Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 150435, 11 pages
http://dx.doi.org/10.1155/2015/150435

Review Article

Challenges and Opportunities for Exploring Patient-Level Data

Pedro Lopes, Luis Bastião Silva, and José Luis Oliveira

Department of Electronics, Telecommunications and Informatics (DETI), Institute of Electronics and Informatics Engineering of Aveiro (IEETA), University of Aveiro, 3810 193 Aveiro, Portugal

Correspondence should be addressed to José Luis Oliveira; jlo@ua.pt

Received 5 May 2015; Accepted 27 August 2015

Academic Editor: Ernesto Picardi

The proper exploration of patient-level data will pave the way towards personalised medicine. To better assess the state of the art in this field we identify the challenges and uncover the opportunities for the exploration of patient-level data through the review of well-known initiatives and projects focusing on the exploration of patient-level data. These cover a broad array of topics, from genomics to patient registries up to rare diseases research, among others. For each, we identified basic goals, involved partners, defined strategies and key technological and scientific outcomes, establishing the foundation for our analysis framework with four pillars: control, sustainability, technology, and science. Substantial research outcomes have been produced towards the exploration of patient-level data. The potential behind these data will be essential to realise the personalised medicine premise in upcoming years. Hence, relevant stakeholders continually push forward new developments in this domain, bringing novel opportunities that are ripe for exploration. Despite last decade's translational research advances, personalised medicine is still far from being a reality. Patients' data underlying potential goes beyond daily clinical practice. There are miscellaneous challenges and opportunities open for the exploration of these data by academia and business stakeholders.

1. Introduction

The widespread collection of patient-level data represents a critical step towards the realization of personalised medicine [1, 2]. These data stem from primary care centres, hospital information systems, clinical trials' cohorts, and administrative platforms. Moreover, they withhold a huge potential that goes beyond daily clinical care [3, 4].

Yet, along with the miscellaneous opportunities to explore patient-level data, this unparalleled growth of patients' digital metadata brings several challenges [5, 6]. Data size, lack of open access, heterogeneity, or the uses of primitive technologies are some of the issues researchers face [7]. In contrast, exploring the potential behind these data will lead to the discovery of new knowledge, essential to improve the current clinical narrative [8, 9].

Although patient-level data from public institutions, such as hospitals or regional/national administration centres, should be easier to access, it is generally locked under primitive technological implementations. This results in closed data silos that hinder scientific and technological evolution. Several large-scale projects already try to commoditize access to these data, whether through policies or through technical standards for data exchanges [10].

Pharmaceutical companies are also responsible for a big chunk of patient-level data [11]. Clinical trials' cohorts generate comprehensive patient datasets whose value for personalised medicine research is immeasurable [12, 13]. Despite this, most of pharmaceutical data are private [14].

It is important to distinguish between private companies' data, which is the basis for internal research and development for new drugs and treatments, from public research datasets, fundamental to advance general scientific research. Although pharmaceutical companies are entitled to keep their results private, policies should be put in place to foster the sharing of clinically relevant results into the public domain.

Dealing with this heterogeneous mixture of private and public patient-level data, tools, standards, and projects is in itself a complex research and development challenge [15]. Ultimately, the entropy in this ecosystem is delaying what should be a swift evolution. Hence, we need to evaluate past and on-going initiatives to better assess and plan the personalised medicine research and development roadmap for the upcoming years [16].

For this matter we established an evaluation framework to analyse the outcomes of existing initiatives, identifying current challenges and uncovering new opportunities. This framework is based on four key pillars: control, sustainability, technology, and science. We assess several components in each of these areas, generating a rather comprehensive study:

(i) the control section focuses on data ownership and access;

(ii) the sustainability topics cover the long-term perspectives for each asset;

(iii) on technology we assess the technical outcomes for each project, where existing;

(iv) at the science level we identify the projects' research areas and their key scientific outcomes.

We present this comprehensive review targeting three key objectives. These were to (1) identify the best initiatives dealing with patient-level data, (2) inspect and study their different features, and (3) evaluate tackled challenges and open opportunities. Furthermore, we shed some light on the current status of public investment into research, where the lack of strict evaluation guidelines brings too much liberty to funded project partners. This research work brings true added value to multiple fields in the scientific domain; from the performance analysis of hospital care [17, 18] to the on-going exploration of pharmaceutical trials data [19], among others [20].

2. Materials and Methods

2.1. Design. This review covers past and on-going large-scale projects. Selected projects' evaluation is based on an assessment framework with four key components: control, sustainability, technology, and science. This design allows us to better understand the projects' outcomes distribution as well as defining an initial categorization for each project. We chose topics for matching criteria in each area based on mappings with existing ontologies, namely, Simple Knowledge Organization System (SKOS) [21] and EMBRACE Data and Methods (EDAM) [22].

At the control level we assess several topics, detailed next.

(i) *Data ownership*: who owns the project data and who decides whether to make data available or not? Available options are *community*, *partner*, or *project*.

(ii) *Data access*: is there open access to the project's data or is it closed to project partners? Available options are *partners only*, *private*, or *public*.

(iii) *Data storage*: are data stored in partners' private repositories or publicly shared with the involved community? Available options are *partners only*, *private*, or *public*.

(iv) *Patient involvement*: are patients engaged in data ownership; that is, can patients control who can use their personal data in the project's systems? Available options are *no* or *yes*.

(v) *Security, privacy, and auditing*: how are security, privacy, and auditing issues dealt with within the project? Available options are *external*, *none*, or *project*.

In this review we also assess the selected projects' sustainability, covering the following areas:

(i) *Business model*: what is the business model behind the data owners? This has implications on what happens beyond each project's scope. Available options are *academia*, *business*, or *undefined*.

(ii) *Data maintenance*: associated with the project's partners' business model, we have to assess what will happen with the collected data when the project finishes. Available options are composed of *stored*, *unpublished*, or *undefined*.

