Protecting Data Generated in Medical Research: Aspects of Data Protection and Intellectual Property Rights

Iryna Lishchuk, Marc Stauch
Institut für Rechtsinformatik
Leibniz Universität Hannover
Hannover, Germany
e-mail: lishchuk@iri.uni-hannover.de, stauch@iri.uni-hannover.de

***Abstract*—This paper investigates legal approaches towards protecting the data generated in medical research. One of the core features of the rules for the processing and sharing of data generated in medical research is their complexity. Thus, data containing personally identifiable information would qualify as personal data and the processing of such data would be subject to the law on data protection. Equally, the generation of data in the course of medical research may involve considerable investment or effort and have an economic or scientific value for the researcher or right holder, including through use in publications, and may well be considered as Intellectual Property (IP). Contractual approaches may also define the rules how the data may be used and shared.**

Keywords-IP rights; data rights; medical data; data curation; personal data; data protection.

I. INTRODUCTION

IT developments in the field of bioinformatics have opened new ways of data procession. The creation of new data, as well as new knowledge, derived out of existing datasets as a result of medical research, may well be considered as an IP and qualify as an object of protection by IP rights [1]. Equally, data generated in medical research, which by one or another parameter may be related to an identifiable natural person, also have the quality of personal data with the resulting protection by the law on data protection.

Innovative genome sequencing techniques are able to process 4 PB data per year (11 TBytes per day), thus reaching the level of Twitter with the processing power of 12 Terabytes per day [2]. Mathematical and computational modeling is used to integrate and interpret the massive amount of data, uncovered in molecular and cell biology [3]. Cancer system biology, which studies how individual components interact to give rise to the function and behavior of the cancerous system as a whole [4], produces a number of data types: molecular data, epigenetic data, clinical data, imaging data, pathology data and other laboratory data.

In the process, the availability of a large amount of data collected in the clinical trials combined with modern data processing techniques have allowed the discovery of new data correlations. For instance, the SIOP 2001/GPOH trial of patients with Nephroblastoma (a malignant tumor arising from the embryonic kidney that occurs in young children, especially in the age range 3–8 years [5]) revealed that whereas 90 % of patients respond to preoperative chemotherapy with tumor shrinkage, in about 10 % the tumor does not shrink, but increases in return, thus making the situation worse [6]. Such discoveries necessitate in-depth research and application of powerful data analytics techniques to identify correlations between negative tumor response and specific characteristics of the non-responding patients.

Thus, advances in data-mining and analytics have made it possible to generate new data and derive new knowledge from existing datasets. This, as well as new methods of differentiating and capturing biological phenomena (including at the micro-level) has led to an exponential growth in available medical data. In principle, such data, recorded in patient or research databases can be of tremendous value when analyzed, in revealing linkages, e.g., between environmental and/or genetic factors and diseases, as well as for comparing patient responses to different treatment therapies. A major advantage too is that such connections can often be identified straight from the records, without the need for further invasive and potentially risky research.

At the same time, as the potential value of health data becomes better understood, efforts to monopolize clinical data by exclusive IP or proprietary rights are also expanding. Copyrights, patent rights, sui generis database rights and the legal regime of undisclosed information may come into consideration, depending, however, on the data – the subject matter of protection. For instance, there are cases when the commercial use of health related data has been asserted under the coverage of database rights [7]. Patentable inventions have also been derived out of the biological material and associated data of the patients and successfully commercialized [8]. The property rights in medical research data may also be claimed under contractual schemes [9]. At some point copyrights may also come to consideration for monopolizing data in medical domain [10].

However, as a precondition for allowing a significant amount of clinical data to be usefully exploited, there is an important initial step required in the form of data curation. In this regard, as we analyze below, most types of IP protection are tailored to protect specific objects that have already passed a certain threshold of maturity (data repositories, confidential information with assignable commercial value, etc.); but, as we discuss, none as such guarantees adequate protection to protect the prior investment made in curating the data.

In what follows, we begin by describing the data curation process in medical research in Section II, explore the complex nature of medical data in terms of law in Section III, proceed to the requirements of data protection for the processing of personal data in Section IV. In Section V, we investigate the potential options of protecting the medical research data by IP

rights and in Section VI then consider their application in the context of a concrete research initiative, namely the EU FP7 project 'CHIC'. Thereafter, contractual approaches towards the government of rights in data are examined in Section VII, before Section VIII concludes by suggesting a potentially more effective approach to protecting researcher investment in curation.

II. Data Curation

The clinical data provided for e-health research usually comprises a large mass of data of multiple data types, formats, words, figures, numerical parameters, abbreviations, etc. Furthermore, even where data is of the same underlying type, this will often have been recorded in different ways – using different clinical concepts and/or measuring systems. This reflects the decentralized, autonomous nature of health care delivery, with different institutions and clinicians often employing different classificatory descriptions and/or record systems.

Data integration is key here, but the format, scope, parameter, structure, context, terminology, completeness, etc., of the individual and heterogeneous data are not standardized, which may affect their quality, and ultimately their interoperability and integration [11]. This could also potentially affect collaboration of the different researchers in this field if they use different semantics and techniques to describe, format, submit, and exchange data.

From a technical standpoint, data integration is still a significant challenge. The curation required to ensure the data relates to and measures the same phenomena with sufficient accuracy to be usable is a large and painstaking task. It includes the problem of dealing with incomplete data fields and cross-checking that various indices were measured and recorded in a similar way (e.g., images were taken using similar equipment, co-morbidities were classified using the same terminology, etc.). In the process the curator may often wish to add metadata to alert the data user to such issues. It is evident too that considerable expertise and skill is required for the task to be performed well: the curator needs to have a real feel and understanding for the subject matter in order to make sensible judgments in resolving various gaps and uncertainties.

In this regard, a starting point in the context of curation may be to see raw data in terms of the 'given', which as yet lacks semantic meaning, with the latter only emerging through the addition of an interpretive context (which also marks the change in state from data into information). It is suggested that the technological development and transformation of raw or incompletely processed data into information (or the uncovering of additional semantic meaning), brought about by the curative process represents a suitable object for IP protection.

At this point a legal challenge arises. On the one hand, an intellectual and/or technical investment made in curating the data and generating new data outcomes may justify an interest of the investors in monopolizing the resulting data as their IP. On the other hand, the data used in medical research originally comes from the patient, which renders such data a potential candidate for protection as personal data. That is so, if the medical data contain personally identifiable information, i.e., it may by some or the other characteristics be linked to the data subject.

