Regulatory Challenges of Data Mining Practices: The Case of the Never-ending Lifecycles of ‘Health Data’

Giovanni Comandè
Sant’Anna School of Advanced Studies, Pisa, Italy, Italy

Giulia Schneider
Università Bocconi, Milan, Italy

Abstract

Health data are the most special of the ‘special categories’ of data under Art. 9 of the General Data Protection Regulation (GDPR). The same Art. 9 GDPR prohibits, with broad exceptions, the processing of ‘data concerning health’. Our thesis is that, through data mining technologies, health data have progressively undergone a process of distancing from the healthcare sphere as far as the generation, the processing and the uses are concerned. The case study aims thus to test the endurance of the ‘special category’ of health data in the face of data mining technologies and the never-ending lifecycles of health data they feed. At a more general level of analysis, the case of health data shows that data mining techniques challenge core data protection notions, such as the distinction between sensitive and non-sensitive personal data, requiring a shift in terms of systemic perspectives that the GDPR only partly addresses.

Keywords

health data – sensitive data – data mining; clusterisation – General Data Protection Regulation (GDPR) – transparency

1 Introduction: From General Considerations to a Case Study

The outburst of information technology as an integrative support of businesses’ activities has come to shape the global economy in new and still relatively
unexplored manners. The enormous amount of data created every second by the ‘technological extensions’ of individuals, has soon shown to be a gold mine for a relentlessly growing field of public and private enterprise collecting, storing, processing and reselling data variously scraped together from different information sources. The combination of different datasets through mining techniques makes it possible to extract from the analysed datasets hidden information as well as productive correlations. This material silently circulates through the hands of various stakeholders that differently incorporate these derivative data in their decision-making processes.

The unpredictability of the consequences of the convergence of wide datasets results in the unpredictability of the legal issues connected to the business of big data.

In this article, we tackle a more general issue of which health data are a paradigmatic case study. Health data can be considered the most special of the ‘special categories’ of data envisaged by Art. 9 of the General Data Protection Regulation. The same Art. 9 GDPR prohibits, with broad exceptions, the processing of ‘data concerning health’. This analysis aims to test the endurance

---


4 Ibid., 40-43.


8 McKinsey Global Institute, *supra* note 1, 10.

9 Regulation *(EU)* 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), Art. 9: ‘Processing of personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing
of the ‘special category’ of health data in the face of data mining technologies and the never-ending lifecycles of (health) data they feed.

Far from being constrained, the processing of health data currently appears to be boosted by various computation analytics falling under the name of data mining technologies. Through data mining technologies, health data have progressively undergone a process of distancing from the healthcare sphere as far as the generation, the processing and the uses are concerned.

More specifically, this article argues that in the current technical environment nearly every type of data circulating in the market is actually or potentially health-inflected. The massive availability of various health-related data has in turn ended up broadening the health-grounded basis of decisions taken far outside the healthcare sphere, feeding in this way spillover uses of health data that drive classifications and, with that, rules of inclusion and exclusion from society.

At a more general level of analysis, the case of health data shows that data mining techniques challenge core data protection notions, such as the distinction between sensitive and non-sensitive personal data, requiring a shift in terms of systemic perspectives that the General Data Protection Regulation only partly addresses.

Drawing upon the scheme of the never-ending lifecycles of health data, this article is structured as follows. Section 1 will provide an insight over the sources of health data and over how health data are collected and processed through data mining techniques (1.1). The processing of health data through data mining techniques generates various types of health-inflected data (1.2). Section 2 goes deeper into the peculiarity of health data mining processing, illustrating how differently health-inflected data are systematized in clusters on the basis of which corporations carry out collective profiling practices (2.1). The clusterisation of different health data points is characterized by technical and legal obscurity (2.2) to which the GDPR provides some remedies. Section 3 will thus ultimately explore the new rules provided by the GDPR with regards to the practice of group profiling (3.1 and 3.2), as well as the challenges raised by the lifecycles of health data to traditional categories of data protection law (3.3). Conclusions will follow in respect to the role of health data as a criterion of inclusion and exclusion from society (4).

---

of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person’s sex life or sexual orientation shall be prohibited"
The Case of Health Data: From Medical to Health-inflected Records

Although many industries have experienced radical changes as a result of the improvement of computational analytics, the spread of big data infrastructure in the healthcare sector currently poses the greatest challenges.

Developments in the information technology field have indeed oriented an outright reorganization of healthcare services around the key asset of electronic health records stored in often decentralized databases. Accordingly, healthcare providers have decreased the capital investment in hardware.

The growing phenomenon of information technology outsourcing has transformed the healthcare industry into an intricate net of contractual arrangements regulating data collection, access and transfer through different sets of platforms, ranging from locked to more collaborative ones.

