Key Ethical Challenges in the European Medical Information Framework

Luciano Floridi[1,2] · Christoph Luetge[3] · Ugo Pagallo[4] · Burkhard Schafer[5] · Peggy Valcke[6] · Effy Vayena[7] · Janet Addison[8] · Nigel Hughes[9] · Nathan Lea[10] · Caroline Sage[11] · Bart Vannieuwenhuyse[9] · Dipak Kalra[12]

Abstract
The European Medical Information Framework (EMIF) project, funded through the IMI programme (Innovative Medicines Initiative Joint Undertaking under Grant Agreement No. 115372), has designed and implemented a federated platform to connect health data from a variety of sources across Europe, to facilitate large scale clinical and life sciences research. It enables approved users to analyse securely multiple, diverse, data via a single portal, thereby mediating research opportunities across a large quantity of research data. EMIF developed a code of practice (ECoP) to ensure the privacy protection of data subjects, protect the interests of data sharing parties, comply with legislation and various organisational policies on data protection, uphold best practices in the protection of personal privacy and information governance, and eventually promote these best practices more widely. EMIF convened an Ethics Advisory Board (EAB), to provide feedback on its approach, platform, and the EcoP. The most important challenges the ECoP team faced were: how to define, control and monitor the purposes (kinds of research) for which federated health data are used; the kinds of organisation that should be permitted to conduct permitted research; and how to monitor this. This manuscript explores those issues, offering the combined insights of the EAB and EMIF core ECoP team. For some issues, a consensus on how to approach them is proposed. For other issues, a singular approach may be premature but the challenges are summarised to help the community to debate the topic further. Arguably, the issues and their analyses have application beyond EMIF, to many research infrastructures connected to health data sources.

Keywords Data ethics · Medical ethics · Ethics of algorithms · Health ethics · GDPR

✉ Luciano Floridi
luciano.floridi@oii.ox.ac.uk

1 Introduction

1.1 The EMIF Project

EMIF, the European Medical Information Framework (2013–2018), was a multidisciplinary research and development project. Its objectives were developing and implementing robust and scalable models to connect health data from a variety of sources across Europe to facilitate large scale clinical and life sciences research.[1] Creation of a common, federated data platform and a governance framework for the identification, assessment and (re)use or repurposing of health data can maximise the scientific research value that can be derived from health data, whilst protecting patient privacy. EMIF is funded[2] through the Innovative Medicines Initiative (IMI),[3] a public private research and development partnership between the European Commission and the European Federation of Pharmaceutical Industries and Associations (EFPIA).[4]

The EMIF Platform provides an efficient integrated information framework for the large-scale re-use of health and life sciences data. The Platform enables data users and data custodians to collaborate throughout the research lifecycle from data discovery to data sharing and data analysis. It enables approved users to analyse securely multiple, diverse, data via a single portal, thereby mediating research opportunities across a large quantity of research data. The EMIF project includes two specific research topics that have helped to guide the development of the Platform: the identification and validation of protective and precipitating factors for conversion to Alzheimer's Disease, and predictors of metabolic complications of obesity. The two clinical research sub-teams have started to publish results from EMIF-supported big data research (for example[5,6]).

The EMIF Platform supports federated analysis, and does not itself hold the research data being queried by its users. A research user transmits an analysis query (in a defined computable form) to multiple connected data sources, which will each execute the query on their own database or on a standardised,[7] mapped, common data model extracted from their original database, and return to the requester only the query results, which might be, for example, as simple as a frequency distribution. Because data originating from multiple sources may have different formats, coding

[1] http://www.emif.eu, viewed 19 January 2018.

[2] Grant number 115372.

[3] The Innovative Medicines Initiative 2016, n.d., viewed 19 January 2018, http://www.imi.europa.eu.

[4] https://www.efpia.eu, viewed 19 January 2018.

[5] Vaudano E, Vannieuwenhuyse B, Van Der Geyten S, van der Lei J, Visser PJ, Streffer J, Ritchie C, McHale D, Lovestone S, Hofmann-Apitius M,Truyen L, Goldman M. Boosting translational research on Alzheimer's disease in Europe: The Innovative Medicine Initiative AD research platform. *Alzheimers Dement. 2015 Sep;11(9):1121-2*.

[6] Roberto G, Leal I, Sattar N, Loomis AK, Avillach P, Egger P, et al. (2016) Identifying Cases of Type 2 Diabetes in Heterogeneous Data Sources: Strategy from the EMIF Project. *PLoS ONE 11(8): e0160648.*

[7] The OMOP Common Data Model https://www.ohdsi.org/data-standardization/the-common-data-model/, viewed 19 January 2018.

and content, the Platform harmonises the data according to accepted semantic standards, in this case via the Observational Medical Outcomes Partnership (OMOP) common data model, to enable the consistent execution of the analysis queries.[8]

This approach, which can be fully or partially automated depending on what is acceptable to each data custodian, enables queries on multiple repositories without having to transfer any subject level data between the parties. In cases where multiple data extracted need more in depth processing (such as linkage), EMIF offers a kind of secure data haven, a Private Remote Research Environment (PRRE), temporarily to host and protect the data. Alternatively, trusted third parties may be contracted to host an extracted research dataset (on a temporary basis) on behalf of a research user and one or more data sources, by mutual agreement.