Against this background, both the economic value of the derived data and the tentative quality of the data as personal data make the data generated in medical research a complex object of legal protection and dictate the type of protection applicable.

III. Complexity of Medical Data in terms of Law

The legal complexity of the data generated in medical research is one of the major factors, which determine the type of protection applicable and the rules governing the use of such data. The medical research data may qualify both as personal data and intellectual property.

Indeed, out of scientific disciplines, medical research (both as sociological research) tends to share significantly less data than others (65% in comparison to 90% in biology or 85% in climatology) [12]. Frequently, this "*low data sharing culture*" is justified by the legal and ethical requirement to protect the privacy of individuals, that is to say data protection [12].

On the other hand, as noted, even where medical data is void of indices, which would render such data personal data in the meaning of data protection law, the aspects of intellectual property also need to be taken into account. If the researcher or research institution, who holds such data in its legitimate possession, considers such data as its "intellectual property" and has an economic interest in exploiting such data for individual gain (e.g., reputation, scientific publication), such qualification of the data may also affect data sharing and determine the circumstances for such data to be shared. It is common in the scientific world that "*Data that a researcher feels could still be exploited for future publications are usually not shared*" [13]. Another practice usual for medical sciences is that the data is no longer protected after the appearance of publications [14]. The legitimate interests of the data holder may also affect the terms and circumstances for such data to be shared. For instance, such data may only be made available to the circles, which may prove a justified scientific interest in the data (e.g., data sharing upon certain conditions inside a research consortium or a limited medical community) [14].

What may also play a role is whether a medical project relates to Big Science, such as physics, Earth and climate science, or Small Science, in particular, small experiments, narrow disciplines [15]. For Big Science data there are "government controlled repositories", which normally govern the use of data as a "public good" [15]. An example is Clinical Trials Registries and Databases, such as registries operated by the National Library of Medicine in the USA [16], the UK Current Controlled Trials [17] and the Japan Pharmaceutical Information Center [18]. However, for Small Science projects, which comprise the majority of data repositories, such pre-determined regulatory frameworks for the handling of data do not exist. The protection practices applied vary

from discipline to discipline and have rather an informal character [15].

In the light of these considerations, for the purposes of choosing and applying adequate protection measures, it is relevant first to ascertain whether the data has a quality of personal data (and is subject to the requirements of the law on data protection); the next questions are whether it has an economic value for the data holder and may be treated as intellectual property (subject to the rules of IP law), or whether such data is considered as a "public good" and must be treated as such.

IV. Data Protection

For legal purposes, the first important question to decide is whether the medical research data contain personally identifiable information. In the meaning of European data protection law, it would be the case if by some or the other characteristics the data may be linked to the data subject. If so, the processing and sharing of such data would be subject to the law on data protection.

Article 2 (a) of the EU Data Protection Directive 95/46/EC (DPD) [19] (which is to be superseded by the General Data Protection Regulation [20] by 25 of May 2018) defines personal data as follows:

"*'personal data ' shall mean any information relating to an identified or identifiable natural person ('data subject'); an identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity;"*.

As is apparent, this is a wide definition, and in principle it may certainly cover some medical data. An example may be a brain image that also shows some of the patient's face; indeed, in the light of modern software, a set of cross-sectional brain images may also qualify – this is if it would be possible with the software to put such images together to reconstruct the face of the patient based on the image parameters. Since health data qualifies as sensitive data [21], the processing of such data is subject to stringent requirements of procession (Article 8 DPD) and must be explicitly legitimized, e.g., by informed consent of the patient (Article 8 (2) (a) DPD) or the national laws that also provide for adequate privacy safeguards (Article 8 (4) DPD) [19].

Medical research is usually conducted on the human body or with the use of clinical data. Blood samples, serum, tissue samples, cells usually constitute material for laboratory examinations, from which the data, used in medical research, is derived. When the laboratory tests are taken in course of medical treatment and/or diagnosis, the patient normally consents to the use of the excised material and data for the purposes of clinical care [9]. However, as a rule such consent does not extend and does not entitle the physician to use such clinical data for research [8]. The use of clinical data for research requires legal justification, which as a rule may be obtained either by informed consent of the data subject or by compatible use of data.

The use of previously collected data for research constitutes secondary use of data. In principle, Article 6 (b) DPD allows secondary use of data subject to specific conditions: "*personal data must be collected for 'specified, explicit and legitimate' purposes (purpose specification) and not be 'further processed in a way incompatible' with those purposes (compatible use).*" [19].

By implication, the compatibility assessment is to be made on a case-by-case basis and in consideration of all relevant circumstances. In particular, the following key factors shall be taken into account:

- "*the relationship between the purposes for which the personal data have been collected and the purposes of further processing;*
- *the context in which the personal data have been collected and the reasonable expectations of the data subjects as to their further use;*
- *the nature of the personal data and the impact of the further processing on the data subjects;*
- *the safeguards adopted by the controller to ensure fair processing and to prevent any undue impact on the data subjects."* [22].

The use of data for scientific research withstands the compatibility assessment as long as the controller implements "appropriate safeguards" and by that ensures "*that the data will not be used to support measures or decisions regarding any particular individuals*" [22]. Such safeguards may be taken in the form of technical and/or organizational measures aimed to ensure functional separation (such as partial or full anonymisation, pseudonymisation, and aggregation of data), privacy enhancing technologies, as well as other measures to prevent the use of data to take decisions or other actions with respect to individuals [22].

From these legal observations it follows that - in simple terms - the use of health data for research must be legitimized: either by the patient´s informed consent or by the law, allowing compatible use of data subject to compatibility assessment and application of appropriate de-identification and security measures. This is also likely to remain the position after the General Data Protection Regulation (replacing the DPD) comes into effect in May 2018 [20]. In such cases, the research conducted subject to adoption of appropriate de-identification and security measures should not cause privacy implications.

It is apparent that by imposing these requirements, the law on data protection aims to protect and safeguard privacy of the individual. "*Data protection rules may be seen as embodying and safeguarding core ethical principles of autonomy, dignity and privacy; they are about making sure that persons remain able to decide how their data will be used and are not exploited or instrumentalised through opaque data processing practices*;" [23]. These matters are essential in order for patients to have trust in medical research and innovative eHealth applications [23].

