health system, as emphasized in CCNE Opinion 129[167] and in its synthesis report for the *États généraux de la bioéthique* (Bioethics Forum).

- *The risk of cross-referencing* – which exists in the United States – of genetic data stored for genealogical purposes in public banks with data collected for legal proceedings.

3.3.4.3 The management of incident data

Genome analysis allows the identification of data that are relevant but not linked to the disease that is the subject of the consultation or the initial research project. The ethical question then arises as to whether or not pathogenic variants should be sought systematically.

These incident data will become extremely frequent, in healthcare and in research projects, and blurring of the distinction between these two areas will only make the question more complex, as clinical research aims at a broad analysis of the genome so as to identify all variants associated with risk. Who should be informed, when and how? How can the person's right not to know be respected? Incident data can be clinically useful and prompt preventive or therapeutic action (actionable variants). These data can also be significant, but may not, in the current state of knowledge, lead to a strategy that can benefit the person. They can also be of uncertain significance. All this is likely to change,

[165] Among the participants in all genetic studies, 79% are of European ancestry, but represent only 16% of the population (GWAS catalogue). Popejoy, AB, Fullerton SM. Genomics is failing on diversity. *Nature* 2016; 538: 161-4.

[166] Martin ar, et al. Hidden 'risk' in polygenic scores: clinical use today could exacerbate health disparities. Biorxiv https://doi.org/10.1101/441261. 11 october 2018.

[167] CCNE Opinion 129, pp.75; Synthesis report of the *États généraux*, pp. 36-50.

which raises the question of subsequent re-examination of the variants. According to the good practices currently accepted in France, the data subject is informed of incidental discoveries if there is a preventive or therapeutic strategy (actionable discoveries)[168]. The frequency of discovery of deleterious mutations affecting actionable genes is currently from one to three percent, but this frequency will only rise. We should not underestimate the distress that the discovery of these incident data can engender not only in carriers of variants, but also in their relatives, who may also potentially share the detected risk. Clinicians therefore need to be trained to interpret and manage these incident data, so as to improve their ability to anticipate and to support the people affected.

It is essential – whatever the clinical or research context – that this question of telling the person concerned of incidental discoveries be anticipated before any prescription or establishment of a data platform, and discussed with the person when the initial consent form is written (see CCNE Opinion 129). Intervention of ethics committees or institutional review boards that approve research projects is desirable.

3.3.5 Risks of loss of sovereignty

Apart from the fact that it concerns not only a given person but also his/her relatives, the specificity of the management of genomic data is due to the particularly great volume of data. This requires infrastructures for storage and for calculations involving the processing of tera-, peta-, and exabytes, and, new skills to exploit and interpret the data. As emphasized by the France Médecine Génomique plan, with its annual capacity of 20 000 exomes and 10 000 genomes, France is lagging well behind those countries able to perform tens of thousands of analyses each year. This poses the question (see below, §3.4.2) of the involvement of the public authorities in the management of these data, and of the risk of a loss of sovereignty. One should not overlook the benefits that private organizations (but also the states where they are found) gain from the information of great value that stems from the processing of the genetic data they collect. Genetics is today integrated in this global phenomenon of a digital and participatory economy, by virtue of its organization and functioning. In the era of the digital economy, whether we wish or not, the sharing of data has become synonymous with commercial exchange and genetic data synonymous with capital[169].

[168] Decree of 27 May 2013 defining the good practice rules applicable to the examination of a person's genetic characteristics for medical purposes.

[169] Corto-Stoeklé H, et al. Le partage des données génétiques : un nouveau capital. *Médecine/sciences* 2018; 34: 735-40.

3.4 What reflections in view of the new problems revealed by these different contexts?

We have seen that data – regardless of the context envisaged (see sections 3.1, 3.2, 3.3) – have become a resource stored in a multiplicity of platforms and repositories or warehouses, sometimes spread throughout the world and accessible to different stakeholders, who will process these data so as to extract new information. When data are collected, one cannot always know what use will be made of them, by whom and when. But data are often collected and disclosed upon the initiative of the data subjects themselves, independently of care or research projects, and this has given rise to the expression the "uberization of health".