The EMIF project is transitioning into a sustainable entity post IMI, with a common data platform supportive of European real-world research. Underpinning this is the work conducted on the platform development, and the experience within Alzheimer's and the metabolic complications of obesity, but for a disease agnostic service provision in the future.

Some key learning from within the EMIF project has been facilitated by specific research use cases, and in particular outlining the current processes involved in conducting real world research. This was illustrative of the inherent challenges with reference to administrative overheads, methodological tensions, and time to an answer.

Of importance for this paper, alongside technical considerations for the EMIF platform development, EMIF has developed a practice-based governance framework, the ethical code of practice (ECoP), to assure the local provenance of source data within a federated network. This is also envisaged to provide more wider guidance, via the ECoP, beyond the auspices of the EMIF project, to the European health data research community.

1.2 The EMIF Ethical Code of Practice

Despite the relative separation between research users and subject level data, it was important that the design and operation of the EMIF Platform be governed by the ECoP to protect the interests of all parties. The goals in developing the ECoP were that the EMIF Platform and its services are used in ways that comply with legislation and various organisational policies on data protection, that EMIF upholds best practices in the protection of personal privacy and information governance, and eventually that EMIF could promote best practices in the conduct of clinical research using health data, for the general (public) interest.

Importantly, EMIF needs to ensure compliance with the European Directive 95/46/EC of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data. Given its imminent

[8] Cunningham JA, Van Speybroeck M, Kalra D, Verbeeck R. (2016) Nine Principles of Semantic Harmonization. AMIA Annu Symp Proc 2016:451–459.

enforcement, it is equally important for the ECoP to comply with the new General Data Protection Regulation 2016/679.

EMIF is not alone in seeking to develop an integration environment that provides research access to collections of data sources in acceptable ways. Several European countries are also building such integrated research capability at national levels. However, EMIF has been the largest scale Europe-wide initiative seeking to do this, and its ECoP may be the most advanced work to date on the governance of a federated big data research infrastructure.

In developing the ECoP, several pre-existing codes and policies were examined during 2014–2015 to evaluate whether component parts of these should be adopted by EMIF. The most relevant examples studied were:

- The IMI Code of Practice on secondary use of medical data in scientific research projects.[9]
- The ENCePP Code of Conduct[10] and checklist.[11]
- UK Medical Research Council (MRC) Policy and Guidance on Sharing of Research Data from Population and Patient Studies.[12]
- Yale University Open Data Access (YODA) Project Procedures to Guide External Investigator Access to Clinical Trial Data.[13]
- The EHR4CR Standard Operating Rules[14] and Consent Model and Trust Model.[15]
- SUMMIT (IMI/115006) Principles for Data Sharing.[16]
- ISO 22221: 2006—Good Principles and Practices for a Clinical Data Warehouse.[17]
- International Committee on Harmonization Topic E 6 (R1) Guideline for Good Clinical Practice (Step 2, 2015).[18]
- The World Medical Association Declaration of Helsinki.[19]

[9] Bahr A and SchlünderI, 2015, Code of practice on secondary use of medical data in European scientific research projects. *International Data Privacy Law,* vol. 5(4): 279–291. http://dx.doi.org/10.1093/idpl/ipv018.

[10] Available from http://www.encepp.eu/code_of_conduct/documents/ENCePPCodeofConduct.pdf.

[11] http://www.encepp.eu/code_of_conduct/documents/Annex2_Checklist.pdf.

[12] https://www.mrc.ac.uk/publications/browse/mrc-policy-and-guidance-on-sharing-of-research-data-from-population-and-patient-studies/.

[13] http://yoda.yale.edu/policies-procedures-guide-external-investigator-access-clinical-trial-data.

[14] http://www.i-hd.eu/i-HD/assets/File/EHR4CR/deliverables/115189_EHR4CR_D1_4%20-%20final%20scenarios%2C%20standard%20operating%20rules%20for%20the%20EHR4CR%20-%20COMPLETE.pdf.

[15] http://www.i-hd.eu/i-HD/assets/File/EHR4CR/deliverables/115189_EHR4CR_D9_17%20-%20The%20EHR4CR%20Consent%20and%20Trust%20Model.pdf.

[16] SUrrogate markers for Micro- and Macro-vascular hard endpoints for Innovative diabetes Tools, http://www.imi-summit.eu

[17] https://www.iso.org/standard/40783.html.

[18] https://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E6/E6_R2__Addendum_Step2.pdf.

[19] https://www.wma.net/policies-post/wma-declaration-of-helsinki-ethical-principles-for-medical-research-involving-human-subjects/.

In practice, none were found directly to target the federated model of big health data research, although EMIF does adhere to the IMI Code, which contains many high-level principles that the ECoP could extend with more operational detail. Many of the instruments focused on principles for bilateral data sharing agreements, and primarily on the scientific validity of a proposed investigation (factors such as statistical power) rather than measures to protect data subject privacy.

