Smichowski, Bruno Carballa; Duch-Brown, Néstor; Martens, Bertin

Working Paper

To pool or to pull back? An economic analysis of health data pooling

JRC Digital Economy Working Paper, No. 2021-06

Provided in Cooperation with:
Joint Research Centre (JRC), European Commission

Suggested Citation: Smichowski, Bruno Carballa; Duch-Brown, Néstor; Martens, Bertin (2021) : To pool or to pull back? An economic analysis of health data pooling, JRC Digital Economy Working Paper, No. 2021-06, European Commission, Joint Research Centre (JRC), Seville

This Version is available at:
https://hdl.handle.net/10419/266523

European Commission

JRC TECHNICAL REPORT

JRC Digital Economy Working Paper 2021-06

To pool or to pull back? An economic analysis of health data pooling

Carballa-Smichowski, B.
Duch-Brown, N.
Martens, B.

2021

Contact information
Name: Bruno Carballa-Smichowski
Address: Inca Garcilaso, 3. 41092 Seville (Spain)
Email: bruno.carballa-smichowski@ec.europa.eu

EU Science Hub
https://ec.europa.eu/jrc

JRC126961

Seville: European Commission, 2021

How to cite this report: Carballa-Smichowski, B., Duch-Brown, N. and Martens, B., *To pool or to pull back? An economic analysis of health data pooling,* JRC Digital Economy Working Paper 2021-06, JRC126961.

Contents

Executive summary 3

Introduction 6

1 The data-pooling dilemma: an economic analysis 8

1.1 A theoretical framework to analyze data-pooling 8

1.2 The baseline data-pooling dilemma 10

1.3 Zero-sum games 18

1.3.1 Equally divided benefits 18

1.3.2 Gains divided according to data endowments 19

1.4 Competing pools 21

1.4.1 Competing pools in non-zero-sum games 22

1.4.2 Competing pools in zero-sum games 23

1.5 Agents' data endowments distribution 23

1.6 Intra-pool negative externalities 28

2 Sub-optimal data-pooling scenarios 34

2.1.1 No data-pooling equilibrium 35

2.1.2 Suboptimal pool size with one equilibrium 37

2.1.3 Suboptimal pool size with multiple equilibria 38

2.1.4 Multiple equilibria with optimal pool size 39

3 Fostering data pooling 41

3.1 Acting on incentives to pool data 42

3.1.1 Providing economic incentives 42

3.1.2 Facilitating hard and soft data interoperability 44

3.1.3 Leveraging on public actors' data 46

3.2 Acting on control rights over data 47

3.2.1 Identifying, diffusing and implementing trust-building mechanisms 47

3.2.2 Mandating data-pooling on the grounds of public interest 50

3.2.3 Developing *de facto* and *de jure* collective control rights over data 52

3.2.4 Enabling posthumous medical data donation 58

4 Conclusions 60

References 63

Authors

Bruno Carballa-Smichowski

Néstor Duch-Brown

Bertin Martens

Abstract

We present a novel generic theoretical framework to analyze the incentives agents have to engage in n-way data sharing or 'data pooling' and the factors affecting those incentives. Based on the results obtained, we provide policy recommendations aimed at fostering health data pooling. Section 1 develops a baseline framework and multiple variations including zero-sum data pooling games, competing pools and intra-pool negative externalities. The section offers analytical solutions and examples to show under which conditions agents decide to pool data. Section 2 illustrates how different factors can lead to sub-optimal data pooling. Section 3 provides policy recommendations to foster data pooling in the health sector and discusses the conditions under which they can be effective.

Executive summary

It is well recognized by scholars (Shaw et al., 2016; Strotbaum et al., 2019) and the European Commission in its Data Strategy (2020) that the exploitation of health data can have a high social impact in two ways (Hummel et al., 2019). First, medical data can contribute to research by paving the way for new hypotheses, treatments, diagnostics and preventive actions. Second, clinical deep-learning-driven treatments and diagnostics such as the delineation of tumors in radiological images or therapeutic decision-making on metastatic breast cancer require large amounts of patient data, clinical cases and background information.

Data pooling (i.e. multilateral data sharing where several agents share their data with each other) is a data sharing modality with high potential for the health sectors of European countries. The envisaged Common European Health Data Space (European Commission, 2020) could constitute a European health data pool. However, the emerging economics literature on data sharing, which focuses on unilateral data sharing, has so far investigated the incentives agents have to join a data pool.

The purpose of this report is twofold. We seek to contribute to the emerging literature on data sharing by investigating the determinants of data pooling under different scenarios in order to show how certain incentive structures and distributions of control rights over data can result in suboptimal welfare outcomes. In other words, we will show how even in presence of economies of scale, scope and quality from data pooling, the structure of incentives and the distribution of control rights over (health) data can lead to not grasping its potential in terms of social welfare benefits and public health. Based on these findings, this report aims at contributing to the Data Strategy by providing a framework of analysis and policy recommendations to foster data pooling with a focus on health data.

We develop a novel framework to study the incentives agents have to share data within a data pool. We show how the following factors, which can have opposite effects in a given (potential) data-pooling situation, can affect an agent's decision on whether and with whom to pool data: the relationship between the economies and costs of data pooling, the relationship between the marginal utility of data-pooling and coordination costs, the existence of a zero-sum game and the associated value sharing rules of the pool, the existence of competing data pools, agents' data endowments distribution and the existence of intra-pool negative externalities. We show how these factors can result in sub-optimal data pooling. On those bases, we derive several policy recommendations on how to foster data pooling with a focus on health data and discuss the conditions for each of them to be pertinent and effective.

Providing economic incentives can foster data pooling if transaction or lump sum costs are high enough in comparison to benefits. In absence of sufficient incentives, there might be no data pooling or a data pool smaller than what policymakers would desire. These incentives can take several forms such as tax breaks, grants, investing in technological infrastructure that lowers the cost of data-pooling, financial support to (health) data cooperatives, attributing subsidies based on the scope of data-pooling or supporting data-pooling projects through grants

In cases in which the optimal data pool is not achieved because of the existence of high lump-sum (e.g., developing a standard to pool data that each agent stores in different formats) and coordination costs (e.g., agreeing on technical and non-technical standards in order to pool data), policies aiming at facilitating hard and soft interoperability are advisable, notably if the data is to be exploited by a third-

party that has to access several datasets held by different data holders. In the particular case of personal health data, the implementation of data donation passes that mirror organ donation passes would allow individuals to easily choose the conditions under which they will let third parties use their health data. In other cases, the development and diffusion of standard organization-to-organization health data-pooling agreements can contribute to develop soft data interoperability.

When public actors are (potential) members of the optimal data pool, leveraging on their data to achieve the optimal data pool configuration is a policy tool that can be used. Public actors can also condition their participation in a data pool to the acceptance of certain members that other pool members do not have incentives to accept.

When intra-pool negative externalities block the constitution of the optimal data pool, policymakers can act by identifying, diffusing and implementing (in cases in which they are part of the public actor is part of the data pool) trust-building data governance mechanisms. The latter should be based on two principles: transparency and accountability. Moreover, three types of trust-building mechanisms that build on these principles and are of particular interest in the case of health data can be fostered. The first one refers to mechanisms leading to agents feeling they are real interlocutors able to shape the system rather than passive stakeholders. Indeed, if agents (notably individuals, which are usually less powerful than organizations) feel they can shape the 'rules of the game' they will be more likely trust other agents with their data, as they know that they will have a say in what can happen if the uses or the provision of the data do not meet their expectations. The second type of mechanisms refers to technical means that minimize the exposure of data. Examples include safe havens, accreditation and technologies allowing exploiting several datasets as one without requiring data holders to migrate their data to a single joint server. The third one is dynamic consent, whereby data holders' preferences regarding how their data can be used and by whom can be exercised as the uses of their data evolve.

When data holders do not have incentives to pool certain datasets for a given purpose, facilitating the use of control rights over data or giving control rights to agents that do have an incentive to the share it within the optimal pool can be a promising solution. Given that in the case of health data these agents are generally individuals and that GDPR gives them portability rights over their personal data, these policies can focus on facilitating the use of (health) data portability (which falls under the current legal framework) and on enlarging the scope of control rights over personal (health) data individuals currently have, which entails modifying the current legal framework, notably GDPR. Policies that fall under the current legal framework include easing the administrative and legal procedures involved in a data steward claiming many individuals' data from data holders, promoting the use of certain standard personal health data licenses, demanding health data holders to implement a homogeneous and user-friendly interphase to allow individuals to exert their data portability rights and public bodies acting as certifiers of both health data cooperatives and data holders that comply with certain standards in terms of facility of data portability, privacy protection and ethical uses of the data. Two types of policies can be put into place to give individuals more control rights over personal data by modifying the current legal framework. The first consist in clarifying the scope of GDPR in order for individuals to be able to have larger control rights over (some) of observed and inferred data that relates to them, eventually under certain circumstances that remain to be specified. The second type of policy consists in creating collective consent and portability rights on relational data (i.e., data that relates to at least two identified or identifiable natural persons), as the current individual-consent based data protection regime hinders data

pooling by individuals when 'personal' data involves several individuals. The latter implies a careful legal and ethical discussion in order to deal with issues such as the articulation with individual control rights over data, who should be allowed to be a member of the collective, the governance of the collective and how do the collectives come about.

Legislating to allow for posthumous medical data donation (PMDD) would allow to pool and exploit data that is currently unused because GDPR does not recognize post-mortem privacy. Doing so would give individuals and data processers a legal ground on which donate and exploit medical data, respectively. In addition to this change in legislation, other implementation policies can be put into place to foster PMDD. Data donor cards could be legally recognized in the same manner as donors' cards are in some countries to donate organs in order to facilitate the use of PMDD.

Finally, when pooling data is in the public interest but data holders have little incentives to pool their data, data pooling can be mandated. Mandated data pooling is an interesting policy option when the data of the optimal data pooled is scattered across several data holders with conflicting interests. This policy can take several forms. It can imply mandating the opening of the data or mandating that some legal persons should authorize access to the data they hold to other specified legal persons only for some specified uses.

Introduction

The digital transformation has brought about an increase in the production of data from all the actors of the health sector. As a result of the digitalization of healthcare systems and the development of new digital practices, increasing amounts of health data being produced from sources as diverse as the clinical care setting, longitudinal studies, clinical trials, surveys, test results, health billing and claims, clinical or biomedical research or patient-reported data. Moreover, data categories not labelled as health data such as GPS location or air pollution data can be used for medical purposes.

It is well recognized by scholars (Shaw et al., 2016; Strotbaum et al., 2019) and the European Commission in its Data Strategy (2020) that the exploitation of health data can have a high social impact in two ways (Hummel et al., 2019). First, medical data can contribute to research by paving the way for new hypotheses, treatments, diagnostics and preventive actions. Second, clinical deep-learning-driven treatments and diagnostics such as the delineation of tumors in radiological images or therapeutic decision-making on metastatic breast cancer require large amounts of patient data, clinical cases and background information.

Given that health data production is scattered across many actors, exploiting its full potential requires them sharing it with each other. However, as the European Commission points out in its Data Strategy (European Commission, 2020, p. 7), data sharing is currently insufficient. The under-use of health-related data represents a considerable opportunity cost for society that translates into societal and individual harms such as misdiagnosis, slower medical innovation, inefficient allocation of resources or harm to patients (Jones et al., 2017). Fostering health data sharing is hence one of the priorities of the Commission 2019-2025, as the creation of European Health Data Spaces to promote better access and exchange of data illustrates[1].

Data sharing is a label covering different realities (Martens et al., 2020) that can be categorized through three non-mutually-exclusive characteristics. First, data can be shared for free, in exchange for a monetary compensation, in exchange for other data or in exchange for a good or service. Second, data can be shared directly (sharing a dataset) or indirectly (sharing a data-driven service). Third, data can be shared multilaterally (several agents sharing their data with each other) or unilaterally (one agent sharing data with other agents that do not share their data with it). The economics literature interested in the determinants of data sharing (Acemoglu et al., 2019; Dosis & Sand-Zantman, 2019; Duch-Brown et al., 2017; Martens, 2020; Martens et al., 2020), and notably its branch focusing on information markets (Admati & Pfleiderer, 1986; Bergemann & Bonatti, 2019; Montes et al., 2019) has been investigating issues related to unilateral data sharing. Little research has been done to this point on the modality of data sharing on which this report focuses: multilateral data sharing or 'data-pooling'.

Data pooling is of particular interest regarding health data, as it constitutes a data sharing modality with high potential for the health sectors of European countries for at least two reasons. First, the benefits of data sharing in the health sector are generally grasped when agents with complementary skills share complementary datasets (e.g., scientists working on the Human Genome Project or the development of a vaccine), which requires multilateral data sharing. Second, contrary to what can be observed in other domains, there are clearly defined *de jure* control rights over health data that allow for a variety of

[1] See https://ec.europa.eu/health/ehealth/dataspace_en

organizational forms of data pooling. Health-data-specific legislation and the entry into force of General Data Protection Regulation (GDPR) in 2018, and article 20 on personal data portability in particular, allow individuals to exert control rights over their health data. As a result, new typologies of horizontally governed health data pools between individuals combining their data for purposes such as accelerating medical research (e.g. Midata, Salus Coop, Myco, Transiscope or Moipatient) have started to emerge in the European Union, which extends the realm of possible uses of pooled data in favor of public health. Organizations can also constitute data pools following a variety of governance systems. For example, the AI firm DeepMind offers healthcare partners such as some NHS trusts an app that analyzes patients' data to alert doctors when patients need attention in order to prevent deaths from acute kidney damage. In doing so, DeepMind's partners pool their data in exchange for a data-driven service that builds on the economies of scale, scope and quality that can arise from pooling data.

The purpose of this report is twofold. We will seek to contribute to the emerging literature on data sharing by investigating the determinants of data pooling under different scenarios in order to show how certain incentive structures and distributions of control rights over data can result in suboptimal welfare outcomes. In other words, **we will show how even in presence of economies of scale, scope and quality from data pooling, the structure of incentives and the distribution of control rights over (health) data can lead to not grasping its potential in terms of social welfare benefits and public health**. Based on these findings, this report aims at **contributing to the Data Strategy by providing a framework of analysis and policy recommendations to foster data pooling with a focus on health data**.

The report is structured as follow. Section 1 does an economic analysis of the determinants of data pooling. It describes the incentives, costs and risks that agents face under different scenarios when deciding whether to pool data and with whom. Section 2 builds on it to identify four types of situations in which these determinants can result in sub-optimal data pooling. With a focus on the particularities of health data, Section 3 describes the types of policies that can be put into place to foster (health) data pooling. The latter are divided into two families: policies acting on incentives to data-pooling (Section 3.1) and policies acting on control rights over data (Section 3.2). Section 4 summarizes the findings with a focus on the policies that can be implemented to foster data pooling and the conditions for each of them to be pertinent and effective.

1 The data-pooling dilemma: an economic analysis

In this section, we develop an analysis of the determinants of data sharing through data pools. After developing a theoretical framework in Section 1.1, we proceed to examine what are the effects of certain key characteristics of data pools on an agent's decision to share data, namely economies and costs of data-pooling (Section 1.2), zero-sum games (Section 1.3), competing pools (Section 1.4), agents' data endowments distribution (Section 1.5) and intra-pool negative externalities (Section 1.6).

1.1 A theoretical framework to analyze data-pooling

In order to develop a theoretical framework to analyze the determinants of data pooling we begin by reminding the main economic characteristics of data.

Data is a non-rival good: its use by one agent does not prevent another one to use it (Jones & Tonetti, 2020). Nonetheless, it is an excludable good. Agents can be excluded from its use through legal means (*sui generis* database protection, trade secret on databases and intellectual property rights on software needed to access it) even in absence of *de jure* property rights over data (Duch-Brown et al., 2017). Data can even be anti-rival in the sense that, when shared, its consumption by multiple agents can improve its use by other agents when there are positive network effects (Prufer & Schottmüller, 2017) or when the use of data makes it possible to improve the quality of the dataset (Carballa Smichowski, 2018). The quality of data refers to the characteristics of a dataset that make it easier to extract meaningful information from it. The meaning of quality is therefore highly dependent on the intended use, since data becomes information in a certain context (Floridi, 2014). This is one of the reasons why a dataset can be very worthy for a certain use but of little interest for others. For example, a dataset about cellphone charging stations in an airport that contains data about how many times cellphones have been plugged in is valuable for managers who want to decide where to place more charging stations based on the number of people that use the existing ones. However, if the dataset does not specify how long each charge lasted, it is of little relevance for companies that develop lithium batteries. It is difficult to list all the properties that constitute quality. In order to illustrate the multidimensional nature the term 'quality' acquires to qualify data, we will retain the following categories of quality employed by Floridi (2014): accuracy, objectivity, accessibility, security, relevancy, timeliness, interpretability and understandability. These dimensions of quality are not meant to be definitive or exhaustive, but rather an indication to the reader of what lies behind the word 'quality'[2]. Moreover, economies of scope in re-use (the same dataset is used for different purposes) and economies of scope in aggregation (two or more data holders achieve more value by pooling their complementary datasets) can emerge when agents decide to share data. Finally, data sharing can also lead to economies of scale when the increase amount of data generated through data sharing allows an agent to obtain a higher value per unit of data. Let us note that in the particular case of data-pooling data economies of scope in data re-use are not necessarily present, as data can be re-used without multilateral data sharing.

2 Other scholars (Batini & Scannapieco, 2006; Olson, 2003; Wang, 1998) have proposed different dimensions of the quality of data. For a good review of the literature on the quality of information see Batini & Scannapieco (2006).

However, agents might refrain from data pooling for various reasons. Data pooling might imply two types of disutilities. The first one refers to different types of costs: marginal or transactions costs of data-pooling (e.g. the cost of renting a server with a price depending on the volume of data stored), fixed costs (e.g. developing a technical infrastructure, maintaining an API, etc.) and coordination costs (all pecuniary and non-pecuniary costs related to coming to an agreement with other data holders before starting pooling data). This can be the case when the pooled data serves the development of a new service or product that generates a revenue that has to be shared between the data contributors. The second type of disutility is negative intra-pool externalities generated by other members of the data pool. Many types of negative externalities can take place such as a risk of privacy breaches, the loss of competitive edge or market share through the use of the data by other members of the pool (i.e. rivalry in the downstream market) or the loss of speed in the innovation race, for example. We will come back to negative intra-pool externalities in Section 1.6. Finally, when there is rivalry over the value created by the data pool, agents might refrain from pooling their data because, although this would create more value, the share of it they will get may not cover the costs of data pooling.

When making a data-pooling decision, agents have therefore to weight the benefits they could capture from the above-mentioned economies of data pooling against the disutilities proper to data pooling and decide whether they will exploit their own data alone or, on the contrary, they will participate of a data pool and exploit the pooled data along with other agents. This situation echoes two types of social dilemmas that have been studied in the common pool resources (CPR) literature (Garner et al., 1990; Ostrom, 2010; Pitt & Schaumeier, 2012). The first one is provision problems. When the disutilitiy related to data sharing are null or close to zero, agents will have incentives to share the data they produce as much as possible. However, they face production and sharing costs. The problem to be answered is therefore how to create incentives for agents to contribute to share data without freeriding. Solutions to these problems include government-funded public goods (e.g. public open data) and data commons in which not-for-profit agents participate. In these data commons, business models do not rely on private appropriation of data. The second type of problems that arises are appropriation problems. When intra-pool negative externalities are significant, or when there is rivalry over the value created through data pooling, an appropriation problem arises. The use (appropriation) of the shared resource (the common data pool) or the value created based on it might diminish the utility of using (appropriating) the shared resource (the data pool) or the jointly created value by another agent. Although the CPR literature studied this problem for rival resources, the underlying logic can be adapted in order to analyze data pooling.

In this paper, we will focus on appropriation problems to study data-pooling dilemma. This approach will allow us to understand the strategic choices most traditional health data collectors such as hospitals or medical research centers face. Indeed, they cannot (either for technical or legal reasons) and have little incentives to (given their non-data-driven business models) increase or improve data collection in order to benefit from data pooling. For these data collectors, health data is most of the times a sub-product of their main activity that can however be pooled with other health data collectors' data to benefit from economies of scope in aggregation, economies of scale and economies of quality. As a result, data-pooling decisions are mostly based on considerations on whether and with whom pooling data with these decisions not affecting the level of production of data. Nonetheless, this logic does not apply to every health data collector. In particular, if health data is collected outside of the context of a treatment (e.g., through a wearable devise or a survey), the technical and legal barriers to collect health data in order to benefit from the economies of data pooling are lowered. In the same vein, health data collectors with a

data-driven business model have more incentives to collect data in order to benefit from the economies of data pooling. The acquisition of FitBit by Google provides a good example.