---

15 Pasquale and Ragone, ibid., 599-600; Pisto, supra note 12, 80, underlying how the traditional closed health information storing methods, based on health information management departments where charts are physically collected and copied, are progressively abandoned.
17 Pasquale and Ragone, supra note 14, 563-564.
18 Hoffmann, supra note 14, 1743.
19 Pisto, supra note 12, 83-84.
Sophistication, openness and interoperability of databases have been thus the premises for a collection of health data on a much wider scale. The management of personal health information by ad hoc businesses bound to healthcare providers by increasingly complex contractual relationships has created a more fluid and less protected information setting. This has led to the radical enlargement of the landscape of health data sources and of health data uses\textsuperscript{20} largely outside the proper healthcare environment.

Such a situation has in turn provided easy grounds for the blossoming of commercial practices carried out by data brokers trading patient data in what has become an outright market of health data.\textsuperscript{21} This market has been radically expanding together with the proliferation of health-related data sources: data directly generated by healthcare institutions has become only a small fragment of a much more complex data environment that variously fosters health data exchanges among commercial parties. Electronic medical records are thus only the top of the iceberg of electronic health records, comprising an incredibly vast amount of data stemming from different sources and variously related to a certain health status or condition.\textsuperscript{22}

\textbf{2.1 Health Data Sources}

In the era of Big Data, companies infer data subjects’ medical conditions or disabilities based on various types of data that are variously related to health conditions. Thus, in the current technical environment there are different types of health data sources, which can be divided into the two main categories of ‘primary health data sources’ and ‘secondary health data sources’ based on their connection to the health sphere.

In this perspective, ‘primary health data sources’ are strictly related to the medical and health domain and produce directly health-inflected data (see Fig. 1). On the contrary, what we define as ‘secondary health data sources’ are at first glance totally uncorrelated to the health sphere and generate data that becomes health-inflected only after being aggregated and processed.


\textsuperscript{21} Pasquale and Ragone, supra note 14, 607.

\textsuperscript{22} T. Rajaretnam, ‘Data Mining and Data Matching: Regulatory and Ethical Considerations relating to Privacy and Confidentiality in Medical Data,’ Journal of International Commercial Law and Technology 9(4) (2014) 294, 302.: “(...), the very concept of health sector specific regulation is flawed because health related or medical inflected data frequently circulates outside of the traditionally recognized health care sector”.

Hence, secondary health data sources produce data which are indirectly — or potentially — health inflected (see Fig. 2).

Beginning with the use of digital technologies in healthcare workplaces, physicians increasingly rely on technical devices in their daily routine practice, such as virtual world environments, scanning technologies and lately also 3-d printers.23

Tele-medicine is a further important source of health-related data: hospitals and other healthcare providers have indeed developed online systems for the real-time processing of patient records, the real-time completion of risk

---

assessments and finally the real-time delivery of clinical decisions all carried out through wireless technologies.²⁴

Pharmaceutical prescriptions are a further important source of health data:²⁵ individual prescriptions filled by pharmacies and containing the patients’ names, the dose of a certain drug as well as the date of the prescription, are constantly mined by data brokers and sold on to drug companies.²⁶

Self-tracking devices and the emerging applications of computation in healthcare additionally broaden the extent of data sources. Nutritional tracking apps, weight management devices, pedometers tracking data subjects’ movements and monitoring their exercise, sleep control apps, all contribute to charting users’ health conditions and thus to ‘quantify’ and ‘hacking’ users’ lives.

Along these lines, social networks have also become a primary source of health-inflected data. In an ‘economics of signalling’²⁷ the amount of health information spread online through the net in the forms of communities and frequently asked questions has been increasing exponentially.²⁸ Social networks have intensified surveillance specifically over health-inflected data and are entering into outright health data sharing agreements with other service providers.²⁹

However, direct health-inflected data are not the only health data that are collected and commercially exploited in the big data markets. Algorithmic probabilistic inferences have indeed widened the health information sources far outside the medical domain.³⁰

---

²⁴ Ibid., 72.
²⁶ Ibid., 75.
³⁰ Hoffmann and Podgurski, supra note 13, 86; Terry, supra note 11, 390-393; Pasquale, supra note 2, 60.
Through the employment of associative and probabilistic analytics, runaway data related to travels or financial transactions, or statistical data referring to crime rates or to house prices in certain areas, become silent and unexpected informants of (current and future) health conditions. Likewise, GPS information regarding access to an emergency room or apparently neutral actions such as online searches or purchases could signal the occurrence of an illness.31

If one thinks about the uncountable possible health relations between the most different aspects and expressions of individuals’ lives and lifestyles, potentially every type of data spread on the web could become health inflected.32

2.2 Data Mining Technologies and the Never-ending Lifecycles of Health-inflected Data

Data stemming from the above-outlined sources are collected and associated through data mining technologies.33 The very feature of data mining techniques lies in their capability to “uncover hidden patterns and subtle relationships in data, and to infer rules that allow for the prediction of future results”.34 Patterns are hypotheses produced by algorithms on the basis of the premises given by a certain set of data. These hypotheses are generated by so-called recognition algorithms,35 which search for variables typically matching with each

---

31 Pasquale, supra note 27, 100-101, giving the example of a woman that had searched online for information regarding multiple sclerosis, subscribing therefore a recommendation service for physicians. That data was collected by the KBM Group, a data analytics and marketing company, that subsequently sent her promotional material for an event for multiple sclerosis sufferers.