Successive drafts of the ECoP were developed during 2014–2016, primarily by a core team of academic, healthcare, pharma industry and legal experts, with periodic wider consultation within the consortium comprising many data custodians, research users and patient organisations across Europe. A further layer of consultation occurred through presentations of the evolving work, and key issues at European and international conferences, and contributions to academic publications.[20,21] The ECoP specifies rules for the appropriate conduct of research users, data providers and any intermediate brokers such as EMIF, when undertaking research using big and/or federated health data sets. It focuses on respectful data use and on the protection of data subject privacy. It is still in an advanced draft status, and it has not yet been published.

In order to provide external validation of the ECoP, and to benefit from additional independent expertise outside the consortium, EMIF convened an Ethics Advisory Board (EAB) in 2015. Members include some of the authors of this article: […].

The Board members reviewed and provided feedback on the EMIF project and the role of the Platform, the federated model of providing query access to data custodians, and on specific details in the ECoP. In particular, EMIF sought advice from the EAB members on several key issues for which there appeared as yet to be no clear consensus on good or acceptable practice within the field.

The most important challenges the ECoP team faced were: how to define, control and monitor the purposes (kinds of research) for which federated health data are used; the kinds of organisation that should be permitted to conduct permitted research; and how to monitor this. More specifically:

- how to define publicly acceptable ("bona fide") research—intended for general (public) interest, with specific challenges around the definition of the general interest, in the context of health and health care, and making research results publicly available, either directly, or indirectly via medical products;
- how to define bona fide research organisation, including the acceptability of commercial versus non-commercial organisations, noting that some data custodians have a reluctance to share data with industry or may not share data at all, being limited to providing results of in-house analyses only;

[20] Lea NC, Nicholls J, Dobbs C, Sethi N, Cunningham J, Ainsworth J, Heaven M, Peacock T, Peacock A, Jones KH, et al. Data Safe Havens and Trust: Toward a Common Understanding of Trusted Research Platforms for Governing Secure and Ethical Health Research. Journal of Medical Internet Research 2016;4(2):e22. https://doi.org/10.2196/medinform.5571.

[21] Kalra, D., Stroetmann, V., Sundgren, M., Dupont, D., Schlünder, I., Thienpont, G., Coorevits, P., and De Moor, G. (2016) The European Institute for Innovation through Health Data. Learning Health Systems, https://doi.org/10.1002/lrh2.10008.

- how to monitor and detect misuse of the EMIF Platform and what form of sanctions can be applied, if any;
- how to balance a need for transparency (and to which stakeholders) about how data repositories have been used and by whom, with the (possibly also commercial) sensitivity about research undertaken during product design and development.

The rest of this paper explores those topics, offering the combined insights of the EAB and EMIF core ECoP team. For some issues we have been able to propose a consensus on how the field should approach an issue. For other issues we believe that it is premature to propose a singular approach and so we have summarised the challenges in a structured way to help the community to debate the topic further. The discussion points are expressed in terms of their application to the EMIF context, but we believe that these issues and their analyses can apply to many other research infrastructures connected to health data sources.

2 Constraints and Good Practices on "Commercial Uses" of Health Data

2.1 Defining Bona Fide Research

There appears to be a generally supportive attitude by individuals towards the use of their health data for scientific research. However, the media, patient organisations, and a number of data custodians regularly express concern about the use of health data by commercial organisations, especially when data are collected through public institutions and using public funds. Frequently the public debate tends to focus on whether the organisations using the data are commercial, rather than on the purposes for which the data are to be analysed, and on how the results are to be used (ref. Wellcome study 2016). This problem is exacerbated when data are shared across different national and sub-national research cultures, which often operate in widely diverse value environments. In Germany for instance, more than 60 universities incorporated a "Zivilklausel" in their constitution, also adopted by the Higher Education Law of several German regions, that outlawed all research cooperation with the military sector, including military medical research. Blanket cooperation and data sharing prohibitions are arguably inconceivable in the UK. In this context, EMIF has attempted to define the concept of *bona fide research* as a way of specifying the kinds of research use that society and data custodians are most likely to find acceptable, and to constrain use of EMIF-brokered and EMIF-facilitated research only to *bona fide research* purposes.

Against this background, the authors have concluded that it is necessary and appropriate to define and constrain the use of EMIF services to "socially acceptable"

forms—what we have termed *bona fide research*[22]—in order to articulate the principles on which EMIF operates and subsequently to address public concerns about the use of health data. This approach seems appropriate, provided it does exercise a clearly-defined discrimination on the suitability of organisations to conduct bona fide research. Three comments may further clarify the issue.

First, defining *bona fide research* as precisely as possible can help in complying with the purpose specification principle in data protection law with regard to the distinction between commercial and non-commercial research.

Second, the problem remains whether the focus here should be on institutions and organisations or on their activities. The former approach has been adopted, for example, in the context of the text data mining (TDM) exception (more on this later) by the European Commission in its legislative proposal of September 14th, 2016 on copyright in the Digital Single Market. It seems easier to manage, but it might prevent some socially beneficial activity (e.g., research done by an insurance company to minimise premium discrimination) and allow others that might be less welcome (e.g., universities carrying out market research for its own courses or for an external company). The latter approach has been adopted by the General Data Protection Regulation. This carves out exemptions for scientific, historical, and health research, adopting a broad notion of research that encompasses activities of public and private entities alike (cf. recital 159). This approach is sensible, but it may be more difficult to manage, especially for a large and complex organisation in which bona fide research is only one of its activities, since it would need to ring-fence the knowledge gathered from conducting the research to those parts of the organisation effecting the permitted purposes.