While the underlying logic of appropriation problems in CPR can be applied to data pooling, there is a major difference between CPR and data. Given that data is a non-rival resource, appropriation rules will rely on *how* can agents appropriate the shared resource (i.e., for which specific uses in downstream markets) rather than on *how much* of the resource they can appropriate[3]. In this regard, the appropriation problem in a data-pooling dilemma presents three similarities with club goods theory. First, like club goods, data is a non-rival and excludable good. Second, clubs can be formed to pool data on voluntary bases. Third, clubs can suffer congestion: as more agents join the club (data pool), intra-club negative externalities (can) become larger. In that sense, we can assimilate the above-mentioned negative externalities of data pooling to congestion.

However, there are three main differences between data pools and club goods. First, in data pools, the level of provision of the shared good (which can be interpreted as increases in the volume or the quality of the pooled dataset) can only be increased if certain agents (data holders) become members. On the contrary, traditional club goods can be enlarged independently of the number of members. Second, in club goods, congestion is a function of how many times the good is used in a certain period because club goods are rival. In a data pool, the existence and level of intra-club negative externalities depends on how (and not how much) certain agents (as some agents might produce no negative externalities) use the shared resource because this resource (data) is non-rival[4]. Third, given the excludable nature of club goods, only club members can benefit from their use. However, agents that are not members of a data pool can benefit from the economies of scope in aggregation, the economies of scale and the economies of quality generated by data pools. Data being non-rival, the members of the data pool can share the pooled data directly (economies of scope in re-use) or indirectly (creating or improving a service by using the pooled data) with non-members. This third difference between club goods and data pools will be of particular importance for policymakers, as we will see in Section 3.

Given the above-mentioned similarities and differences between data pools, on the one side, and club goods and CPR, on the other side, in the following lines of this section we will develop a novel theoretical framework that takes elements from both of these literatures.

1.2 The baseline data-pooling dilemma

In this section, we will develop the main features of theoretical framework to understand the logic of data pooling, which we will translate into what we will call hereafter the baseline data-pooling dilemma. This term translates the determinants of the strategic choice an agent has to make in order to decide

[3] In zero-sum-games the appropriation problem of how much of the common resource is appropriated can also exist: although the resource itself (data) is non-rival, the value it creates (e.g. the revenues generated by a data-driven new diagnostic technique co-developed between the members of the data pool) is rival. We will come back to this case in Section 1.3.

[4] In zero-sum-games the appropriation problem of how much of the common resource is appropriated can also exist: although the resource itself (data) is non-rival, the value it creates (e.g. the revenues generated by a data-driven new diagnostic technique co-developed between the members of the data pool) is rival. However, contrary to what happens in club goods, in which exclusion is difficult, agents cannot unilaterally decide to overuse the common resource. We will come back to this case in Section 1.3.

whether and with whom pooling its data in a baseline scenario in which the only factors to be taken into account are the economies of data-pooling, transaction costs, fixed costs and coordination costs. We will add more layers of complexity to this scenario throughout Section 1 in order to analyze how each of the possible features of data pools will affect agents' data-pooling decisions, namely zero-sum games (Section 1.3), competing pools (Section 1.4), agents' data endowments distribution (Section 1.5) and intra-pool negative externalities (Section 1.6).

Every agent holds a data endowment. There is therefore no production cost: we assume that the cost has been paid in a past period by the data holder. For each agent, being a member of the data pool implies a payoff that is a function of the collective value of the data pool. The capacity to appropriate this shared value, symbolized by factor a[5], depends on the agent's attributes (position in the market, technological capacity, efficiency, etc.). Depending on the nature and the intended use of the data, it can be altered through an agreement between the members of the pool (e.g. if it entails distributing the monetary value created by a service that requires accessing the pooled data) or not (e.g. if members use the pooled data for different commercial and non-commercial purposes).

Being a member of the pool also implies different types of costs evoked above. First, agents face marginal costs of data sharing (transaction costs, API running costs, renting servers space, etc.). Second, they face fixed costs such as interoperability costs. This lump-sum cost can be reduced either by choosing an easy-to-adopt standard/architecture or by sharing the cost of a common data-sharing infrastructure. In the latter case, this lump cost will depend negatively on the number of members of the pool. Third, they face coordination costs to constitute the data pool, which depend positively on the number of members and on their attributes as coming to a data-pooling agreement can be easier between certain types of agents. We will omit negative externalities at this stage and analyze them in Section 1.6.

The utility agent i receives when it shares data (U_i) is what we will call hereafter the 'baseline data-pooling dilemma'. Equation 1 summarizes it.

Equation 1: General form of the utility function of the basic data-pooling dilemma

$$U_i = a_i . f\big(X(s)\big) - t_i(x_i) - l_i - u_i(s)$$

Where:

a_i is agent i's appropriation factor. This factor translates how much of the total value created by the data pool is appropriated by agent i. When there is non-rivalry on the jointly created value (e.g., if the value translates in each member of the pool getting a better diagnosis) $\sum_i^S a_i > 1$. If, on the contrary, there is rivalry on the jointly created value (e.g., if the pooled data results in a drug patent generating revenue for

[5] If the pooled data results in a unique monetary value created by the pool, then the game is a zero-sum-game and therefore $\sum_{i=1}^{S} a_i = 1$. We will study this case in Section 1.3.

the members of the pool), then $\sum_i^s a_i = 1$. In the latter case, we are in presence of a zero-sum game. We will examine this case in detail in Section 1.3.

f() is the function that relates positively the size of the data pool to a level of value that can be created with it. This function can be linear, concave or convex depending on the existence and the nature of economies of scale, economies of scope and data quality improvements derived from data sharing.

X is the total amount of quality-homogeneous data shared by all members of the pool (i.e. the level of provision of the shared resource). The amount of data takes into consideration both the volume of the dataset and its quality.

x_i is the total amount of quality-homogeneous data shared by agent i.

$t_i()$ is agent i's data-sharing transaction cost or marginal cost function, which depends positively on the amount of data shared by agent i

l_i is the lump-sum cost of data sharing paid by agent i[6]

u_i *()* is agent i's coordination costs function, which depends on the amount of members the pool has and their attributes

s is the number of members of the pool

It should be noted that all of these variables could be interpreted as both monetary or non-monetary costs and benefits. For example, for an individual, marginal cost function $t_i()$ could be the time and effort implied in authorizing a health data steward to share his/her personal health data with a third party.

Data quantities x_i and *X* are notated as volumes of quality-homogeneous data. Any reference to the volume of data we will make hereafter should therefore be understood as units of quality-homogeneous data. This choice of notation accounts for the fact that the increase in value brought about by data depends on both its value and its quality. Hence, if two agents contribute with the same volume of data to the pool but each dataset has a different degree of quality, they will not contribute to creating the same amount of value for the data pool. As mentioned above, what constitutes the quality of data depends on its intended use (Floridi, 2014). Hence, the value of a quality-homogeneous datum is contextual to data pool's intended uses.

The first part of the equation shows the payoffs agent i will receive if it decides to join the pool and share its data with the rest of the members. It depends on function *f()* translating the economies of scope, scale and/of data quality improvements derived from pooling data and agent i's capacity to use that collective value to produce a positive payoff for itself, translated by factor *a*.

The second part of the equation represents the costs of data pooling: transaction or marginal costs (function *t*), fixed costs (*l*) and coordination costs (function *u*). The latter depend positively on the number of members of the pool and depends on their attributes, as coordination to set up a data sharing

[6] If members have to set up a common data-sharing infrastructure, they can share its cost. In that case, *l* will depend negatively on the number of members. For the sake of simplicity, we have not included this variation in Equation 1.

agreement might be easier between certain types of agents such as agents not competing in the same downstream markets.

Moreover, the following hypotheses will be made throughout the report when analyzing data-pooling dilemmas:

- **H1 – Consented entry**: agents cannot be forced to join a data pool

- **H2 – Free exit**: every agent can decide to leave a pool without costs whenever it wants

- **H3 – Exclusivity**: agents can only share a dataset with a single pool

- **H4 – Unicity of data endowments**: when an agent decides to share data, it shares its entire data endowment.

H1 allows us to understand the data-pooling choices agents would make without coercion. Based on this understanding, in Section 3.2.2 we analyze under which circumstances mandatory data pooling would be a suitable policy. **H1 does not mean that agents can freely join any pool. This will depend on the admission rules of each pool**. While some pools let any agent to join it without restrictions, some others might have tighter admission rules such as certain qualified members granted with admission rights accepting new members or every member agreeing upon accepting a new one.

The remaining hypotheses simplify the analysis. They can be dropped to reflect more accurately particular cases. H2 eliminates switching costs between data pools. However, they might exist, for example if a firm signs a contract according to which it has to pay other members of the data pool a fee in case it exists it before the end of the agreed-upon term. H3 eliminates situations in which an agent might participate in more than one data pool simultaneously with the same dataset. This could be the case if there is little or no rivalry between the data pools. In this case, each data pool could be analyzed separately using our framework. Otherwise, it is likely that pools will demand exclusivity to their members, which is translated by H3 in our framework. H4 eliminates the possibility of agents sharing a certain amount of a dataset with one pool and the rest with other pools. While technically feasible, this situation is unlikely, as we consider the amount of data shared by an agent in a data pool (x_i) to be the minimal amount of data required for the intended use the data pool has.

Having presented the underlying hypotheses of the data-pooling dilemma and the equation of its baseline form, we shall now proceed to assess the impact of different relations between the benefits and costs of data pooling on agents' decisions on data pooling. In order to do that, we will show under which conditions an agent with admission rights will decide to admit a new member to the pool (which increases the amount of data pooled) and what the determinants of the optimal pool size, and hence of the amount pooled data.

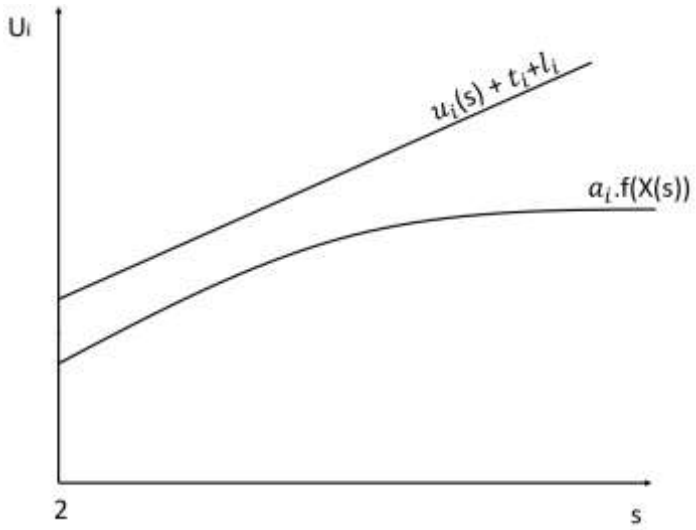
Intuitively, the stronger the benefits of data pooling are, the stronger the incentives to admit new members are. This can be easily appreciated by deriving U_i in Equation 1 with respect to *s* and finding the conditions under which the result would be positive, which results in the following expression.

$$a_i . f'\big(X(s)\big) > u'_i(s)$$

As shown above, agent i will admit a new member as long as the marginal utility brought about by the new member's data it can appropriate is higher than the coordination costs involved in admitting a new member. Hence, the stronger the economies of scope, scale and/or quality of data-pooling are, the more members the pool will admit. The optimal pool size will therefore depend on the comparison between the marginal benefit (economies of scope, economies of scale and/or economies of quality) and the marginal cost (coordination cost) of data pooling, as well as on the level of lump and transaction costs. Depending on the nature of economies of scope, quality and scale, the nature of coordination costs and the level of fixed and marginal costs of data sharing, four possible outcomes are possible.

The first possible outcome is that no pool will be formed. This would happen if, for every value of *s* above 2 (i.e. the minimum number of agents needed for data-pooling to take place), the gain from data-pooling appropriated by an agent i holding admission rights (the share a_i of the joint value *f()*) is lower than the sum of lump, transaction and coordination costs ($t_i()$, l_i and $u_i()$, respectively) born by agent i. For this to be the case, lump sum and coordination costs (l_i and u_i respectively) have to be higher than the appropriable benefit from data pooling ($a_i f()$) for s=2 and the marginal cost of data pooling have to be higher than the marginal benefit of it. Figure 1 illustrates it.

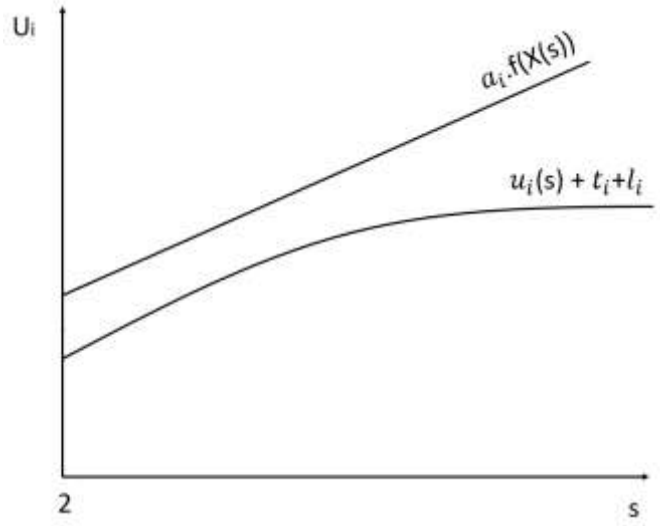
Figure 1: Example of economies and costs of data-pooling resulting in no pool formation



The second outcome is that a single pool with as many members as possible will be formed. This would happen if, for every value of *s* above 2, the gain from data-pooling for agent i holding admission rights ($a_i f()$) is higher than the sum of the lump, transaction and coordination costs it bears. For this to be the case, lump sum and coordination costs (l_i and u_i respectively) have to be lower than the appropriable

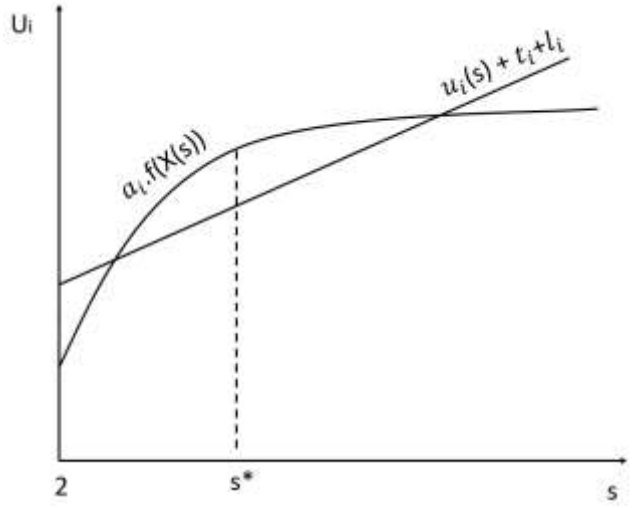
benefit from data pooling ($a_i.f()$) for s=2 and the marginal cost of data pooling have to be lower than the marginal benefit of it. Figure 2 illustrates it.

Figure 2: Example of economies and costs of data-pooling resulting in a single pool with the maximum number of members



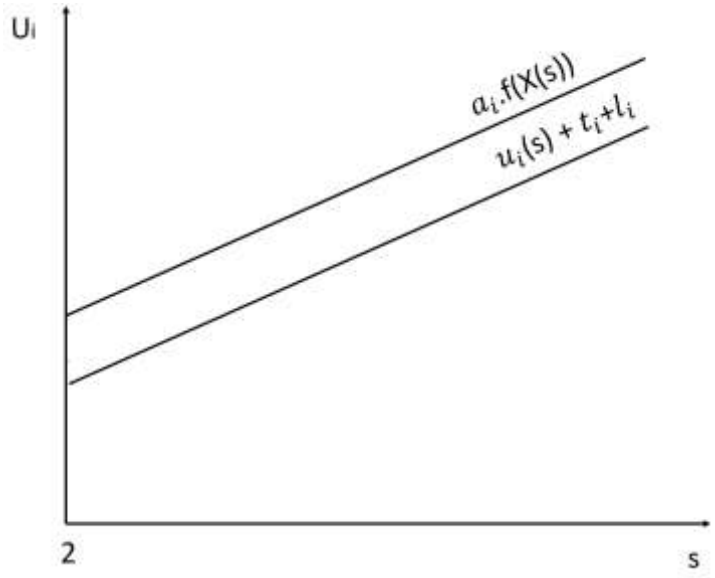
The third possible outcome is that at least a single pool with a number of members lower than the maximum will be formed. This would happen if, for certain values of *s* between 2 and a number lower than the total number of agents, the benefits from data-pooling appropriated by agent i holding admission rights ($a_i.f()$) is higher than the sum of lump, transaction and coordination costs it bears. Figure 3 illustrates it.

Figure 3: Example of economies and costs of data-pooling resulting in a pool with a restricted number of agents forming



The fourth possible outcome is that any possible number of members is an equilibrium. This would happen if, for every value of *s* above 2, the benefits of data-pooling appropriated by agent i holding admission rights ($a_i.f()$) is higher than the sum of lump, transaction and coordination costs it bears and the marginal benefit it obtains from it equals the marginal coordination cost of adding a new member. Figure 4 illustrates it.

Figure 4: Economies and costs of data-pooling resulting in any number of members being an equilibrium



We have presented the four possible outcomes illustrated by Figures 1, 2, 3 and 4 from the perspective of a single agent granted with admission rights. However, regardless of the governance structure of the pool, all the members have to agree to participate in it (cf. H1). As a result, **for any given number of members *s** that maximizes the utility of the agent(s) holding admission rights, all the admitted members' utility has to be higher within the pool than outside of the pool**. In other words, for every agent *j* being part of the optimal data pool.

$$U_j(s^*) > 0 \qquad \text{with } j \in s^*$$

We have so far presented the theoretical framework, the baseline data-pooling dilemma and the effects of the interplay between the economies and costs of data pooling on the optimal pool size under the baseline dilemma. We shall now introduce three features that add layers of complexity to the data-pooling dilemma and analyze their impact on the decision agents have to make on whether and with whom sharing data. Although we will continue using more complex versions of Equation 1 to obtain analytical results, the introduction of these features will need adopting a game theoretic approach.

The first feature we will introduce is zero-sum games. In zero-sum games, the total value produced by the pooled data is divided between the members of the pool. This can happen if the output of data pooling is a finite and divisible resource, typically money (e.g. the licensing of a patent or the earnings produced by a service developed using the pooled data) or if agents compete in the same market. For example, if agents are all medical service providers and compete with each other. In zero-sum games, the appropriation problem is more closely related to the CPR framework as, although the shared resource itself (the data pool) is not rival, the outputs it produces is. On the contrary, in non-zero-sum games such as the one presented in the baseline dilemma with $\sum_i^s a_i > 1$, each agent can appropriate value without that implying another agent appropriating less value. This is typically the cases in ecosystems in which agents use a shared data pool to produce or improve non-competing services.

The second feature we will introduce is competing pools. In a competing pools' scenario, the social value created through data-pooling (economies of scale, economies of scope, and/or improvement of data quality) will not only depend on the size and quality of its data pool, but also on that of other competing pools'. On the contrary, in a non-competing pools' scenario such as the baseline data-pooling dilemma, the value created through data pooling depends solely on the size of the data pool.

The third feature we will introduce is intra-pool negative externalities. In this case, the presence of a certain type of agent in the pool creates negative externalities that are suffered by one or more members of the pool at different levels. For example, when employers enter a health data pool, workers may benefit if it improves occupational health services, but might face a disutility if the pooled data can be used against them in selection interviews. Another example of intra-pool negative externality is the entry of an insurer in a health data pool to which some of its customers belong. In this case, the pooled data can be used by insurers to raise some customers' insurance policies based on insights drawn from the pooled data.

In order to isolate the effect of each of these features on the data-pooling decision, we will study them one by one. For every feature, we will show under which conditions an agent will decide to admit a new member to the pool, which increases the amount of data shared. This will determine the optimal pool size *s** in cases in which the benefits from data-pooling are not superior to the disutilities this involves for every possible number of members, as illustrated above by Figure 3.

Given that these features are not mutually exclusive, their possible combinations result in 8 types of data sharing attribution problems summarized in Table 2.