However, when talking about protecting medical research data it is essential to distinguish the primary goal of such protection. In this respect it must be noted that the purpose and meaning of the law on data protection is to protect

privacy of the individual, and not to do with the economic or exploitation interests in the data itself. Therefore, when legal protection is sought to protect economic interests of the data holder, the law on data protection would not fulfill that objective. The requirements of the law on data protection must rather be taken into account as a necessary means of protecting privacy and rights of the data subjects.

V. Potential IP Protection

In contrast to the law on data protection, which serves to protect privacy rights of the individuals, the IP law aims to reward and protect the creators - either authors or inventors – for their intellectual or economic input into society.

A. Data as Protectable Subject Matter

When we consider the data produced in medical research, such as measurements, experiments, outcomes of data analytics, etc., in the context of IP law, we can observe that, as a rule, such data do not automatically fall into the category of IP protected objects. In the absence of legal protection applicable directly, alternative protection mechanisms are frequently sought, such as: copyrights, sui generis database rights under IP law; or through the application of the legal regime of undisclosed information, an aspect of competition law. In addition, contractual mechanisms may be used to address proprietary interests in data. However, the application of these forms of protection may often be problematic. For instance, copyright may not arise in the absence of creative input or proprietary claims in data may be challenged due to the questionable legal nature of property rights in data [24]. More generally, IP law would normally not protect the data as such. Instead, a requirement for IP protection is that added value produced from the data. Examples may be a creative scientific work covered by copyright, an industrially applicable invention in the patent law or commercial value of the information protectable as know-how by competition laws.

This also fits with the underlying motivation for IP protection, which is to motivate an author or inventor, by rewarding them for their intellectual activity (here, in extracting value from the data). In contrast, raw data, which is void of such intellectual input does not constitute a protectable IP and as a matter of policy should be kept free for public use.

The applicability of the IP laws in relation to medical research data is considered in more detail below.

B. Copyright and Related Rights

Clinical data comes for the most part from clinical trials, laboratory results, medical examinations, etc. An example of the clinical research data is shown in Figure 1 [25]. Such data is usually expressed in some numeric parameters, figures, words, combinations of such items. The representation of clinical data in this format is suitable and useful for digital data processing. However, the isolated items, be they words, keywords, syntax, figures or mathematical concepts as such, will not attract copyright. According to the Court of Justice of European Union (CJEU), items, "*considered in isolation, are not as such an intellectual creation of the author who employs them."* [26]. In order to be protected by copyright, the data must constitute the expression of the original author´s creativity, which is only present when "*through the choice, sequence and combination of those words that the author may express his creativity in an original manner and achieve a result which is an intellectual creation*" [26].

The protection of clinical data by copyright may under circumstances be acceptable for the medical reports, written by the physician or the patient, insofar as the expression of original creativity is achieved [10]. However, for isolated datasets, especially where (as is desirable) the curator follows a standardized procedure, it seems much less likely that sufficient originality exists for copyright purposes.

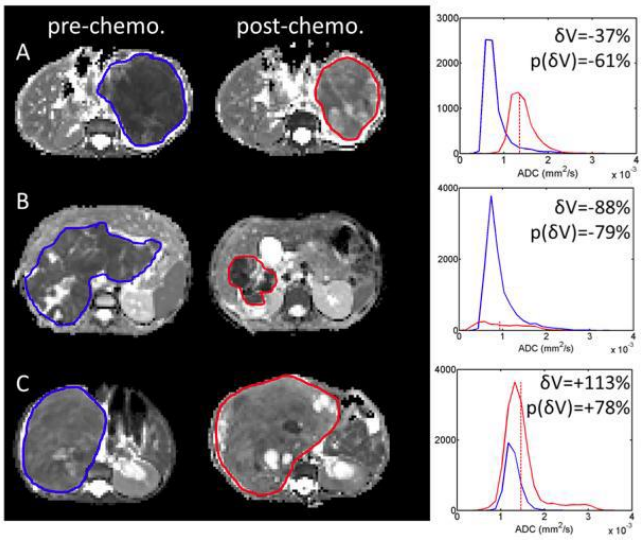

Figure 1. DWI and ADC mapping of nephroblastoma from different patients before and after pre-operative chemotherapy. Presented at the annual meeting of the British Chapter of the ISMRM, September 2012, provided by Prof. Kathy Pritchard-Jones from UCL. Copied from CHIC Deliverable D2-2 "Scenario based user needs and requirements" [25].

As may be seen from the image, some data is presented in visual form and is represented by images. However, medical images are normally produced by technical means (such as X-Ray, Ultrasound, etc.) and lack the creativity – an indispensable pre-requisite for copyright. A similar standard of copyright and requirement of original creativity applies to photographic works as well. According to Recital 16 Directive 2006/116/EC [27], a photographic work is protected by copyright, if it is original. A work "*is to be considered original if it is the author's own intellectual creation reflecting his personality*". Other criteria such as merit or purpose are not relevant for copyright. According to the CJEU decision in the case C 145/10 REC of Eva-Maria Painer [28], copyright protects pictures taken by an individual, exercising free and creative choices, thus stamping a picture with his personal touch. It follows that

only pictures taken by an individual expressing some level of the author´s personality and creativity may be protected by copyright. On the other hand, images, generated automatically, will lack the necessary creativity. Since the images, produced in medical domain, are normally taken automatically and the process of recording is mostly completely managed by technical means, such images normally do not express creativity and do not attract the protection by copyright, respectively.

Apart from the rights considered so far, in the field of copyright *senso strictu*, there are a number of other emerging rights granted as a response to relevant investment. These rights are normally provided to the person, who invests in producing the protectable information. Such rights are referred to as related rights. Protection by related rights does not necessarily link to the intellectual creation (as the case is with traditional copyright), but rather to the economic investment.

The major rationale for protection by related rights tends to shift between intellectual creation and the investment of resources required [29]. A mixture of artistic creation and investment attracts exclusive rights to performers in fixations of their performances. The economic investment constitutes a major factor, which renders exclusive rights to phonogram producers in their phonograms, to the film producers in respect of first fixations of their films, to broadcasting organizations in fixations of their broadcasts [30].

However, the number of related rights as of now is rather limited (mostly to those, indicated above). Therefore, attaching added value to the data enriching, post-processing, modification, etc., does not constitute the kind of investment protectable by related rights.