Third, qualifying research as *bona fide*, *socially acceptable*, effectively stating which kind of research one considers appropriate, avoids addressing the question whether EMIF should be equally open to research conducted by commercial for-profit companies and by publicly funded not-for profit organizations.

With this framework and the aforementioned provisions in mind, the *characterization* of *bona fide research* supported and endorsed by all the authors of this article is the following:

> Research qualifies as bona fide whenever its ultimate goal is to discover new knowledge intended for the general interest in health and to be made publicly accessible (e.g., published in scientific journals or disseminated through digital media) without undue delay.

2.2 New Knowledge Intended for the General Interest in Health

This characterisation deserves some comments to be clarified.

[22] For some helpful definitions of *bona fide research* already available please see: http://www.nshd.mrc.ac.uk/data/data-sharing/meta-data-repository/bona-fide-research/.

First, this is indeed a characterisation and not a definition. It does not provide necessary and jointly sufficient conditions, but rather a guideline for an intelligent and sensible handling of personal data.

Second, the characterisation avoids using the more common phrase "intended for the public good" because it is notoriously difficult to determine (let alone predetermine) exactly what is intended by a specific use of health data, and what the public good of it may be. Note that the literature on health data usually refers to "public good of health knowledge" not just the "public good". Furthermore, as it is well known, in economics, "public good" is a technical term to refer to a good that is both non-excludable (individuals cannot be effectively excluded from its use) and non-rivalrous (use by one individual does not reduce availability to other individuals). This is not the sense in which "public good" should be understood in the context of EMIF-related research, hence the preference for the use of "general interest" instead. This is still a technical expression, but one which is more common in legal contexts, and of which authoritative interpretations exist by the highest courts in Europe (e.g., CJEU and ECHR). In the same vein, the GDPR uses the term "public interest" to qualify tasks the performance of which may require certain forms of data processing without consent or to delineate the scope of "archiving purposes" for which specific exemptions apply (e.g., Art. 6, al.1 (e); Art. 9, al.2 (i) and (j); Art. 14, al.5 (b), etc.).

Third, referring to the "general interest" as the ultimate goal that should orient any *bona fide research* has the further advantage of focusing on the purposes for which the data are being used (although such a notion of purpose should not be linked necessarily with the purpose specification principle occurring in EU data protection law). Such a purpose-centred approach enables one to exclude all uses of EMIF services that do not intend to attain the general interest, e.g. mere market research.

Finally, the qualification of the goal as "ultimate" may seem redundant but it is actually intended here to allow research that, for example, improves a method, or the understanding of how medical research works (e.g. from a sociological perspective), which then, only in a next step, may lead to, for example, better drugs or a better way to develop medical research (although in both cases one could argue that they are captured by "knowledge" and "public interest" if they are suitably widely interpreted). We are aware that there may be a problem with some kinds of research exemption insofar as this may presuppose some kind of idealisation or typical notion of research that is very much based on basic science or medicine, while in fact there is also sociological, ethical, legal etc. research that might equally benefit from EMIF data (e.g. a meta-study on profiles of people who volunteer for research) and which should not be excluded, even though its "benefits" may be more diffuse and indirect.

The characterisation of *bona fide research* specified above is intended to exclude uses such as market research and intelligence gathering that might be exploited for targeted sales purposes. Recent research in the UK, yet to be published, also highlights public concern about the use of health data that leads to discriminatory (e.g. life insurance) practices or which informs cuts in health services.[23] Data custodians and public attitudes indicate that these are the least

[23] https://www.herc.ac.uk/get-involved/citizens-jury/.

acceptable uses of population health data. Even when faced by strong counter-arguments, usually formulated in terms of creation of wealth or improvement of security measures, we remain in favour of excluding such uses. Admittedly, the definition of *bona fide research* brings about some possible indeterminacy. However, with a clear focus on functions and purposes of the organisation, rather than on its nature (see below 2.3 on the notion of commercial research and profits), it should be possible to deal with such an indeterminacy in ways that are satisfactory both scientifically and ethically. The orienting aim in every decision is to ascertain whether data obtained through EMIF services are used in the general interest and specifically the general interest in health. From this perspective, it is right to exclude some purposes. More specifically, the following broad areas of data uses may easily be problematic:

(1) Market research and intelligence gathering. If data could eventually also be used for unacceptable (or indeed illegal) non-health purposes—such as using the data to make insurance or employment decisions that affect the data subject, e.g. denying employment to people who participated in a study, or increase their health premiums—then such uses should always be excluded explicitly.
(2) Military (or defence) research or applications focused on weapons development. Although attitudes differ quite widely in Europe on the appropriateness for academics to be involved with weapons research, any use of EMIF data should be strictly non-military in all its aspects and derivations. However, it is important to distinguish military weapons-related research from research seeking to derive knowledge to improve the health and care of military personnel, for example if injured in field situations or having been exposed to biological weapons. It needs to be remembered that military research may be medical in nature, such as healthcare innovations that originated in the military environment e.g. in acute trauma surgery, and also in the domain of public health e.g. pandemic epidemiology. Military health systems sometimes care for families of military personnel and veterans.
(3) Algorithms training. This is a growing area of difficult issues and—in light of the potential benefits but also risks involved and the fast-developing technologies at stake—it is likely that additional requirements set by EMIF or the requirements at the users' institution may not be sufficient. In this case too, the ultimate goal of serving the general interest of the public in health remains an essential guide.