This temporal and geographic disruption between the data and the data subject, defined above in different contexts, prompts three-fold ethical reflection on our relation to our data: consent, the international dimension, and our own disclosure of the data.

3.4.1 Individual consent and collective trust: what changes?

We have seen that consent is one of the legal cornerstones of the processing of personal health-related data, but not the only one, except for genomic data for which express consent is always necessary (apart from the particular case dealt with in article L 1131-1 of the public health code).

Even when consent is required, it can assume various forms, essentially to enhance the efficacy of the research, if the person in charge of treatment has, at the outset of the project, made all useful provisions to optimize the information given to the person and the protection of that person's individual rights.

Some people propose a vision that contrasts with this individualistic conception of the relation of the person to his/her data and consent. Personal health-related data come from the most private sphere, but also, when pooled, they become components of a network of information useful for the public good. This network constitutes a common asset associated with collective protection of private life[170] and is justified by two things:

- in a network, the data are not independent – personal data can also reveal information concerning another individual, as intervention on some data will alter other data. The

[170] Regardless of the service provider (public or private, even the state itself) in charge of this protection, the provider would be obliged to take appropriate measures to guarantee data security collectively and to avoid unfortunate consequences for people or ethical breaches. See also Pierre Bellanger: Les données personnelles : une question de souveraineté. *Le Débat* 2015 No. 183, pp. 14-25.

data become "relational" and so no longer correspond to purely individualistic concerns: their use should not depend on their holder's wishes alone;
- protection of the network is better adapted than protection of the data of an individual, of a link in a chain. To be effective[171], protection from an individualistic viewpoint would severely limit the sharing of data, unless the data are anonymized, in which case data treatment would be less effective.

Health data cannot be protected by a liberal approach concentrated on maximizing individual liberties (by means of consent or data ownership). Rather, it is necessary to adopt a more community-minded or collective outlook, which could restrict certain individual liberties, in the name of the public interest and the common good.[172]

The notion of public interest also inspires the claim of a "right to science"[173] which, because of the benefit it affords the population and because of the need to promote progress and innovation, would question whether data processing requires consent, which could be considered as an excessive obstacle[174]. Others arguing for less demanding control consider that individual consent may no longer be required if there is a strong probability that data processing helps improve the health of the person and of the community (reciprocity principle), when the risk of harm is low (proportionality principle).

A third conception has been proposed. It is intermediary between the individual vision and the collective vision: that of interactive and relational autonomy, in which the person manages his/her data but is integrated in a community, which implements a collective project likely to evolve and which protects the person. This conception is also based on the "dynamic" nature of the data, which circulate and so are no longer localized or proprietary, but relational.[175]

These divergent approaches show that what is at the heart of the debate is the evolution of the relation between the individual and the collective, between the increased autonomy of everyone and the protection required by the generalized use of technologies in-

[171] Joly Y, et al. Are data sharing and privacy protection mutually exclusive? *Cell* 2016; 167: 1150.

[172] Bourcier D, de Filippi P. Vers un droit collectif sur les données de santé. *Revue de droit sanitaire et social*, 2018; pp. 444-56.

[173] Knoppers BM, Thorogood AM. Ethics and Big Data in health. *Current Opinion in Systems Biology* 2017, 4: 53-7.

[174] "We must remember who gets left behind when consent is required”. Taylor P. When consent gets in the way. *Nature* 2008; 456: 6.

[175] The model of a personal data ecosystem centered on the person reveals an egocentric sociological approach, while networks produce circulation of traces, which are partial, ephemeral, and shared. Boullier has proposed the concept of "*habitèle*", the human ability to carry our affiliations like a new envelope that humans create themselves and which is made of data not centered on their ego, but on their situational commitments and on issues that they must resolve (see Dominique Boullier, Sociologie du numérique, 2016; Armand Collin).

volving the treatment of big data. Progress in one reinforces the other. Their reconciliation is not seamless and generates tensions because individual interest and public interest do not necessarily coincide, at least in the short term. The individual, collective, and relational conceptions can all three contribute to the search for a point of equilibrium, which constantly needs to be redefined because it is continuously challenged by technological progress and by changing lifestyles (see RECOMMENDATION No. 3).