At the technology level we identified the technological outcomes from the studied projects, where available.

(i) *Technological outcomes*: are there (or will there be) any relevant technical outcomes from the project? Available options are *yes*, *only scientific*, *too soon to know*, or *undefined*.

(ii) *Technology*: what are the main technological outcomes of each project? This includes *database*, *framework*, *infrastructure*, *library*, *standards*, *virtual machine*, *web services*, or *undefined*.

At last, we inspected the key scientific outcomes for each project, evaluating their areas of impact.

(i) *Field of research*: it is the fields of research with results that will have direct application to improve patient-level data exploration. These include *EHR*, *epigenomics*, *genomics*, *metabolomics*, *pharmacogenomics*, *phenomics*, *proteomics*, *transcriptomics*, and *other*.

(ii) *Area of interest*: similarly to the field of research, we identified the technological areas of scientific interest that were studied in the project. Available options are *analytics*, *annotation*, *data integration*, *data visualization*, *ontology*, *semantic analysis*, *text-mining*, and *other*.

2.2. Inclusion and Exclusion Criteria. We searched for large-scale international projects in literature and general listings. From there, the inclusion criteria for this review were as follows:

(i) is on-going or finished after January 1st, 2011;

(ii) is sponsored mainly by the NIH, IMI, or the European Commission;

(iii) includes partners from both academia and the business sector;

(iv) must focus on rare diseases, pharmacy or have direct patient involvement;

(v) must have public published results.

For all identified projects, we reviewed titles, funding information, references, and available publications to better assess if the projects appeared to meet all inclusion criteria. If insufficient information was available to make a confident decision, we contacted key project partners to disclose further details.

3. Results

This review provides an overview of the different attempts at improving the exploration of patient-level data. This section details the projects' evaluation according to our framework, including a tabular and visual comparison of their distinct features. From this evaluation we identify the main challenges and opportunities for future research endeavours.

3.1. Projects. Our initial dataset was extracted from the online project databases of three major funding agencies: USA's National Institutes of Health (NIH), European Commission (EC), and the Innovative Medicines Initiative (IMI) [23–25]. After a comprehensive filtering and selection process, 16 projects met our inclusion criteria (Table 1).

On a first glance we can quickly assess that the selected projects' domains and goals are heterogeneous, with the access or use of patient-level data being one of the few common threads. There is also an obvious bias towards European projects, as the European Commission continues to be a strong proponent of research, namely, on the life sciences and medical areas.

3.2. Feature Comparison. In this section we explore the projects' evaluation results according to the several pillars of our evaluation framework.

3.2.1. Control. From Figure 1, highlighting the control pillar, we can conclude that there is real diversity in the projects being assessed regarding who controls the data. The notable exception concerns the patient involvement (Figure 1(d)). Although patients play a fundamental role in the research workflow, patients and patient advocacy groups are seldom considered as partners. As the other charts in Figure 1 show, data are equally distributed, owned, and stored by partners, the project, and the public domain. However, if we make a more basic categorization between open (public or community) and private (project or partner), the division is steeper.

3.2.2. Sustainability. Our sustainability review entails better prospects for future data exploration. As Figure 2(b) highlights, the majority of projects already do or plan on doing active data maintenance. This implies that data collected within the project's scope will be stored for future use. Even if the access is limited, keeping these data alive opens good prospects for future endeavours. About half the evaluated projects will continue to provide their results to academia and some will focus on creating a business to sustain their research work once the project finishes (Figure 2(a)).

3.2.3. Technology. At the technological level, all evaluated projects already produced public results. As expected from the heterogeneous project goals, there is an assorted amount of technical outcomes. Figure 3 highlights the current trend, where services and databases are the focus of produced work. Next, infrastructure development is also a key area in selected projects, although they were more relevant for projects started before 2011 (Figure 3(A)). These particular results are of particular relevance for our review. We can infer that there is already proper effort put towards creating infrastructures for research. Hence, we should move our focus to the better exploration of existing resources, namely, with the creation of additional frameworks, standards, and services.

3.2.4. Science. As shown in Figure 4, we find greatest variety of project features at the scientific level. Figure 4(a) chart presents the various fields of research for projects started before 2011 (Figure 4(a)(A)) and after 2011 (Figure 4(a)(B)). In these, genomics is evidently important. Although the results are biased due to the selected projects' domain, there is a clear influence of genomics, pharmacogenomics, and biobanking at the patient-level domain (EHR). Nevertheless, as shown in Figure 4(a)(B), the miscellaneous omics research fields continue to be of interest and EHR interest is growing.

Figure 4(b) also validates the fundamental role of data integration in the various research fields. Nowadays, data integration expertise must be a vulgar commodity for life sciences and medical related research projects. More importantly, Figure 4(b)(B), for projects started after 2011, the differences in the fields of analytics, ontologies, text-mining, and semantic analysis are staggering. This reveals the growing significance of semantic web related technologies, as they complement analytics, ontologies, and text-mining features.

3.3. Challenges and Opportunities. With this evaluation we identified several challenges and opportunities. Challenges relate to data discovery, access, acquisition, and ownership. This brings several opportunities to deploy future solutions that fully explore the enormous amounts of patient-level data, using technological paradigms that projects are already supporting.

3.3.1. Challenges. There is a clear dichotomy regarding data. Patient-level data is a very specific use case for exploration. While there are too many data scattered throughout multiple stakeholders, they are wildly difficult to obtain. The outcome of this is that, in the end, there is not enough data to generate statistically meaningful conclusions. Hence, we cannot discover or infer new knowledge because there is no access to a minimal amount of patient data. Along with distribution, data heterogeneity arises as a key challenge for exploring patient-level data. As shown in Figure 3, there are already several projects dealing with creating new and improving existing data standards for data sharing. However, these are far from being widely adopted throughout international stakeholders. Bioinformatics and pharmacogenomics projects also face these challenges [41]. Nevertheless, for these

Table 1: List of evaluated projects.