32 See R. Rubinstein, ‘Big Data: the End of Privacy or a New Beginning?’, *International Data Privacy Law* 3(2) (2013) 74 ff.; Terry, supra note 11, 386. As has been stated, corporations use a huge amount of “non-traditional” third party data sources, such as consumer buying history, to predict a life insurance applicant’s health status with accuracy comparable to a medical exam”. See D. Robinson and H. Yu-A. Rieke, ‘Civil Rights, Big Data, and Our Algorithmic Future’, September 2014, available online at http://centerformediajustice.org/wp-content/uploads/2014/10/Civil-Rights_Big-Data_Our-Future.pdf, 6.


other on account of association rules and link analyses generating secondary data. This secondary data derives either from direct matches, from probabilistic links or predictive guesses: a data subject could be said to have diabetes based on a clinical record (direct match), to be diabetes-exposed for his family history (probabilistic link), or to be diabetes-concerned based on his dietary habits or his sedentary lifestyle (predictive guess).

In this light, the categorization made by the OECD in the Report on Data-Driven Innovation for Growth and Well-being appears to well suit the case of big health data. The report distinguishes between provided data, data stemming from individuals’ direct and voluntary actions; observed data, data involuntarily shared by data subjects and hence collected without the data subjects’ direct knowledge; derived data, that can be logically derived from existing combined data, and finally inferred data, that cannot be deduced logically from the available provided and observed data but that is obtained using intelligent systems relying on probability-based analytic tools. Inferred data is in turn assembled for the identification of new patterns on the basis of which new

36 With regard to predictive analyses, the literature has distinguished between different types of predictions, including ‘preferential predictions’ regarding individuals’ preferences, ‘consequential’ ones, related to the consequences of certain behaviours, and finally ‘pre-emptive’ ones, directly used to influence the future options of an individual. See L. Moerel and C. Prins, Privacy for the Homo Digitalis: Proposal for a New Regulatory Framework for Data Protection in the Light of Big Data and the Internet of Things (Kluwer: Alphen aan den Rijn, 2016) 29-30.


38 Inferred and derived data pose some problems with regards to their accurateness and truthfulness: it is indeed evident how the more a certain health status is indirectly attributed and inferred through probabilistic analyses, the greater the likelihood of analytical mistakes is. T.Z. Zarsky, ‘Desperately Seeking Solutions: Using Implementation-based Solutions for the Troubles of Information Privacy in the Age of Data Mining and the Internet Society’, Maine Law Review 56(1) (2004) 14, 23-25; Mayer-Schonberger and Cukier, supra note 1, 43. With regard to the problem of accuracy, literature distinguishes between material and procedural accuracy. Procedural accuracy refers indeed to whether a company takes the necessary measures to warrant the integrity of data in the moment of the collection of the same data. This is very difficult to achieve in a situation where thousands of data are collected automatically from a very vast range of sources. Conversely, material accuracy refers to an even more delicate feature that collected data should possess, namely its correspondence to reality that is, in other terms, whether the data collected and further processed provide a correct representation of facts. For a broader assessment on the issue see Moerel and Prins, supra note 36, 69.
data are generated with increasingly weaker connections to the actual health conditions of data subjects.\textsuperscript{39}

The ceaseless collection and extrapolation of data, together with the mainly automated nature of data mining processing, creates a generative cycle of health data, in the spiral of which always newly generated health data is produced. This data has a growingly weak relationship to the actual health condition of the data subject to which the data refers. All these data, as combined and associated to each other, are finally arranged into “clusters” of homogeneous groups in which the mined databases are ultimately structured.\textsuperscript{40} The generative nature of data mining processes enables to extract new data from the formed clusters, thus triggering potentially infinite circles of further data processing based on always newly discovered information patterns and variables (see Fig. 3).\textsuperscript{41}

3 Health Data Uses

The empowerment of computational correlations techniques, has radically changed the spectrum of data processing possibilities and thus also of the uses made of derived health-inflected data. The increasingly open design of health IT systems together with the empowerment of computational analyses has rapidly induced a shift from static to more dynamic and generative data aggregation methods\textsuperscript{42} in that way opening up new frontiers for health data uses.\textsuperscript{43}