3.4.2 The international dimension and the question of national sovereignty

The CERNA report[176] on sovereignty in the digital era and on mastering our choices and values reiterates that digitization profoundly alters the problem posed by the control of health data, since it potentially evades geographic frontiers and changes time scales. Digitization allows almost instantaneous exchanges via the internet and data storage unlimited in time. It can therefore establish significant correlations between data, by evading the national sovereignty of states.

France cannot control the scientific and medical innovations that are anticipated to stem from data processing, or the evolution of its healthcare system, unless it makes every effort to meet the following challenges:

- the technological and human challenge of the storage, security, and exploitation of a constantly increasing volume of health data (see § 3.3.5 on genomic data). There is also the risk of losing the guarantee of data management that respects ethical principles (see recommendation No. 10). In France, the creation of the Health Data Hub, which should be announced in the next health law and will be initiated in the first half of 2019, is a first response.
- the challenge of research in fundamental mathematics, referred to above, must ensure a high technological level, which guarantees national and European independence for the use of data and their application in healthcare.

This question is of particular concern in the United States and, more recently, in China, whose financial investments, which are disproportionately large compared with those of European countries, benefit their private companies (GAFAM and BATX), which are technologically dominant. In China, there is an intent to set limits to the use of data and a law on the protection of personal data came into effect on 1 June 2017. One of the notable dimensions of this law is the obligation to store "important data" and personal

[176] CERNA is a commission set up by Allistene with oversight on research in the digital sciences and technologies. The report (in French) is available at https://www.allistene.fr/files/2018/10/55708_Opinion-Souverainete-CERNA-2018.pdf

data on local servers on Chinese territory. This results in tensions between the expression of nationalism applied to data, the internationalization of research programs, and the interests of large pharmaceutical groups[177].

In a world where large private providers increasingly rival states and take on functions that until recently were the subject of a national monopoly (see abovementioned CERNA report), the institution's guarantee of access to research data (the importance of which we stressed in section 3.2.1.2) assumes that this same institution has enough control over operations concerning these data to ensure that its guarantee is effective and inspires trust. We referred above to the investment of the American giants in hosting data in France, which could prompt other offers of effective solutions for their exploitation, thus clearly posing the question of sovereignty.

3.4.3 New e-health practices, outside care pathways and without precise regulation

The widespread use of social media and the internet as sources of information or access to health services, plus the practice of the quantified self, incites people themselves to divulge their data: search for health information, visiting online patient communities, online sales of medication, online medical advice, making medical appointments online, direct offers of teleconsultations, and electronic medical services. A report by the Conseil national de l'ordre des médecins (CNOM; French Medical Association)[178] has dubbed this tendency the "uberization of health". There is also the apparently harmless transmission of data while surfing the internet or using connected objects (messaging, social media, visiting websites) and data collected by connected objects (physiological information, geolocation, purchases, places frequented, lifestyles). These data only secondarily become health-related when their cross-referencing provides a person's health parameters or lifestyle choices. To take an example, the assiduous frequentation of a specialized healthcare establishment (determined by geolocalization) and the repeated purchase of certain products or dietary choices (identified by credit card payments) show that a person is suffering from a disease and may be undergoing treatment.

[177] *Cyranosky d. China's crackdown on genetic breaches could deter data sharing.* Nature *15 november 2018; and morgane tual: en chine, une loi controversee sur les donnees personnelles et la cybersecurite. Le monde - pixels, 1 june 2017.*

[178] The CNOM observes an accelerated trend towards the "uberization of health", through online offers that correspond to unregulated electronic commerce and which tend to reduce medical practice to a simple electronic service against payment, via platforms of the market sector. The CNOM has asked whether the state should continue to produce normative regulatory texts applied to the practice of medicine using digital means while also allowing unregulated digital offers to prosper on the e-health market, and has demanded the introduction of regulation of digital health offers to ensure respect of ethical principles in health (*Rapport télémédecine et autres prestations médicales électroniques*, February 2016).