Table 1: Types of data-pooling attribution problems

	Non-competing clubs	Competing clubs
Non-zero-sum game	Data-driven improvements in quality or productivity for all the members (without intra-club negative externalities)	Competing data ecosystems to develop several uses (without intra-club negative externalities)
	Data-driven improvements in quality or productivity for all the members (with intra-club negative externalities)	Competing data ecosystems to develop several uses (with intra-club negative externalities)
Zero-sum game	Development of a monetized service facing little or no competition (without intra-club negative externalities)	Rent-seeking competition (without intra-club negative externalities)
	Development of a monetized service facing little or no competition (with intra-club negative externalities)	Rent-seeking competition (with intra-club negative externalities)

Let us note that in presence of zero-sum-games or intra-pool negative externalities, intra-pool competition and intra-pool cooperation (data-pooling) coexist. In these cases, we are therefore in presence of intra-pool coopetition.

1.3 Zero-sum games

In zero-sum games, the total value produced by the pooled data results in a finite and divisible (rival) resource, typically money. In terms of the equations used in this paper, this implies that $\sum_i^s a_i = 1$. When deciding whether to join a pool or accepting a new member, and agent has therefore to assess what share of the pool's benefit from data sharing it will get. This, in turn, implies the pool having a rule on how to divide the benefits among its members. Given that an infinite number of rules can be imagined, we will analyze the effect of zero-sum games on the data sharing decision using two repartition rules that are particularly interesting to describe real cases. The first one is an equal division of the pool's benefits or "egalitarian bargaining solution" (Kalai, 1977). This rule is more likely to be applied in pools with a more horizontal governance, which can be either the result of the nature of the agents that participate in it (e.g. not-for-profit organizations) or the fact that agents have similar levels of bargaining power. The second rule reflects the opposite situation: benefits are divided proportionally to each agent's data endowments. This rule implies a governance scheme in which agents use their bargaining power (the data they can bring to the pool).

1.3.1 Equally divided benefits

In order to analyze the effect a zero-sum game in which agents divide the pool's utility based on their share of data has on an agent's data sharing decision, we modify the utility function presented in Equation 1 to obtain Equation 2.

Equation 2: Utility function of data pooling in a zero-sum game with equally divided benefits

$$U_i = \frac{\varphi_i\left(x_i + x_{\bar{\imath}}(s)\right)}{s} - t_i(x_i) - l_i$$

Where $X = \left(x_i + x_{\bar{\imath}}(s)\right)$. In other words, the pool's data is equal to the data held by agent i (x_i) plus the data held by other members of the pool ($x_{\bar{\imath}}$).

Value appropriation factor a_i has been replaced by the pool's data divided by the number of members it has.

Function φ_i is the function that relates positively the size of the data pool to a level of value that can be created with it net of coordination costs. It therefore combines functions *f()* and $u_i(s)$ into one function in order to simplify calculations.

If we derive U_i in respect to *s* and we look for the condition under which it would be larger than zero we obtain Equation 3.

Equation 3: Marginal utility of admitting a new member to the data pool in a zero-sum game with equally divided benefits

$$\varphi_i' > \frac{\varphi_i\left(x_i + x_{\bar{\imath}}(s)\right)}{s}$$

Equation 3 shows that**, in a zero-sum game in which the benefits of data-pooling are divided equally, agents with admission rights will admit new members as long as the net marginal utility of adding a new member is superior to each member's appropriated utility, the latter being an equal share of the pool's utility for each member**. In other words, agents with admission rights will only admit new members that will bring enough quality-homogeneous data to the pool for their appropriated value (the pool's average value per member) to increase.

In accordance with H1, the utility of being a member of the pool of each member *j* belonging to the resulting optimal data pool membership s* has to be positive. In other words,

$$\mathrm{U_j}(s^*) > 0 \qquad \text{with j} \in s^*$$

1.3.2 Gains divided according to data endowments

In order to analyze the effect of a zero-sum game in which agents divide the pool's utility based on their share of data on an agent's data sharing decision, we modify the utility function presented in Equation 1 into Equation 4:

Equation 4: Utility function of data pooling in a zero-sum game with benefit divided proportionally to agents' data endowments

$$U_i = \varphi_i\left(x_i + x_{\bar{\imath}}(s)\right).\frac{x_i}{\left(x_i + x_{\bar{\imath}}(s)\right)} - t_i(x_i) - l_i$$

Like in Equation 2, value appropriation factor a_i has been replaced by the share of the pool's data held by agent i.

If we derive U_i in respect to *s* and we equate it to zero se obtain Equation 5:

Equation 5: Marginal utility of admitting a new member to the data pool in a zero-sum game with benefits divided proportionally to agents' data endowments

$$\frac{x_i}{\left(x_i + x_{\bar{\imath}}(s)\right)^2}.x_{\bar{\imath}}(s)'.\left[\varphi'_{i_{x_{\bar{\imath}}}}.\left(x_i + x_{\bar{\imath}}(s)\right) - \varphi_i\left(x_i + x_{\bar{\imath}}(s)\right)\right]$$

The first term, $\frac{x_i}{\left(x_i+x_{\bar{\imath}}(s)\right)^2}$ is agent i's share of the total square data of the pool.

The second term, $x_{\bar{\imath}}(s)'$, is other agents' (quality-homogeneous) data addition to the pool's data.

The third term, $\varphi'_{x_{\bar{\imath}}}$, is other agent's marginal contribution to the pool's value creation.

The fourth term, $\left(x_i + x_{\bar{\imath}}(s)\right)$, is the pool's total data.

The fifth term, $\varphi\left(x_i + x_{\bar{\imath}}(s)\right)$, is the utility created by the pool by sharing data.

This expression is equal to zero if either one of the following are true.

- **Agent i has no data or data without quality.** The term $\frac{x_i}{\left(x_i+x_{\bar{\imath}}(s)\right)^2}$ is null only if x_i=0, which means that agent i has no data. In that case, it would not have any incentive to admit a new member to the pool, as it would not be in any pool in the first place since its lack of data would give it no claim to any pool's payoff.

- **Other agents have no data**. If $x_{\bar{\imath}}(s)' = 0$ the entire expression is null. This would entail that other agents have not data or have data with no quality and would be therefore not admitted into a pool.

- **The marginal utility of admitting a new member is equal to the pool's mean value added per quality-homogeneous datum**.

Indeed, the term $\varphi'_{i_{x_{\bar{i}}}}.\left(x_i + x_{\bar{i}}(s)\right) - .\varphi\left(x_i + x_{\bar{i}}(s)\right) = 0$ if:

$$\varphi'_{i_{x_{\bar{i}}}} = \frac{\varphi\left(x_i + x_{\bar{i}}(s)\right)}{\left(x_i + x_{\bar{i}}(s)\right)}$$

This expression shows that, **in a zero-sum game in which the benefits of data-pooling are divided proportionally to their data endowments**, **agents with admission rights will admit a new member as long as the net marginal utility the pool would obtain from that admission is at least equal to the mean net utility of a quality-homogeneous datum in the pool**. In other words, agents with admission rights will not admit a member with an amount of data that will result in a lower mean value added per datum, as this would imply that each agent would appropriate a smaller share of value.

In accordance with H1, the utility of being a member of the pool of each member *j* belonging to the resulting optimal data pool membership s* has to be positive. In other words,

$$\mathrm{U_j}(s^*) > 0 \qquad \text{with j} \in s^*$$

It should be noted that this rule on the division of the value created by pooling data requires being able to assess the marginal contribution of an additional member's dataset, which might not be always possible. In general terms, the smaller the dataset, the lower its marginal contribution will be. If this marginal contribution is too small, it might be difficult to assess.

1.4 Competing pools

In a competing pools' scenario, the social value created by a pool through data-pooling (economies of scale, economies of scope and/or improvement of data quality) will not only depend on the size and the quality of its data pool, but also on that of other competing pools'. More precisely, in this scenario the more unequally all-pools' data is distributed in favor of a certain pool, the higher the utility of that pool will be. This happens when there are positive network effects that attract agents to the largest data pool.

This competition can have two distinct dynamics depending on whether it is a zero-sum game or not. If pools compete in a zero-sum game, the competitive dynamic describes rent-seeking competition: pools compete to obtain a certain amount of value and, the more data a pool has in respect to other pools, the more likely it will be for it to obtain that value or the larger the share of the value it will obtain. This is the case of a consortium pooling data in order accelerate the innovation process (genomics mapping, development of a new treatment based on artificial intelligent, etc.) and obtain a patent and/or be the first one in the market to commercialize a product or service. Pools competing in a non-zero-sum game describe situations such as competing data ecosystems. Indeed, while each ecosystem creates value without detriment to other pools', the larger the amount of data an ecosystem has in respect to other competing ecosystems, the more value it can create.

Therefore, as the analytical development will show, the effect of competing pools on data sharing decisions have to be analyzed by distinguishing zero-sum and non-zero-sum games.

1.4.1 Competing pools in non-zero-sum games

In order to arrive to an analytical expression of under which conditions will an agent decide to share data with a new pool member in a competing pools scenario, we first need to insert the notion of uneven distribution of data between pools into the utility function, which can be done in several ways. As shown below by Equation 6, we have chosen to do it using a measure of inter-pool data distribution inspired in the Herfindahl–Hirschman Index (HHI), which measures the concentration of market shares.

Equation 6: Utility function of data pooling in a competing pools non-zero-sum game scenario

$$U_i = a_i . \varphi_i \left(\frac{\left(x_i + x_{\bar{\imath}}(s)\right)^2}{\left(x_i + x_{\bar{\imath}}(s)\right)^2 + \sum \hat{X}_{\bar{\imath}}^2} \right) - t_i(x_i) - l_i$$

Where $\hat{X}_{\bar{\imath}}$ is the total quality-homogeneous data held by a pool other than agent i's.

As Equation 6 shows, the pool's utility depends on its squared data divided by all pool's squared data. With this measure of inter-pool pool data concentration (as opposed to, for example, a simple share of all pools' data), given pool A's data endowment, the way in which the rest of the pools' total endowments is divided between them will affect pool A's utility. For example, if pool A has 5 units of quality-homogeneous data and other pools sum up 8, pool A's utility will be higher if these 8 units are divided between 8 pools (each holding 1 unit) than if it is held by a single pool.

If we derive U_i in respect to s and we look for the condition under which it is positive we obtain Equation 7.

Equation 7: Marginal utility of admitting a new member to the data pool in a competing pools non-zero-sum game scenario

$$\sum \hat{X}_{\bar{\imath}}^2 > 0$$

This means that agents with admission rights will admit new members to the pool as long as other pools (hence other agents, which bring data to pools and can constitute one-agent-pools) have data. In other words, **in a competing pools non-zero-sum game scenario, there will be a tendency for agents to converge to a single pool with the highest possible number of members**, as admitting new members will only increase the total value obtained by the single pool and the share of the total pool's data agents obtain does not depend on the number of members but on their capacity to exploit the common value created by the pool.

In accordance with H1, all agents *j* will join the single optimal pool of *s** members as long as the utility of doing so is higher than the associated costs. In other words,

$$\mathrm{U_j}(s^*) > 0 \qquad \text{with j} \in s^*$$

Moreover, agents' *j* utilities associated with joining the pool have to be superior to that of associating any competing pool willing to admit them.

1.4.2 Competing pools in zero-sum games

As we have just seen, when pools compete in a zero-sum game there will be a tendency for them to converge to a single pool due to the fact that this will provide the members of the pool with the fullest possible share of utility or the highest probability of obtaining it without having to divide it between each other.

On the contrary, if pools compete in a non-zero-sum game scenario, the larger the pool is, the more the social value created through data sharing will have to be divided between members. Hence, the intra-pool utility distribution rule (equally divided benefits, division according to data endowments or any other rule) will prevail over the tendency towards converging to a single pool. As a result, in the equilibrium there will not necessarily be only one pool. The optimal pool size will depend on the factors that determine it in a zero-sum game which have been shown in Section 1.3. In other words, when the features 'competing pools' and 'zero-sum game' coexist, the latter gets the upper hand in determining the optimal number of members of each pool.

1.5 Agents' data endowments distribution

In the previous analyses, agents' data endowments have affected data sharing behaviors in two ways. In zero-sum games, an agent's data endowment will determine whether the pool accepts it as a member of not: it will do so only if the net marginal utility the pool would obtain from that admission is at least equal to the mean net utility of a member (if the utility is divided equally between members) or a datum in the pool (if the utility is divided according to data endowments). In competing pools, the more concentrated data is *at the pool level,* the more the pools that concentrate much data will benefit from pooling data and vice versa.

As seen above, Equation 3 and the last term of Equation 5 give us the conditions under which admission-rules-holder agent *i* will decide add a new member to the pool when the pool's utility is divided equally and when it is divided according to each member's data endowment, respectively. In both cases there is no inter-pool competition.

In both cases agent *i* has to assess the other agent's quality-homogeneous data endowment in respect to the pool's in order to decide who will be admitted as a member. Agent *i* will decide to admit a new member if it has more quality-homogeneous data than the average quality-homogeneous data endowment per member (if benefits are divided equally) or if it has enough quality-homogeneous data for the utility per datum of the pool not to fall if it joins the pool.

As a corollary, we can deduce that **in a zero-sum-game scenario without inter-pool competition, the more uneven quality-homogeneous data endowments are distributed between agents, the less data pooling there will be between large data holders and small data holders**. Indeed, the larger the gap between agents' data endowments, the stronger the weight of the (negative) effect of having to divide data will be in respect to the (positive) effect of creating value from pooling data when assessing whether to admit a member into the pool. On the contrary, if every agent had the same quality-homogeneous data endowment, the data-pooling decision will depend solely on the comparison between the benefits (economies of scale, scope, and/or quality) and the costs (lump costs, marginal costs and coordination costs) of pooling data, as seen in Section 1.2. The case of Myriad Genetics illustrates this point. This company was granted patents that gave it the possibility to be the sole testers for the BCRA1 and BCRA2 genes in the United States for more than a decade. As a result, Myriad was able to collect large databases of mutations and other clinical information that was costly and difficult to replicate by competitors. Consequently, it stopped contributing to public databases and started protecting the new and larger database it created based on these patents with trade secret (Sichelman & Simon, 2016).

Moreover, if the game is zero-sum and there is inter-pool competition, the distribution of data endowments between agents will have another effect. If two or more agents decide to form a pool, this will affect other pools' (including single-member pools) utility, as it will impact inter-pool data concentration. More precisely, the higher the agent-level data concentration is (i.e. the more uneven the distribution of data endowments is), the higher the pool-level data concentration will be. Indeed, agents that have considerably more quality-homogeneous data than the average will decide to pool their data only with agents with a similar quality-homogeneous data endowment or not to pool data at all in order to benefit from inter-pool data concentration. Whether they will decide to pool data at all will depend on how much of the pool's data value they will have to share with other members according to the value distribution rule. As a result, **if the game is zero-sum and there is inter-pool competition, agents with lower quality-homogeneous data endowments will tend to form pools together to countervail the negative effect of agents with high data endowments forming their own pools (either single-member pools or pools between agents with high data endowments) on their utility.**

Let us illustrate this point with an example. There are four agents (A_1, A_2, A_3 and A_4) in a zero-sum-game with pool competition in which the value produced by the pool is divided according to each agent's data endowment. The utility function of every agent is the following one:

$$U_i = \left(\frac{\left(x_i + x_{\bar{\imath}}(s)\right)^2}{\left(x_i + x_{\bar{\imath}}(s)\right)^2 + \sum \hat{X}_{\bar{\imath}}^2} \right) \cdot \frac{x_i}{\left(x_i + x_{\bar{\imath}}(s)\right)}$$

For the sake of simplicity, we will assume there are no costs associated to data sharing.

Let us now calculate the payoffs for each agent in each non-redundant possible pool configuration with three different quality-homogeneous data endowments ranging from a purely even distribution to a high concentration of quality-homogeneous data endowments. The following tables show the results. Agents within the same parenthesis belong to the same pool. For example, (A_1); (A_2, A_3, A_4) means agent A_1 does not pool data (it belongs to a single-member pool) and that agents A_2, A_3 and A_4 form a pool.

Table 2: Payoffs for different pool configurations if all agents have a data endowment equal to 1

Pool configuration code	Pool configuration	Agent			
		A1	A2	A3	A4
A	(A_1); (A_2); (A_3); (A_4)	0,25	0,25	0,25	0,25
B	(A_1); (A_2, A_3, A_4)	0,10	0,30	0,30	0,30
C	(A_1, A_2); (A_3, A_4)	0,25	0,25	0,25	0,25
D	(A_1, A_3); (A_2, A_4)	0,25	0,25	0,25	0,25
E	(A_1, A_2, A_3); (A_4)	0,30	0,30	0,30	0,10
F	(A_3); (A_1, A_2, A_4)	0,30	0,30	0,10	0,30
G	(A_1); (A_2, A_3); (A_4)	0,17	**0,33**	**0,33**	0,17
H	(A_1); (A_2); (A_3, A_4)	0,17	0,17	**0,33**	**0,33**
I	(A_1, A_2); (A_3); (A_4)	**0,33**	**0,33**	0,17	0,17
J	(A_1, A_2, A_3, A_4)	0,25	0,25	0,25	0,25

NB: The highest payoffs for each agent are highlighted

As Table 3 shows, although pool configurations G, H and I are preferred by some agents, they are not Nash equilibriums, as for each of these configurations there is at least one agent that has incentives to deviate to another pool configuration. On the contrary, pool configurations A (no pool), J (a single pool), C and D (two pools with two agents) are Nash equilibriums. These very different configurations share the particularity of leading to even inter-pool data distributions.

Table 3: Payoffs for different pool configurations if agents A_1 and A_2 have 10 units of data each and agents A3 and A4 have 5 units of data each

Pool configuration code	Pool configuration	Agent			
		A1	A2	A3	A4
A	(A_1); (A_2); (A_3); (A_4)	0,40	0,40	0,10	0,10
B	(A_1); (A_2, A_3, A_4)	0,20	0,40	0,20	**0,20**
C	(A_1, A_2); (A_3, A_4)	0,40	0,40	0,10	0,10
D	(A_1, A_3); (A_2, A_4)	0,33	0,33	0,17	0,17
E	(A_1, A_2, A_3); (A_4)	0,38	0,38	0,19	0,04
F	(A_3); (A_1, A_2, A_4)	0,38	0,38	0,04	0,19
G	(A_1); (A_2, A_3); (A_4)	0,29	0,43	**0,21**	0,07
H	(A_1); (A_2); (A_3, A_4)	0,33	0,33	0,17	0,17
I	(A_1, A_2); (A_3); (A_4)	**0,44**	**0,44**	0,06	0,06
J	(A_1, A_2, A_3, A_4)	0,33	0,33	0,17	0,17

NB: The highest payoffs for each agent are highlighted

As shown in Table 4, with a data endowment distribution in which agents A_1 and A_2 have twice more data than agents A_3 and A_4, agents A_1 and A_2 prefer a pool configuration in which they share their data but A_3 and A_4 do not (pool configuration I). However, if agents A_1 and A_2 share their data, agents A_3 and A_4 will have incentives to form a pool, which would result in the Nash equilibrium: pool configuration C. In this configuration, neither agents with large data endowments (A_1 and A_2) nor agents with small data endowments (A_3 and A_4) have incentives to enter or leave their pools.

Table 4: Payoffs for different pool configurations if agent A_1 has 15 units of data, A_2 has 10, A_3 has 5 and A_4 has 1

Pool configuration code	Pool configuration	Agent			
		A1	A2	A3	A4
A	(A_1); (A_2); (A_3); (A_4)	**0,64**	0,28	0,07	0,00
B	(A_1); (A_2, A_3, A_4)	0,47	0,33	0,17	0,03
C	(A_1, A_2); (A_3, A_4)	0,57	0,38	0,05	0,01
D	(A_1, A_3); (A_2, A_4)	0,58	0,21	**0,19**	0,02
E	(A_1, A_2, A_3); (A_4)	0,50	0,33	0,17	0,00
F	(A_3); (A_1, A_2, A_4)	0,56	0,37	0,04	**0,04**
G	(A_1); (A_2, A_3); (A_4)	0,50	0,33	0,17	0,00
H	(A_1); (A_2); (A_3, A_4)	0,62	0,28	0,08	0,02
I	(A_1, A_2); (A_3); (A_4)	0,58	**0,38**	0,04	0,00
J	(A_1, A_2, A_3, A_4)	0,48	0,32	0,16	0,03

NB: The highest payoffs for each agent are highlighted

As seen in Table 5, agents' preferences regarding the optimal pool configuration differ. However, there is a Nash equilibrium. Agent A_1, which has considerably more data than the rest, favors pool configurations in which it does not share data, the exception being configuration I, which is preferred over two configurations in which it does not share data: B and G. However, configuration I is not a Nash Equilibrium: in case agents A_1 and A_2 decide to form a pool, agents A_3 and A_4 would prefer forming a pool themselves and hence switch to configuration C. Nevertheless, the latter is not a Nash equilibrium either, because agent A_1 has an incentive not to share data with any other agents and hence switch to pool configuration B. In this pool configuration in which agent A_1 does not share data and agents A_2, A_3 and A_4 form a pool, none of them has incentives to change the pool configuration. Pool configuration B is therefore the Nash equilibrium in this case.