Project	Start	End	URL	Description
BBMRI	2008	2011	http://bbmri.eu/	BBMRI connects researchers, biobankers, patient advocacy groups, and pharmaceutical research companies to foster a quicker discovery of new treatments [26]. Their strategy is based on the enrichment and harmonization of biobanks.
BioMedBridges	2012	2015	http://www.biomedbridges.eu/	BioMedBridges' goal is to launch a shared e-infrastructure for biological and biomedical data.
BioSHaRe-EU	2010	2015	https://www.bioshare.eu/	BioSHaRe-EU partners are working to ensure the development of harmonized measures and standardized computing infrastructures.
BRIDGEtoData	2011	—	http://www.bridgetodata.org/	BRIDGEtoData aims to be an online reference platform describing population healthcare databases for use in epidemiology and health outcomes research.
DDMoRe	2011	2016	http://www.ddmore.eu/	The Drug Disease Model Resources (DDMoRe) project aims to establish a universal standard framework for modelling drugs and diseases [27, 28].
EHR4CR	2011	2014	http://www.ehr4cr.eu/	EHR4CR partners built, validated, and deployed a Europe-wide innovative technological platform to reuse EHRs data for clinical research purposes [29].
ELIXIR	2010	2018	http://www.elixir-europe.org/	ELIXIR project's goal is to coordinate the collection, quality control, and archiving of large amounts of biological data [30].
EMIF	2012	2018	http://www.emif.eu/	EMIF's goal involves the creation of an innovative and connected patient registry catalogue that will enable researchers and pharmaceutical companies to search for patient-level data based on the databases' digital fingerprints [31].
ESGI	2011	2015	http://www.esgi-infrastructure.eu/	ESGI's goal is to integrate and standardise current and emerging technologies, providing access to infrastructures so that a broad group of European researchers can use the new technologies.
eTRIKS	2012	2017	http://www.etriks.org/	eTRIKS' objective is to address knowledge management gaps by building a sustainable translational research informatics/knowledge management platform and to provide additional sustainable services.
EU-ADR	2008	2012	https://bioinformatics.ua.pt/euadr/	EU-ADR project aimed developing a unique computerized system to detect adverse drug reactions (ADRs), supplementing spontaneous reporting systems [32].
EURenOmics	2012	2018	http://www.eurenomics.eu/	EURenOmics work is based on rare kidney diseases, where the project seeks to establish more accurate diagnoses strategies and improve clinical care.
Euro-BioImaging	2010	2014	http://www.eurobioimaging.eu/	Euro-BioImaging's main work covered the improvement of existing research infrastructures on a large scale.

Table 1: Continued.

Project	Start	End	URL	Description
GEN2PHEN	2008	2013	http://gen2phen.org/	GEN2PHEN was created to unify human and model organism genetic variation databases towards increasingly holistic views into Genotype-to-Phenotype (G2P) data and to link this system into other biomedical knowledge sources via genome browser functionality [33].
NeurOmics	2012	2018	http://rd-neuromics.eu/	NeurOmics' research objectives feature the study of neurodegenerative and neuromuscular diseases in an attempt to explore Omics technologies to improve diagnosis, treatments, and general patient care.
OMOP	2008	2013	http://omop.org/	OMOP's goal was to design experiments testing a variety of analytical methodologies in a range of data types to look for drug impacts, going towards a complete database analysis standard [34].
Oncotrack	2011	2016	http://www.oncotrack.eu/	Oncotrack deploys several methods for systematic next generation oncology biomarker development [35, 36].
OpenPHACTS	2011	2014	http://www.openphacts.org/	OpenPHACTS works with the integration of a relevant and continuously expanding subset of distributed heterogeneous data sources into one "virtual resource," via the creation of a semantic interoperability layer [37].
RD-Connect	2012	2018	http://rd-connect.eu/	RD-Connect will launch an integrated platform connecting databases, registries, biobanks, and clinical bioinformatics for rare diseases research [38].
Sentinel	2008	—	http://www.fda.gov/Safety/FDAsSentinelInitiative/default.htm	Sentinel is a USA-based electronic system that will transform FDA's ability to track the safety of drugs, biologics, and medical devices [39, 40]. This initiative aims to develop and implement a proactive system that will complement existing systems that the FDA has in place to track reports of adverse events linked to the use of its regulated products.