It is with regard to these data collected and stored outside the care relationship that are posed with the greatest acuity questions concerning confidentiality, the security of personal data, their processing and fate. Whereas these sites now have to respect the obligations imposed by the GDPR, one may wonder about the value of the information delivered to the data subject and of the consent obtained, and about the respect of ethical principles affecting private life. This is illustrated by the CNIL's recent sanction against Google for not meeting the requirement for clarity and accessibility in its confidentiality policies[179].

The large private providers that dominate this market act within an international framework. They are essentially located in the United States (GAFAM), even though China is tending to become a major stakeholder in this field (BATX). While the United States have an attachment to individual rights comparable to that of Europeans, they nonetheless have a different conception of the protection of personal data and of the consent to their use[180]. To summarize, the US conception can be called 'contractualist', whereas the European conception is 'personalist'[181].

While European regulations are designed to avoid the possibility that the accompanying rules are broken, the logic of the large providers remains that of an ever-improving offer of goods and personalized services, which is based on the exploitation of a constantly increasing amount of personal data. Users are advised of their rights but, in most cases, when they want to benefit from the functionalities on offer, they give their consent without considering all the consequences.

The general public's trust, however, is weak, as shown by reactions when, a few months ago, the large-scale pirating of data by Facebook was revealed[182].

Yet the public and the large providers share the common interest that their relations are based on mutual trust. The providers are aware of this because they know that their development model assumes the cooperation of the greatest number. To preserve this

[179] A CNIL sub-commission announced on 21 January 2019 a 50 million euro fine against Google LLC in application of the GDPR, for lack of transparency, insufficient information, and absence of valid consent for the personalization of publicity.

[180] Note that the GDPR must be respected if these stakeholders treat the data of European Union citizens, or target European Union citizens.

[181] In the United States, the protection of personal data is based on the logic of contractual freedom and so is rooted in the idea that consent expressed by the holder of the data suffices, even if it appears largely formal. In Europe, on the other hand, protection is deemed to correspond to public freedoms and is supported by legal texts designed to guarantee the protection of private life, starting from the principle that imbalance between the contracting parties and the impenetrability of the system are such that formal consent is not sufficient.

[182] *See le monde-pixels. D. Leloup and m. Untersinger. Faille facebook : des donnees de 29 millions de comptes recuperees par les pirates. 12 october 2018.*

trust, they set up internal ethics committees and do not hesitate to include particularly qualified luminaries from outside the company[183]. These initiatives are clearly useful and the reflections thus undertaken contribute to the necessary debate on the ethical use of personal data. They lead to precautions and procedures that are brought to the public's attention, but it is nonetheless clear that the initiatives are intended to augment corporate efficiency, with the underlying idea that these guarantees will lead to a membership sufficient to continue to obtain the greatest possible amount of personal data. The proliferation of dedicated ethics committees, even though they can be endowed with means comparable to those of the companies that set them up, should not damage the maintenance of a more fundamental ethical reflection, which can only be achieved by an independent authority free of any utilitarian objective. The great value of dedicated ethics committees is that they are a favored means for providers to demonstrate their loyalty in the manner they apply to their clients, notably European, the protective rules of the GDPR. Beyond the scope of application of the GDPR, it is up to the providers to demonstrate in concrete and verifiable terms the importance they accord to the protection of data confidentiality.

Loyalty, while obviously necessary, is not enough to form a lasting relation of trust. It also appears necessary to inform the public and to raise their awareness of everything concerning the disclosure of personal health data. The public notably must know how to configure and use their connected objects such that access is provided only to information necessary for their own needs. In this regard, patient groups can be effective intermediaries and the large internet providers involved in healthcare would be well advised to associate such groups with the actions by which they seek to ensure the trust of users **(see RECOMMENDATION No. 1**).

[183] See for example: Facebook lance un centre de recherche consacré à l'éthique de l'intelligence artificielle, by Morgane Tual - Le Monde (Pixels) 20 January 2019. This institute, funded by Facebook and developed in the Technical University of Munich, is intended to be "independent", emphasizes Facebook. https://www.lemonde.fr/pixels/article/2019/01/20/facebook-lance-un-centre-de-recherche-consacre-a-l-ethique-de-l-intelligence-artificielle_5411861_4408996.html

GENERAL CONCLUSION

In view of the fundamental issues at the heart of the digital sciences and technologies, the impact of which on the future of humanity will probably be considerable, the CCNE expresses its strong support for the development of innovation in this field and affirms the importance of vigilance regarding the protection of the basic rights and individual liberties of people when these technologies are used in health. The CCNE deems it necessary to reiterate the invariance of certain ethical benchmarks on which are based respect of the person. The human subject, who is also represented by his/her digital health data, cannot be instrumentalized. The CCNE wishes to analyze how these basic ethical benchmarks can help meet the challenge posed by the complexity and international dimension of the digital applications of big data in healthcare.