The previous three areas should not be interpreted as exhaustive, nor as excluding all non-health research. The latter may still be considered valid research based on health data especially when this concerns studies into the effectiveness of a drug to decide whether it should be included in a health plan. The reasonable concern is that similar studies may end up harming the very people whose data are being used to develop them. And indeed, people should be able to participate in research without having to fear that the outcome may harm them or their group. Nevertheless, as shown by the National Institute for Health and Care Excellence (NICE) in the UK, for example, these types of effectiveness studies

can and should be part of the licensing regime, while safeguarding the patients whose data are being used. The fundamental point remains that, in similar cases, data are used in "socially beneficial" ways, to decide an equitable way to allocate resources and distribute advantages and costs amongst a solidarity group.

2.3 New Knowledge Intended to be Made Publicly Accessible

The incorporation of research results into a product that is then made available (sold) to health systems is a way of translating research results into practice and eventually making them available to people who need it. If this is the case, then EMIF may not always require that the actual results of the analyses performed be published in a scientific outlet. However, making research results public is essential, since *bona fide* research is linked to the idea of general interest. So "productisation" alone, especially in the form of patents, may not always be sufficient to satisfy the characterisation of *bona fide* research. This does not mean that concerns about "productisation" are always valid *prima facie* as such, but it does mean that more often than not mere productisation cannot count as making knowledge publicly accessible—for the simple reason that the knowledge that went into the product cannot always be inferred from the product—and that in such cases publication should be mandatory, at least dissemination in some additional form. On the other hand, the "publication" does not have to be in a peer-reviewed journal. It is a matter of public access to available knowledge. Some form of dissemination stating whether and how results of the analyses were used may suffice. Demanding a form of feedback or public acknowledgement may also be appropriate, as well as the availability of a "public service" for possible future improvements. If "productisation" involves a patent, as it is often the case, then a publication could still be made available, given that a fair balance between public benefit and monopoly is exactly what the patent system tries to achieve. Indeed, the general interest may be better served by both the availability of the product and the sharing of the knowledge, so one should strive to make both available. Of course, there could be exceptions, e.g. if the product is a software programme (e.g., an app for some wearable digital system). As a good example consider that in the BRAINS (a database with brain scans) project, a short report is required on how the data were used, but that is only partly (or to a small degree) for the benefit of the public. The purpose is mostly to prove the value-for-money to funders. Finally, in setting best practice standards one should avoid facilitating the perpetuation of unpublished negative results.

2.4 Legal Compliance is a Prerequisite

As a matter of principle, there is no connection between the existence of data subject consent (e.g. cohort study data versus extracted hospital EHR data) and how EMIF defines or applies *bona fide* rules because consent precedes logically and determines legally any uses of data that EMIF should make possible. A clear example of such law-abiding use of personal data is given by techniques of pseudonymisation. Moreover, since one should take into account the processing of big data, it is possible

that differential privacy-techniques could be useful in this context (Roth and Work 2014). As *bona fide* research is the kind of research that meets ethical requirements, EMIF does not (have to) add anything over and above that. However, since EMIF seeks to protect the interest of all parties connected to its data federation (data custodians and data users), and the reputation of EMIF as a trusted research platform, EMIF's ECoP does place an obligation and expectation on parties involved in a data sharing interaction to verify themselves that the intended research complies with any applicable data subject consent and with any necessary research ethics approvals. This implies that only data that have the appropriate consent and approval can be put into the original source database, and they can only be used for the uses within the existing consents' and approval parameters.

3 Constraints in Relation to Type of Organisations

3.1 Defining Bona Fide Research Organisations

As a further level of assurance to the public and to data custodians, we have supplemented the above restrictions of use for *bona fide* research with a restriction of use only by *bona fide research organisations*, which we have characterised thus:

> any organisation appointed or accredited or funded to undertake bona fide research, and/or which has made public its commitment to adhere to recognised research governance principles.

Non-acceptable kinds of organisation follow as a consequence of the characterisation of *bona fide* research. Rather than a list of such kinds, it is preferable to focus on their functions and their ways of using data. Still, some illustrations of what representative kinds of organisation are not acceptable would be helpful for explanatory and illustrative purposes. The examples may also help to gain public acceptance.