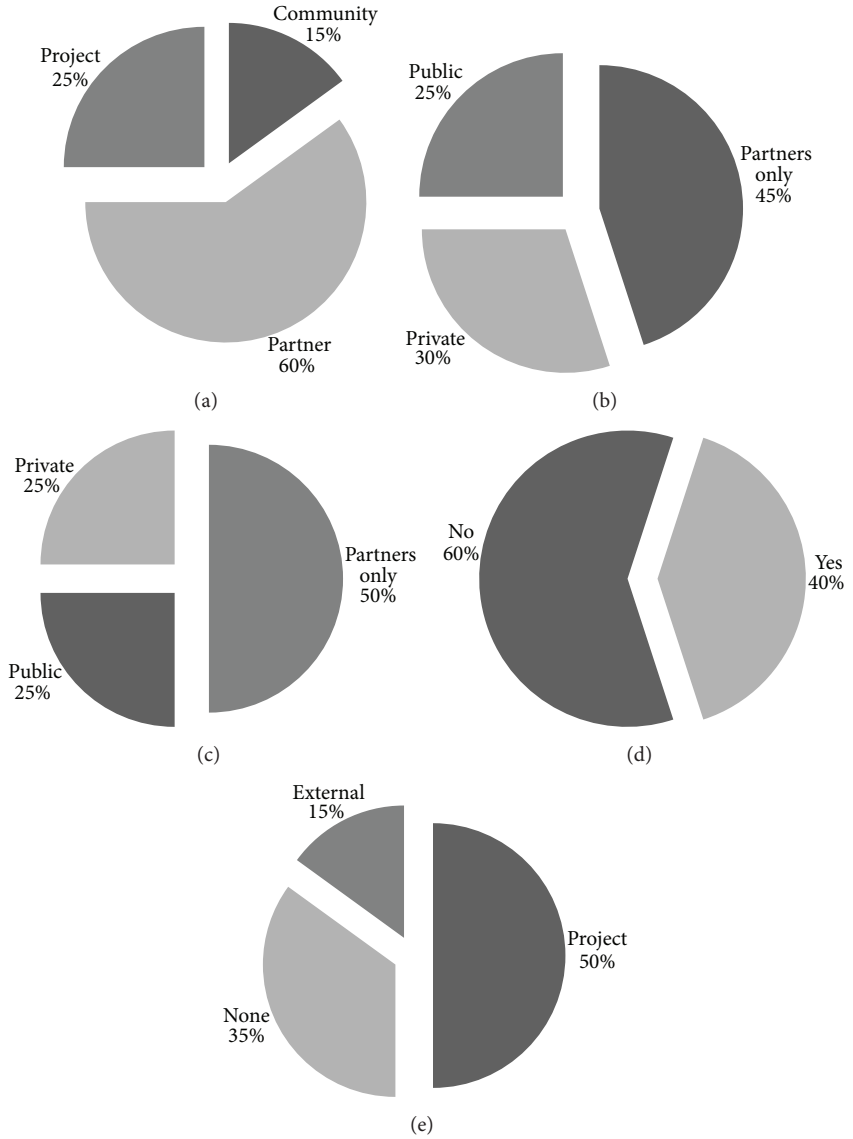

Figure 1: Data control evaluation breakdown charts. Charts summarizing evaluation results for the control section of the proposed evaluation framework. (a) Data ownership; (b) data access; (c) data storage; (d) patient involvement; (e) security, privacy, and auditing.

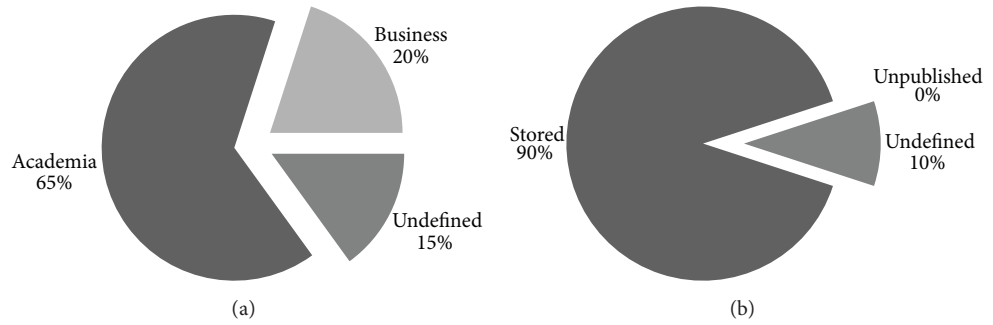

Figure 2: Data sustainability evaluation breakdown charts. These two charts feature the tracked sustainability topics in the proposed evaluation framework. (a) Business model and (b) data maintenance.

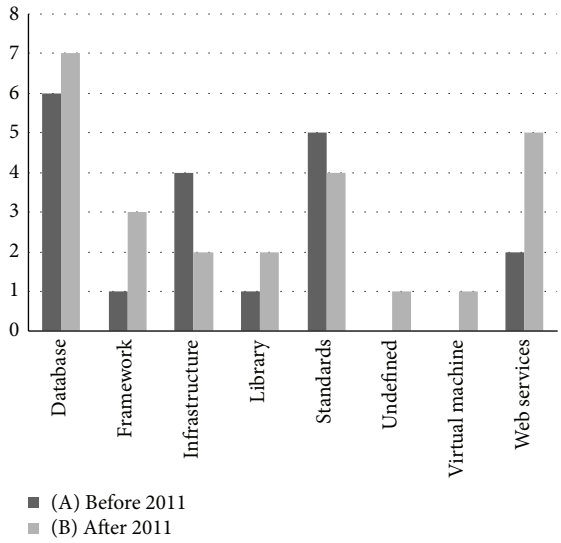

Figure 3: Technology outcomes' evaluation evolution breakdown chart. This chart features the key technological outcomes across the various projects, as assessed according to the proposed evaluation framework. To better understand the results' evolution over time, project evaluation results are divided between projects started before the year 2011 (A) and after the year 2011 (B).

there are already adequate standards for data storage and exchange [42–45].

In the same vein, data translation also arises as a complex challenge for researchers. In addition to the obvious sense (translating data between multiple languages [46]), there is the data translation from a low-level free text data to structured information [47, 48]. Clinicians' reports traditionally include their notes in free text. These notes must be mapped to a shared domain, elevated from simple text to meaningful structured knowledge. Again, the growing relevance of text-mining and semantic web technologies, as highlighted before, is visible.

Data discovery, access, and acquisition are typical problems that can be solved by improving existing technologies and by focusing on their widespread adoption. Unlike these, data ownership is a much more complex issue. Dealing with data ownership involves tackling issues related with government's policies, stakeholders' interests, and projects' internal guidelines. In an ideal scenario, all patient-level data should be available for research purposes. This should be particularly enforced in publicly funded projects. Yet, this does not happen. As seen in Figure 1, projects' data ownership, storage, and access resort to closed solutions. In most cases, data are privately held, or at most, shared to project partners. Moreover, where data are shared publicly to researchers, access restrictions are in place.

3.3.2. Opportunities. Great challenges leverage great opportunities. From our review, we believe there is room for improving how we explore patient-level data and how we can use it to further improve research and development towards personalised medicine. As Figure 4 highlights, on-going projects are already solving important technological challenges.