This example illustrates the conclusions developed above**. In a competing pools zero-sum-game scenario, the more uneven quality-homogeneous data endowments are distributed among agents, the more concentrated pools will be**. In that respect, competition between pools countervails the tendency agents have to pool data between agents with similar data endowments in zero-sum-

games, as inter-pool competition will foster data pooling between agents with heterogeneous data endowments in order to offset the negative effect on their utilities of some agents and/or pool concentrating a large share of the data.

1.6 Intra-pool negative externalities

We have so far analyzed cases in which the negative effects of data-pooling either do not depend on other agents (lump costs, marginal costs, coordination costs) or depend on other agents' data endowments (dividing the pool's utility in the case of zero-sum games, pool-level data concentration in the case of competing pools). We shall now introduce a case in which there are intra-pool negative externalities. This means that some agents' utilities diminish if an agent with certain attributes (for example, a direct competitor in a downstream market when the members of the pool are two insurance companies) becomes a member of the pool. Hence, intra-pool negative externalities are non-anonymous (for each agent, the level of the negative externality depends on other members' attributes) and heterogeneous (not all members suffer the same level of negative externalities). Moreover, given that these externalities will only take place if other members use the share data in a certain way, intra-pool negative externalities are not only non-anonymous and heterogeneous as explained before, but also a random variable.

Intra-pool negative externalities are more common when agents are competitors in the same downstream market(s) and the shared data could be used to gain a competitive advantage in that/those market(s). Given that data is an experience good (Koutroumpis & Leiponen, 2013), negative externalities can result from the inherent imperfect information agents have regarding potential uses of data. Agents might overestimate the potential negative externalities generated by other members of the pool using the pooled data, as they cannot assess the impact of it before it is done. Even in presence of perfect information, intra-pool negative externalities arise because data-pooling is in itself an incomplete contract (Dosis & Sand-Zantman, 2019). Indeed, when deciding to pool data, agents are making an ex-ante contract of finite length that cannot include provisions for all possible ex-post uses of the data by other members. Moreover, some uses of the data by other agents might be impossible to attribute to a particular agent (unobservable) or unenforceable. Finally, if certain uses of the pooled data by other members of the pool are irreversible agents might prefer not to share data at all in order to avoid excessive uncertainty regarding intra-pool negative externalities. This is all the more so if they cannot threaten to retaliate in order to deter other members of the pool from generating them.

The level of negative externalities also depends on the governance of the pool, as some forms of governance make certain negative-externalities-generating uses of the pooled data by some agents less likely. Members can set up rules that can filter the entry of new members that could create strong negative externalities for (some of) the members of the pool. In general terms, the more shared the governance of the pool is, the smaller the incertitude of negative externalities will be. For example, if all members have veto rights over admitting a new member, they will expect less intra-pool negative externalities than if one member of the pool that might have different incentives decides on admission. Members can also set up rules to constraint the uses they can do of the common data pool in order to limit negative externalities. The success of these rules in limiting negative externalities will depend on the capacity to observe deviations from the rules and the capacity members have to punish them (penalties

decided by the members, retaliation capacity of some members in the downstream market or other markets, etc.). However, restrictive rules aiming at lowering intra-pool negative externalities might render the pool small and the exploitation of the shared data pool limited, which would result in a low social welfare. Therefore, in presence of negative externalities, agents have to find a balance in the setting up of the governance rules that allows them to exploit and enlarge the data pool as much as possible without considerably hindering the entrance to the pool.

In order to formalize negative externalities, we can introduce them by adding a term to Equation 1, which results in Equation 8.

Equation 8: General form of the utility equation with intra-pool negative externalities

$$U_i = a_i f\big(X(s)\big) - t_i(x_i) - l_i - u_i(s) - \int_0^{\infty} g_i\big(c(s_{\bar{\imath}})\big)\, dc$$

We have added a term to Equation 1 that shows that the negative externality is equal to a random variable *c* which depends on the identity of other members of the pool ($s_{\bar{\imath}}$) and is distributed according to density function g_i.

Intra-pool negative externalities will affect data pooling decisions in that they will make certain pool configurations less likely and hence limit data pooling. Let us illustrate this with an example.

There are three agents in a non-competing pools non-zero-sum game in which the utility function is equal to the natural logarithm of the pool's data and every agent has a quality-homogeneous data endowment of 10. There are no costs associated to data sharing. Table 6 below shows the payoffs for each agent in every possible pool configuration.

Table 5: Payoffs for each possible pool configuration with a utility function equal to the natural logarithm and three agents with a data endowment of 10 each

Pool configuration code	Pool configuration	Agent		
		A1	A2	A3
A	(A_1); (A_2); (A_3)	2,3	2,3	2,3
B	(A_1, A_2); (A_3)	3,0	3,0	2,3
C	$(A_1; A_3)$; (A_2)	3,0	2,3	3,0
D	(A_2, A_3); (A_1)	2,3	3,0	3,0
E	(A_1, A_2, A_3)	**3,4**	**3,4**	**3,4**

NB: The highest payoffs for each agent are highlighted

In this case, as expected from the conclusions made above, the Nash equilibrium consists in the three agents forming a single pool (pool configuration E), as they all maximize their payoffs in that situation.

Let us now see what happens if there are negative externalities between agents A_2 and A_3 translated by the fact that each of them expects to lose 2 points of utility if they join a pool in which the other one is present. Table 7 below presents the payoffs for each agent in each possible pool configuration in presence of this negative externality.

Table 6: Payoffs for each possible pool configuration with negative externalities between A2 and A3

Pool configuration code	Pool configuration	Agent		
		A1	A2	A3
A	(A_1); (A_2); (A_3)	2,3	2,3	2,3
B	(A_1, A_2); (A_3)	3,0	**3,0**	2,3
C	$(A_1; A_3)$; (A_2)	3,0	2,3	**3,0**
D	(A_2, A_3); (A_1)	2,3	1,0	1,0
E	(A_1, A_2, A_3)	**3,4**	1,4	1,4

NB: The highest payoffs for each agent are highlighted

As shown in Table 7, in presence of these intra-pool negative externalities, converging to a single pool is not a Nash equilibrium, as both agents A_2 and A_3 would be better off in pool configurations in which they do not share the same pool. Configurations B and C, in which either agents A2 or A3 forms a pool with A1 and the other one does not share data, are Nash equilibria.

Table 8 summarizes the main findings of Section 1.

Table 7: Impact of different features on data-pooling decisions

Category	Feature	Impact on data-pooling decisions
Economies and costs of data sharing	Relationship between economies of data-pooling and lump, transaction and coordination costs	An agent will be admitted to a data pool only if the economies of data-pooling agents with admission rights can appropriate are higher than the sum of lump, transaction and coordination costs admitting it imply and if admitted agents obtain a net positive utility from joining the data pool.
	Relationship between the marginal utility of data sharing and coordination costs	An agent holding admission rights will admit a new member as long as the marginal utility brought about by the new member's data it can appropriate is higher than coordination costs of data pooling and if admitted agents obtain a net positive utility from joining the data pool. If the economies of data-pooling are higher than the sum of lump, transaction and coordination costs, the relationship between the marginal benefits and the marginal coordination will result in: • **One pool with the highest possible number of members** if the marginal cost of data-pooling decreases in respect to the marginal benefit of data-pooling as the number of members (s) increases • **One or more pools with a number of members lower than the maximum** if the marginal cost of data pooling is equal to the marginal benefit from it for a number of members of the pool located between

		2 and the number of possible members. • **An undetermined number of pools with an undetermined number of members** if the benefits of data-pooling appropriated by the agents with admission rights is higher than the sum of lump, transaction and coordination costs and the marginal cost of data-pooling they bear is equal to the marginal benefit from it for a number of members of the pool located between 2 and the number of possible members. In every case, admitted agents will have to obtain a net positive utility from joining the data pool.
Zero-sum games	Zero-sum games with pool's value equally divided	Agents with admission rights will admit new members as long as the net marginal utility of adding a new member is superior to the pool's average utility per member and if admitted agents obtain a net positive utility from joining the data pool.
	Zero-sum games with pool's divided proportionally to agents' data endowments	Agents with admission rights will admit a new member as long as the net marginal utility the pool would obtain from that admission is at least equal to the mean net utility of a datum in the pool and if admitted agents obtain a net positive utility from joining the data pool.
Competing pools	Competing pools in non-zero-sum games	Tendency for agents to converge to a single pool with the highest possible number of members
	Competing pools in zero-sum games	Agents with admission rights will admit a new member as long as the net marginal utility they would obtain from that admission is higher than the pool's average utility per

		member if benefits are divided equally or the mean net utility of a datum in the pool if the benefits are divided proportionally to each agents' data endowments. In every case, admitted agents have to obtain a net positive utility from joining the data pool.
Agents' data endowments distribution	Data endowments distribution in a zero-sum-game	The more unevenly data endowments are distributed among agents, the less data-pooling there will be between large data holders and small data holders
	Data endowments distribution in a zero-sum-game with competing pools	The more unevenly data endowments are distributed among agents, the more likely data-pooling between the agents that do not have the highest concentration of data will be
Intra-pool negative externalities		The stronger negative externalities between certain agents are, the less likely it will be for them to pool data within the same pool.

2 Sub-optimal data-pooling scenarios

In Section 1 we have developed a framework that allowed us to determine incentive structures and control rights over data can result in different data pools configurations. Depending on what is the policy objective, these factors might lead to sub-optimal data-pooling scenarios. Indeed, policymakers could for example decide to maximize social welfare. In this case, the optimal pool configuration would be that one for which the sum of all agents' utilities is the highest. Another option would be to maximize the total amount of data pooled. In this case, the optimal pool configuration would be that for which the sum of the data endowments of pools with at least two members is the highest. Policymakers could also try to maximize consumer welfare. This would entail that the optimal pool configuration would be that one for which the agent that represents consumers has the highest utility. Policymakers could also seek to enhance data-driven innovation and prefer therefore certain pool configurations in which certain agents (for example, research labs) are present and the amount of quality-homogeneous data pooled in the pools in which they are present is the highest.

This list of possible policy objectives is not exhaustive. Regardless of the content of all the possible policy objectives, we can imagine three types of optimizations that could translate policy objectives in terms of the theoretical framework developed in Section 1: optimizing by maximizing social welfare, optimizing by maximizing total data pooled and optimizing by favoring a particular data pool configuration**.** Regardless of the criteria used to determine which is the optimal data pool, a data pool configuration is defined by a) the members of the pool; b) which data they pool (which entails considerations on the quality of the

data); c) how much data they pool (the volume) and d) for which uses. **In cases in which agents' incentives lead them to a pool configuration different from the one that would result from a benevolent social planner that would optimize in any of the possible ways evoked above, policy intervention would be justified**.

In this section we will show how the determinants of data-pooling analyzed in Section 1 can lead to the four possible cases in which there might be a mismatch between agents' incentives and the optimum: a no data sharing equilibrium, suboptimal pool size with one equilibrium, suboptimal pool size with multiple equilibria and multiple equilibria with optimal pool size.

2.1.1 No data-pooling equilibrium

Economies of data-pooling too small in relation to the cost of data-pooling, the existence of inter-pool or intra-pool competition (negative externalities and zero-sum games), and the repartition of data endowments between agents might result in every agent finding optimal not to share data despite the economies it implies.

Let us illustrate this with an example. Suppose there are 4 agents (A1, A2, A3 and A4) in a zero-sum game with no inter-pool competition. Agents divide the pool's utility equally. Each agent has a data endowment of 1 and the following utility function, which supposes there are no costs related to data sharing for the sake of simplicity.

$$U_i = \frac{\left(x_i + x_{\bar{\imath}}(s)\right) + 5}{s}$$

Table 8: Payoffs for each possible pool configuration resulting in a scenario resulting in no data sharing

Pool configuration code	Pool configuration	Agent			
		A1	A2	A3	A4
A	(A_1); (A_2); (A_3); (A_4)	**6,0**	**6,0**	**6,0**	**6,0**
B	(A_1); (A_2, A_3, A_4)	**6,0**	2,7	2,7	2,7
C	(A_1, A_2); (A_3, A_4)	3,5	3,5	3,5	3,5
D	(A_1, A_3); (A_2, A_4)	3,5	3,5	3,5	3,5
E	(A_1, A_2, A_3); (A_4)	2,7	2,7	2,7	**6,0**
F	(A_3); (A_1, A_2, A_4)	2,7	2,7	**6,0**	2,7
G	(A_1); (A_2, A_3); (A_4)	**6,0**	3,5	3,5	**6,0**
H	(A_1); (A_2); (A_3, A_4)	**6,0**	**6,0**	3,5	3,5
I	(A_1, A_2); (A_3); (A_4)	3,5	3,5	**6,0**	**6,0**
J	(A_1, A_2, A_3, A_4)	2,3	2,3	2,3	2,3

NB: The highest payoffs for each agent are highlighted

As Table 9 shows, only in pool configuration A (no data sharing) every agent gets the highest payoff and hence this pool configuration is the only Nash Equilibrium. Let us note that if agents did not have the same data endowments and the division of the pool's payoff was done according to each agent's share of data the result would be the same.

The case of Finland's Act on the Secondary Use of Health and Social Data illustrates how when the economies of data pooling are not sufficient in relation the corresponding costs the result is a lack of data pooling. This law forces public and private holders of certain types of health data to share it with a public body that, in turn, acts as a trusted intermediary that can grant third parties 'data permits' to use data in the public interest (research, education, statistics, etc.). The fact that the government decided to mandate and organize by itself data-pooling results from the fact that at least some large health data holders did not benefit much from the economies of data-pooling while pooling data would imply costs or even negative intra-pool externalities too high for them to decide to pool their data. In terms of the developments made in Section 1, the appropriation factor a of at least some key agents was too low in

comparison to the costs and eventual intra-pool negative externalities of data-pooling for them to gain from pooling data.

2.1.2 Suboptimal pool size with one equilibrium

In order to illustrate how the nature of payoffs, negative externalities and costs related to data pooling developed above can result in a suboptimal pool size with one equilibrium, let us develop a simple example. Suppose there are three heterogeneous data-holding agents named A_1, A_2 and A_3 and that policymakers want to maximize social welfare or total data pooled, which is equal to the sum of each agent's utility plus the externalities generated by the use agents make of the pooled data. Given the utility function of each agent, the utility of each possible pool configuration for each of the 3 agents can be determined. Table 10 below summarizes them.

Table 9: Utilities for each member and social welfare given all the possible pool configurations in a scenario with suboptimal pool size and one equilibrium

Pool configuration code	Pool configuration	Agents' payoffs (U_1, U_2, U_3)	Intra-pool welfare	Social welfare
A	(A_3); (A_1, A_2)	(**2**, 3, 0)	5	6
B	(A_2); (A_1, A_3)	(-1, 0, -1)	0	2
C	(A_1); (A_2, A_3)	(0, 2, **1**)	3	2
D	(A_1, A_2, A_3)	(1, **10**, **1**)	12	16

NB: The highest payoffs for each agent are highlighted

As Table 10 shows, pool configuration D, in which the three agents pool their data within the same pool, is the one that implies the largest data pool and the largest social welfare. Agent A_2 has a strong preference for that pool and agent A_3 has a weak preference for it, as it is indifferent between pool configurations C and D. However, agent A_1 has a strong preference for pool configuration A, as its utility is the largest in that pool. If agent A_1 decides to form that pool, the other agents cannot form a pool in which they will be better off. Indeed, in pool configuration C agent A_2 is worst off than in pool configuration A. Therefore, there is only one Nash equilibrium that is not the one that maximizes social welfare: pool configuration A.

The above-mentioned case of Myriad Genetics is a good illustration of a suboptimal pool size with one equilibrium. The optimal data pool configuration from a social and consumer welfare perspective would be for Myriad Genetics to contribute to the public database, as long as it would accelerate and increase the insights that can be drawn from larger datasets on clinical information. In other words, this would be

the case if the economies of scale and scope in exploiting genetic data were not fully exploited without Myriad Genetic's data beig part of the data pool. However, Myriad Genetics had incentives to be the only agent to benefit from its (larger) dataset, as this would allow it to monetize it more than if it shared it with other agents. As a result, the resulting pools (the public databases) had a suboptimal size.

2.1.3 Suboptimal pool size with multiple equilibria

Let us now examine the following data-pooling game in which policymakers want to maximize social welfare. We suppose again that there are three heterogeneous data-holding agents named A_1, A_2 and A_3. Given the utility function, the utility associated to each possible pool configuration for each of the three agents can be determined. Once the externalities generated by each of the possible uses of the pooled data associated to each pool configuration are added to the intra-pool welfare, we obtain the social welfare. Table 11 below summarizes the payoffs of each agent, the intra-pool welfare and the social welfare associated to each pool configuration.

Table 10: Utilities for each member and social welfare given all the possible pool configurations with suboptimal pool size and multiple equilibria

Pool configuration code	Pool configuration	Agents' payoffs (U_1, U_2, U_3)	Intra-pool welfare	Social welfare
A	(A_3); (A_1, A_2)	(0, 3, 0)	3	6
B	(A_2); (A_1, A_3)	(**2**, 0, **1**)	3	1
C	(A_1); (A_2, A_3)	(0, 2, **1**)	3	3
D	(A_1, A_2, A_3)	(**2**, **10**, -1)	11	15

NB: The highest payoffs for each agent are highlighted

In this case, pool configuration D is again the one with the largest number of members and therefore the largest data pool. It is also the one that creates the highest social welfare. However, it is the least preferred pool configuration by agent A_3. Moreover, as long as A_2 and A_1 cannot compensate A_3 with a share of their payoffs, agent A_3 will not choose to join it, as doing so would result in a negative utility for it. Therefore, pool configuration D is not an equilibrium. Pool configuration A is not an equilibrium either. Indeed, agent A_1 would be better off if it decided to form a pool with agent A_3 instead (pool configuration B) and agent A_3 would be better off if it decided to form a pool with agent A_1 (pool configuration B) or with agent A_2 (pool configuration C). Finally, pool configurations B and C are both a Nash equilibrium. If pool configuration B is constituted, neither agent A_1 nor agent A_3 will have incentives to form an alternative pool. Similarly, if pool configuration C is constituted, neither agent A_2 nor agent A_3 will have an incentive to form an alternative pool.

This simple example illustrates how multiple sub-optimal data pools in terms of social welfare can be constituted given agents' attributes. The possible pools can differ in terms of social welfare, in terms of the distribution of value among agents and in terms of the total amount of data pooled. In this example, the two possible pools provide the same level of intra-pool welfare but different levels of social welfare. However, none of them maximizes social welfare.

It should be noted that, in this example, the socially optimal pool configuration D could be reached if agents can distribute the payoffs among each other. For example, if A_2 shares 3 units of its payoffs with A_3 if it agrees to join pool configuration D, then the latter would be the only Nash Equilibrium. Every agent would be better off and hence no public intervention would be needed to reach the socially optimal pool.

2.1.4 Multiple equilibria with optimal pool size

If the social optimum is not only defined by the sum of agents' utilities but, in turn, by a particular pool configuration, there could also be a mismatch between agents' incentives and the social optimum. This could happen of course if there is one equilibrium in which agents decide to pool data that it is not the social optimum. It could also happen if there are multiple data-pooling equilibria (i.e. multiple pool configurations) that result in the same level of social welfare without any of them being the configuration preferred by the policymaker. Let us illustrate this case with an example.

There are four agents (A_1, A_2, A_3 and A_4) in a zero-sum game with no inter-pool competition. The pool's payoff is divided equally between its members. Every agent has an endowment of quality-homogeneous data of 1 and the following utility function:

$$U_i = \frac{\left(x_i + x_{\bar{\imath}}(s)\right)^4 - 1}{s}$$

The social optimum is the pool configuration in which agents A_2, A_3 and A_4 share data between each other (pool configuration B in Table 12 below), as only their combined datasets can lead to the development of a much-needed vaccine. If the four agents share data, this would not be a social optimum, as this could give the market leader, A_1, a dominant position in other vaccines' markets and hence lead to price increases. Table 12 below shows the payoffs associated to the different possible pool configurations.