A grey area left unspecified by the characterisation above is represented by "dual nature" organisations, which today would include many commonplace institutions such as public universities. According to the characterisation of what counts as a *bona fide* organisation, it is not a requirement that *bona fide* research is the primary business of that organisation, or that all of the research undertaken by that organisation and that is unrelated to EMIF data is published. It is also not a requirement that the organisation is publicly funded. Pharmaceutical companies are not the only commercial entities that might become EMIF users: medical device manufacturers, the insurance industry, health clubs and private healthcare providers might also be EMIF commercial users conducting *bona fide* research. Conversely, publicly funded bodies such as universities sometimes undertake commercially sponsored research, for example as a consultancy. They also occasionally spin out publicly-funded research into a company to commercially exploit the results. The objective in proposing and using the characterisation above is to apply sensible and informed discernment on the basis of the *function* of the organisation rather than its nature, i.e., whether it is a public body or not, or a not-for-profit or a for-profit organisation. It is crucial to include all and only the "right" (intended) types of organisation. For

example, excluding science journalism trying to scrutinise the way medical research is done should be seen as an unwelcome form of censorship.

The definition provides the right approach, as it is, at least for the time being, sufficiently clear and broad. However, given the ramifications of GDPR, *bona fide* organisations must have a Data Protection Officer, who is overlooking the responsible processing and management of personal data and, at least in certain cases, the equivalent of an Ethics Advisory Board or at least access to such, which may be an external Ethics Committee.

3.2 The Discussion on Commercial v. Non-commercial, Non-profit v. for-Profit

Given the previous clarifications, it may be tempting to emphasise and use—as the basis for regulating access to data sources connected through EMIF services—the previous two characterisations of *bona fide* research and *bona fide* research organisations as if they simply overlapped with the distinction between commercial vs. non-commercial research and non-profit vs. for-profit organisations. This overlap may even be defensible to public scrutiny and public opinion. But it would be too simplistic and in the end incorrect. The criterion of commercial versus non-commercial research, or for-profit versus non-profit organisations, is actually not very helpful or pertinent, as also research carried out for commercial reasons, or by commercial organisations, can be (and usually is) very beneficial for society as a whole. A very similar discussion took place in the context of the 2014 expert group report on standardisation in the field of text and data mining (Hargreaves et al. 2014), where the question arose whether a copyright exception should be introduced for text and data mining (hereafter 'TDM') research purposes, and if so, whether this exception should cover only non-commercial or also commercial research. The conclusion of the expert group was that such distinction would slow down innovation and be very difficult in practice (p. 67: "Moreover, as we have argued in the economics section of this report, it does not make sense from a strictly economic point of view to distinguish between the commercial and the non-commercial. […] A TDM exception applying to all scientific researchers, commercial and non-commercial, would avoid most of these problems and would represent a huge improvement on the status quo."). The text of the proposed Directive on Copyright in the Digital Single Market (European Commission, COM (2016) 593) is still under discussion, but the European Parliament (European Parliament, JURI Draft Report, March 2017) and the Council seem to endorse the definition of "research organisation" suggested by the Commission (Council Presidency Compromise, September 2017). Under that definition, only entities operating on a non-for-profit basis, or reinvesting all the profits in their scientific research, or acting pursuant to a public interest mission recognised by a Member State, qualify for the TDM exception (together with cultural heritage institutions). Although recital 10 explains that these organisations also benefit from the exception when they engage in public–private partnerships, the concept of "research organisation" is more limited than what we envisage in the definition of "*bona fide* research organisation" for EMIF. Nevertheless, the proposed Directive does not explicitly limit the exception to TDM for non-commercial research

(contrary to, for instance, the TDM exception in Section 29 of the UK Copyright, Designs and Patents Act 1988, introduced in 2014).

In our view, the distinction between commercial and non-commercial research is not the pivotal issue, given the purpose-centred approach indicated above, nor does the general for-profit nature of the organisation matter much, if the ultimate goal for which EMIF data are being used remains *bona fide* research as characterised above. For example, a company may develop bona fide research for the sake of improved public relations, to enhance their reputation, or in view of commercial benefits that may occur only indirectly.

3.3 Benefit Sharing and Exclusivity

Indeed, two related considerations are more crucial. One is *benefit sharing* (BeSh). The public opposes commercial interests when organisations appear to be exploiting public funds or infrastructures (including data in this case), and when the BeSh arrangement is not fair. That is why it is important to demand a fair BeSh scheme if commercial and for-profit entities have access. Some of the research may already have benefit sharing obligations under the national legal framework under which it was conducted. In these cases, it would be up to the parties to ensure that these obligations are fulfilled and carried over into any collaboration that comes out of their EMIF engagement. Furthermore, it may be preferable to encourage such an approach and add a clause that recommends parties to consider appropriate BeSh arrangements (Vayena and Tasioulas 2016).

The other issue is *exclusivity*. Inevitably, as parties move towards commercialisation or productisation, a degree of exclusivity (e.g. through a patent) may be required or demanded. This may be problematic. Consider a partner making data initially available through EMIF, then brokering a connection that could lead to a product, but with the condition that the original study/data is no longer made available to other researchers or EMIF partners (this could be in order to gain a time advantage e.g. if more than one team work in that direction). In this case, we recommend, as a solution, a "perpetuity" clause stating that, once research data have been made available through EMIF, they should be accessible to all authorised research users irrespective of any specific research in progress or undertaken. The data should only be withdrawn from access if the grounds for making them available have changed (e.g. if an ethical approval or consent is reversed).