The circulation of sensitive health information outside its “natural” scope of action, urgently requires a reassessment of the balance between the benefits and the risks of the use of aggregated electronic health records.\textsuperscript{44}

\begin{itemize}
\item \textsuperscript{40} Ibid., 23. “The algorithms thus make several initial divisions into groups and inform the user of the strengths of correlations between the defined variables within every group and of the possible overlaps between the suggested clusters. After examining these options, the miner can decide on the level of acceptable accuracy and request that the software divide the database into the final grouping”.
\item \textsuperscript{41} Zarsky, \textit{supra} note 34, 9.
\item \textsuperscript{42} Cohen, \textit{supra} note 3, 86.
\item \textsuperscript{43} For a more general historical reconstruction of the processing of aggregated data, see D.J. Solove, ‘A Taxonomy of Privacy’, \textit{University of Pennsylvania Law Review} 154(3) (2006) 477, 505-518.
\end{itemize}
The healthcare uses\textsuperscript{45} encouraged by open health information systems, not only in terms of research stimulation\textsuperscript{46} but also in terms of epidemic tracking\textsuperscript{47} and of monitoring medicine flows must be indeed balanced against the non-healthcare uses that health data are undergoing: free running health data spread in the digital net have soon become of key significance in the context of data markets governed by data brokers, such as credit bureaus, analytics firms, direct marketers, list brokers, affiliated and others.\textsuperscript{48} Aggregated health data have in recent times become one of the most profitable sources of business competitiveness and efficiency.\textsuperscript{49}

\textsuperscript{45} Literature has widely reflected on non-commercial uses of health data, first of all, for the promotion of high-quality care, see B.R. Furrow, ‘Data Mining and Substandard Medical Practice: the Difference between Privacy, Secrets, and Hidden Defects’, \textit{Villanova Law Review} 51(4) (2006) 803. Stressing the enhancement of pharmacovigilance systems through the massive collection of data, Pasquale, \textit{supra} note 44, 697.


\textsuperscript{48} Pasquale, \textit{supra} note 2, 32.

3.1 From Individual Health Profiling to Health “Pigeonholing”

Automated algorithmic computation power has come to strengthen companies’ profiling and tracking capabilities. Data mining techniques have transformed the countenance of health data processing, which is increasingly shifting from individual tracking to collective profiling practices, cataloguing data subjects on the basis of selected features commonly shared among a group. Data subjects are thus treated as “deconstructed” entities, variously fitting into always newly formed clusters.

The clusters originating from the collected and processed health-inflected data are outright models in which data subjects can be temporary or long lastingly trapped. These models have the effect of “pigeonholing” individuals into predetermined categories formed on the basis of data generated not only by direct matches but also by probabilistic links and predictive analytics. This means that algorithmic computation may be systemizing and segregating individuals on the basis of health scores determined by mere proxies, i.e. data only probabilistically related to the clustered data subject. This raises thoughtful concerns regarding the accuracy of these same models.

50 As has been noticed, the enhancement of analytical tools used for the processing of data regarding individuals’ traits and behaviour has enormously enriched companies “digital dossiers”. On the issue see D. Solove The Digital Person, Technology and Privacy in the Information Age (New York: NYU Press, 2012), 32; See also Pasquale supra note 27, 116-117, talking about the “rise of the personal prospectus”.

51 “(...) The gathered information is analysed possibly with the addition of data purchased on the now-vibrant data market, to create a profile of the individual, or of a group of individuals fitting certain criteria”. See Zarsky, supra note 38, 22, emphasis added. On the issue see also Solove, supra note 50, 36; E. Singer, ‘The Measured Life’, Mit. Tech. Rev. (2011), available online at https://www.technologyreview.com/s/424390/the-measured-life/.


53 Talking about a shift from merely ‘individual’ profiling to ‘racial’ and ‘ethnic’ profiling, Cohen, supra note 39, 167.


56 Proxies are factors, that when applied in concert, are correlated with the discriminatory factor beyond a specific threshold. So N.P. Terry, ‘Big Data Proxies and Health Privacy Exceptionalism’, Health Matrix 24 (2014) 65, available online at http://scholarlycommons.law.case.edu/healthmatrix/vol24/iss1/6, 77-79.
The so formed clusters become the basis of critical judgements about data subjects carried out in the context of various business operations, ranging from direct advertisements and marketing,\textsuperscript{57} to credit issuing and employment.\textsuperscript{58} In all these sectors, decision-making processes increasingly rely on scoring charts determining differential economic treatments.\textsuperscript{59} The aftermaths of this processing scenario are even greater if one considers the fact that the formed clusters interact with a potentially infinite, and constantly regenerating, health-related information landscape. This generates a “negative spiral”,\textsuperscript{60} deterministically affecting individuals’ room of free choice and action.\textsuperscript{61}

Against this bedrock, it appears that the issue of health privacy has become strictly related to the much wider and complex matter of inclusion and exclusion from society,\textsuperscript{62} along the lines of undercover stigmatization and discrimination courses\textsuperscript{63} resulting from the process of identification, classification and

---


\textsuperscript{58} Pasquale, \textit{supra} note 2, 15.

\textsuperscript{59} The systematic collection of personal data has, for example, stirred the practice of personalised pricing. The White House, ‘Big Data and differential prices’, issued on February 2015, available online at https://www.whitehouse.gov/sites/default/files/docs/Big_Data_Report_Nonembargo_v2.pdf, 8 ff.