Table 11: Payoffs for the non-redundant possible pool configuration in a scenario resulting in multiple equilibria

Pool configuration code	Pool configuration	Agents' payoffs				Intra-pool welfare	Social welfare
		A1	A2	A3	A4		
A	(A_1); (A_2); (A_3); (A_4)	0,00	0,00	0,00	0,00	0	0
B	(A_1); (A_2, A_3, A_4)	0,00	**0,11**	**0,11**	**0,11**	0,33	1,5
C	(A_1, A_2); (A_3, A_4)	0,09	0,09	0,09	0,09	0,36	1
D	(A_1, A_3); (A_2, A_4)	0,09	0,09	0,09	0,09	0,36	1
E	(A_1, A_2, A_3); (A_4)	**0,11**	**0,11**	**0,11**	0,00	0,33	0,75
F	(A_3); (A_1, A_2, A_4)	**0,11**	**0,11**	0,00	**0,11**	0,33	0,75
G	(A_1); (A_2, A_3); (A_4)	0,00	0,09	0,09	0,00	0,18	0,5
H	(A_1); (A_2); (A_3, A_4)	0,00	0,00	0,09	0,09	0,18	0,5
I	(A_1, A_2); (A_3); (A_4)	0,09	0,09	0,00	0,00	0,18	0,5
J	(A_1, A_2, A_3, A_4)	0,10	0,10	0,10	0,10	0,4	1,25

NB: The highest payoffs for each agent are highlighted

As Table 12 shows, there are three Nash equilibria: pool configurations B, E and F[7]. In each of these pool configurations, three agents maximize their payoffs by forming a data pool and excluding the other agent, which has a null payoff. Although each agent is indifferent between the three Nash equilibriums in which it maximizes its utility, the social optimum, pool configuration B, might not be chosen.

In this case, contrary to the previous one, even if agents can distribute the payoffs between each other, the optimal pool configuration B cannot be reached. This justifies public intervention.

3 Fostering data pooling

We have seen how different factors affecting agents' data-pooling incentives (Section 1) can result in sub-optimal data pools (Section 2). In this section, we will present policies that can be implemented to foster data pooling with a focus on health data, on which the examples given will focus. However, most of these policies (the exception being those relating to posthumous medical data donation) can be applied to other sectors.

Some considerations regarding the scope and the applicability of the policy recommendations should be made before starting developing on them. First, it should be noted that we will focus on how to reach optimal data pools from a societal perspective as described in Section 2. This does not imply that data pools or larger data pools are always preferable, as there can be welfare losses from data-pooling such as anti-competitive behaviors (Lundqvist, 2018), negative data externalities or information asymmetries (Martens et al., 2020). Second, in coherence with the approach taken in Section 1, we will focus on appropriation problems. Consequently, we will describe situations in which already-existing data is not being pooled in the optimal pool (which includes situations in which data is not pooled at all) and exclude policy recommendations aiming at fostering the production of more and/or higher-quality health data. However, we will take into account some aspects of provision problems when they can affect the applicability of the policy recommendations evoked. In particular, we will point out under which conditions some of the policies proposed might lower incentives to data generation. Third, also in coherence with Section 1, we will focus on data pooling, a type of data sharing defined by multilateral data sharing between at least two agents. Hence, these policies do not exhaust the range of policies that can be implemented to encourage data sharing, as the Joint Research Centre reports on business-to-business data sharing (Martens et al., 2020) and business-to-government data sharing (Martens & Duch-Brown, 2020), which focus on unilateral data sharing, illustrate.

The policy recommendations we will present should be understood as a toolbox for policymakers wishing to foster data pooling. By evaluating each suboptimal data-pooling scenario, policymakers can recur to one or more of these policies depending on the types of data, agents and incentives involved. Most of these policies are in fact complementary. For example, policies aiming at increasing data interoperability (Section 3.1.2) can only increase the impact of a wider scope of data portability (Section 3.2.3) in terms of data pooling. For the sake of clarity, we will divide our policy recommendations into two categories: those that imply acting on agents' incentives to pool data (Section 3.1) and those that entail acting on agents' control rights over data (Section 3.2).

[7] We have only included these 3 equilibria to illustrate without multiplying the number of redundant possible pool configurations. There is therefore a fourth pool configuration that is an equilibrium: (A2); (A1, A3, A4).

3.1 Acting on incentives to pool data

3.1.1 Providing economic incentives

As we have seen in the previous sections, if transaction or lump-sum costs are high enough in comparison to benefits, there might be no data-pooling (cf. Figure 1) or the resulting data pool might be smaller than what policymakers would desire (cf. Table 10 and Figure 3). In those cases, policymakers can create economic incentives for agents to decide to form the optimal data pool. In order to do so, policymakers can act on the costs and benefits of data pooling in several ways.

They can recur to **tax breaks or grants that would apply to certain agents pooling data** under conditions that can be more or less strict depending on the policy objective. For example, certain pharmaceutical or e-health companies reluctant to share data could benefit from a tax break if they contribute with a certain amount of data to a European Health Data Space. In a more complex version of this policy, the extent of the tax break or the grant could increase if the volume, the quality and/or the scope of the sharing (in terms of which agents can access the pool data and for what purposes) increases. The obtention of the grant or tax break could even be conditional on public access to the dataset (Price & Nicholson, 2014). Inversely, taxation can be used to penalize the absence of data pooling in a similar way to the taxation of vacancy in housing markets.

Another way of reducing data-pooling costs consist **in investing in technological infrastructures for health data pooling** such as IHAN[8]. This is an undergoing policy that the European Union is already pursuing by financing the development of data spaces and through research projects that created data sharing infrastructures such as H2020 DECODE[9] or My Health My Data (MHMD)[10]. These infrastructures reduce lump-sum costs because agents can use an existing data-pooling infrastructure instead of investing in developing one.

Policymakers can also reduce lump sum and transaction costs by **financially supporting health data cooperatives**[11]. Health data cooperatives are organizations that act as third parties between data holders (typically individuals) and data users such as researchers. These organizations pool data holders' data and asks them what kind of health data they want to pool (blood analysis results, genetic data, etc.), for which purposes do they want their data to be used (broad medical research, development of a specific treatment, etc.) and with which type of agents (non-for-profit public researchers, for-profit researchers, public bodies, etc.) do they want to share it. Examples of such cooperatives include Midata, Salus Coop, Transiscope, Moipatient, Patients Like Me and CureTogether. The resulting data pools are then shared with third parties in such a way that respects individuals' stated preferences. Health data cooperatives help to bring about the economies of scope and scale of data pooling by allowing

[8] See https://ihan.fi/

[9] https://decodeproject.eu/

[10] http://www.myhealthmydata.eu/

11 We will use the term "data cooperative" hereafter because it is the most widespread one. However, the reader should bear in mind that they do not necessarily have the legal status of a cooperative.

individuals to easily share data that data holders (hospitals, e-health firms, etc.) might not have incentives to share with third parties such as non-for-profit research labs. They do so by reducing the transaction and lump-sum costs of data-pooling for individuals: once individuals give their consent to pool the data and the conditions of its use, the health data cooperative takes care of the technical and legal procedures required to reclaim the data from data holders and structuring it in a common data base that will be shared with third parties. Moreover, individuals play a role in the governance of the cooperative, which reduces the expected intra-pool negative externalities evoked in Section 1.6 by increasing the transparency of the use of the pooled data. For example, individuals can have a vote in deciding on the ethical chart of such cooperatives.

Members of health data cooperatives benefit from their membership in two ways. The main one is the non-pecuniary reward of contributing to the public good members have. Many times the motivation individuals have to do such a contribution stems from the fact that they or people close to them suffer from a particular condition to which treatment their data could contribute. Members of health data cooperatives also benefit from data-driven health services provided by the cooperative based on the analysis of the pool's data. For example, MiData has an app called MirtrendS that allows patients with multiple sclerosis (a disease that has different effects in different patients) to record their symptoms and obtain personalized therapies, something that "in today's clinical practice only possible to a limited extent or not at all"[12]. PatientsLikeMe illustrates how these two types of benefits patients obtain from participating in a health data cooperative can be combined. PatientsLikeMe is a website in which people with life-changing illnesses share their health data (symptoms, treatments, etc.), obtain individual-level graphical health profiles, customized reports and are matched with other patients with similar clinical and demographic characteristics. This helps them improving their outcomes by learning from other patients' experience in dealing with the disease and obtaining aggregated data reports based on their data, as corroborated by a study (Wicks et al., 2010). The aggregated data collected by PatientsLikeMe, in turn, can be used to advance medical research. For example, it has been used to study the outcomes of off-label prescribing, which remains understudied because of a lack of pertinent data (Frost et al., 2011).

However, given that individuals are sensitive about allowing health data cooperatives to do a commercial use of their data, the latter tend to have fragile business models until they reach a certain critical mass of data that would allow for the restricted commercial uses of the pooled data to finance operating costs (Hafen, 2019). Hence, financial support of the initial development of health data cooperative from public bodies can result in an increase in health data pooling. This financial would not necessarily require an exclusive budget but rather an explicit targeting of (health) data cooperatives within existing public financial support to digital initiatives such as centers of excellence (Jacquart et al., 2018), subsidies, or research grants[13]. Nonetheless, the long-run economic sustainability of health data cooperatives has to be taken into account when financing their development, especially when selling data is not part of their business model. For example, PatientsLikeMe selectively sells aggregated data from its 250.000 users to industrial partners, which has helped it expanding since its foundation in 2005. In the case of health data cooperatives that do not want to sell data as part of their revenue model, their long-run sustainability will be more fragile, although not impossible to reach. Research partnerships or membership fees could (partially) substitute revenues from selling data if a critical scale is reached.

[12] See https://www.midata.coop/en/offer/mitrends/

[13] We will develop on data cooperatives and the other channels through which they can foster data-pooling in Section 3.2.3

Finally, policymakers can act on the benefits of data pooling to create incentives for agents to share data when they are insufficient. This can take the form of **subsidizing data pooling**. Certain agents could benefit from a subvention if they contribute with a certain amount of data to a European Health Data Space. The extent of the subvention could increase if the volume, the quality and/or the scope of the sharing (in terms of which agents can access the pool data and for what purposes) increases. Another way of acting on benefits consists in financing research projects that require health data pooling. Public actors can set data-pooling conditions that might exceed the timespan of the project in order for agents to benefit from the financing of a research project. Finally, in order to make data-pooling lasting, subsidies can be granted to agents that decide to apply open licenses to their databases so that they can be re-used and pooled with other data in the future. This last option is particularly interesting when databases do not contain personal data and include data with a long lifespan such as anonymized genetic data. In these cases, data re-use would be maximized over time with an open license.

In order to be efficient, these policies have to target the agents that refrain from pooling certain type of health data because of a lack of economic incentives, which requires taking two major policy design considerations. First, targeting might pose a legal challenge, since instruments such as taxation and subventions cannot be granted on case-by-case bases. Therefore, the design of these taxes and subventions has to be carefully tailored to target scenarios (and not particular agents) in which the beneficiaries of the tax break or the subvention were unwilling to pool data because of lack of incentives. Second, economic incentives in return for data pooling should not be given to individuals. As the literature on blood donation (e.g. Sandel, 2013) and, more broadly, on repugnant markets (Ambuehl et al., 2015; Roth, 2007) has shown, when economic incentives are given to foster transactions that are not economically motivated, agents tend to do less transactions instead of more because the economic incentive undermines the non-economic motivations. Finally, in order to be efficient, the positive welfare effect of all of these measures has to be larger than their fiscal cost.

3.1.2 Facilitating hard and soft data interoperability

One of the major barriers to data sharing in general is the lack of interoperability in the broad sense of the term, as we will show in this section. In the terms of the theoretical framework developed in the previous sections, a lack of interoperability translates into higher lump-sum costs (e.g. developing a standard to pool data that each agent stores in different formats) and higher coordination costs, as agents have to agree on technical and non-technical standards in order to pool data in the first place. Lowering these costs would therefore encourage data pooling, notably by third parties that exploit data scattered across multiple data holders.

Interoperability in the broad sense of the term implies both 'hard' and 'soft' interoperability and each of the dimensions of these types of interoperabilities requires policies of different natures to increase interoperability**. In the case of health data, the most relevant dimensions of hard or 'technical' interoperability are standard protocols for data sharing, standard formats to store health data, standard terminologies and cross-institutional patient identification numbers** (Albrecht et al., 2016; Henriksen et al., 2018; Strotbaum et al., 2019). When these lack, pooled data is hard to compare and evaluate and therefore of little use (Gay & Leijdekkers, 2015). This the case with many

types of health data today. For example, "Henriksen et al. identified 423 distinct devices of 132 different brands which are designed for recording physical activity, either in general or for specific needs [13]. Similar findings were reported by Wiesner et al. [42] who surveyed participants at a large running event." (Strotbaum et al., 2019, p. 224). In order to tackle this issue, **policymakers can act as coordinators by co-developing European technical and metadata standards with health data holders**. In this vein, Australia's Government Australian Institute of Health and Welfare has developed METeOR, a national repository that contains national health metadata. Such an endeavor would imply replicating in the health sector the undergoing work of the European Commission with the INSPIRE Directive with geographical data. Similarly, the European Union could work on developing European technical and metadata standards in order to foster their adoption across countries. Once developed, these standards can be gradually imposed through legislation to ensure their widespread adoption. The legal imposition of such standard could take the form of a directive that mirrors in the health sector what the European PSS2 directive has done in terms of harmonization of payment products, infrastructures and technical standards in the payment services sector.

Soft interoperability covers all the non-technical aspects of data interoperability (Nedovic-Budic & Pinto, 2001). In the case of health data, this refers mainly to standard agreements for data sharing by organizations and individuals. When an organization decides to share data with a third party, it usually writes a contract specifying the conditions of data sharing (for free or not, for which intended uses, intellectual property of the results of the use of the data, etc.). The writing of these contracts entails negotiations that can take time and hinder data sharing (Bellivier et al., 2015), especially when it comes to sensitive data such as medical data. In terms of the development, we did in Section 1, this implies a high coordination cost factor that can result in suboptimal data-pooling scenarios. While standard contracts (material transfer agreements) are common in the case of biobanks that share biosamples and information (Bellivier et al., 2015), they are not widespread in the case of health data sharing. In order to lower these costs, policymakers can work on **developing with health data holders standards agreements for data-pooling that cover the most typical scenarios in which data-pooling is still not widespread** in the same way as agencies such as the CIRAD have done it in the case of agricultural research data[14]. These scenarios can be defined in terms of the types of health data and agents involved and the intended uses. Similarly, when an individual decides to share his or her medical data, many possibilities arise: whether it can be shared at all (which might depend on the sensitivity of the health data involved); who (if anyone) else it concerns (typically family members, as in the case of genetic data); who it can be shared with; what form it can be shared in (fully identifiable, de-identified, pseudonymised and fully anonymized); who (if anyone) needs to give consent for any other type of sharing in the future (Shaw, 2019), and for which purposes can data be used (research, commercial uses, etc.)[15]. All these possibilities can be difficult to assess and understand for individuals, which represents a barrier to data-pooling by individuals. In order to lift this barrier, authors such as Schapranow, Brauer and Plattner (2017) have proposed the development of "**data donation passes**". Mirroring existing organ donation passes, they would allow individuals to easily choose the conditions under which they will let third parties use their health data. These passes can even have a real-time

[14] See https://www.cirad.fr/

[15] We have added the 'purpose' variable to Shaw's data-donation metadata list.

dimension managed by individuals through a smartphone application. The use of standard iconography such as the icons used to describe Creative Commons licenses and easy-to-understand terms and conditions can be included to render health data pooling by individuals easy and intelligible.

3.1.3 Leveraging on public actors' data

In Section I we have analyzed data-pooling decisions by agents that seek the maximization of the benefits they can obtain from it. However, public actors do not fall into this category, as they can act in the interest of reaching the optimal data pool rather than in order to maximize their own utility. When the Nash equilibrium differs from the optimal data pool (cf. Section 2), public actors can act in several manners in order to provide incentives for other actors to choose to constitute the optimal pool.

One of the ways in which public actors can do so is to act on repartition rules in the case of zero-sum games. We have analyzed cases in which benefits are split equally between the members of the data pool (Section 1.3.1) and cases in which gains are divided according do agents' data endowments (Section 1.3.2). However, regardless of the repartition rule negotiated between the members of the pool, **public actors can decide to lower their share of the value created by the data pool in favor of that of agents that should be included in the optimal pool but do not have enough incentives to join it**. **Moreover, public actors can condition their participation to the implementation of repartition rules that favor agents that should be in the optimal data pool but do not have enough incentives to do so under the repartition rules that would apply if the public actor does not intercede**.

Another leverage that public actors can use is negotiating admission rules**. They can condition their participation in a data pool to the acceptance of certain members that other pool members do not have incentives to accept when doing so would result in an optimal data pool being constituted**. This strategy is particularly interesting when the existence of intra-pool negative externalities prevents one or more actors to be part of the optimal data pool and all the members of the pool would benefit from it even in presence of the agents that generate the negative externalities. This strategy can only work if three conditions are met: i) by conditioning their participation in a data pool to the acceptance of certain members, public actors eliminate the first best (a suboptimal data pool that includes the public actor but not all the members of the optimal data pool) other members have and the optimal data pool it seeks is their second best, which implies public actor's dataset to be of high interest to other agents; ii) the public actor is legally allowed to pool the data and iii) the public actor is not bound to share the data with any other actors other than those that belong the optimal data pool. For example, this strategy would not work if the public actor holds data under open licenses.

Both strategies are all the more effective when public actors hold data that is particularly valuable (either because of its volume, quality or uniqueness) for the potential members of the optimal data pool as, as shown in Section 1, this generates incentives for other agents to form a pool with it and, hence, it gives it a larger bargaining power to act on admission rules or repartition rules. For this reason, when possible, the creation of coalitions of data-holding public actors that can bargain as one entity is

an interesting policy option for public actors to be able to effectively leverage on their data to reach optimal data pools.

3.2 Acting on control rights over data

We have so far presented different policies that can be implemented to reach optimal data pools by acting on agents' incentives. In this section, we will present other types of policies that are based on altering agents' control rights over data. This can be done within the current legal framework (either by applying certain laws or through soft law) or by creating new legislation.

3.2.1 Identifying, diffusing and implementing trust-building mechanisms

In Section 1.6 we introduced the notion of intra-pool negative externalities (represented by random variable *c)* to capture how agents' expectations about how the data they will pool might be used by the other members of the pool or third parties that would share the data with might create disincentives to pool data. Outside of the economics literature, this problem is typically referred to as a problem of lack of trust between the parties. It is common for public and private actors that see an interest in pooling their data to refrain from doing it or to pool less data than what the policymaker would prefer because of a lack of trust. This issue is particularly pressing in the health sector, in which the expected negative externalities are considerably strong as medical data is often sensitive data. Hence, when acting on agents' incentives, trust-building mechanisms can be impactful in providing agents with the pecuniary and non-pecuniary incentives required for them to choose to form the optimal data pool.

Trust-building mechanisms, in turn, are a key component of the governance of pooled data, a term we can define as 'the multitude of actors and processes that lead to collective binding decisions" (Van Asselt & Renn, 2011, p. 431), in this case around a specific resource: a pooled database and, in some cases, services built on it. In this sense, as mentioned in Section 1.1., the concept of governance of a pooled dataset can be assimilated to that of the governance of a shared resource in the commons literature (Coriat, 2015; Gardner et al., 1990; Hess, 2008; Ostrom, 1990). While there is no one-size-fits all data governance structure that can guarantee trust between agents pooling data, some principles and good practices have been identified, notably in the health sector. In this section, we will present them**. Trust-building mechanisms can be used by policymakers to foster data pooling in two ways: by researching, gathering and advocating the adoption of good trust-building practices in data pools by other actors and by applying them in situations in which the public actors are a stakeholder in a data pool**. A good example of the former is the Understanding Patients Data initiative, a British organization working with charities, patient groups and the NHS on research, advocacy and communication to build and diffuse the responsible share of patient data[16]. The application of these good practices by the public actor itself in turn, is particularly interesting in the health sector, in which it is common for public actors to be data holders.

[16] See https://understandingpatientdata.org.uk

Scholars from various disciplines have defined trust in multiple ways (Seppänen et al., 2007). With the advent of the information revolution, a subfield of e-trust, understood as the study of trust in digital contexts and/or involving artificial agents, has emerged. For the purpose of this report, we will use a broad definition of trust: "a relationship in which an agent (the trustor) decides to depend on another agent's (the trustee) foreseeable behaviour in order to fulfil his expectations" (Taddeo & Floridi, 2011, p. 1). In the case of data pools, agents have two types of expectations regarding other members of the pool behavior: those involving the availability and the quality of the data provided and those on how the data will be used by each member of the pool (types of analyses made with the data, commercial uses, sharing the data with third-parties, etc.). Hence, when constituting a data pool, agents have to make commitments on their contribution to the constitutions and maintenance of the shared resource and on how they will use it, the former being easier to observe than the latter. In order for those commitments to be held, **effective trust-building-mechanisms have to be built on two intertwined principles: transparency and accountability** (Krutzinna et al., 2019a; Sorbie, 2019; Vayena et al., 2018).