C. Sui Generis Database Right

As a rule, clinical institutions, participating in medical research, manage and maintain their clinical data in clinical data repositories. Some clinical institutions manage their clinical information and store the results of clinical trials using project-tailored data management systems. For example, the CHIC project utilises an Ontology-based Clinical Trial Management Application (ObTiMA) [31]. Other institutions prefer data management systems specific to their medical activities. Against this background, an option of protecting the clinical data under the umbrella of sui generis database rights comes into consideration first.

The legal protection of databases is provided for by the Directive 96/9/EC of 11 March 1996 on the legal protection of databases (the Database Directive) [32]. Such protection is granted in recognition of the fact that constructing a database requires "*investment of considerable human, technical and financial resources"* [32]. The Directive 96/9/EC aims to reward and protect such investment by providing the maker of a database with a sui generis data base right that places him in a position to prevent unauthorized access and copying of the database contents, which he compiled. In this regard, Article 7 Database Directive states:

"*Member States shall provide for a right for the maker of a database which shows that there has been qualitatively and/or quantitatively a substantial investment in either the obtaining, verification or presentation of the contents to prevent extraction and/or re-utilization of the whole or of a substantial part, evaluated qualitatively and/or quantitatively, of the contents of that database*." The object of protection in terms of the Database Directive is a 'database' meaning "*a collection of independent works, data or other materials arranged in a systematic or methodical way and individually accessible by electronic or other means*" [32].

Databases are given their own *sui generis* right of protection for the *"blood, sweat and tears that go into producing a database"* [33]. Consequently, as we have just seen, the Database Directive demands that *"there has been qualitatively and/or quantitatively a substantial investment in either the obtaining, verification or presentation of the contents"* [32]. The type of investment required can be time, financial resources, personnel, or technical means invested, or indeed any other "sweat of the brow"-type resource, as distinct from creative, intellectual efforts.

The CJEU is very strict in its understanding that the investment must be made to *obtain* the contents. A database that is a mere spinoff/by-product from another investment/activity (such as scientific data resulting from research) does not typically qualify for protection under the Database Directive's *sui generis* regime. There must additionally be a further substantial investment in obtaining, verifying or presenting the data [34].

In other words, the CJEU demands that the investment be made specifically to *"seek out existing independent materials and collect them in the database"* [35]. An investment in *"the creation of materials which make up the contents of a database"* [35] is deemed insufficient. As a result, creators of data rarely enjoy a sui generis right of protection for any non-original database constructed out of that data – so-called "single source databases" [36] – unless there is also a substantial investment in the verification or presentation of the contents.

"*Verification*" is understood to mean steps taken to ensure the information is reliable. As with the requirement of "obtaining" data, an investment in verifying information during the information's creation is excluded [34].

"*Presentation*" is defined as the way data is structured and made accessible to others, so that the creation of an index or the design of a user interface can all be seen to fulfill the requirements of an investment in the presentation of the contents [34].

Finally, the investment must also be of a "*qualitatively and/or quantitatively*" substantial nature [32]. The Database Directive does not define "*substantial*" and neither has the ECJ ruled on the matter. However, the Preamble of the Directive indicates that, "*as a rule, the compilation of several recordings of musical performances on a CD (...) does not represent a substantial enough investment to be eligible*

under the sui generis right" [32]. Member States generally adopt a low level approach to the requirement, and the Advocate General has taken the same stance [34].

As regards the quantitative and/or qualitative qualification, these are understood to mean investments quantifiable and not-quantifiable, respectively, such as money on the one hand and intellectual effort on the other [37].

In fact, additional substantial investment is often present in the case of data resulting from clinical trials. Such data must first undergo an extensive verification process before it can be used in research and entered into a database. Importantly, the data verification process is subsequent and separate from the obtaining/creation of the original data, as otherwise it would be excluded from protection.

Accordingly, protection by the sui generis right can be considered as a plausible option for clinical data repositories, provided the given repository satisfies the above criteria. As regards the scope of the database right, it would protect the collected data from being copied as a whole or in substantial part, evaluated "*qualitatively and/or quantitatively*" and either copied in one action or step by step [32].

Provided the clinical data repository qualifies as a database in the meaning of Database Directive and the clinical institution holds the sui generis database rights, the institution may stipulate the terms of using the repository contents as a whole, grant the rights of use under contractual license, prevent and enforce the unauthorized extraction/reutilization of the repository contents as a whole or in substantial part. The rights holder may thereby leverage how the contents of its repository may be used, whether the data items may be extracted (downloaded) and in what form or quantity, whether the data may be transferred to external parties or whether the data processing may only be done on its premises.

At the same time, this sui generis protection applies to the contents of the repository as a whole or in substantial part and may apply separately and irrespective of protectability of data items by other rights, such as copyrights. Article 7 (4) makes this explicit, saying that the database right: "*shall apply irrespective of eligibility of the contents of that database for protection by copyright or by other rights. Protection of databases [….] shall be without prejudice to rights existing in respect of their contents*".

Thus, the holder of the repository may manage the use of the repository contents as a whole. However, the use of separate data items in the repository may remain governed by the terms, stipulated by the data providers and/or holders of rights in such items. For instance, the access rights to the datasets, handled as confidential, may require signing of non-disclosure agreement (NDA) and the use of such data may be limited and be subject to technical protection measures, etc.

In this regard, we consider further the options of protection, which potentially may apply to separate datasets, next.

D. Know-how

Because of the high sensitivity of health related data (and the potential harm to the patient's interests in privacy, dignity and autonomy from disclosure), clinical data in the medical treatment domain is managed under the rules of professional medical secrecy and subject to fiduciary duties. Similarly, as was discussed in Section IV, the data, so far as individual patients may be identified from it, will be subject to data protection rules. In this regard, a plausible option (fitting well with such privacy-based considerations) for protecting the research investment made in collecting or curating clinical data may be to invoke the legal regime of know-how (or undisclosed information). This is, especially so after such data leaves the medical domain and enters the domain of clinical research (where not necessarily all parties are bound by the rules of professional secrecy).

Protection of undisclosed information is provided by Section 7, Article 39 et seq. TRIPS Agreement [38] and the Directive 2016/943 on the protection of undisclosed know-how (the Trade Secret Directive) [39]. The legal regime of know-how enables natural and legal persons, who are in legitimate possession of valuable information, to prevent such information "*from being disclosed to, acquired by, or used by others without their consent in a manner contrary to honest commercial practices."* [38]. Unfair practices for these purposes would include the acquisition of information via "*unauthorised access to, appropriation of, or copying of any documents, objects, materials, substances or electronic files…. containing the trade secret or from which the trade secret can be deduced*" [39]; violation of contractual duties, breach of confidentiality obligations, inducement to breach, etc. [38].