More than any other big data, those concerning health are at the crossroads of the individual and the collective, of the personal and the general, of the public and the private. In this regard, what underpins ethical reflection must again be enunciated and reaffirmed, but this does not exclude making changes in the way we see this reflection, in the light of the use of these big data and of their appropriation by the people. Thus, this is how the notion of private life fitted a static definition, conceived of as resisting any intrusion, whereas current ways of living are evolving towards a society of networks where the personal sphere is constantly being redefined as a function of the relations that the individual weaves with the environment. This does not mean, as some have suggested, the "end of private life", but rather what Antonio Casilli views as "*a shift from the personalization of private life to a 'collective negotiation' within a framework in which autonomy and freedom are respected by design"*[184]*.*

The treatment of data will certainly be a factor in diagnostic and therapeutic improvements and, without doubt, in improvements in health. Can everyone benefit from these advances, when equality of access to care is only imperfectly ensured? Through the exploitation of big data, can another form of relation and solidarity be created, to the benefit of the community, of society? The digital revolution is opening up immense perspectives, but the rapidity of technological changes upsets our benchmarks. It makes us see several new risks, which were voiced during the *États généraux de la bioéthique* (Bioethics Forum) and which prompt us to seize the opportunities offered by these technological changes:

[184] Il s'agit d'une sortie « *de la logique de la personnalisation de la vie privée, pour entrer dans celle d'une « négociation collective » inscrite dans un cadre dans lequel les autonomies et les libertés sont respectées by design* » Antonio Casilli. Quatre thèses sur la surveillance numérique de masse et la négociation de la vie privée. Jacky Richard and Laurent Cytermann. Étude annuelle 2014 du Conseil d'État "Le numérique et les droits fondamentaux", La Documentation Française, pp.423-434, 2014, études et documents, Conseil d' État.

- automated exploitation of data in a care pathway could weaken the listening and dialogue that underpin the relation between the patient and care providers. This invites us to reflect upon the protection of individuals and the strengthening of the relation of trust that could be achieved if the automation of certain tasks frees up time to promote the human relation;

- undifferentiated treatment of big data risks erasing the particularities of health-related data and applying to them a pure market logic. But relevant use of applications developed for the needs of this economic activity can help measure health parameters in real life. While this utilization questions the role given to the right to not know, it can also help empower patients and inform carers;

- lack of understanding among the general public about the processing of personal health-related data and the risks of unwanted disclosure[185] threatens individual rights, but it is a potent stimulus to developing quality information and promoting ethical behavior in all stakeholders;

- profiling, made possible by the cross-referencing of the most diverse personal data, threatens the principle of solidarity which underpins the sharing of risks, the foundation of the funding of our health system, and carries a risk of discrimination against the most vulnerable people. But it also provides the means to identify more precisely the factors of inequality, with a view to remedying this inequality;

- globalization, which affects researchers and all public and private stakeholders involved in healthcare, poses major questions on how to preserve our sovereignty, and urges us to make the effort needed to grow our know-how and the infrastructures required for the storage and treatment of an ever-increasing amount of data.

[185] Particularly expressed during the *États généraux de la bioéthique* (Bioethics Forum). See CCNE synthesis report pp 38-50.

LIST OF RECOMMENDATIONS[186]

The CCNE proposes recommendations that pass on and extend numerous recent private and public initiatives. These initiatives express the certainty that a new balance must be found in the ethical approach to the effects of the digital sciences and technologies, without limiting the expected benefits, but equally without weakening the principles that underpin the quality of being human and human relations. These recommendations outline a possible ethical response to the main questions and tensions raised by the exploitation of big data in healthcare. They are listed below according to the principles that we feel are essential to sustain in the face of these developments, and that the *États généraux* and CCNE Opinion 129 have also emphasized: ensure the autonomy and protection of the person, and the person's right to make choices and decisions that concern him/her; respect individual freedom without compromising solidarity and public interest; in research, permit the acquisition of new knowledge that will benefit the health of everyone, without risking deviation.