\textsuperscript{60} Pasquale and Citron, \textit{supra} note 55, 12.


\textsuperscript{62} Moerel and Prins, \textit{supra} note 36, 24.

\textsuperscript{63} Scoring modelling has indeed become one of businesses’ primary tools for increasing revenue or profits from specific transactions: these practices have been also appointed
comparative assessment of groups analytically generated on account of their expected health value or risk (see Fig. 4).64

As this image shows, data mining techniques search for correlations existing among different data points. These correlations result in classifying clusters. Predictions regarding the data subjects’ health status are then made from these clusters. Based on these variously inferred predictions, new clusters are formed directly serving decision-making purposes. New correlations can be inferred from these decision-making clusters, with that newly activating the entire circle.

3.2 The Obscurity of Health Data Processing

Health data mining techniques share a specific feature, the obscurity of the criteria governing the algorithms processing health data.65 Algorithms determining the clusterisation of data subjects on the basis of health-inflected data are obscure respectively for structural features, technical remedies, and ultimately, for legal measures that render algorithms inaccessible to external enquiry most

---


65 In more general terms the risks of discrimination in data mining techniques were emphasized in technical literature for the first time by D. Pedreschi, S. Ruggieri and F. Turini. ‘Discrimination-aware Data Mining’ (2008), Proceedings of the 14th ACM SIGKDD international conference on knowledge discovery and data mining, available online at http://dl.acm.org/citation.cfm?id=1401959, 560-568.
of the time. The generative nature of machine-learning capabilities makes it difficult to spot which algorithm has carried out the single data processing. The criteria used for the classification and for the subsequent decision-making are non-interpretable, and hence nearly impossible to put under human (and judicial) review.

Moreover, technical anticircumvention measures most of the time impede the detection of how information is used and how the decisions about these data are carried out.

With regards to the legal tools chosen by businesses to protect their algorithms, businesses refuse to disclose their processing criteria mostly leveraging on trade secret protection. Trade secrecy is more convenient for businesses in respect to patent protection, which is, as is known, structurally based on the disclosure of the features of the protected technical devices.

Trade secrets are however not the only legal tools obscuring data mining processes: aggregated data can find legal shield under the sui generis database right and through copyright in the case the selection and arrangement of the

---

66 Pasquale supra note 2, 29; see also Cohen, supra note 3, 122.
69 Pasquale, supra note 67, 245. Due to trade secret protection it is thus difficult to ascertain if the processes carried out are correct or, to the contrary, somehow compromised with regards to inputs or outputs. Pasquale and Citron. supra note 55, 10.
71 This right provides however static protection to a precise set of collected data: it does not cover the previous activity of collection and selection of the data and the subsequent phase of analysis of the same data. The European Court of Justice has expressly excluded from database protection secondary data as aggregated in so-called “spin-off databases”. European Court of Justice C-203/2002, The British Horseracing Board Ltd and others v. William Hill Organization Ltd. Given that in the case of big health data it is nearly impossible to distinguish between primary and secondary data and that the most valuable
databases satisfy the originality requirement.\textsuperscript{72} Copyright can also be invoked for the protection of software processing the data.\textsuperscript{73}

Finally, other tools of protection are to be found outside the intellectual property framework, through recourse to contractual non-disclosure agreements. End-user licence agreements are indeed extensively used to restrict access to sensitive information.\textsuperscript{74}

Hence, the combination of proprietary and contractual tools obscure ranking and rating schemes, making it difficult for a third party to detect businesses’ decision-making benchmarks\textsuperscript{75} and thus arduous for wronged parties to pursue their cause.\textsuperscript{76} The ubiquitous collection and processing of health-inflected information significantly weakens the protection of data subjects’ rights. In the next section, we quickly sketch the answers the \textit{GDPR} provides.

\section{The \textit{GDPR}'s Regulatory Answers}

\subsection{“Automated Decision-making” under the \textit{GDPR}}

The General Data Protection Regulation takes into consideration the practice of collective profiling under Art. 22 \textit{GDPR} concerning “automated individual decision-making, including profiling”. The practice of profiling is very broadly defined as:

\begin{quote}
and used health data are the inferred-\textit{i.e.} secondary generated- ones, database protection is to be considered only an eventual and optional protection measure. On the issue see A. Wiebe, ‘Protection of Industrial Data: a New Property Right for the Digital Economy?’ \textit{Journal of Intellectual Property Law and Practice} 12(1) (2017) 62-71.
\end{quote}

\begin{quote}
\end{quote}

\begin{quote}
However, it must be recalled how recital 11 of the Directive 2009/24/EC of 23 April 2009 on the protection of computer programs expressly affirms that copyright protection of computer programs should not extend to “the ideas and principles which underlie any element of a program”. Copyright protection impedes indeed the reproduction of a certain computer program, but not the detection of its underlying functioning criteria. So Drexl, et al., supra note 70, 6.
\end{quote}

\begin{quote}
Pasquale and Citron, supra note 55, 20-22; Hoffmann and Podgurski, supra note 14, 505-506.
\end{quote}

\begin{quote}
\end{quote}

\begin{quote}
Mattioli, \textit{ibid.} 71, 544.
\end{quote}
any form of automated processing of personal data consisting in the use of personal data to evaluate certain personal aspects relating to a natural person in particular to analyse or predict aspects concerning the data subjects’ performance at work, economic situation, health, personal preferences or interests, reliability or behaviour, location or movements, where it produces legal effects concerning him or her or similarly significantly affects him or her.\textsuperscript{77}