There is huge potential behind the combination of data available worldwide. Yet, we need to develop and disseminate new technologies that improve how relevant entities collect, store, and share patient-level data.

As data integration is already commonplace, to obtain real advances in this domain we must see worldwide patient-level data as a whole, and not as single detached data silos. Although we already have the technology to accomplish this, stakeholders must unite efforts to make this holistic view a reality.

At the technical level, opportunities arise that demand the creation of new software and new standards. Likewise, at a policy level, we must improve existing guidelines and policies to better cover data sharing and ownership and ethics issues.

New data management standards should promote better (and easier) ways to access and share data. This will promote knowledge discovery and enable the integration and interoperability among patient-level data silos throughout the world. Likewise, going from patient-level data to summary-level data, and vice-versa, should be a simple straightforward process with the latest text-mining and semantic web tools.

Ideally, new software will empower collaboration and sharing among patients and clinicians. These should promote ease of access to patient information and enhance the communication process among clinicians. Furthermore, new tools are required to enhance data ownership controls, facilitating how patients, clinicians, or researchers express who has access to relevant personal data. More importantly, a combination of policies and guidelines should be put in place to foster the active involvement of patients in clinical care.

Despite the great opportunity for creating new standards and software, these assets alone are not enough to change the current scenario. New politics and guidelines, stemming directly from key worldwide stakeholders, must be disseminated to all interested parties. Moreover, with adequate support from governmental agencies (regional, national, and international), projects and their internal partners will proactively work towards implementing these new guidelines.

4. Discussion

As this review reveals, there is room for change in the exploration of patient-level data. However, we must take in account that these results are biased and strict. This is an ever-expanding field with lots of partners, projects, and companies working in this subject.

While we tried to be comprehensive, this review has obvious limitations. Namely, identifying each project's features and technical/scientific outcomes was a complex task. Once the projects finish, little to no effort is put into maintaining an accurate dissemination summary and rarely the projects results are assessed a couple years after each project's conclusion.

4.1. The Growing Relevance of Genomics Data. The core focus of this review revolves around projects dealing with

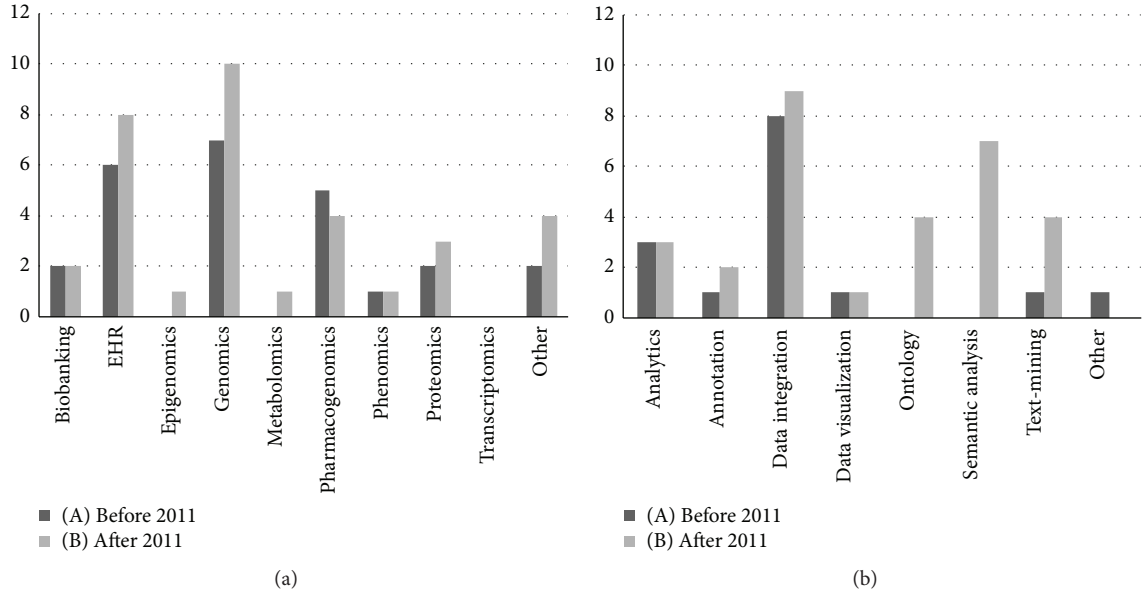

Figure 4: Science outcomes' evaluation evolution breakdown charts. Charts summarizing the various scientific research topics covered across the various projects assessed with the proposed evaluation framework. (a) Field of research; (b) area of interest. To better understand the results' evolution over time, project evaluation results are divided between projects started before the year 2011 (A) and after the year 2011 (B).

patient-level data stemming from electronic patient records. However, as shown in Figure 4(a), the quantity and quality of projects interacting with patient databases focused on genomics data are growing [49]. Furthermore, next generation sequencing technologies streamline the generation of huge patient datasets [50].

In a sense, patient sequencing data are patient-level data. Projects, such as 1000 Genomes [51] or Genome of the Netherlands [52], are trying to sequence large numbers of individuals to better understand existing genotype-phenotype relationships and uncover new ones.

In the long term, these data will be included in clinical patient registries. They may even be part of the electronic patient record. At this stage, clinicians will require new tools to adequately exploit the true value behind these data. In summary, this is a whole new field of exploration for personalised medicine and patient-level data research that cannot be ignored [53].

4.2. Implications for Future Research. As detailed in previous sections, the various opportunities highlight the room for improvement in this domain. Assessing the projects' timing evolution we identify that the focus on sharing, dissemination, and patient control is of growing relevance in the field.

The creation of new technical standards and data sharing policies will be fundamental for future research. Moreover, these topics are emerging in current project calls. Thus, they are becoming a stepping-stone for future research and infrastructure initiatives.