In order to be protectable, the relevant information should have the quality of protectable subject matter. The Trade Secret Directive, both as Article 39 TRIPS Agreement accord protection to information, which:

"(a) is secret in the sense that it is not, as a body or in the precise configuration and assembly of its components, generally known among or readily accessible to persons within the circles that normally deal with the kind of information in question;

(b) has commercial value because it is secret; and

(c) has been subject to reasonable steps under the circumstances, by the person lawfully in control of the information, to keep it secret." [39].

At the same time, one weak point of protecting clinical data as know-how is that the know-how protection across Europe is not that well harmonized with varying data objects considered as protectable know-how and the laws, which accord such protection, ranging from IP laws to competition laws [40].

The newly adopted Trade Secret Directive is intended to harmonize the national laws in relation to know-how protection and in many aspects repeats the provisions of the TRIPS Agreement: in particular, it relates to the protectable subject matter and requirements for protection (Article 2),

acts of unlawful acquisition, use and disclosure of information (Article 4), availability of legal remedies against the unlawful acquisition, use and disclosure of trade secrets (Article 6 et seq), etc. With respect to protection of medical research data as know-how, it may also be queried how far the Trade Secret Directive would improve the protection for data, the preparation of which consumed much effort, but which for one or another reason may not reach the level of protectable know-how. Here the key obstacles in applying know-how protection to the clinical data, processed for research, relate to the need (in order to be protected) for such data to be secret, subject to confidentiality measures and have economic value.

First, to satisfy the criterion of secrecy, the information, sought to be protected, must be accessible to a limited number of persons only. The use of such information must be subject to confidentiality measures. The application of confidentiality measures means that the data must be stamped as "Confidential" and the sharing of such data must be contingent upon non-disclosure obligation and observation of the confidentiality measures. Disclosure of such datasets without due confidentiality measures might compromise the regime of secrecy so that protection would be forfeited. As regards the requirement of economic value of know-how, this will be considered to be present if through publication, the research investment and competitive standing of the entity doing the work would be undermined [41].

In relation to the volumes of clinical data made available for research, this requirement, besides being at odds with the underlying data sharing culture of academic research, would create further workload. The data, subject to the regime of confidentiality, must first be strictly identified. The confidentiality mark would need to be attached to individual data items and any use and disclosure of such data to any third party must be subject to the latter signing a non-disclosure agreement (NDA). This preservation of the confidentiality mark, conclusion of NDA and control over handling such data as confidential would present another challenge.

Against these considerations, the protection of clinical data under the legal regime of know-how might, in principle, be possible in relation to some defined amount of data, but hardly offers a feasible solution, when protection of large amounts of data, processed in medical research is sought. It also may operate against the ethos of openness, if optimal use is to be made of the data by the research community, exploiting the full potential of available datasets.

VI. Application of IP Regimes to Data Curation in CHIC

A. Background

The research project "Computational Horizons In Cancer (CHIC): Developing Meta- and Hyper-Multiscale Models and Repositories for In Silico Oncology", is an ICT research project in the clinical domain [42]. CHIC develops clinical trial driven tools and services within a secure infrastructure, which facilitate the creation of multiscale cancer hyper-models (integrative models) by technical means. These composite multiscale constructs of models (hyper-models or integrative models) are intended to synthesize and imitate the biological processes, which occur in course of tumor progression, at several temporal and spatial levels (molecular, cellular, etc.) at once.

In this context too, the study of how individual cancer components interact with each other has led to the generation of different types of data, such as: molecular data, epigenetic data, clinical data, imaging data, pathology data and other laboratory data [43]. These different data types are assembled in order to systematically explore and formalize them in mathematical models.

Subsequently, the models are developed and validated against clinical data either taken from the literature or provided by the clinical partners [44]. The data management systems, used by the clinical partners, differ. Whereas the integration of data from data management system ObTiMA is harmonized, the data from individual clinical data repositories need to be adapted to the requirements of the project. The use of divergent data management systems by the clinical institutions leads to the situation that the data, collected from different sources, is not inter-operable with each other and mostly cannot be used for research as such. The clinical data also needs to be post-processed by the modelers so that it fits into the set of parameters, which the models recognize and can utilize as an input for running the simulations. Such data curation is a very important step because the inputs, outputs and descriptions of processes, simulated by the models, need to be standardized into the set of parameters, acceptable and usable by all cancer models.

B. Applicability of IP Regimes to Project Data Curation

The clinical data, which after the necessary de-identification enters the domain of CHIC, is placed and stored in the CHIC clinical data repository. The CHIC data repository hosts data categorized per data type: imaging data (DICOM etc.), descriptive/structural data (age, sex, etc.), other files (histological reports), links (to other data repositories) etc. The datasets for each type are accessible individually so that the data corresponding to the model parameters may be chosen. The fact that the repository is built "based on the experience already accumulated during the implementation of other data repositories" should be sufficient to prove the requisite investment in "either the obtaining, verification or presentation" of its contents [32]. Against this background, the database right in the CHIC clinical data repository is likely to be granted.

Protection of the CHIC data repository by the sui generis database rights would be accorded to the maker of the database. In the meaning of the Database Directive, the maker of a database is seen as "the person who takes the initiative and the risk of investing", but excluding subcontractors [32]. Thus, the party, who constructed the CHIC repository, would be in a position to manage the use of the repository, such as

by allocating the access rights to the project parties or external parties, to define the rights of use (access only, modification, download, etc.), to divide the repository into sections and define different regimes of uses depending on the data stored therein, etc. Grant of the sui generis protection would also entitle the right holder to enforce his rights, once unauthorized copying of the repository contents on the large scale has occurred.