The provision appears to refer to the case of a wide range of algorithms employed in various sectors, such as the ones of recommendation systems, credit and insurance risk assessments, computational advertisements and social networking.

Art. 22 of the GDPR does not entail a specific prohibition of the so defined practice of automated profiling, affirming instead that “the data subject shall have the right not to be subject to a decision” carried out automatically. The wording appears quite ambiguous. It remains unclear whether it is a general provision – thus an implicit prohibition – or rather a right that has to be individually actioned by the interested subject. If this is the case, it is equally unclear how an individual should be aware that an automated processing of his or her data is going on and thus ask not to be subjected to the decision taken on the basis of the connected automated processing.\textsuperscript{78}

An individual enactment of the right appears even more difficult in light of the above-stressed generative nature of the algorithmic data processing through which always newly generated data are used to form clusters forming the basis of always different decisions.

By stating that:

\begin{quote}
the controller (...) should implement technical and organizational measures (...) that prevent, inter alia, discriminatory effects on natural persons based on racial or ethnic origin, political opinion, religion or beliefs,
\end{quote}

\textsuperscript{77} See Recital 71 of the General Data Protection Regulation, emphasis added. Through reference to the “legal effects” of the processing, the European legislator appears to expressly refer to those cases in which the processing of sensitive (health) data have legal consequences on the data subjects’ life. For a critical assessment on this see Goodman and Flaxman, \textit{supra} note 30, 3.

trade union membership, genetic or health status or sexual orientation, or that result in measures having such an effect.

recital 71 GDPR appears to implicitly admit the legitimacy of automated processing and at the same time shows to recognize the discriminatory threats underlying these profiling practices. Art. 22.4 GDPR poses a precise prohibition of profiling practices, but only in respect to the “special categories of data” listed in Art. 9 GDPR, among which health data is also included.

The different regulatory solutions given in respect to profiling regarding sensitive data (expressively prohibited), and non-sensitive data (implicitly admitted), need to be questioned in light of the intertwining between sensitive and non-sensitive data. This stems from the never-ending lifecycles of health data extracting information from non-sensitive data that is highly sensitive.79

4.2 Transparency Remedies

The GDPR has introduced some data protection measures specifically aimed at shedding light over the technical features of data mining processing and at spotting the risks underlying the same processing activities.

In this light, the provision of Art. 35 GDPR appears to be of particular interest for it requires the controller to carry out a “Data Impact Assessment”, consisting in an “assessment of the impact of the envisaged processing operations on the protection of personal data”. As has been observed by a strand of the literature, the function of the Data Protection Impact Assessment is that of ex-ante mapping the possible threats to data subjects’ rights originating by a certain processing activity. This is done through a “systematic evaluation of personal aspects related to natural persons which is based on automated processing including profiling, and on which decisions are based that produce legal effects concerning the natural person or similarly significantly affect the natural person”.80

Moreover, the GDPR has acknowledged the need to mitigate purely technical algorithmic processing, through the establishment of “appropriate safeguards, which should include specific information to the data subject and the


80 Art. 35, 2 lett. a) of the General Data Protection Regulation.
right to human intervention”. Accordingly, articles 13-15 GDPR entail a “right to explanation”, consisting in the data subject’s right to be informed about the “existence of automated decision-making, including profiling, referred to in Art. 22(1) and (4), and at least in those cases, meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject”. The so defined right is a direct expression of the principle of accountability enshrined in Art. 5.2 of the GDPR. The same accountability ratio underlies the obligation born by controllers under Art. 30 of the GDPR, imposing the redaction of an “internal record of processing” that is to “be made available to the supervisory authority on request”.

4.3 Health Data Mining Challenges to Data Protection: Some Open Issues

It has been so far demonstrated how the process of collection of health data mostly happens outside the strict healthcare sector and how the process of processing health-inflected data mostly ends up affecting decisions that are not at all connected to the healthcare sector. This peculiar processing environment is very difficult to enshrine and thus properly regulate. The GDPR has reformed the data protection framework by introducing provisions intended to be more adherent to the current technological landscape. However, also under the reformed framework, effectiveness of protection appears very difficult to reach. In this perspective, it can be reasonably affirmed that health data mining techniques call for a rethinking of many fundamental assumptions of data protection law. Just a few considerations will suffice to illustrate this.