4 Oversight of Bona Fide Constraints

4.1 Auditing and Enforcing these Bona Fide Constraints

If the reputation of EMIF, the trust of data custodians, as well as the trust of the authorities and the public, substantially hinge upon access to EMIF services that is limited to *bona fide* research organisations and for *bona fide* research purposes, EMIF will be expected to assure itself and others that these conditions are met. In

principle, data sharing requests should be supported by approval of an ethical oversight body (at the requesting institution or obtained by the data custodian prior to undertaking the research investigation). Beyond that, however, the question still remains as to the further responsibility of EMIF. Some tensions arise here partly due to whether EMIF has an obligation to monitor and investigate the internal activities of EMIF users (such as pharmaceutical companies), and partly due to the feasibility of such an obligation given cost issues, since large scale monitoring will be expensive.

Self-declaration is insufficient to enforce *bona fide* constraints. Something beyond pure self-declaration should be implemented, to avoid loss of credibility. Our view is that EMIF should have dedicated staff for this screening. A 'light' and complementary solution would be to set up a notice-and-action system, whereby EMIF users themselves can signal "inappropriate" organisations if they become aware of them (cf. flagging systems commonly used on social media). An intermediate solution, also compatible with the work of dedicated staff, would be to allow EMIF users to 'rate' other users on the basis of the interactions they had with each other (cf. rating systems of auctioning websites like eBay). However, given that research groups may be in competition, peer rating might be motivated by interests that are not visible to EMIF. In both cases, signalling and rating would need to be independently assessed. In addition to peer mechanisms, data protection authorities and their strengthened powers on the basis of GDPR will play a major role in this context. Finally, there should also be obligations on the other end, that is, an appeals process against the decision, to avoid the danger of inadvertently creating unfair market advantage or monopoly to the detriment of an organisation that is "substantially similar" to one that has been allowed to participate.

It is not realistic to try to prevent a large organisation from sharing direct access to a research dataset, or indirectly sharing the research results, with departments and staff that are not conducting *bona fide* research, such as marketing departments. One should simply require that this is not done by contract when signing up to be an EMIF user, or perhaps by means of a regular self-declaration. One may then require periodically some form of evidence that the research results have only been used for permitted purposes. As a general strategy, one should regard the organisation as a whole as doing *bona fide* or *non-bona fide* research with EMIF data, otherwise the implementation would become unrealistic. It does not seem feasible to screen every contract, but it may be important to add a clause to the Terms and Conditions and the contract for use of the EMIF Platform, as just indicated. Note that the focus remains on the nature of the specific activity, not on the nature of the organisation (although the two are easily connected, it is the priority given to the former that matters, see above). The sanction would be the (temporary or permanent) exclusion of the whole organisation from having access to EMIF. Adding a clause to the effect that the data cannot be used for anything else than the intended purpose would mean that, if something goes wrong, the wrongdoers will be found to have violated the contract. Finally, in order to enforce *bona fide* constraints, terms and conditions of the EMIF platform shall make reference to GDPR provisions, such as Art. 40 on Codes of Practice and Art. 42 on certifications. In addition, Art. 55 on the powers of supervisory authorities should be taken into account. Whether or not EMIF can fully

police enforcement mechanisms, this activity will be complemented by the enforcement mechanisms of GDPR. On this basis, one can imagine further safeguards via external auditors, ethics certificates, and so forth.

Even if there is no routine provision of evidence, EMIF should retain powers to investigate any concerns about the use made of EMIF services by one of its user organisations, such as the authority to appoint an external auditor and a requirement that the organisation support (and pay for) the investigations of such an auditor. Such power is crucial. To cite an example: within the ethics management system of the Bavarian construction industry, a company has to undergo an audit every 3 years and pay for it in order to renew their ethics certificate. This is essential.

A range of sanctions can be applied by EMIF to a research organisation found to be in breach of its obligation only to conduct *bona fide* research. One may consider suspending temporarily or permanently the license to use EMIF services. One may also envisage publishing a list of usage breaches, partly for transparency to the public, partly to act as a deterrent, and partly to 'name and shame'. The problem is both legal and ethical. From an ethical point of view, EMIF should adopt 'closed' sanctions, such as a warning, or a temporary suspension—and in case of serious and repeated breaches, a permanent revoking—of the license to use EMIF services. However, from a legal point of view, 'naming and shaming' can be problematic, because issue of legal liability for reputational damage may arise. One should only 'name and shame' if one proceeds in accordance with fair trial requirements (the party should be given a reasonable period to react and restore the breach, the decision should be open for appeal, etc.). We advocate in the Terms and Conditions of the EMIF Platform the possibility of some kind of mediation for serious cases whereby EMIF users accept the authority of an existing or ad hoc panel of mediators and are willing to bear the costs. To give an analogy, the ethical 'sanctions' would be comparable to the peer review system, which is good at many issues but less good at spotting fraud and ultimately relies on the honesty of the partners. It is preferable to acknowledge clearly this reliance than hide it behind grand-sounding but unenforceable provisions. Of course, legal sanctions remain a possibility but they are a different issue (e.g. enforcement of current legal rules on data protection). A question to be addressed is how transparent the ethical sanctions should be. At minimum EMIF could make public (e.g. via its web site) the action it is taking when a breach is identified, without naming the wrongdoer. This would have the further benefit of signalling that unethical uses will be identified and sanctioned.