Transparency refers to "the possibility of accessing information, intentions or behaviours that have been intentionally revealed through a process of disclosure" (Turilli & Floridi, 2009, p. 105). When designing a data pool, it is therefore crucial for agents to agree on what information should be disclosed to whom and how (timeliness of the disclosure, format of the disclosure, completeness of the information, etc.) regarding both the data and its (intended) use for all the stakeholders to trust each other. The backlash against the NSH care.data scheme in the United Kingdom illustrates how shortcomings in transparency can lead to the failure of a project based on personal health data sharing (Sterckx et al., 2016). It is also important for information not only being available but also intelligible (Laurie et al., 2015). Given the complexity of the analyses that can be made with data, the number of parties involved with different types and degrees of responsibility and the possible future implications of the use of health data, information has to be made intelligible for each stakeholder to be able to use it.

Accountability, in turn, implies decision-makers being able to provide subjects with justifications and explanations for their decisions in order for decision-subjects to be able to sanction those decisions if they do not judge them adequate (Binns, 2018). This implies, on the one hand, that any relevant decision taken by a decision-maker regarding the use of the pooled data or a connected project (development of a treatment based on the data, research outputs, a new service, etc.) has to be justified to the other parties and their decision-makers identified, which in turn requires transparency mechanisms being put into place. On the other hand, accountability requires a formal and/or informal sanctions system to exist. As shown by the literature on commons (Ostrom, 1990), in order for a health data pool to survive as sustained cooperation between agents over time, sanctions should be i) graduated depending on the seriousness and the context of the deviation from the rule ii) plausible and credible and iii) have low monitoring costs.

In addition to the two key trust-building principles of transparency and accountability, some specific mechanisms especially suited to build trust between the stakeholders of a health data pool have been identified by researchers. The first one refers to **mechanisms leading to agents feeling they are real interlocutors able to shape the system rather than passive stakeholders** (Durante, 2015; Floridi, 2019). Indeed, if agents feel they can shape the 'rules of the game' they will be more likely trust other agents with their data, as they know that they will have a say in what can happen if the uses or the provision of the data do not meet their expectations. This is of particular importance when individuals are

involved as data holders, as they tend to feel less powerful than organizations in shaping the decision-making process. In order to render stakeholders real interlocutors, workshops and conversations aimed at understanding their attitudes, motivations and expectations on the health data pool with the stakeholders prior to its establishment are of outmost importance. For example, this is what the Academy of Medical Sciences did with patients, the public and healthcare professionals to device how could patient data be used in future technologies (Castell et al., 2018).

A second type of mechanism that can be put in place to foster trust in health data pool refers to **technical means by which the exposure of data can be minimized. Examples include safe havens, accreditation and technologies allowing to exploit several datasets as one without requiring data holders to migrate their data to a single joint server** (Laurie et al., 2015). An example of the latter is the Dutch initiative Personal Health Train[17], a technology that allows healthcare researchers and companies to work with various data sources located in different servers without migrating the data and respecting each data holder's authorizations of use of their data. By minimizing the possibilities of the data being used for other purposes and other agents than those agreed upon when setting up the data pool, these techniques reinforce trust (which, in terms of the development of Section 1, reduces the expected value of random variable *c*), which in turn encourages data-pooling.

A third mechanism to foster trust in data-pooling by increasing transparency and accountability is **dynamic consent, whereby data holders' preferences regarding how can their data be used and by whom can be exercised as the uses of their data evolve**. This is particularly interesting in the case of research projects requiring patients' consent to use their data. As research progresses and new types of uses of patient data emerge, dynamic consent allows patients to reveal new preferences or reconsidered their previous preferences in light of new information and proposals. This practice reinforces trust by different channels (Kaye et al., 2015), as it has been studied by researchers working on biobanks (Whitley et al., 2012). First, it enhances transparency and accountability in the use of the data because patient data can be tracked across research studies (and, more broadly, across other non-research-related uses of the data) to remove bias and erroneous identification. Second, it requires setting up an operational control system throughout the lifecycle of the data, which allows auditing the uses of the data and generate an early warning system of potential security breaches. Third, it allows contacting data holders that decided to pool their data for their opinions on controversial issues regarding the use of the data that could have not been imagined when the decision to share the data was made. Indeed, if a new use of the data that was not envisioned when the consent was given emerges during research (the logic can be extended to non-research-related uses of the data) emerges, because it is difficult and lengthy to re-contact every patient, ethic comities and national regulatory bodies tend to represent their interests, which might not be shared by every patient. Dynamic consent software, in turn, makes it easy to contact each participant, provide intelligible information on the intended new uses of the data and collect their preferences regarding them. In that manner, individuals can trust that their consent will not be broad enough for the members of the data pool to use it in ways they had not envisioned when giving it, which raises the likeliness of individuals deciding to pool their data. Moreover, in the specific case of research requiring recruiting patients, dynamic consent can foster optimal health data pools by raising the amount of pooled data through the enhancement of the recruitment process (Kaye et al., 2015).

[17] See https://pht.health-ri.nl/

Indeed, the costs of recruitment into research are high and participation rates relatively low although patients are usually willing to donate their data for research purposes. Dynamic consent software allows to automatically select participants willing to be involved whenever a new research request takes place. This facilitates and lowers the costs of the recruitment process, which should result in larger data pooling by individuals.

3.2.2 Mandating data-pooling on the grounds of public interest

In certain cases, the optimal data pool is the one that would allow acting in the public interest (developing a treatment for a rare disease, reducing dramatically the cost of a vaccine, etc**.) but agents do not have an incentive to pool their data**. For example, following on the developments made in Section 1, if the economies brought about by data-pooling face decreasing returns to scale in terms of monetization, agents might choose not to share as much data as they could even if doing so is in the public interest. **In this case, a public actor can force agents to share certain data with each other or with a third party when doing so is in the public interest**. Doing so does not imply a transfer of all the control rights the data holders have on the datasets, but rather a set of authorizations for one agent to use other agents' data for a certain purpose. Such an authorization-based approached to data sharing on the grounds of public interest is an established governance mechanism in health research (notably in the case of public actors such as biobanks; cf. Aitken et al., 2016; Capps, 2012) but not widespread in other contexts in which the use of health data generated by diverse actors (companies producing wearables, hospitals accumulating numerous tests results as a by-product of their activity, etc.) can be in the public interest. Mandating data pooling is particularly interesting when at least one of the following non-excluding conditions are met:

i) The data is scattered across several data holders holding rights to exclude other agents from its use
ii) Data holders' interests regarding data-pooling differ (Sorbie, 2019)
iii) The optimal data pool is difficult or costly to (re)produce without data holders' involvement

When conditions i) and ii) are met, the ´"tragedy of the anti-commons" (Heller, 1998) emerges: when too many parties have exclusion rights over a resource (the optimal data pool's dataset), the latter will tend to be underused. Mandated data pooling can be a solution to this problem, as the case of DNA sequence data illustrates. In 1996, the Bermuda Principles reached by scientists and policymakers "required that all DNA sequence data generated by the Human Genome Project (HGP) be released to the public just twenty-four hours after generation" (Contreras, 2014, p. 2). This was necessary to ensure that geographically dispersed laboratories that would take months or even years to publish data (if they had decided or were bound to do so) would commit to contributing to and benefit from the acceleration of scientific research brought about by the economies of scale of genetic data efforts and the minimization of duplication of efforts. Indeed, the Bermuda Principles allowed stopping working in parallel on producing the same data and focusing on optimizing their respective tasks based on a timely updated common dataset.

The implementation of forced data pooling on the grounds of public interest can take place on a case-by-case basis or by default. In the first case, whenever an optimal data pool in the public interest cannot be reached, the public actor can intervene to force agents to pool their data on the bases of the public interest, as envisioned by the French Member of Parliament Cédric Villani in his report on artificial intelligence[18]. This mechanism would be similar to compulsory licensing of drug patents under FRAND licenses. When doing so, policymakers should consider whether the possible uses of the data by the members of the data pool would entail a disincentive for data holders to generate more data. If, for example, policymakers decide to mandate the pooling of data with competitors, this might result in the agent being forced to pool data producing less data in the future. On the contrary, if the mandatory data pooling involves agents that are not in rivalry with the data holder or the uses of the mandated pooled data that create rivalry can be credibly precluded in the terms of the license, the latter should continue producing the data in the future. Given that assessing the potential uses of the data on a case-by-case basis might be challenging without accessing the data first, policymakers could mandate some sensitive types of data being shared with a regulatory agency, which would not disclose it for a certain period of time (Sichelman & Simon, 2016) and then assess whether and with whom it might mandate data-pooling. Alternatively, policymakers can mandate the pooling of certain types of datasets by default. The above-mentioned case of Finland's Act on the Secondary Use of Health and Social Data illustrates this approach. A more restrictive version of this approach is France's Digital Republic Act of 2016 *(Loi pour une République Numérique*), which gives municipalities the right to demand private operators to open their data if one of the following conditions is met:

1. The private actor benefits from a public service delegation contract
2. The private actor's activity relies on at least 23 000 € of public subsidies
3. The private actor holds energy consumption data generated by public infrastructure
4. In certain cases, if there is jurisprudence data involved.

Although mandated data sharing on the grounds of public interest is a powerful policy tool to unleash the potential of health data pooling, it should be used with caution for several reasons. First, the very notion of 'public interest' remains a contested notion both legally and ethically (Sorbie, 2016). We shall not develop on this, as it falls outside of the scope of this report. However, it is important to stress that the contestability of the concept does not entail that it should not be used when forced data-pooling can contribute to the public interest, but rather that the latter should be, in as much as possible, explained and justified by the legislator and reflect societies' ethical choices. Second, before implementing forced data sharing, some conditions that justify this measure to be the most efficient one to reach the optimal pool should be met:

a) The data cannot be easily (re)produced or obtained by other means than forcing the parties to pool it
b) It has not been possible to reach voluntary agreements for the parties to pool the data or the incentives required to do so imply a cost too high to justify them being deployed

[18] Villani, C. (2017). « Donner un sens à l'intelligence artificielle. Pour une stratégie nationale et européenne » Available at : https://www.aiforhumanity.fr/pdfs/9782111457089_Rapport_Villani_accessible.pdf

c) Forcing parties to pool their data will not result in a significant disincentive to produce more data in the future
d) Pooling the data does not violate current laws (e.g., privacy laws, national security laws, etc.)

Moreover, even when justified in terms of public interest, mandated data pooling raises the question of the distribution of the benefits and costs of mandated data pooling. On the benefits side, the policy question raised by mandated data-pooling is whether and if so how should the benefits of the mandated data pool be distributed between members, especially when there is rivalry among the members over the jointly created value (cf. Section 1.3). On the cost side, given that data-pooling implies costs for the members that in some cases would not have decided on their own to pool their data, policymakers have to decide whether and if so how to divide these costs among members, on the one side, and between members and the public actor, on the other side.

3.2.3 Developing *de facto* and *de jure* collective control rights over data

We have seen how different factors can affect data holders' willingness to pool data (Section 1) in ways that result in sub-optimal data-pooling scenarios (Section 2). Although different in nature, these factors translate data holders' lack of incentives to pool data with certain agents for certain purposes. This is typically the case with some types of medical research such as orphan diseases, for which the lack of commercial interest results in the abandoning of valuable venues of research. As a result, scientific health research data is underused (Tempini & Del Savio, 2019).

A promising solution to this problem consists in agents that do have incentives to pool the data in the optimal way to have control rights over the data. In the healthcare sector, the above-mentioned (cf. Section 3.1.1) data cooperatives have been developing with that purpose in the past years. Examples include Midata, Salus Coop, Transiscope, Moipatient, Patients Like Me and CureTogether. Some have already proven to contribute to the public interest (Blasimme et al., 2018). For example, A Patiens Like Me authored study of the drug lithium-carbonate has been published in *Nature Biotechnology* (Wicks et al., 2011) and in 2016 and 2018 members of the OpenAPS project, which "collects measurements of blood glucose levels from all community members using an artificial pancreas system" (Strotbaun et al, 2019), submitted two studies to the meeting of the American Diabetes Association (D. Lewis et al., 2016; D. M. Lewis et al., 2018)

The benefits of data cooperatives arise from the fact that they develop effective mechanisms for the agents with incentives to pool their data in the public interest (for example, patients suffering from a rare disease) to effectively make a use of their personal data in that manner. Indeed, these organizations facilitate the tedious decision-making process required for individuals to make truly informed decisions about each data processing operation, which can hinder data pooling from individuals even when they wish their data to be pooled and processed for a given purpose. In doing so, they seek "striking a balance between granular control and day-to-day delegation to representatives" (Tempini & Del Savio, 2019, p. 15). Moreover, data cooperatives also operate as facilitators in that they lift barriers imposed by the legal complexity of privacy policy. The latter can reach hundreds of pages and be "written in legalese", and hence "it is unreasonable to expect that an individual can read and comprehend them and give a truly informed consent" (Purtova, 2017, p. 17). Finally, it is important to point out that data cooperatives

are not merely intermediaries that reduce the non-pecuniary transaction costs involved in reclaiming personal data from data holders and pooling it with other organizations. They are also a vehicle for collective decision-making on the uses of datasets in which individuals play a major role in the governance. As such, as seen in Section 3.2.1., they create the necessary trust for individuals to pool sensitive personal (health) data and authorizing it being used by third parties. Moreover, because they are typically the result of the self-organization of collectives of patients in the interest of contributing to solving health issues through data pooling, their growth is generally welfare improving.

However, many factors hinder the unleashing of the potential of health data cooperatives, including the fact that they are relatively new entities looking in search for their own organizational and business models**. In this section, we will evoke policies that can contribute to lifting the main barriers to the development of health data cooperatives and, more generally other organizations and legal vehicles for individuals to exert collective control rights over (personal) data**. We will avoid mentioning policies that overlap with other policy recommendations present in this report, such as those meant to enhance soft and hard interoperability (Section 3.1.2) or those aiming at strengthen data cooperatives' business models (Section 3.1.1). We will begin with the policy recommendations that do not require adapting the current individual-rights-based legal framework (Section 3.2.3.1) and would hence support the development of *de facto* collective control rights over data. Then we will develop on policies that would imply creating new collective *de jure* control rights over data (Section 3.2.3.2).

3.2.3.1 Within the current legal framework

Within the current legal framework, data cooperatives and other data stewards function as aggregators of individual control rights over data that RGPD and legislation related to health information give to individuals. Each member of the data cooperative has to use his/her data portability rights to reclaim his/her data from multiple data holders (hospitals, private practitioners, companies that sell wearables, etc.) and then authorize a third party (the data cooperative) to pool the data with other individuals' for certain purposes that might differ depending on the type of data involved. This process could be facilitated in several manners.

First, policymakers could ease the administrative and legal procedures involved in a data steward such as a data cooperative claiming many individuals' data from data holders while remaining in the current legal framework of individual data portability (Jacquart et al., 2018).

Second, they can foster the use of certain standard personal health data licenses that, just like Creative Commons licenses, could be used by individuals, data holders and data processors to easily understand which datasets can be used for which purposes. In order to do so, policymakers can work with existing health data cooperatives in designing these licenses. The idea of these licenses has been put forward by scholars (Maurel, 2014) under different names such as "privacyleft" (Saint-Aubin, 2012) or "Privacy Commons" (Mercier, 2014) and experimented by health data cooperatives that try to understand which terms and conditions regarding the use of data are the more relevant for individuals to decide to pool their data depending on the type of data involved. For example, after months of working with individuals willing to share their health data for research purposes, the Spanish health data cooperative Salus Coop has designed a standard data-sharing license that takes into

account the terms and conditions that emerged as necessary for individuals to allow the sharing of their health data[19].

Third, policymakers can demand health data holders to implement a homogeneous and user-friendly interphase to allow individuals to exert their data portability rights. The Blue Bottom is a good example of it. In 2010, a simple blue bottom allowing individuals to view, download and transmit their personal health data from data holders' (laboratories, practitioners, hospitals, insurers, etc.) websites was implemented for the first time in the United States by the Department of Veterans Affairs. Since 2014, more than 500 organizations signed an engagement[20] to adopt the Blue Button. Since then, more than 100 million Americans have been able to easily download their health data to transmit it to third parties. Until today, almost 3 million Medicare beneficiaries, Veterans and active-duty military personnel have accessed their health information using the Blue Button[21]. Finally, in conjunction with a European Blue Button, as mentioned in Section 3.1.2, policymakers can develop a single access point at the European level for individuals to easily manage the available and downloaded health data, the authorizations of use of their personal health data given to different data cooperatives and other data processers throughout the entire lifecycle of data. By mandating that health data cooperatives and other data processers become interoperable with this unique access point, policymakers can avoid the current identity fragmentation problem (i.e., personal data being scattered across a myriad of data holders and user interphases) being reproduced at a second-level with individuals having to deal with different data cooperatives they might be part of individually, each having its own user interphase and operational procedures.

Finally, policymakers can act as certifiers of both health data cooperatives and data holders that comply with certain standards in terms of facility of data portability, privacy protection and ethical uses of the data (Jacquart et al., 2018). This could work as a trust-reinforcing mechanism that would foster the choice of health providers allowing individuals to easily migrate their data, hence creating economic incentives for data holders to compete on data portability. This would in turn facilitate the development of health data cooperatives and other data stewards. A similar process has been put in place by the United Kingdom's National Health Service, which opened a website destined to the general public that lists "trusted health and wellbeing apps"[22].

3.2.3.2 Enlarging the scope of personal data protection and creating a legal framework for collective portability and consent

As mentioned above, under the current legal framework (notably GDPR) data cooperatives are based on individual data portability and individual consent to the use of personal data. **However, this legal framework only gives individuals the legal capacity to control a fraction of their personal data, which hinders the potential that data cooperatives and other data stewards can unleash in terms of fostering data pooling**. This is due to the fact that "the rapid evolution of data collection and analysis technology may create ambiguous borderline cases in the definition of personal data"

[19] See https://www.saluscoop.org/licencia

[20] See https://www.healthit.gov/patients-families/pledge-info

[21] See https://obamawhitehouse.archives.gov/blog/2015/10/01/celebrating-5-year-anniversary-blue-button-open-health-data

[22] https://www.nhs.uk/apps-library/

(Duch-Brown et al., 2017, p. 16). In particular, the current legal framework is ill suited to give individuals control over several non-excluding types of data: observed data, inferred data and relational data.

Regarding the first type of data, doubts remain on whether only volunteered data (e.g., data on how many kilometers a person walked typed by the individual in an app) is to be considered personal data in terms of GDPR, and hence be eligible for data portability. It is yet unclear if or under which conditions can observed data (e.g., data on how many kilometers a person has walked obtained through GPS location) or inferred data (a health score based on how much does the person walk per day) are legally considered to be personal data (Bayamlioğlu et al., 2018). Clarifying the scope of GDPR in terms of observed and inferred data in such a manner that gives individuals a solid legal ground to exert (individual) control rights over more of 'their' data could hence foster the potential of (health) data cooperatives and other data stewards.

Finally, because GPDR is built on individual consent, it is ill equipped to provide individuals with control rights over 'relational data', a term we will use to refer to data that relates to at least two identified or identifiable natural persons. As several authors have pointed out (Article 29 Working Party, 2014; Gniady, 2007; Goldman, 2005) "medical data is rarely just about one individual but often relates to others, who may be harmed as a result" (Harbinja, 2019, p. 100). The classic example of relational health data is the genetic data of an individual, which can reveal sensitive information about other members of his/her family and even harm them as a result. For example, genetic data about one individual can reveal a hereditary disease about a family member to an insurance company or an employer that might use it as a basis for discrimination. More generally, data used for profiling purposes revealed by one individual can be used to obtain information about other individuals that might have even not revealed their personal data, generating so negative externalities (Acemoglu et al., 2019; Fairfield & Engel, 2015). This is all the more so because data is increasingly gathered about large and undefined groups and analyzed on the basis of patterns and group profiles (Taylor, Floridi, et al., 2017). In addition, although generally anonymized, aggregated data is becoming increasingly re-identifiable (de Montjoye et al., 2012; Ohm, 2009) and hence portable in the sense of GDPR .