Despite the scale of on-going projects, they will not cover every possible topic. Technological developments in analytics tools, text-mining, ontologies, semantic web, data visualisation, integration, and interoperability, originating from distinct areas, must be brought to patient-level exploration.

The semantic web arises as a ground breaking paradigm to foster the intelligent integration of structured information. Sustained by state-of-the-art standards such as RDF, OWL, SPARQL, and LinkedData, semantic web promotes better strategies to express, infer, and make knowledge interoperable.

Latest advances in the area cover the research and development of new algorithms to further improve how we collect data, transform data into meaningful knowledge assertions, and publish connected knowledge. To further improve this, we must rely on the latest text-mining technologies. Elevating clinical text data to abstract knowledge or mapping the best matching ontologies to patient datasets require advanced text-mining solutions.

The combination of these strategies, semantic web, text-mining, and ontologies will pave the way towards interoperable scientific knowledge. These technologies will foster data integration and interoperability, enabling an effortless connection between heterogeneous distributed knowledge, obtained from patient-level data. Hence, the foundation of translational research, where multiple technical research areas collide, will be even more meaningful in the future.

4.3. Impact. Although this review had the main goal of covering the scientific results, we cannot ignore additional fundamental questions surrounding large-scale projects.

Hence, we must discuss the privacy policies applied to research-oriented datasets, the creation of businesses sustained by public funding, or the lack of publicly visible project evaluation outcomes.

The general community perceives that there is a huge amount of public funds being poured into research projects in all areas. Still, the outcomes of these projects are not as public as desired. There is an underlying sense of fulfilment in investing on research, especially in fields related with life sciences, such as rare diseases treatments, pharmaceutical research, or any other relevant omics field: IMI, EC, and NIH are funding science.

Figure 1(b) highlights that only a quarter of studied projects expect to provide their data publicly to the general research audience. Data access restrictions are too common on research. Large investments, with public funds, are being applied to clinical drug trials, patient registries development, and next generation sequencing technologies. Yet, the majority of research outcomes will not be made available to the public. And, despite pharmaceutical companies financial involvement in IMI projects, the expected profit outcome from these projects will definitely surpass invested money. Patient-level data, obtained with public research funds, which have the potential of being fundamental to create new knowledge, are not available to the research community as they are closed behind complex privacy policies and never-ending access restrictions.

Likewise, Figure 2 charts show that there are several projects whose future sustainability will rely on implementing a profit-oriented business model. Hence, we must ask, again, how can public funds, applied to research projects, be used to create self-sustainable companies? These companies will sell products, software or data, created with research funds stemming from public investment.

At last, there is a great difficulty in finding projects details and their respective evaluation results. It is as if the IMI, EC, and NIH projects lists are difficult to access and lack essential project details on purpose. The general audience cannot find out how projects are evaluated, their assessment results and, more importantly, their visible outcomes. Despite having concluded that most project results are private, the projects' evaluation should be public. Furthermore, it should be supported by a clear long-term plan that assessed the proper use of public funds to actually advance research. Finished projects should be evaluated in multiple timespans, not just when the deadline is reached. Evaluating projects 2, 5, or 10 years after their finish date would improve the understanding of how successful was the large sum of invested money.

The reality is that IMI, EC, and NIH are funding projects that have the liberty to create for-profit businesses and, more importantly, the liberty to apply public funds to the most diverse research tasks, whether they are directly related to the expected project results.

5. Conclusions

This review provides an overview of different initiatives that try to properly explore patient data. We limited our study to research and development projects in the recent past. We established base criteria to evaluate on-going initiatives. This resulted in the identification of several opportunities for future developments, namely, (1) bringing distributed data together by putting more advanced sharing and integration at clinicians' fingertips; (2) focus on text-mining and semantic web technologies to create real knowledge from distributed and heterogeneous data; and (3) pressuring stakeholders for stricter project evaluations that will foster a quicker evolution pace. The lack of well-established and widely adopted solutions covering these areas represents a major roadblock for the adequate exploration of patient-level data. However, if future projects consistently adopt these overarching goals, personalised medicine will be one step closer.

More importantly, in addition to the research-specific evaluation outcomes, we must highlight the strange patterns behind large-scale project funding. Although IMI, NIH, and EC provide intensive financial support for research, what we witness is that the money is being used to create for-profit businesses and closed research datasets. Furthermore, funding agencies lack clear evaluation frameworks that properly assess the success of public investment into large-scale research.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

The research leading to these results has received funding from the European Community (FP7/2007–2013) under Reference no. 305444, the RD-Connect project, from EU/EFPIA Innovative Medicines Initiative Joint Undertaking (EMIF Grant no. 115372), and from the QREN "MaisCentro" program, Reference CENTRO-07-ST24-FEDER-00203, the Cloud Thinking project.

References

[1] A. Miles, M. Loughlin, and A. Polychronis, "Evidence-based healthcare, clinical knowledge and the rise of personalised medicine," *Journal of Evaluation in Clinical Practice*, vol. 14, no. 5, pp. 621–649, 2008.

[2] M. A. Hamburg and F. S. Collins, "The path to personalized medicine," *The New England Journal of Medicine*, vol. 363, no. 4, pp. 301–304, 2010.

[3] A. Harvey, A. Brand, S. T. Holgate et al., "The future of technologies for personalised medicine," *New Biotechnology*, vol. 29, no. 6, pp. 625–633, 2012.

[4] P. Coorevits, M. Sundgren, G. O. Klein et al., "Electronic health records: new opportunities for clinical research," *Journal of Internal Medicine*, vol. 274, no. 6, pp. 547–560, 2013.

[5] G. H. Lyman and N. M. Kuderer, "The strengths and limitations of meta-analyses based on aggregate data," *BMC Medical Research Methodology*, vol. 5, article 14, 2005.