Apart from the protection of the repository contents as a whole by sui generis database rights, the items in the repository may also enjoy protection in their own right. Since the clinical data repository deals with highly sensitive information (meaning that already for that reason, access to the data is strictly limited), application of the legal regime of know-how to some data items at least may be an option. As we saw above, for this, the data items selected for know-how protection, must be identified, the access and use of such data be limited to a defined number of people only, and the management of such data be subject to confidentiality measures. In the case of CHIC, the regime of secrecy may be applied to the data by marking it as "Confidential" and making the disclosure of such data subject to the non-disclosure obligation. From the technical perspective, the confidentiality mark would then need to be placed and borne by the data throughout the whole research process so that the data marked as "confidential" at the time of input comes out marked "confidential" at output. This would present an additional workload, but is implementable. Also, disclosure of such data items to the CHIC parties subject to the non-disclosure obligation would not present a significant obstacle, because the project parties are bound by the contractual relations within the project. The factual use of data within the project may also be managed by technical measures, such as granting or denying access rights, rights of use and extraction, and limiting the data processing to within the technical infrastructure of CHIC. Whereas the application of such contractual and technical confidentiality measures to the clinical data in CHIC may be feasible, in how far such technical and confidentiality measures may be implemented in other medical research projects may be questionable.

By contrast, copyrights and related rights offer less plausible options for protecting the clinical data in CHIC. As noted above, the clinical data in CHIC is represented by technical data from clinical trials, which is composed from different parameters. As observed in Section III, isolated items are not protectable by copyright. Copyright will fail for lack of creativity expressed in such data. Equally, the investment, deployed in curating the data for CHIC, does not qualify as investment protectable by related rights.

However, in the case of CHIC, the exploitability of clinical data under the umbrella of IP rights is limited by the restraints of data protection and research ethics. Whereas for the lifetime of the CHIC project, the de-identification of clinical data was ensured and clinical research ethically approved, the exploitation of the data beyond the scope of the project might be possible, if the adequate legal and security framework would be set up.

C. Related Studies

Indeed, the legal mechanisms offered by IP rights are widely used now by the players in the healthcare sector to support the claims and protect the investment they might have in the data. The database rights and the legal regime of know-how are the tools that suit these interests best and are used by the holders of clinical data most.

One example is deCODE. In the case of deCODE, a Health Sector Database, initially built to hold centralized health records of the population of Iceland [45], migrated into the genetics research database. By application of modern genomics techniques to the data (120,000 research participants), it allowed to find genetic sequences associated with diseases [7]. In consideration of the relatively small population of Iceland, access to a large amount of data allowed deCODE to find itself in a position to be able to predict the genetic dispositions to diseases of about 200,000 living and 80,000 deceased Icelanders, who have not consented to participate in the research [7]. Apart from the privacy considerations (which go beyond the scope of legal analysis presented in this paper), the case of deCODE allows us to infer that centralization of a large amount of clinical data in one database combined with modern IT solutions allows to retrieve new correlations and exploit the added value under the coverage of database rights (which may not always be in compliance with the principles of data protection law).

A similar example is the case of NIVEL. NIVEL, the Netherlands institute or health services research, has built a primary care database, which "*uses routinely recorded data from health care providers to monitor health and utilisation of health services in a representative sample of the Dutch population*." [46]. NIVEL obtains the data under contractual arrangements with general practitioners. Under the application of double de-identification measures [47] and giving the patients the possibility to opt-out, NIVEL uses itself and allows the use of data for clinical research.

The legal regime of confidentiality is another legal measure, which is often applied to preserve the secrecy of clinical data. Where the use of data in the domain of healthcare services is subject to the obligation of professional secrecy [19] [20]), the secrecy of data, or certain datasets, may be maintained by contractual mechanisms for the data to leave the healthcare sector (and enter the research domain). The application of confidentiality measures is typical for the data derived in clinical trials. Article 39 (3) TRIPS calls for protecting the data collected in clinical trials for the pharmaceutical products "*which utilize new chemical entities, the submission of undisclosed test or other data, the origination of which involves a considerable effort*" [38]. For the purposes of making the results of clinical trials public (either in scientific literature or clinical trial registries and databases), the legal regime of undisclosed information and contractual arrangements are often applied to preserve the

secrecy of certain datasets against undesirable disclosure [48]. This approach is often used by the pharmaceutical industry.

VII. Contractual Approaches

A. Contractual Approaches

Insofar as the IP regimes for protecting the data, produced in medical research projects fail, one further method for regulating rights in data may be by contractual relations.

Such relations exist at different levels. Thus, research projects are normally conducted by educational or research institutions and the research is typically done by research associates. Usually, the researchers do their work on the materials of the institution and achievement of scientific results in dependent position belongs to their employment obligations. In such circumstances, the researcher receives remuneration for the work he does, the institution acquires the ownership and also the exploitation rights over the achieved results, provided the agreement does not foresee otherwise [9].

Students or PhD students, who produce some research results under a membership relation to the university, do not have an obligation to create scientific works and are not obliged to pass ownership in their results to the university. In this constellation, the respective student owns the results of his work. In contrast, the PhD students, who are bound to the university by employment relations and do the research by order of the university, fall under the regulation of ownership in employment, considered above. Thus, the ownership over research results, achieved by a PhD student in an employee position, would normally pass to the institution [9].

In other cases, where the researchers perform some work as freelancers or sub-contractors, the question who acquires what rights in the results of the performed work depends on the contract [49].

Secondly, at an institutional level, third party funded projects and the rights in research results are typically governed by a contract between the sponsor and relevant project partner institution(s). The sponsor is typically interested to exploit the project results and grants the funding in exchange for acquisition of the ownership and exploitation rights over the research results. This model normally does not cause problems in practice [9]. The research institutions are bound by these contractual relations and it is their obligation to procure the ownership over the research results from the personnel, whom they engage into the project, and to ensure that the rights in research results are passed to the sponsor free from third party claims.

However, some research agreements are formulated in another way. For example, an agreement may provide that research results shall be the ownership of the party "*carrying out the work generating such results"*. The like provision may cause legal problems in practice. Let us consider the application of this rule in relation to the results of simulations done in a research project such as CHIC.

As we saw, in the context of that project, the simulations, which produce the data outputs, are executed by the models, developed by the modeling parties. Based on this provision, (a) the modeling parties, who developed the simulation software and (b) the clinical parties, who provided the clinical data for running the simulations may each claim rights in data outputs.

a) Modeling parties: by interpreting the above contract rule broadly, the modeling parties, who have developed the simulation software, may argue that they carried out the work generating the model, which produces the data, and shall own the rights in data, generated by the model, respectively. However, on a narrow interpretation, the modelers carried out the work generating the model, and not the data, calculated by the model, and shall own copyrights in the model code, and not in the data outputs from the model, respectively.

b) Clinical parties: may also claim rights in the results of simulations, since they provided the data, which the models used as an input to compute the data outputs. The counter-argument of the clinical parties may be that software models are used as a tool for data processing and do not give the modelers any rights in the data outputs themselves. An analogy with the use of Microsoft word for writing a PhD thesis, which does not confer on Microsoft any rights in the PhD thesis itself may support this argument.