- **Ensuring the person's autonomy and right to make choices and decisions that concern him/her**

RECOMMENDATION No. 1 (2.1.3 and 3.4.3)

Everyone has a right to intelligible, accurate, and fair information on the treatment and fate of his/her data, whether or not consent is required. This information, effective application of which must be evaluable, is adapted to each context and is regularly updated. It must also be easily accessible, particularly when it concerns the most vulnerable people, who must receive, directly or via their legal representatives if need be, the appropriate incentives to allow them to exercise their rights.

If this information requirement is to be respected and effective, the CCNE considers that all our fellow citizens must be educated about the specificities of digital technologies and about their associated advances and risks, both for their own sake and for society's, so that they can make responsible use of their personal data.

[186] The section references in brackets refer to those parts of the text that fully develop the arguments that led to the formulation of the different recommendations.

RECOMMENDATION No. 2 (2.1.3)

Given the fast pace of scientific and technological innovations and the changes that they bring in the collection and use of health-related data, the CCNE considers that it is necessary periodically to evaluate the effective implementation of legal measures, so as to verify the sustained efficacy of the personal data protection system these measures introduce.

RECOMMENDATION No. 3 (3.2.1.1)

In this constantly changing context, it is necessary reflect upon the notion of consent to the collection and treatment of big data in a field where there is heterogeneity among stakeholders and practices evolve. This reflection should concern the purpose of consent and consent is collected, so as to ensure a lasting balance between the respect of human rights and changing uses. This reflection should foster public debate on ethical recommendations and enable periodic updating of the law.

RECOMMENDATION No. 4 (2.3.2 and 3.1.2.2)

The relevance of decisions concerning a care pathway taken with assistance from the algorithmic treatment of big data depends on the quality of those data, on the absence of bias in their selection, and on the rigor of the methodology used for their treatment. The CCNE considers that the results obtained can only be evaluated and approved by human guarantees, which are a condition of the accountability of stakeholders. These guarantees should be given by independent authorities.

In research, it is necessary to preserve the genetic diversity of data subjects when selecting the data processed, so that the results obtained are not skewed by insufficient representation of non-European populations.

RECOMMENDATION No. 5 (2.3.3 and 3.2.1)

The CCNE considers that health professionals, during their initial training and throughout their career, should undertake suitable training in digital technologies, in the ethical principles that govern data collection and treatment, in the means implemented to respect these principles, and in the risks of bias should they not be respected.

Experts in the management and analysis of big data (data scientists) and researchers must be sufficiently knowledgeable about the ethical questions raised by these technologies to be able to safeguard the protection of basic rights and individual liberties.

RECOMMENDATION No. 6 (2.3.3)
The multiplication of websites and applications that, outside the care pathway, give advice on improvements in lifestyle and well-being, poses the question of the rigor with which these websites and apps collect, interpret, and treat health-related data. The CCNE considers that these websites and apps should be evaluable, as should the quality of the information delivered to users, so as to avoid insufficiently rigorous approaches harming people's behavior and health.

RECOMMENDATION No. 7 (3.1.2.2)
The CCNE reiterates that the trust at the heart of the care relationship requires the preservation of a direct personal relation between the health professional and the patient. The patient cannot be reduced to a set of data to be interpreted, rendering unnecessary listening to the patient and the taking into account of the patient's personal circumstances. Useful as they are in helping in the diagnosis and in guiding treatment, data cannot replace dialogue.zzz

The use by health professionals of recent technologies must also aim to free up time for listening to and speaking with the patient, by simplifying the collection of relevant information. The use of such technologies should enable patients to be further empowered as stakeholders in the care pathway, by allowing them to appropriate their data, a sine qua non of a fully responsible attitude.