[6] P. M. Coloma, M. J. Schuemie, G. Trifirò et al., "Combining electronic healthcare databases in Europe to allow for large-scale drug safety monitoring: the EU-ADR Project," *Pharmacoepidemiology and Drug Safety*, vol. 20, no. 1, pp. 1–11, 2011.

[7] C. Tudur Smith, P. R. Williamson, and A. G. Marson, "Investigating heterogeneity in an individual patient data meta-analysis

of time to event outcomes," *Statistics in Medicine*, vol. 24, no. 9, pp. 1307–1319, 2005.

[8] K. A. Broeze, B. C. Opmeer, F. van der Veen, P. M. Bossuyt, S. Bhattacharya, and B. W. J. Mol, "Individual patient data meta-analysis: a promising approach for evidence synthesis in reproductive medicine," *Human Reproduction Update*, vol. 16, no. 6, Article ID dmq043, pp. 561–567, 2010.

[9] H. Xu, Z. Fu, A. Shah et al., "Extracting and integrating data from entire electronic health records for detecting colorectal cancer cases," in *Proceedings of the AMIA Annual Symposium*, pp. 1564–1572, Washington, DC, USA, October 2011.

[10] S. Marceglia, P. Fontelo, and M. J. Ackerman, "Transforming consumer health informatics: connecting CHI applications to the health-IT ecosystem," *Journal of the American Medical Informatics Association*, vol. 22, no. 1, pp. e210–e212, 2015.

[11] B. Wieseler, N. Wolfram, N. McGauran et al., "Completeness of reporting of patient-relevant clinical trial outcomes: comparison of unpublished clinical study reports with publicly available data," *PLoS Medicine*, vol. 10, no. 10, Article ID e1001526, 2013.

[12] P. Nisen and F. Rockhold, "Access to patient-level data from GlaxoSmithKline clinical trials," *The New England Journal of Medicine*, vol. 369, no. 5, pp. 475–478, 2013.

[13] D. E. Johnson, "Fusion of nonclinical and clinical data to predict human drug safety," *Expert Review of Clinical Pharmacology*, vol. 6, no. 2, pp. 185–195, 2013.

[14] B. N. Sampat and F. R. Lichtenberg, "What are the respective roles of the public and private sectors in pharmaceutical innovation?" *Health Affairs*, vol. 30, no. 2, pp. 332–339, 2011.

[15] C. Daniel, E. Albuisson, T. Dart, P. Avillach, M. Cuggia, and Y. Guo, "Translational bioinformatics and clinical research informatics," in *Medical Informatics, e-Health*, A. Venot, A. Burgun, and C. Quantin, Eds., Health Informatics, pp. 429–461, Springer, Paris, France, 2014.

[16] A. P. Abernethy, A. Ahmad, S. Y. Zafar, J. L. Wheeler, J. B. Reese, and H. K. Lyerly, "Electronic patient-reported data capture as a foundation of rapid learning cancer care," *Medical Care*, vol. 48, no. 6, pp. S32–S38, 2010.

[17] A. J. Sutton, D. Kendrick, and C. A. C. Coupland, "Meta-analysis of individual- and aggregate-level data," *Statistics in Medicine*, vol. 27, no. 5, pp. 651–669, 2008.

[18] K. R. Olsen and A. Street, "The analysis of efficiency among a small number of organisations: how inferences can be improved by exploiting patient-level data," *Health Economics*, vol. 17, no. 6, pp. 671–681, 2008.

[19] K. K. Jain, "Personalised medicine for cancer: from drug development into clinical practice," *Expert Opinion on Pharmacotherapy*, vol. 6, no. 9, pp. 1463–1476, 2005.

[20] A. E. Cuellar and P. J. Gertler, "Strategic integration of hospitals and physicians," *Journal of Health Economics*, vol. 25, no. 1, pp. 1–28, 2006.

[21] A. Miles, B. Matthews, M. Wilson, and D. Brickley, "SKOS core: simple knowledge organisation for the web," in *Proceedings of the 5th International Conference on Dublin Core and Metadata Applications (DC '05)*, pp. 3–10, Madrid, Spain, September 2005.

[22] J. Ison, M. Kalaš, I. Jonassen et al., "EDAM: an ontology of bioinformatics operations, types of data and identifiers, topics and formats," *Bioinformatics*, vol. 29, no. 10, pp. 1325–1332, 2013.

[23] Innovative Medicines Initiative I, IMI Ongoing Projects, 2015, http://www.imi.europa.eu/content/ongoing-projects.

[24] Publications Office of the European Union (OP), "CORDIS Projects and Results," 2015, http://cordis.europa.eu/projects/home_en.html.

[25] National Institutes of Health (NIH), "NIH Awards," 2015, http://www.report.nih.gov/award/index.cfm.

[26] H.-E. Wichmann, K. A. Kuhn, M. Waldenberger et al., "Comprehensive catalog of European biobanks," *Nature Biotechnology*, vol. 29, no. 9, pp. 795–797, 2011.

[27] F. Mentré, M. Chenel, E. Comets et al., "Current use and developments needed for optimal design in pharmacometrics: a study performed among DDMoRe's european federation of pharmaceutical industries and associations members," *CPT: Pharmacometrics & Systems Pharmacology*, vol. 2, no. 6, pp. 1–2, 2013.

[28] L. Harnisch, I. Matthews, J. Chard, and M. O. Karlsson, "Drug and disease model resources: a consortium to create standards and tools to enhance model-based drug development," *CPT: Pharmacometrics & Systems Pharmacology*, vol. 2, no. 3, article e34, 3 pages, 2013.

[29] A. El Fadly, B. Rance, N. Lucas et al., "Integrating clinical research with the Healthcare Enterprise: from the RE-USE project to the EHR4CR platform," *Journal of Biomedical Informatics*, vol. 44, supplement 1, pp. S94–S102, 2011.