This example shows that such contractual formulation may create legal uncertainty: first, with respect to qualifying simulation outputs as research results and, second, with respect to identifying the project party, who owns or holds the exploitation rights over such results. Unclear contractual formulations may give rise to potential legal disputes if the one or the other party would like to appropriate the data, achieved in the result of simulations for itself, and would seek to interpret the agreement in its favor.

A successful example showing how the contractual mechanism can be used to balance the rights of research participants against researchers´ rights is the case of PXE International (Pseudoxanthoma Elasticum (PXE)). PXE International is a research foundation, which represents the interests of individuals and families living with PXE, promotes and invests into PXE research [50]. When engaged into genetics research and the gene associated with the PXE disease was discovered and patented, PXE International managed to negotiate economic rights in the patent (i.e., deciding on the licensing strategy, sharing royalties, co-defining the prices) in exchange for the contribution of tissue and data that it made into the research [7] [51].

VIII. Conclusions

As we have seen, there are various ways in which the activity of curating clinical datasets could benefit from IP protection. Thus, collecting, arranging the data into a repository and making it suitable for use may render the investment, deployed in collecting and presenting the data, protectable by sui generis database rights. Similarly, the

generation of research data and adoption of additional confidentiality and security measures to keep this data secret to the broader community may render such data protectable as know-how.

However, the present approach that seeks to maintain (commercial) data confidentiality by keeping data secret leads to a fragmented research environment, and reduces the chances for greater data interoperability to be achieved. Here the law - aided by technology should aim to encourage greater openness, while assuring appropriate incentives and rewards for skilled curation. This could, e.g., take the form of an officially endorsed mechanism or system for measuring and tagging changes produced in a given data set (or the merging of several data sets) resulting from curation efforts, as the reward-trigger. At the same time, as another crucial policy element, the law needs – especially in the case of the curation of sensitive health data – to ensure that privacy and other interests of patients and research subjects are and remain adequately protected.

In particular, it will here be necessary to take account of (and compensate for) the knock-on effects of IP changes, where data-holders are no longer (also) motivated by commercial considerations to keep their data secure and confidential. This concern is all the greater here since the activities of data sharing and curation being encouraged, also by their nature present enhanced risks to personal privacy. The point of curation is precisely to uncover new connections and patterns in data that help generate robust inferences (usable – for good or ill) about the relevant data subjects. Accordingly, it is submitted that any system for rewarding investment in data curation should also require (as a condition for such rewards) that the data curator takes every appropriate measure to counterbalance the associated enhanced risks to privacy.

ACKNOWLEDGMENT

The research leading to these results has received funding from the European Union Seventh Framework Programme FP7/2007-2013 under grant agreement No 600841.

REFERENCES

[1] I. Lishchuk and M. Stauch, "Options for Protecting Medical Data by IP Rights," in Proc. GLOBAL HEALTH 2016, The Fifth International Conference on Global Health Challenges, Venice, 2016, pp. 29-34.

[2] J. Eils, "Strategy of sequencing the whole genome in clinical practice," presented at The Eighth International Conference on eHealth, Telemedicine, and Social Medicine eTELEMED 2016, April 24 - 28, 2016 - Venice, Italy.

[3] A. Popel and P. Hunter, "Systems biology and physiome projects," Wiley Interdiscip. Rev. Syst. Biol. Med.1:153–58, 2009.

[4] T. Deisboeck, M. Berens, A. Kansal, S. Torquato, A. Stemmer-Rachamimov, and E. Chiocca, "Pattern of self-organization in tumour systems: complex growth dynamics in a novel brain tumour spheroid model. Cell Prolif." 34:115–34, 2001.

[5] Children's Cancer Research Fund, Types of Childhood Cancer, Nephroblastoma <http://www.childrenscancer.org/main/wilms_tumor_nephroblastoma/> 02.05.2017.

[6] CHIC Deliverable No. D2.2 Scenario based user needs and requirements <http://chic-vph.eu/uploads/media/D2-2_Scenario-based_user_needs_and_requirements.pdf> 02.05.2017.

[7] D. M. Gitter, "Informed Consent and Privacy of De-Identified Information and Estimated Data, Lessons from Iceland and the United States in an Era of Computational Genomics," (Published Conference Proceedings style), in Proc. ALLDATA 2016, The Second International Conference on Big Data, Small Data, Linked Data and Open Data (includes KESA 2016), Lissabon, 2016, pp. 7-12.

[8] Moore v. Regents of University of California, Supreme Court of California, July 9, 1990, 51 Cal. 3d 120.

[9] H.-D. Lippert, „Wem gehören Daten, die im Rahmen von Forschungsprojekten gewonnen werden?"/"To whom belongs the data generated in research projects?" in: Geistiges Eigentum: Schutzrecht oder Ausbeutungstitel?/in: Intellectual Property: Protection right or title to exploit, Springer, Volume 5, 2008, pp. 359-369.

[10] Haimo Schack, "Zur Rechtfertigung des Urheberrechts als Ausschliesslichkeitsrecht"/"On justification of copyright as exclusive right," in: Geistiges Eigentum: Schutzrecht oder Ausbeutungstitel?/in: Intellectual Property: Protection right or title to exploit, Springer, Volume 5, 2008, pp. 124-140.

[11] European Commission, "E-Health Task Force Report – Redesigning health in Europe for 2020," Luxembourg 2012, ISBN 978-92-79-23542-9

[12] C. Tenopir, et al, "Data Sharing by Scientists: Practices and Perceptions," 2001, in: PLoS ONE, Vol. 6, No. 6, S. 1-21.

[13] KE (Knowledge Exchange), "Sowing the Seed: Incentives and Motivations for Sharing Research Data," a Researcher's Perspective, 2014<www.knowledge-exchange.info/Default.aspx?ID=733> 02.05.2017.

[14] J. Ludwig, „Zusammenfassung und Interpretation/Summary and Interpretation," Publ. in: Heike Neuroth, Stefan Strathmann, Achim Oßwald, Regine Scheffel, Jens Klump, Jens Ludwig (Hg.): Langzeitarchivierung von Forschungsdaten. Eine Bestandsaufnahme. Boizenburg: Werner Hülsbusch, 2012, pp.295-310.