- **Respecting individual freedom without compromising solidarity and public interest**

RECOMMENDATION No. 8 (2.2.3)
The advent of precision medicine fosters marked progress in the prevention, diagnosis, and treatment of diseases. But the individualization of the risk it implies can jeopardize sharing in health funding mechanisms and also runs the risk of deviation towards discriminatory profiling, notably for economic reasons. The CCNE considers that health stakeholders should be particularly vigilant regarding the preservation of our values of equality, fairness, and solidarity.

RECOMMENDATION No. 9 (2.3.1)
The CCNE stresses the need to ensure that people without access to digital technologies, for economic reasons, for example, or because of difficulty in understanding how these technologies work, benefit, like others, from advances in healthcare and are neither penalized nor discriminated against in their access to care.

- **Permitting the acquisition in research of new knowledge that benefits the health of everyone, without running the risk of deviation.**

RECOMMENDATION No. 10 (3.1.2.3)

In confronting the technological challenge that the storage, sharing, and treatment of big data in healthcare pose to national and European sovereignty, the CCNE recommends the development of national and interconnected shared platforms. Open according to modalities that need to be defined for public and private stakeholders, these platforms must enable France and Europe to preserve their strategic autonomy and not lose control of the asset constituted by data, while prioritizing controlled sharing, which is indispensable to the efficacy of care and medical research.

The CCNE emphasizes the need to make a major effort in scientific research so as to be able to ensure, with a high level of competence, technological improvements in data treatment, while respecting ethical principles.

RECOMMENDATION No. 11 (3.2.1.1)

The CCNE considers that in research the ethical imperative must be adapted to each particular situation, so as to justify a relation of trust between the holders of the data and those who have access to and treat these data. It is essential that the holder of the data be informed of how the supervisory authority acts as a trusted third party. A threefold ethical requirement must therefore be met:
(1) rigorous and transparent evaluation of the relevance of the research that justifies access to data, which must contribute to the benefit of everyone and to enhancing healthcare knowledge;
(2) sharing of information on progress in research with the participants, according to modalities adapted to different contexts;
(3) ensuring data security and traceability, and the absence of malicious use of these data.

RECOMMENDATION No. 12 (3.2.2)

The CCNE considers it necessary to expedite the sharing of health data for the purposes of research. It considers notably that researchers should have access to data collected on the internet or social media by platforms whose governance is controlled, for research protocols whose purposes are strictly defined, while respecting the rights of people who consent to provide their data.

APPENDICES

APPENDIX 1

CCNE members who participated in the working group

Gilles Adda
Thomas Bourgeron
Laure Coulombel (rapporteur)
Pierre Delmas-Goyon (rapporteur)
Jean-Marie Delarue (CCNE member until December 2017)
Claude Delpuech
Pierre-Henri Duée
Claude Kirchner
Martine Le Friant
Jean-Pierre Mignard
Francis Puech
Dominique Thouvenin (CCNE member until December 2016)
Bertrand Weil (CCNE member until December 2017)

APPENDIX 2

Experts consulted :

Charles Auffray (European Institute for Systems Biology and Medicine)

Thomas Bourgeron (Professor of Genetics, Université Paris 6, Institut Pasteur)

Mathieu Cunche (Inria - Institut national de recherche dédié aux sciences du numérique)

Victor Demiaux (advisor to the President of the Commission nationale de l'informatique et des libertés [French Data Protection Authority])

Mathieu Galtier (OWKIN)

Jean-Gabriel Ganascia (University Professor, Université Paris 6, LIP6 Laboratory, chairman of the CNRS Ethics Committee (COMETS)

(xxx confidential) (former legal advisor to Google France)

Erwann Le Pennec (Adjunct Professor in the Applied Mathematics Department at the École polytechnique)

Frédérique Le Saulnier (Data protection officer at the Institut national de la santé et de la recherche médicale)

Alain Livartovski (Institut Curie, Medical Information Department)

Sophie Narbonne (Commission nationale de l'informatique et des libertés [French Data Protection Authority])

Gilles Wainrib (OWKIN and Assistance Publique-Hôpitaux de Paris)

OPINION 130

COMITÉ CONSULTATIF NATIONAL D'ÉTHIQUE
POUR LES SCIENCES DE LA VIE ET DE LA SANTÉ