[30] L. C. Crosswell and J. M. Thornton, "ELIXIR: a distributed infrastructure for European biological data," *Trends in Biotechnology*, vol. 30, no. 5, pp. 241–242, 2012.

[31] M. Gottwald, "How can the innovative medicines initiative help to make medicines development more efficient?" in *Re-Engineering Clinical Trials: Best Practices for Streamlining the Development Process*, p. 55, Elsevier, 2014.

[32] P. M. Coloma, M. J. Schuemie, G. Trifirò et al., "Combining electronic healthcare databases in Europe to allow for large-scale drug safety monitoring: the EU-ADR Project," *Pharmacoepidemiology and Drug Safety*, vol. 20, no. 1, pp. 1–11, 2011.

[33] A. J. Webb, G. A. Thorisson, and A. J. Brookes, "An informatics project and online 'Knowledge Centre' supporting modern genotype-to-phenotype research," *Human Mutation*, vol. 32, no. 5, pp. 543–550, 2011.

[34] M. J. Schuemie, R. Gini, P. M. Coloma et al., "Replication of the OMOP experiment in europe: evaluating methods for risk identification in electronic health record databases," *Drug Safety*, vol. 36, no. 1, pp. S159–S169, 2013.

[35] M. Elsner, "OncoTrack tests drugs in virtual people," *Nature Biotechnology*, vol. 29, no. 5, article 378, 2011.

[36] D. Henderson, L. A. Ogilvie, N. Hoyle, U. Keilholz, B. Lange, and H. Lehrach, "Personalized medicine approaches for colon cancer driven by genomics and systems biology: oncoTrack," *Biotechnology Journal*, vol. 9, no. 9, pp. 1104–1114, 2014.

[37] L. Harland, "Open PHACTS: a semantic knowledge infrastructure for public and commercial drug discovery research," in *Knowledge Engineering and Knowledge Management*, A. ten Teije, J. Völker, S. Handschuh et al., Eds., pp. 1–7, Springer, Berlin, Germany, 2012.

[38] R. Thompson, L. Johnston, D. Taruscio et al., "RD-Connect: an integrated platform connecting databases, registries, biobanks and clinical bioinformatics for rare disease research," *Journal of General Internal Medicine*, vol. 29, no. 3, pp. S780–S787, 2014.

[39] M. A. Robb, J. A. Racoosin, R. E. Sherman et al., "The US food and drug administration's sentinel initiative: expanding the horizons of medical product safety," *Pharmacoepidemiology and Drug Safety*, vol. 21, no. 1, pp. 9–11, 2012.

[40] B. M. Psaty and A. M. Breckenridge, "Mini-sentinel and regulatory science—big data rendered fit and functional," *The New England Journal of Medicine*, vol. 370, no. 23, pp. 2165–2167, 2014.

[41] J. L. Oliveira, P. Lopes, T. Nunes et al., "The EU-ADR Web Platform: delivering advanced pharmacovigilance tools," *Pharmacoepidemiology and Drug Safety*, vol. 22, no. 5, pp. 459–467, 2013.

[42] R. Stevens, S. Jupp, J. Klein, and J. Schanstra, "Using semantic web technologies to manage complexity and change in biomedical data," in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC '11)*, pp. 3708–3711, IEEE, Boston, Mass, USA, August-September 2011.

[43] M. Kohl, "Standards, databases, and modeling tools in systems biology," in *Data Mining in Proteomics*, pp. 413–427, Springer, Berlin, Germany, 2011.

[44] R. C. Jiménez and J. A. Vizcaíno, "Proteomics data exchange and storage: the need for common standards and public repositories," in *Mass Spectrometry Data Analysis in Proteomics*, vol. 1007 of *Methods in Molecular Biology*, pp. 317–333, Humana Press, 2013.

[45] C. M. Machado, D. Rebholz-Schuhmann, A. T. Freitas, and F. M. Couto, "The semantic web in translational medicine: current applications and future directions," *Briefings in Bioinformatics*, vol. 16, no. 1, pp. 89–103, 2015.

[46] P. Pecina, O. Dušek, L. Goeuriot et al., "Adaptation of machine translation for multilingual information retrieval in the medical domain," *Artificial Intelligence in Medicine*, vol. 61, no. 3, pp. 165–185, 2014.

[47] S. T. Rosenbloom, J. C. Denny, H. Xu, N. Lorenzi, W. W. Stead, and K. B. Johnson, "Data from clinical notes: a perspective on the tension between structure and flexible documentation," *Journal of the American Medical Informatics Association*, vol. 18, no. 2, pp. 181–186, 2011.

[48] D. Rebholz-Schuhmann, A. Oellrich, and R. Hoehndorf, "Text-mining solutions for biomedical research: enabling integrative biology," *Nature Reviews Genetics*, vol. 13, no. 12, pp. 829–839, 2012.

[49] S. C. Schuster, "Next-generation sequencing transforms today's biology," *Nature Methods*, vol. 5, no. 1, pp. 16–18, 2007.

[50] E. R. Mardis, "The impact of next-generation sequencing technology on genetics," *Trends in Genetics*, vol. 24, no. 3, pp. 133–141, 2008.

[51] M. Via, C. Gignoux, and E. G. Burchard, "The 1000 Genomes Project: new opportunities for research and social challenges," *Genome Medicine*, vol. 2, article 3, 2010.

[52] D. I. Boomsma, C. Wijmenga, E. P. Slagboom et al., "The Genome of the Netherlands: design, and project goals," *European Journal of Human Genetics*, vol. 22, no. 2, pp. 221–227, 2014.

[53] J. S. Ware, A. M. Roberts, and S. A. Cook, "Next generation sequencing for clinical diagnostics and personalised medicine: implications for the next generation cardiologist," *Heart*, vol. 98, no. 4, pp. 276–281, 2012.