4.2 Transparency of Access Requests

A final question that arises, is whether data custodians should be informed about every request for access to data they made available through EMIF, and if so, with which level of detail. In this context, a fair balance needs to be struck between commercial sensitivity arguments invoked by data users, who may wish to conduct research confidentially (even if the results may later be published), and

data custodians who may need to confirm adherence to any constraints they have placed on the permitted uses of their data.

One scenario, favoured by some of the EMIF data custodians we have consulted, is that an audit report extracted from the query log is always accessible to authorised individuals within the data custodian organisation. This would allow inspection of which organisations have executed queries each day, on which categories of their data and with what parameters, and if any disclosure controls were applied. Since this would reveal the research areas being investigated by each research user, this access would need to be governed by a confidentiality agreement, with penalties for breach along the same lines as those described for research users above.

A second scenario, that might be acceptable to some data custodians, is for a filtered overview of analysis activity to be provided regularly to data custodians, with a more detailed audit log extract only produced if a concern is raised. Such a filtered overview report can be defended as the more usable approach. Because of this need for filtering, the system should make it technically possible to indicate which constraints are attached to uses of certain datasets (otherwise the data custodian will run the legal risk of non-compliance). Online repositories such as SSRN, for example, frequently make public user statistics like "this paper has been seen/downloaded/cited X times" etc. Something similar could be envisaged. It would be sufficient to indicate to users that there is interest, and maybe patterns in that interest, as opposed to having detailed and explicit information about who is interested in what. In short, a balance should be struck between the commercial sensitivity of some data users and the protection of data custodians. With a caveat: transparency is not the one-size-fits-all solution. For example, in addition to a trusted third-party audit (EMIF positions itself as a trusted third-party to audit query activity on behalf of the data custodians, instead of allowing the data custodians direct insight into the queries being run on their data), it is possible that some technical solutions will help us in this context, such as zero-knowledge proofs.

This is an area where there is a need for wider consultation with data custodians and research users on the appropriate level of activity transparency and the mechanisms for protecting commercial sensitivity, to balance the legitimate and understandable interests on both sides.

Acknowledgements The research leading to these results has received support from the Innovative Medicines Initiative Joint Undertaking under EMIF Grant Agreement No. 115372, resources of which are composed of financial contribution from the European Union's Seventh Framework Programme (FP7/2007–2013) and EFPIA companies' in kind contribution.

References

European Commission, Proposal for a Directive of the European Parliament and of the Council on copyright in the Digital Single Market—COM(2016)593, September 14, 2016.

Hargreaves, I., Lucie, G., Christian, H., Peggy Valcke & Bertin, M. 2014. Standardisation in the area of innovation and technological development, notably in the field of Text and Data Mining—Report from the Expert Group, European Commission—DG RESEARCH, 2014. http://ec.europa.eu/research/innovation-union/pdf/TDM-report_from_the_expert_group-042014.pdf.

Roth, Aaron, & Work, Cynthia. (2014). The algorithmic foundations of differential privacy. *Foundation and Trends in Theoretical Computer Science, 9*(3–4), 211–407.

Vayena, E., & Tasioulas, J. 2016. The dynamics of big data and human rights: The case of scientific research. Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences 2016 Dec 28;374(2083). pii: 20160129. https://doi.org/10.1098/rsta.2016.0129.

Affiliations

Luciano Floridi[1,2] · Christoph Luetge[3] · Ugo Pagallo[4] · Burkhard Schafer[5] · Peggy Valcke[6] · Effy Vayena[7] · Janet Addison[8] · Nigel Hughes[9] · Nathan Lea[10] · Caroline Sage[11] · Bart Vannieuwenhuyse[9] · Dipak Kalra[12]

Effy Vayena
effy.vayena@hest.ethz.ch

[1] Oxford Internet Institute, University of Oxford, 1 St Giles, Oxford OX1 3JS, UK

[2] The Alan Turing Institute, British Library, 96 Euston Road, London NW1 2DB, UK

[3] TUM School of Governance, Technical University of Munich, Arcisstrasse 21, 80333 Munich, Germany

[4] Law School, University of Turin, Lungo Dora Siena 100, 10153 Turin, Italy

[5] School of Law, University of Edinburgh, David Hume Tower, George Square, Edinburgh EH8 9JX, Scotland, UK

[6] IMEC-CiTiP-KU Leuven, Sint-Michielsstraat 6, 3000 Louvain, Belgium

[7] Health Ethics and Policy Lab, Department of Health Sciences and Technology, Swiss Federal Institute of Technology, ETH Zurich, Auf der Mauer, 17, Zurich, Switzerland

[8] Biogen Idec Limited, Innovation House, 70 Norden Rd, Maidenhead SL6 4AY, UK

[9] Janssen Research and Development, Beerse, Belgium

[10] Institute of Health Informatics, University College London, London, UK

[11] CMAST bvba, Georges Van Dammeplein 57, 9140 Temse, Belgium

[12] The European Institute for Innovation Through Health Data, Building 3, 5th Floor, De Pintelaan 185, 9000 Ghent, Belgium