[15] J. Reichman and P.F. Uhlir, "A Contractually Reconstructed Research Commons for Scientific Data in a Highly Protectionist Intellectual Property Environment," 2013 <http://scholarship.law.duke.edu/cgi/viewcontent.cgi?article=1283&context=lcp> 02.05.2017.

[16] U.S. National Institutes of Health, registry and results database of publicly and privately supported clinical studies of human participants conducted around the world <www.clinicaltrials.gov> 03.05.2017.

[17] BioMed Central Ltd, ISRCTN registry: a primary clinical trial registry recognised by WHO and ICMJE that accepts all clinical research studies <www.controlled-trials.com> 03.05.2017.

[18] JAPIC Clinical Trials Information <www.clinicaltrials.jp> 03.05.2017.

[19] 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, OJEU No L 281 /31, 23.11.95.

[20] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), OJEU, Volume 59 4 May 2016.

[21] Article 29 Data Protection Working Party, Advice paper on special categories of data ("sensitive data"), Ref. Ares(2011)444105 - 20/04/2011.

[22] Article 29 Data Protection Working Party, Opinion 03/2013 on purpose limitation, adopted on 2 April 2013, 00569/13/EN WP 203.

[23] CHIC Deliverable No. D4.4 Whitepaper - Recommendations for an amended European legal framework on patients' and researchers' rights and duties in E-health related research.

[24] B. J. Evans, "Much Ado about Data Ownership", Harvard Journal of Law & Technology, Vol.25, Number 1 Fall 2011.

[25] CHIC, Deliverable D2-2 "Scenario based user needs and requirements", <http://chic-vph.eu/uploads/media/D2-2_Scenario-based_user_needs_and_requirements.pdf> 03.05.2017.

[26] CJEU, Judgment of 16 July 2009, Case C 5/08, Infopaq International A/S v Danske Dagblades Forening, Ref. 45.

[27] Directive 2006/116/EC on the term of protection of copyright and certain related rights (codified version), OJEU, L 372/12, 27 December 2006.

[28] CJEU, Judgment of 7 March 2013, Case C 145/10 REC, Eva-Maria Painer v. Standard VerlagsGmbH, Axel Springer AG, Süddeutsche Zeitung GmbH, Spiegel-Verlag Rudolf Augstein GmbH & Co. KG, Verlag M. DuMont Schauberg Expedition der Kölnischen Zeitung GmbH & Co. KG.

[29] H. Zech, "Information als Schutzgegenstand," Tübingen, 2012, ISSN: 0940-9610 (Jus Privatum).

[30] Directive 2001/29/EC of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society, OJEU L 167/10 - L 167/19, 22.6.2001.

[31] H. Stenzhorn, et al, "The ObTiMA system - ontology-based managing of clinical trials," Stud Health Technol Inform. 2010;160(Pt 2):1090-4.

[32] Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases, OJEU, L 77/20 - L 77/28, 27.3.96.

[33] J. A. Bovenberg, "Property Rights in Blood, Genes & Data: Naturally Yours?" p. 159, 2006.

[34] E. Derclaye, "The Legal Protection of Databases," pp. 92 et seq, 2008.

[35] CJEU, Case C-203/02 The British Horseracing Board Ltd and Others v William Hill Organization Ltd., para 42.

[36] European Commission, DG Internal Market and Services Working Paper – First evaluation of Directive 96/9/EC on the legal protection of databases, 2005, p. 14<http://ec.europa.eu/internal_market/copyright/docs/databases/evaluation_report_en.pdf> 03.02.2017.

[37] CJEU, Case C-338/02 Fixtures Marketing Ltd v Svenska Spel AB, para 28.

[38] Agreement on Trade-Related Aspects of Intellectual Property Rights, the TRIPS Agreement, Annex 1C of the Marrakesh Agreement Establishing the World Trade Organization, Marrakesh, Morocco, 15 April 1994.

[39] Directrive (EU) 2016/943 of 8 June 2016 on the protection of undisclosed know-how and business information (trade secrets) against their unlawful acquisition, use and disclosure, OJEU L157/1, 15.06.2016.

[40] Hogan Lovells International LLP, "Report on Trade Secrets for the European Commission – Study on Trade Secrets and Parasitic Copying (Look-alikes), MARKT/2010/20/D," 2011.

[41] K. Lodigkeit, Intellectual Property Rights in Computer Programs in the USA and Germany, Peter Lang GmbH, 2006, pp. 98-101.

[42] Computational Horizons In Cancer (CHIC): Developing Meta- and Hyper-Multiscale Models and Repositories for In Silico Oncology <http://chic-vph.eu/project/> 03.02.2017.

[43] C. Coveney, J. Gabe, and S. Williams, "The sociology of cognitive enhancement: medicalisation and beyond," Health Sociol. Rev., 20 (2011), pp. 381–393.

[44] J. Dejaegher, L. Solie, S. De Vleeschouwer, and S. W. Van Gool, "Dendritic Cell Vaccination for Glioblastoma Multiforme: Clinical Experience and Future Directions," In G. Stamatakos and D. Dionysiou (Eds): Proc. 2014 6th Int. Adv. Res. Workshop on In Silico Oncology and Cancer Investigation – The CHIC Project Workshop (IARWISOCI), Athens, Greece, Nov.3-4, 2014 (www.6thiarwisoci.iccs.ntua.gr), pp.14-18. (open-access version), ISBN: 978-618-80348-1-5.

[45] A. Abbott, "Icelandic database shelved as court judges privacy in peril," Nature, vol. 429, p. 118, May 13, 2004, doi:10.1038/429118b.

[46] NIVEL, databases and panels <https://www.nivel.nl/en/databases-and-panels> 30.05.2017.

[47] S. Gutwirth, R. Leenes, P. De Hert, „Data Protection on the Move. Current Development in ICT and Privacy/Data Protection," 2016, p.101 et seq.

[48] Joint Position on the Disclosure of Clinical Trial Information via Clinical Trial Registries and Databases <www.ifpma.org/clinicaltrials> 30.05.2017.

[49] C. Reed and J. Angel, "Computer Law: The Law and Regulation of Information Technology," 6th ed, 2007, p. 352 et seq.

[50] PXE International <https://www.pxe.org/about-pxe-international> 30.05.2017.

[51] P. Smaglik, "Tissue donors use their influence in deal over gene patent terms," NATURE, Vol. 407, 19,10, 2000, p. 821.