The fact that "in the context of the modern information practices no personal data remains strictly personal" (Purtova, 2017, p. 73) and the promising emerging practices in terms of collective control rights over data that data cooperatives illustrate (but do not exhaust)[23] call for an evolution of the current legal framework to acknowledge forms of collective control

[23] There are some interesting examples of communities of interests of individuals shaped around data that, lacking *de jure* collective rights over it, have devised actions to exert *de facto* collective control rights over data. Unable to control how Waze used data on their territories to direct drivers through them, inhabitants from cities such as Lieusaint (France) or Cornebarrieu (France) have coordinated to signal fake events (traffic jams, accidents, etc.) on Waze in order for the platforms' algorithms to direct drivers away from these cities (Signoret, 2019). In the context of social networks, a group of teenagers has devised a system to share their Instagram accounts in order to avoid individual profiling based on the data generated by the individual use of the social network (Ng, 2020). In 2009, following a controversy over a policy change that was perceived by users as giving Facebook unchecked power over their data, the platform allowed them to give feedback and vote on some new terms of service that included the allowed uses of user data. However, given that the turnout (0.3%) was below the minimal threshold that Facebook had set for the new terms of service to be binding (30%), Facebook users' collective rights over their data did not materialize (Robertson, 2018).

rights over 'personal' data along and in articulation with individual control rights (Taylor, Floridi, et al., 2017). Although no standard term has emerged to cluster contributions on the idea of collective control rights over data, many authors have employed the term "group privacy" to refer to it (Helm, 2016; Mittelstadt, 2017; Taylor, van der Sloot, et al., 2017). Coherently, they see data protection law as the logical area of law in which these rights should be rooted when communities of interests around data exist, for example regarding genetic groups (Hallinan & de Hert, 2017)[24]. More precisely, in order to encourage the development of (health) data cooperatives, two types of rights that mirror those granted by GDPR at the individual level could be created: collective consent and collective data portability. Indeed, **these two rights would constitute the legal tools that collectives of individuals having the incentives to constitute optimal data pools could rely on to be able to seamlessly do so. Moreover, by recurring to collective consent and collective data portability, conflicts between individuals affected by the use of data can be avoided and hence the constitution of such data pools fostered.**

To our knowledge, collective consent has been put forward for the first time by Bygrave and Shartum (2009), who employ the term "to denote consent exercised on behalf of a group of data subjects but without these persons individually approving each specific exercise of that decisional competence. In other words, it denotes a mechanism involving collective conferral or withdrawal of consent which is binding on all of the group members, even when some of these disagree with the particular decision" (Bygrave & Schartum, 2009, p. 169). Other authors such as Lindberg (2018) have pursued research on this emerging legal concept. Moreover, some basic principles should apply for collective consent to be compatible with individual consent:

a) Collective consent can be conferred and withdrawn by the collective
b) Individuals can freely confer and withdraw their consent to a collective
c) Individual withdraw of consent by an individual from a collective should not entail the loss of membership of the collective

Collective portability, a concept put forward by authors such as Maurel (2019), Messaud (2018) or Carballa Smichowski (2019), would derive directly from collective consent to the extent that, once a collective of individuals is constituted as a legal entity, it could exert its right to data portability in the same way as individuals can do it today regarding 'their' personal data. Designing collective consent and portability is beyond the scope of this report. However, it is interesting to point out at this stage that, as studied by Hallinan and De Hert (2017) regarding genetic communities, introducing group privacy into GPDR (which is more ambitious than merely introducing collective consent and data portability) presents few obstacles, the exemption being the concept of 'data subject', which currently excludes all but natural

[24] Interestingly, it is in the health sector that considerations regarding collective control rights over data as part of data protection law have been made. In 2004, the Article 29 Working Party mentioned in reference to genetic data that a "legally relevant social group can be said to have come into existence namely, the biological group, the group of kindred as opposed, technically speaking, to one's family" (Article 29 Working Party, 2014, p. 9).

persons. However, the implementation of collective consent and data portability would require taking into solving several issues to order for them to be effective. In the rest of this section we will lay out the main ones identified by Bygrave & Schartum (2009).

The first point to clarify is who should be allowed to be a member of the collective. As put by Ruhaak (2020), "as a general rule, the group should be made up of people affected by the decisions and not be made up by people not affected by the decision. That being said, in some cases those affected by the decision to share data may not be able to form part of the collective (for instance because they have not yet been born) and a representative might be elected to advocate on their behalf". This membership logic matches the emerging collectives that individuals constitute around health data cooperatives. Indeed, the latter are usually constituted by groups of patients that share a (rare) disease on which they want to foster research by pooling their data, family members or patients that share the use of a wearable device that generates health data pools about them. This membership logic might be channeled by new legal persons (e.g., the above-mentioned data cooperatives or data trusts) as well as existing ones that already promote or safeguard common interests related to a group of individuals' relational data such as advocacy groups and patients' associations. Moreover, collectives should be able to evolve in their membership over time and depending on the intended uses of the data. Indeed, "the groups contributing to and affected by data, or the groups for whom the data is relevant are not always static and their boundaries shift" (Purtova, 2017, p. 77); they "are usually dynamic entities: they come in an endless number of sizes, compositions, and natures, and they are fluid" (Taylor, Floridi, et al., 2017, p. 16). In an era of "calculated publics" (groups constructed by algorithms with the aim of influencing behavior, cf. Gillespie, 2014) "where almost everyone is constantly being grouped and regrouped, unaware, by data analytics" (Taylor, van der Sloot, et al., 2017, p. 279) collectives having control rights over data should have a variable geometry.

The second point refers to the governance of the collective. While, as seen in Section 3.2.1, there is no one-size-fits all best data governance model, collective consent and portability will require each collective having a clear and legally-binding decision-making process that makes explicit who decides on what within the collective and according to which process. This decision process should be able to find an equilibrium between three types of interests conflicts that may emerge (Hallinan & de Hert, 2017): first, individuals and the collective; second, collectives and data controllers; third, between collectives for which interests over a given dataset might overlap. Moreover, the questions on whether and, if so, under which circumstances (some) decision-making bodies should be determined by external agents such as the State remains to be answered.

The third point is how collectives come about. Currently, these collectives emerge from a common action by a group of data subjects with a common interest. However, it could be imagined that the collective would be initiated by the data collector when trying to obtain data from a collective of individuals that are affected by its use, or even constituted *a priori* through legislation in the same manner as in many countries companies that exceed a certain number of employees are legally bound to have unions representing workers' collective interests.

3.2.4 Enabling posthumous medical data donation

In Section 2 we have seen how the different mechanisms affecting data-holding agents' incentives to pool data (cf. Section 1) can result in sub-optimal data-pooling scenarios. One of the ways in which this can be solved is by 'giving' control rights over data to agents that do have incentives to pool it within an optimal data pool. This option, which we have analyzed in terms of agents constituting new control rights over data through data pools' governance mechanisms (Section 3.2.1), mandated data-pooling (Section 3.2.2.) and measures that foster collective data control rights over 'personal' data (Section 3.2.3) is particularly interest when it comes to posthumous data donation.

Once deceased, individuals' health data falls into a legal blind spot that renders its pooling extremely difficult. While posthumous organ donation is a well-established practice in many countries across the European Union, (Cronin & Price, 2008; D. Price, 2000; Van Ypersele, 2009), **posthumous medical data donation is not**. As a result, existing health data that could be pooled to create value is not exploited because, even when data holders have the intention to do so, they usually do not have the required legal consent. This represents "a huge opportunity cost and has a negative effect of advancements in research" (Harbinja, 2019, p. 98), as studies have shown (Jones et al., 2017) along with the fact that most individuals are willing to donate personal data for research purposes in the public interest (Skatova et al., 2014). Lifting this blind spot entails creating new control rights that allow for posthumous medical data donation (PMDD). In terms of the theoretical development made in Sections 1 and 2, these policies imply favoring the emergence of new players and their respective data endowments that have more incentives to pool data within the optimal data pool. In this section, we will succinctly present the main legal and organizational changes that policymakers can implement to encourage PMDD.

One of the main barriers to PMDD is the lack of a common legal approach on the matter across the European Union. The main common European framework to deal with personal data, GDPR, does not cover the data of the deceased. However, some Member States such as France in its Digital Republic Act of 2016 (*Loi pour une République Numérique*) or Hungary (Castex et al., 2018) have created legislation to address the processing of personal data (including, but not limited to medical data) of deceased persons. Hence, **one of the main policies that can be put in place to foster PMDD consists in recognizing post-mortem privacy at the European level in accordance with GDPR**. "Consent would, therefore, need to be freely given, informed and unambiguous, by a statement or by a clear affirmative action, whereby an individual signifies agreement with PMDD" (Harbinja, 2019, p. 109).

Practically, such a legislation could translate into a European data advance directive in which individuals would legally state while in life how their personal health data can be used once they pass away (Shaw, 2019). **Important implementation questions remain to be answered in order to implement this type of directive. One of them is whether, following article 4(11) of the GDPR, this consent should cover specific or broad uses**. The latter option is favored by researchers in the emerging field of PMDD (Krutzinna et al., 2019b; Shaw et al., 2016) mainly because all the possible specific uses cannot be listed in advance, especially when technological progress opens the door to unimagined uses of data (Shaw, 2019). Even if perfect foresight was possible, it would entail a lengthy process that could discourage individuals from donating their data postmortem. Given that individuals cannot be consulted to consent to unforeseen specific uses once diseased, broad consent is a good option. For example, as

early as 2013, the Finish BioBank Act[25] allowed biobanks to obtain broad consent for the use of the newly collected data. By recurring to broad consent, researchers were able to avoid asking each patient for his/her consent every time for every project that required using a patient's sample. In that manner, they can avoid the problem of re-consenting, which is "costly and time-consuming, and difficulty in locating people can result in high drop-out rates" (Kaye et al., 2015, p. 1). **Another non-excluding option consists in trusting (some) of the data to Posthumous Data Guardians (PDGs) to which open consent can be given to decide on behalf of the deceased person** (Shaw, 2019; Sorbie, 2019)**.** These third-party data stewards could take different legal forms such as data trust (Delacroix & Lawrence, 2019; Royal Society, 2017) or the above-mentioned data cooperatives. Moreover, third-party data stewards could contribute to solving another problem inherent to PMDD: conflicting familial interest. As shown in the previous section, even personal health data such as genetic data can be relational. As a result, certain uses of a deceased person's health data can harm his or her family, for example by revealing a hereditary disease, which could result in discrimination by third parties (Harbinja, 2019). If third-party stewards represent both the interests of the family and the deceased individual (hence being a legal vehicle for collective rights over data, cf. Section 3.2.3), they can be a solution to this issue. Another option would consist in excluding relational health data from a PMDD regime, as proposed by Krutzinna, Taddeo and Floridi (2019b).

In terms of implementation, several organizational principles can be put in place to foster PMDD if a legislation allowing it is put in place. First, as mentioned in Section 3.1.2, **the questions addressed to individuals willing to donate their health data post mortem should simple and intelligible but also comprehensive**, which implies covering at least the following items: whether the data can be shared at all (which might depend on the sensitivity of the health data involved); who (if anyone) else it concerns (typically family members, as in the case of genetic data); who it can be shared with; what form it can be shared in (fully identifiable, de-identified, pseudonymised and fully anonymized); who (if anyone) needs to give consent for any other type of sharing in the future (Shaw, 2019), and for which purposes can data be used (research, commercial uses, etc.)[26]. Second**, data donor cards could be legally recognized in the same manner as donors cards are in some countries to donate organs in order to facilitate the use of PMDD** (Harbinja, 2019). Third, **a register of data donors could be constituted to avoid seeking consent from families if the individual has expressed his or her wishes regarding data donation** (Caldicott, 2013; Shaw et al., 2016). Fourth, in order to foster the adoption of PMDD, **individuals can be asked to agree to PMDD when applying for a driving license, signing up at a general practitioner practice (Sorbie, 2019) or voting for the first time,** as many countries do regarding organ donation.

[25] https://www.openaccessgovernment.org/finland-a-framework-genetic-research/68395/

[26] We have added the 'purpose' variable to Shaw's data donation metadata list

4 Conclusions

In Section 1 we showed how the following factors, which can have opposite effects in a given (potential) data-pooling situation, can affect an agent's decision on whether and with whom to pool data: the relationship between the economies and costs of data-pooling, the relationship between the marginal utility of data-pooling and coordination costs, the existence of a zero-sum game and the associated value sharing rules of the pool, the existence of competing data pools, agents' data endowments distribution and the existence of intra-pool negative externalities. In order to encourage data-pooling, policymakers should in as much as possible assess how these factors are affecting agents' data-pooling decisions in order to put into place policies to achieve the optimal data pool (i.e., a list of agents pooling certain types of data for specified uses) according to their policy objectives (maximizing the volume of pooled data, increasing total welfare, increasing consumer welfare, etc.). Once the optimal data pool has been determined for a given situation, different policies can be put into place depending on which factors are impeding its realization.

Providing economic incentives to data-pooling can foster data pooling if transaction or lump sum costs are high enough in comparison to benefits. In absence of sufficient incentives, there might be no data-pooling (cf. Figure 1) or a data pool smaller than what policymakers would desire (cf. Table 10 and Figure 3). These incentives can take several forms such as tax breaks, grants, investing in technological infrastructure that lowers the cost of data-pooling, financial support to (health) data cooperatives, attributing subsidies based on the scope of data-pooling or supporting data-pooling projects through grants. In order for these policies to be efficient, the positive welfare effect of all of these measures has to be larger than their fiscal cost. Moreover, it is important that these policies do not target individuals but rather organizations. Indeed, when economic incentives are given to foster transactions that are not economically motivated such as health data donation by individuals, agents tend to do fewer transactions instead of more because the economic incentive undermines the non-economic motivations.

In cases in which the optimal data pool is not achieved because of the existence of high lump-sum (e.g., developing a standard to pool data that each agent stores in different formats) **and coordination costs** (e.g., agreeing on technical and non-technical standards in order to pool data), **policies aiming at facilitating hard and soft interoperability are advisable, notably if the data is to be exploited by a third-party that has to access several datasets held by different data holders**. In the case of health data, fostering hard interoperability implies developing and encouraging the adoption of standard protocols for data sharing, standard formats to store health data, standard terminologies and cross-institutional patient identification numbers. Soft interoperability, in turn, refers mainly to standard agreements for data sharing by organizations and individuals in the case of health data. In the particular case of personal health data, the implementation of data donation passes that mirror organ donation passes would allow individuals to easily choose the conditions under which they will let third parties use their health data. In other cases, the development and diffusion of standard organization-to-organization health data-pooling agreements can contribute to develop soft data interoperability.

When public actors are (potential) members of the optimal data pool, leveraging on their data to achieve the optimal data pool configuration is a policy tool that can be used. In the case of zero-sum-games, public actors can decide to lower their share of the value created through data pooling to increase that of an agent that does not have enough incentives to join the optimal data pool. Public actors can also condition their participation in a data pool to the acceptance of certain members that other pool members do not have incentives to accept. This strategy is particularly interesting when the existence of intra-pool negative externalities prevents one or more actors from being part of the optimal data pool and all the members of the pool would benefit from it even in presence of the agents that generate the negative externalities. In order for strategies based on leveraging on public actors' data to work, the latter should hold data that is valuable to the other members of the optimal data pool and be legally able to decide on its allocation.

When intra-pool negative externalities block the constitution of the optimal data pool, policymakers can act by identifying, diffusing and implementing (in cases in which they are part of the public actor is part of the data pool) **trust-building data governance mechanisms**. The latter should be based on two principles: transparency and accountability. Moreover, **three types of trust-building mechanisms that build on these principles and are of particular interest in the case of health data can be fostered**. The first one refers to **mechanisms leading to agents feeling they are real interlocutors able to shape the system** rather than passive stakeholders. Indeed, if agents (notably individuals, which are usually less powerful than organizations) feel they can shape the 'rules of the game' they will be more likely trust other agents with their data, as they know that they will have a say in what can happen if the uses or the provision of the data do not meet their expectations. The second type of mechanisms refers to **technical means that minimize the exposure of data. Examples include safe havens, accreditation and technologies allowing exploiting several datasets as one without requiring data holders to migrate their data to a single joint server**. The third one is **dynamic consent**, whereby data holders' preferences regarding how their data can be used and by whom can be exercised as the uses of their data evolve.

When data holders do not have incentives to pool certain datasets for a given purpose, facilitating the use of control rights over data or giving control rights to agents that do have an incentive to the share it within the optimal pool can be a promising solution. Given that in the case of health data these agents are generally individuals and that GDPR gives them portability rights over their personal data, these policies can focus on facilitating the use of (health) data portability (which falls under the current legal framework) and on enlarging the scope of control rights over personal (health) data individuals currently have, which entails modifying the current legal framework, notably GDPR. Policies that fall under the current legal framework include easing the administrative and legal procedures involved in a data steward claiming many individuals' data from data holders, promoting the use of certain standard personal health data licenses, demanding health data holders to implement a homogeneous and user-friendly interphase to allow individuals to exert their data portability rights and public bodies acting as certifiers of both health data cooperatives and data holders that comply with certain standards in terms of facility of data portability, privacy protection and ethical uses of the data. Two types of policies can be put into place to give individuals more control rights over personal data by modifying the current legal framework. The first consist in **clarifying the scope of GDPR in order for**

individuals to be able to have larger control rights over (some) of observed and inferred data that relates to them, eventually under certain circumstances that remain to be specified. The second type of policy consists in **creating collective consent and portability rights on relational data** (i.e., data that relates to at least two identified or identifiable natural persons), as the current individual-consent based data protection regime hinders data-pooling by individuals when 'personal' data involves several individuals. The latter implies a careful legal and ethical discussion in order to deal with issues such as the articulation with individual control rights over data, who should be allowed to be a member of the collective, the governance of the collective and how do the collectives come about.

Legislating to allow for posthumous medical data donation would allow to pool and exploit data that is currently unused because GDPR does not recognize post-mortem privacy. Doing so would give individuals and data processers a legal ground on which donate and exploit medical data, respectively. In addition to this change in legislation, other implementation policies can be put into place to foster PMDD. **Data donor cards** could be legally recognized in the same manner as donors' cards are in some countries to donate organs in order to facilitate the use of PMDD. In accordance with GDPR, these cards should be established based on individuals' freely given, informed and unambiguous consent. Moreover, as mentioned above, this consent should follow from **intelligible and comprehensive questions regarding the post mortem use of medical data addressed to individuals. A register of data donors could be constituted** to avoid seeking consent from families if the individual has expressed his or her wishes regarding data donation. Finally, **individuals can be asked to agree to PMDD when applying for a driving license, signing up at a general practitioner practice or voting for the first time,** as many countries do regarding organ donation.

Finally, when pooling data is in the public interest but data holders have little incentives to pool their data, data pooling can be mandated. Mandated data pooling is an interesting policy option when the data of the optimal data pooled is scattered across several data holders with conflicting interests. This policy can take several forms. It can imply mandating the opening of the data or mandating that some legal persons should authorize access to the data they hold to other specified legal persons only for some specified uses. It can be by default (e.g., mandating data pooling of certain types of data by default) or on a case-by-case basis. It can also imply an economic compensation to the original data holder for the provision of the data or not. In order for mandated data-pooling to be justified as the most efficient policy to reach the optimal data pool, several conditions should be met: a) the data cannot be easily (re)produced or obtained by other means than forcing the parties to pool it; b) it has not been possible to reach voluntary agreements for the parties to pool the data or the incentives required to do so imply a cost too high to justify them being deployed; c) forcing parties to pool their data will not result in a significant disincentive to produce more data in the future; d) pooling the data does not violate current laws (e.g., privacy laws, national security laws, etc.).

References

Acemoglu, D., Makhdoumi, A., Malekian, A., & Ozdaglar, A. (2019). *Too much data : Prices and inefficiencies in data markets*. National Bureau of Economic Research.

Admati, A. R., & Pfleiderer, P. (1986). A monopolistic market for information. *Journal of Economic Theory*, *39*(2), 400-438.

Aitken, M., Jorre, J. de S., Pagliari, C., Jepson, R., & Cunningham-Burley, S. (2016). Public responses to the sharing and linkage of health data for research purposes : A systematic review and thematic synthesis of qualitative studies. *BMC medical ethics*, *17*(1), 73.

Albrecht, U.-V., von Jan, U., Pramann, O., & Fangerau, H. (2016). Kapitel 7. Gesundheits-Apps im Forschungskontext. *Chancen und Risiken von Gesundheits-Apps (CHARISMHA)*.

Ambuehl, S., Niederle, M., & Roth, A. E. (2015). More money, more problems ? Can high pay be coercive and repugnant? *American Economic Review*, *105*(5), 357-360.

Article 29 Working Party. (2014). *Working Document on Genetic Data*. https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2004/wp91_en.pdf

Batini, C., & Scannapieco, M. (2006). *Data quality. Data-centric Systems and Applications*. Springer.