Although the GDPR affirms, in consistency with the principle of data minimization, that “personal data shall be (...) adequate, relevant and limited in relation to the purposes for which they are processed”, the current processing reality contrarily suggests that we are in an era of outright data maximization where data is exploited for the generation of other data. This data appears to

81 Ibid., recital 71.
82 Ibid., Art. 13, 2 lett. f). Emphasis added.
83 Ibid. On the debate surrounding these articles of the GDPR and for their systemic interpretation, see Malgieri and Comandé, supra note 78.
have increasingly weaker factual connections to the data subject and to his/her actual health conditions.\(^8^5\)

In the face of the massive data transfers happening in the backyards of companies’ exchange agreements, also the purpose limitation and proportionality principles appear to be weakened in an information environment where data are differently used and combined for decision-making occurring in totally different contexts from the ones in which the data were originally collected. In this regard, it is interesting to notice that the GDPR has adopted a more flexible approach to the purpose limitation rule, allowing, under Art. 6, secondary uses and thus the further processing of the data originally collected for a specific purpose when the secondary use and further processing is carried out for a purpose that is “compatible” with the purpose of collection.\(^8^6\) It is evident that the compatibility test provides a \textit{de facto} exemption to the principle of purpose limitation, allowing greater flexibility of processing for secondary purposes.\(^8^7\)

As far as consent is concerned, it should be considered that the present processing capabilities render it very difficult to quantify the actual amount of the data being processed based on a single given consent.\(^8^8\) In this light, the task of discerning \textit{to whom} the consent is given and \textit{for which} processing activities becomes a very hard one.\(^8^9\)

\(\text{---}\)

85 As it has been very well expressed, “while in the past personal data were primarily a by-product of the purpose for which they were collected (such as in the example of booking a flight), this is no longer the case today whereby data are increasingly collected precisely in order to determine which services may be of interest to start providing”. Moerel and Prins, \textit{supra} note 36, 44.

86 See Art. 5.1 of the General Data Protection Regulation: “Personal Data shall be b) collected for specified, explicit and legitimate purposes, \textit{and not further processed in a manner that is incompatible with those purposes}”. Emphasis added. It is worth to stress that the compatibility test has been presented by the legislator in the negative form, with a double negation (no further processing if that processing is incompatible), with that providing even greater flexibility, given the fact that the data controller must not prove the compatibility of the secondary processing, but must rather show the absence of incompatibility.

87 Moerel and Prins, \textit{supra} note 36, 44.

88 On the issue see European Commission General Data Protection Regulation: IAB UK Response to Ministry of Justice Call for Evidence, issued on September 2012, available online at https://www.iabuk.net/sites/default/files/EC%20Data%20Protection%20Rules%20-%20IAB%20UK%20Response%20to%20MoJ%20Call%20for%20Evidence.pdf, 2: “We believe explicit consent is difficult to implement in practice in a digital environment and may place a significant burden on businesses and a cumbersome online experience for users.

Furthermore, the highly sophisticated technological environment renders it difficult to spot who the actual data controllers and data processors are. Against the backdrop of the changed processing reality, the distinction between data controllers that decide on “the purposes and means” of the processing and all the other processing subjects involved, such as data processors, recipients and third parties, has become very difficult to detect. As it has been stated, the assumption that “a controller is always identifiable and accountable and that it is up to him to decide whether to attribute the data processing to a data processor or to other parties, does not stand any more in contemporary processing environments”. During most of the processing activities, more than one data controller and processor are usually involved. The Regulation has clearly acknowledged this, through the introduction of the scheme of “joint controllers” to be found “where two or more controllers jointly determine the purposes and means of the processing of personal data”.

The peculiarities of health data mining techniques shutter the borders of data protection categories also from a further perspective, directly related to the generative potential of algorithms: as illustrated above indeed, the correlative patterns discovered by computational analytics and the secondary predictive data stemming from these, appear to blur the borders between non-sensitive and sensitive personal data with the connected regulatory distinctions. This creates great legal uncertainty, given that potentially every type of data, both of a personal or non-personal nature, is likely to become the source of very sensitive health information. In this regard, the endurance of the prohibition posed under Art. 9 of the GDPR with reference to the processing of “special categories of data” such as “data concerning health or data concerning a natural person’s sex life or sexual orientation”. The actual scope of this prohibition evidently

---

92 Ibid., 186.
93 So Art. 24 of the General Data Protection Regulation. See Guarda, supra note 108, 8, stating that “the particular case of co-data controllers on the same treatment” is to be found “when the choices on purpose, method, tools and security measures for the treatment are related to multiple subjects”.

needs to be better clarified in light of the sensitive-signalling potential of apparently neutral and hence non-sensitive personal data.94

Finally, it must be observed how the collective profiling activities carried out over probabilistically-inferred data seem to impair the same notion of data subject, traditionally intended as the physical individual to which a certain piece or set of data is to be referred to. In consequence of the growing classification patterns through which an individual with certain features is likely to bear the same decision-making outcomes of subjects with same or similar profiles, the latitude of the same notion of data subject should equally be reconsidered. Data subjects are not only- or not any more- the generator of specific data from which new data can be derived, but also the subjects to whom secondary data generated by other similar subjects are applied for the purposes of decision-making.95 Consequently, it appears that the notion of data subject has lost its previous strictly individual dimension, and has rather ended up assuming a collective scope;96 the risks related to the processing of a certain dataset are not any more related to the single data subject but rather to the category in which the single data subject has been systematized. As the never-ending lifecycles of health data suggest, this systematization mostly occurs based on data that has only an indirect connection to him/her health status.

Thus, under the umbrella of algorithmic computational analysis of health data, the data subject has become the one who is either included in or excluded from models governing businesses’ decision-making.97

94 Comandè and Malgieri, supra note 39, 23.
95 See Comande, supra note 52, 169.
96 L. Kammourieh, ‘Group Privacy in the Age of Big Data,’ in L. Taylor, L. Floridi and B. Van der Sloot (eds.), Group Privacy-New Challenges of Data Technologies (Springer: Basel, 2016) 37-66. The collective impact of the mass-processing of data enabled by data mining technologies, has been underlined especially by that literature that stresses the need of considering the interests involved in the data processing: “(...) law-makers will also need to define limits where data processing is not legitimate at all in light of the collective interests that are increasingly at stake”, so Moerel and Prins, supra note 36, 18. Emphasis added. In the same sense, see Spina, supra note 7, 248, 252, stating that “the market failures associated with forms of Big Data encourage us to look at privacy risks and at privacy as not only an individual right but a collective interest”, emphasis added.
Conclusions: Health Data as Drivers of Classification

The blossoming of data mining techniques has changed not only the way health data are collected, but also the way in which health data are generated and processed. In the current IT infrastructure, sensitive health information does not originate only from medical records any more. As this article has illustrated, the sources of sensitive health-related information have enormously expanded through algorithmic processing combining different sets of data, spotting correlation patterns and inferring probabilistic forecasts stemming from collected and secondary produced data. The new algorithmic processing infrastructure can detect sensitive health conditions from apparently neutral data, such as a simple online search or a GPS location.

The consequence of this is that nearly every piece of data circulating in the hands of private or public processors entails- either itself or through combination with other data- reference to an actual or potential health condition that will further constitute the factual basis of decisions taken far outside the healthcare field. As a matter of fact, directly or indirectly generated health data end up being processed and used for purposes that are very different from medical ones.

This technical reality has cherished and stirred the enactment of collective profiling practices based on scoring models- clusters, in more technical terms- determining individuals’ social opportunities. The collective scope of algorithmic processing is thus leading to the creation of outright social categories, on the basis of which credit issuers, insurances, advertisers and other service providers adjust their activities. In this light, health data as collected and systematized through clusters have become the hidden rules governing corporations’ decisions. As outlined above, technical stigmatization schemes determine a differential treatment between those who have been grouped as healthy, wealthy and generally risk-averse and those who, to the contrary, have not.

The code underlying the functioning of these algorithmic clusters and thus the rationale behind discriminatory practices, remains mostly obscure in virtue of both technical and legal measures. The secrecy outcomes of both intellectual property measures and contractual arrangements could be nonetheless mitigated by some of the newly introduced data protection measures, such as those regarding the right to explanation and data impact assessment. Although not addressing the problems of the technical obscurity of

98 Pasquale, supra note 27, 121-122.
99 Ibid., 125.
algorithmic processing, these measures are aimed at shedding some light on data processing mechanisms, with the purpose of enabling data subjects to track the decision-making processes that have been carried out based on the processed health data.

As has been outlined, the current health data environment— as far as both the data generation and processing methods are concerned— greatly challenges traditional data protection principles and categories. This does not, however, mean that data protection measures have lost their overall regulatory efficacy in respect to the phenomenon of big health data. On the contrary, by greatly emphasizing the principles of transparency and accountability, the new provisions of the General Data Protection Regulation show how, in the present data scenario, control over personal health data means knowledge over the sources and types of the risks that could stem from the processing of health data.\textsuperscript{100} In other terms, it means knowledge over the legal consequences of using health data as benchmarks for decision-making in fields far outside the healthcare sector.

The mapping of these risks and legal consequences however remain still a challenging regulatory task in the face of the never-ending lifecycles of health data occurring in the shadows of algorithmic analytics.

From a more general perspective, the case of health data triggers some thoughts on the systemic perspectives under which data protection law must currently be interpreted. From an objective standpoint, it is evident how data protection law is no longer a mere end in itself, but, on the contrary, the bastion for the protection of many other fundamental rights, first of which is the right not to be discriminated. Conversely, from a subjective standpoint, the assembly of data points to models on the basis of which individuals sharing similar features are judged. This calls for a reconsideration of the scope of data protection as a merely individual right; issues deserving further research.