Bayamlioğlu, E., Baraliuc, I., Janssens, L., & Hildebrandt, M. (Éds.). (2018). *Being profiled : Cogitas ergo sum : 10 Years of Profiling the European Citizen*. Amsterdam University Press. https://doi.org/10.2307/j.ctvhrd092

Bellivier, F., Benhamou, F., Cornu, M., & Noiville, C. (2015). Collections muséales et collections biologiques : De la conservation à l'accès. *Le retour des communs. La crise de l'idéologie propriétaire, propriétaire*, 197.

Bergemann, D., & Bonatti, A. (2019). Markets for information : An introduction. *Annual Review of Economics*, *11*, 85-107.

Binns, R. (2018). Algorithmic accountability and public reason. *Philosophy & technology*, *31*(4), 543-556.

Blasimme, A., Vayena, E., & Hafen, E. (2018). Democratizing health research through data cooperatives. *Philosophy & Technology*, *31*(3), 473-479.

Bygrave, L. A., & Schartum, D. W. (2009). Consent, proportionality and collective power. In *Reinventing data protection?* (p. 157-173). Springer.

Caldicott, F. (2013). Information : To share or not to share. Information Governance Review. *Information: To share or not to share*.

Capps, B. (2012). The public interest, public goods, and third-party access to UK Biobank. *Public Health Ethics*, *5*(3), 240-251.

Carballa Smichowski, B. C. (2018). The value of data : An analysis of closed-urban-data-based and open-data-based business models. *Sciences Po Cities and Digital Technology Chair N°01/2018*.

Carballa Smichowski, B. C. (2019). Alternative Data Governance Models : Moving Beyond One-Size-Fits-All Solutions. *Intereconomics*, *54*(4), 222-227.

Castell, S., Robinson, L., & Ashford, H. (2018). *Future data-driven technologies and the implications for use of patient data*. https://acmedsci.ac.uk/file-download/6616969

Castex, L., Harbinja, E., & Rossi, J. (2018). Défendre les vivants ou les morts? *Reseaux*, *4*, 117-148.

Contreras, J. L. (2014). Constructing the genome commons. *Governing Knowledge Commons*, *99*, 112-113.

Coriat, B. (2015). *Le retour des communs : & la crise de l'idéologie propriétaire*. Éditions Les Liens qui libèrent.

Cronin, A. J., & Price, D. (2008). Directed organ donation : Is the donor the owner? *Clinical Ethics*, *3*(3), 127-131.

Delacroix, S., & Lawrence, N. D. (2019). Bottom-up data Trusts : Disturbing the 'one size fits all'approach to data governance. *International Data Privacy Law*, *9*(4), 236-252.

de Montjoye, Y.-A., Wang, S. S., Pentland, A., Anh, D. T. T., & Datta, A. (2012). On the Trusted Use of Large-Scale Personal Data. *IEEE Data Eng. Bull.*, *35*(4), 5-8.

Dosis, A., & Sand-Zantman, W. (2019). The ownership of data. *Available at SSRN 3420680*.

Duch-Brown, N., Martens, B., & Mueller-Langer, F. (2017). The Economics of Ownership, Access and Trade in Digital Data. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2914144

Durante, M. (2015). The democratic governance of information societies. A critique to the theory of stakeholders. *Philosophy & Technology*, *28*(1), 11-32.

European Commission. (2020). *Communication A European Strategy for Data*. COM(2020) 66.

Fairfield, J. A., & Engel, C. (2015). Privacy as a public good. *Duke LJ*, *65*, 385.

Floridi, L. (2014). Big Data and information quality. In *The philosophy of information quality* (p. 303-315). Springer.

Floridi, L. (2019). *The logic of information : A theory of philosophy as conceptual design*. Oxford University Press.

Gardner, R., Ostrom, E., & Walker, J. M. (1990). The Nature of Common-Pool Resource Problems. *Rationality and Society*, *2*(3), 335-358. https://doi.org/10.1177/1043463190002003005

Gay, V., & Leijdekkers, P. (2015). Bringing health and fitness data together for connected health care : Mobile apps as enablers of interoperability. *Journal of medical Internet research*, *17*(11), e260.

Gillespie, T. (2014). The relevance of algorithms. *Media technologies: Essays on communication, materiality, and society*, *167*(2014), 167.

Gniady, J. A. (2007). Regulating direct-to-consumer genetic testing : Protecting the consumer without quashing a medical revolution. *Fordham L. Rev.*, *76*, 2429.

Goldman, B. R. (2005). Pharmacogenomics : Privacy in the era of personalized medicine. *Nw. J. tech. & intell. prop.*, *4*, 83.

Hafen, E. (2019). Personal data cooperatives–a new data governance framework for data donations and precision health. In *The Ethics of Medical Data Donation* (p. 141-149). Springer, Cham.

Hallinan, D., & de Hert, P. (2017). Genetic classes and genetic categories : Protecting genetic groups through data protection law. In *Group Privacy* (p. 175-196). Springer.

Harbinja, E. (2019). Posthumous Medical Data Donation : The Case for a Legal Framework. In J. Krutzinna & L. Floridi (Éds.), *The Ethics of Medical Data Donation* (Vol. 137, p. 97-113). Springer International Publishing. https://doi.org/10.1007/978-3-030-04363-6_6

Heller, M. A. (1998). The tragedy of the anticommons : Property in the transition from Marx to markets. *Harvard law review*, 621-688.

Helm, P. (2016). Group privacy in times of big data. A literature review. *Digital Culture & Society*, *2*(2), 138-151.

Henriksen, A., Mikalsen, M. H., Woldaregay, A. Z., Muzny, M., Hartvigsen, G., Hopstock, L. A., & Grimsgaard, S. (2018). Using fitness trackers and smartwatches to measure physical activity in research : Analysis of consumer wrist-worn wearables. *Journal of medical Internet research*, *20*(3), e110.

Hess, C. (2008). Mapping the new commons. *Available at SSRN 1356835*.

Hummel, P., Braun, M., & Dabrock, P. (2019). Data Donations as Exercises of Sovereignty. In *The Ethics of Medical Data Donation*. Springer Open.

Jacquart, G., Medjek, S., & Molins, M. (2018). *Pilote mes infos. Synthèse, enseignements, actions.* FING. http://mesinfos.fing.org/publications/

Jones, C. I., & Tonetti, C. (2020). Nonrivalry and the Economics of Data. *American Economic Review*, *110*(9), 2819-2858.

Jones, K. H., Laurie, G., Stevens, L., Dobbs, C., Ford, D. V., & Lea, N. (2017). The other side of the coin : Harm due to the non-use of health-related data. *International Journal of Medical Informatics*, *97*, 43-51.

Kalai, E. (1977). Proportional solutions to bargaining situations : Interpersonal utility comparisons. *Econometrica: Journal of the Econometric Society*, 1623-1630.

Kaye, J., Whitley, E. A., Lund, D., Morrison, M., Teare, H., & Melham, K. (2015). Dynamic consent : A patient interface for twenty-first century research networks. *European journal of human genetics*, *23*(2), 141-146.

Koutroumpis, P., & Leiponen, A. (2013). Understanding the value of (big) data. *2013 IEEE International Conference on Big Data*, 38-42.

Krutzinna, J., Taddeo, M., & Floridi, L. (2019a). An Ethical Code for Posthumous Medical Data Donation. In *The Ethics of Medical Data Donation* (p. 181-195). Springer, Cham.

Krutzinna, J., Taddeo, M., & Floridi, L. (2019b). Enabling posthumous medical data donation : An appeal for the ethical utilisation of personal health data. *Science and Engineering Ethics*, *25*(5), 1357-1387.

Laurie, G., Ainsworth, J., Cunningham, J., Dobbs, C., Jones, K. H., Kalra, D., Lea, N. C., & Sethi, N. (2015). On moving targets and magic bullets : Can the UK lead the way with responsible data linkage for health research? *International journal of medical informatics*, *84*(11), 933-940.

Lewis, D., Leibrand, S., & #OpenAPS Community. (2016). Real-World Use of Open Source Artificial Pancreas Systems. *Journal of Diabetes Science and Technology*, *10*(6), 1411-1411. https://doi.org/10.1177/1932296816665635

Lewis, D. M., Swain, R. S., & Donner, T. W. (2018). Improvements in A1C and Time-in-Range in DIY Closed-Loop (OpenAPS) Users. *Diabetes*, *67*(Supplement 1), 352-OR. https://doi.org/10.2337/db18-352-OR

Lindberg, A. (2018). *Alternatives for the Individual Consent in Data Protection Law*.

Martens, B. (2020). Data Access, Consumer Interests and Social Welfare : An Economic Perspective. *Consumer Interests and Social Welfare: An Economic Perspective (May 18, 2020)*.

Martens, B., de Streel, A., Graef, I., Tombal, T., & Duch-Brown, N. (2020). Business-to-Business data sharing : An economic and legal analysis. *EU Science Hub*.

Martens, B., & Duch-Brown, N. (2020). *The economics of Business-to-Government data sharing*.

Maurel, L. (2014, septembre 1). Une gouvernance en communs des données personnelles est-elle possible ? - *S.I.Lex* -. https://scinfolex.com/2014/09/01/une-gouvernance-en-communs-des-donnees-personnelles-est-elle-possible/

Maurel, L. (2019). *Contre le pouvoir des plateformes, établir une portabilité sociale des données?*

Mercier, S. (2014, mars 12). *Biens communs et données personnelles : Il nous faut inventer !* http://www.bibliobsession.net/2014/03/12/biens-communs-et-donnees-personnelles-il-nous-faut-inventer/

Messaud, A. (2018, octobre 9). *Régulation des contenus : Quelles obligations pour les géants du Web ?* La Quadrature du Net. https://www.laquadrature.net/2018/10/09/regulation-des-contenus-quelles-obligations-pour-les-geants-du-web/

Mittelstadt, B. (2017). From individual to group privacy in big data analytics. *Philosophy & Technology*, *30*(4), 475-494.

Montes, R., Sand-Zantman, W., & Valletti, T. (2019). The value of personal information in online markets with endogenous privacy. *Management Science*, *65*(3), 1342-1362.

Nedovic-Budic, Z., & Pinto, J. K. (2001). Organizational (soft) GIS interoperability : Lessons from the US. *International Journal of Applied Earth Observation and Geoinformation*, *3*(3), 290-298.

Ng, A. (2020, février 4). *Teens have figured out how to mess with Instagram's tracking algorithm*. CNET. https://www.cnet.com/news/teens-have-figured-out-how-to-mess-with-instagrams-tracking-algorithm/

Ohm, P. (2009). Broken promises of privacy : Responding to the surprising failure of anonymization. *UCLA l. Rev.*, *57*, 1701.

Olson, J. E. (2003). *Data quality : The accuracy dimension*. Elsevier.

Ostrom, E. (1990). *Governing the commons : The evolution of institutions for collective action*. Cambridge university press.

Pitt, J., & Schaumeier, J. (2012). Provision and Appropriation of Common-Pool Resources without Full Disclosure. In I. Rahwan, W. Wobcke, S. Sen, & T. Sugawara (Éds.), *PRIMA 2012 : Principles and Practice of Multi-Agent Systems* (Vol. 7455, p. 199-213). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-32729-2_14

Price, D. (2000). *Legal and ethical aspects of organ transplantation*. Cambridge University Press.

Price, W., & Nicholson, I. I. (2014). Black-box medicine. *Harv. JL & Tech.*, *28*, 419.

Prufer, J., & Schottmüller, C. (2017). *Competing with big data*.

Purtova, N. (2017). Do Property Rights in Personal Data Make Sense after the Big Data Turn? : Individual Control and Transparency. *Journal of Law and Economic Regulation*, *10*(2).

Robertson, A. (2018, avril 5). *Facebook used to be a democracy—But nobody voted*. The Verge. https://www.theverge.com/2018/4/5/17176834/mark-zuckerberg-facebook-democracy-governance-vote-failure

Roth, A. E. (2007). Repugnance as a Constraint on Markets. *Journal of Economic perspectives*, *21*(3), 37-58.

Royal Society. (2017). *Data management and use : Governance in the 21st century—A British Academy and Royal Society project*.

Ruhaak, A. (2020, Février). *When One Affects Many : The Case For Collective Consent*. Mozilla Foundation. https://foundation.mozilla.org/fr/blog/when-one-affects-many-case-collective-consent/

Saint-Aubin, T. (2012). *Design your privacy : Pour une licence de partage des données personnelles*. InternetActu.net. http://www.internetactu.net/2012/06/22/design-your-privacy-pour-une-licence-de-partage-des-donnees-personnelles/

Sandel, M. J. (2013). Market reasoning as moral reasoning : Why economists should re-engage with political philosophy. *Journal of Economic Perspectives*, *27*(4), 121-140.

Schapranow, M.-P., Brauer, J., & Plattner, H. (2017). The data donation pass : Enabling sovereign control of personal healthcare data. *Proceedings of the 2017 international conference on Health Informatics and Medical Systems (HIMS'17)*.

Seppänen, R., Blomqvist, K., & Sundqvist, S. (2007). Measuring inter-organizational trust—A critical review of the empirical research in 1990–2003. *Industrial marketing management*, *36*(2), 249-265.

Shaw, D. M. (2019). Defining Data Donation After Death : Metadata, Families, Directives, Guardians and the Route to Big Consent. In *The Ethics of Medical Data Donation* (p. 151-159). Springer, Cham.

Shaw, D. M., Gross, J. V., & Erren, T. C. (2016). Data donation after death : A proposal to prevent the waste of medical research data. *EMBO reports*, *17*(1), 14-17.

Sichelman, T. M., & Simon, B. M. (2016). The Pathologies of Data-Generating Patents. *BIG DATA, HEALTH LAW, AND BIOETHICS (I. Glenn Cohen et al., eds., 2017 Forthcoming)*.

Signoret, P. (2019, janvier 15). *Lieusaint contre Waze : Comment une ville française se bat contre les itinéraires intelligents*. Numerama. https://www.numerama.com/tech/455455-lieusaint-contre-waze-comment-une-ville-francaise-se-bat-contre-les-itineraires-intelligents.html

Skatova, A., Ng, E., & Goulding, J. (2014). *Data Donation : Sharing Personal Data for Public Good?* https://doi.org/10.13140/2.1.2567.8405

Sorbie, A. (2016). Conference report : Liminal spaces symposium at IAB 2016 : What does it mean to regulate in the public interest. *SCRIPTed*, *13*, 374.

Sorbie, A. (2019). Medical Data Donation, Consent and the Public Interest After Death : A Gateway to Posthumous Data Use. In *The Ethics of Medical Data Donation* (p. 115-130). Springer, Cham.

Sterckx, S., Rakic, V., Cockbain, J., & Borry, P. (2016). "You hoped we would sleep walk into accepting the collection of our data" : Controversies surrounding the UK care.data scheme and their wider relevance for biomedical research. *Medicine, Health Care and Philosophy*, *19*(2), 177-190. https://doi.org/10.1007/s11019-015-9661-6

Strotbaum, V., Pobiruchin, M., Schreiweis, B., Wiesner, M., & Strahwald, B. (2019). Your data is gold–Data donation for better healthcare? *it-Information Technology*, *61*(5-6), 219-229.

Taddeo, M., & Floridi, L. (2011). The case for e-trust. *Ethics and Information Technology*, *13*(1), 1-3.

Taylor, L., Floridi, L., & van der Sloot, B. (2017). Introduction : A new perspective on privacy. In *Group Privacy* (p. 1-12). Springer.

Taylor, L., van der Sloot, B., & Floridi, L. (2017). Conclusion : What do we know about group privacy? In *Group Privacy* (p. 225-237). Springer.

Tempini, N., & Del Savio, L. (2019). Digital orphans : Data closure and openness in patient-powered networks. *BioSocieties*, *14*(2), 205-227.

Turilli, M., & Floridi, L. (2009). The ethics of information transparency. *Ethics and Information Technology*, *11*(2), 105-112.

Van Asselt, M. B., & Renn, O. (2011). Risk governance. *Journal of risk research*, *14*(4), 431-449.

Van Ypersele, C. (2009). Organ Transplantation : Ethical, Legal and Psychosocial Aspects. Towards a Common European Policy. *Clinical Kidney Journal*, *2*(1), 96-96. https://doi.org/10.1093/ndtplus/sfn170

Vayena, E., Haeusermann, T., Adjekum, A., & Blasimme, A. (2018). Digital health : Meeting the ethical and policy challenges. *Swiss medical weekly*, *148*, w14571.

Wang, R. Y. (1998). A product perspective on total data quality management. *Communications of the ACM*, *41*(2), 58-65.

Whitley, E. A., Kanellopoulou, N., & Kaye, J. (2012). Consent and research governance in biobanks : Evidence from focus groups with medical researchers. *Public Health Genomics*, *15*(5), 232-242.

Wicks, P., Vaughan, T. E., Massagli, M. P., & Heywood, J. (2011). Accelerated clinical discovery using self-reported patient data collected online and a patient-matching algorithm. *Nature Biotechnology*, *29*(5), 411-414. https://doi.org/10.1038/nbt.1837

List of figures

Figure 1: Example of economies and costs of data-pooling resulting in no pool formation 14
Figure 2: Example of economies and costs of data-pooling resulting in a single pool with the maximum number of members 15
Figure 3: Example of economies and costs of data-pooling resulting in a pool with a restricted number of agents forming 15
Figure 4: Economies and costs of data-pooling resulting in any number of members being an equilibrium 16

List of tables

Table 2: Types of data-pooling attribution problems 18
Table 3: Payoffs for different pool configurations if all agents have a data endowment equal to 1 25
Table 4: Payoffs for different pool configurations if agents A_1 and A_2 have 10 units of data each and agents A3 and A4 have 5 units of data each 26
Table 5: Payoffs for different pool configurations if agent A_1 has 15 units of data, A_2 has 10, A_3 has 5 and A_4 has 1 27
Table 6: Payoffs for each possible pool configuration with a utility function equal to the natural logarithm and three agents with a data endowment of 10 each 30
Table 7: Payoffs for each possible pool configuration with negative externalities between A2 and A3 31
Table 8: Impact of different features on data-pooling decisions 32
Table 9: Payoffs for each possible pool configuration resulting in a scenario resulting in no data sharing 36
Table 10: Utilities for each member and social welfare given all the possible pool configurations in a scenario with suboptimal pool size and one equilibrium 37
Table 11: Utilities for each member and social welfare given all the possible pool configurations with suboptimal pool size and multiple equilibria 38
Table 12: Payoffs for the non-redundant possible pool configuration in a scenario resulting in multiple equilibria 40

List of equations

Equation 1: General form of the utility function of the basic data-pooling dilemma 11
Equation 2: Utility function of data pooling in a zero-sum game with equally divided benefits 18
Equation 3: Marginal utility of admitting a new member to the data pool in a zero-sum game with equally divided benefits 19
Equation 4: Utility function of data pooling in a zero-sum game with benefit divided proportionally to agents' data endowments 20
Equation 5: Marginal utility of admitting a new member to the data pool in a zero-sum game with benefits divided proportionally to agents' data endowments 20
Equation 6: Utility function of data pooling in a competing pools non-zero-sum game scenario 22
Equation 7: Marginal utility of admitting a new member to the data pool in a competing pools non-zero-sum game scenario 22
Equation 8: General form of the utility equation with intra-pool negative externalities 29

GETTING IN TOUCH WITH THE EU

In person

All over the European Union there are hundreds of Europe Direct information centres. You can find the address of the centre nearest you at: https://europa.eu/european-union/contact_en

On the phone or by email

Europe Direct is a service that answers your questions about the European Union. You can contact this service:

- by freephone: 00 800 6 7 8 9 10 11 (certain operators may charge for these calls),

- at the following standard number: +32 22999696, or

- by electronic mail via: https://europa.eu/european-union/contact_en

FINDING INFORMATION ABOUT THE EU

Online

Information about the European Union in all the official languages of the EU is available on the Europa website at: https://europa.eu/european-union/index_en

EU publications
You can download or order free and priced EU publications from EU Bookshop at: https://publications.europa.eu/en/publications. Multiple copies of free publications may be obtained by contacting Europe Direct or your local information centre (see https://europa.eu/european-union/contact_en).

The European Commission's science and knowledge service

Joint Research Centre

JRC Mission

As the science and knowledge service of the European Commission, the Joint Research Centre's mission is to support EU policies with independent evidence throughout the whole policy cycle.

EU Science Hub
ec.europa.eu/jrc

@EU_ScienceHub

EU Science Hub - Joint Research Centre

EU Science, Research and Innovation

EU Science Hub