Exploratory study
2/2017

Biomedical knowledge production in the age of big data

Analysis conducted on behalf of the Swiss Science and Innovation Council SSIC

—

Professor Sabina Leonelli, Philosophy of Science, University of Exeter, UK

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

**Schweizerischer Wissenschafts- und Innovationsrat
Conseil suisse de la science et de l’innovation
Consiglio svizzero della scienza e dell’innovazione
Swiss Science and Innovation Council**

The Swiss Science and Innovation Council

The Swiss Science and Innovation Council SSIC is the advisory body to the Federal Council for issues related to science, higher education, research and innovation policy. The goal of the SSIC, in conformity with its role as an independent consultative body, is to promote the framework for the successful development of the Swiss higher education, research and innovation system. As an independent advisory body to the Federal Council, the SSIC pursues the Swiss higher education, research and innovation landscape from a long-term perspective.

Der Schweizerische Wissenschafts- und Innovationsrat

Der Schweizerische Wissenschafts- und Innovationsrat SWIR berät den Bund in allen Fragen der Wissenschafts-, Hochschul-, Forschungs- und Innovationspolitik. Ziel seiner Arbeit ist die kontinuierliche Optimierung der Rahmenbedingungen für die gedeihliche Entwicklung der Schweizer Bildungs-, Forschungs- und Innovationslandschaft. Als unabhängiges Beratungsorgan des Bundesrates nimmt der SWIR eine Langzeitperspektive auf das gesamte BFI-System ein.

Le Conseil suisse de la science et de l'innovation

Le Conseil suisse de la science et de l'innovation CSSI est l'organe consultatif du Conseil fédéral pour les questions relevant de la politique de la science, des hautes écoles, de la recherche et de l'innovation. Le but de son travail est l'amélioration constante des conditions-cadre de l'espace suisse de la formation, de la recherche et de l'innovation en vue de son développement optimal. En tant qu'organe consultatif indépendant, le CSSI prend position dans une perspective à long terme sur le système suisse de formation, de recherche et d'innovation.

Il Consiglio svizzero della scienza e dell'innovazione

Il Consiglio svizzero della scienza e dell'innovazione CSSI è l'organo consultivo del Consiglio federale per le questioni riguardanti la politica in materia di scienza, scuole universitarie, ricerca e innovazione. L'obiettivo del suo lavoro è migliorare le condizioni quadro per lo spazio svizzero della formazione, della ricerca e dell'innovazione affinché possa svilupparsi in modo armonioso. In qualità di organo consultivo indipendente del Consiglio federale il CSSI guarda al sistema svizzero della formazione, della ricerca e dell'innovazione in una prospettiva globale e a lungo termine.

About the author

Sabina Leonelli is Professor in the Philosophy and History of Science at the University of Exeter, where she co-directs the Exeter Centre for the Study of the Life Sciences and leads the "Data Studies" research strand. Her research, currently funded by a European Research Council award, focuses on the philosophy of data-intensive science, especially the methods, outputs and social embedding of open and big data, and the ways in which they are redefining what counts as research and knowledge. She published over fifty papers in high-ranking journals within biology as well as the philosophy, social studies and history of science, and is the author of *Data-Centric Biology: A Philosophical Study* (2016, Chicago University Press). She has been invited to present her work to scholars and policy-makers around the world, most recently in venues such as the American Association for the Advancement of Science, the World Conference on the Future of Science, the Open Science EU Presidency Conference and the World Science Forum.

The research underpinning this report was funded by the European Research Council (grant award 335925), the Leverhulme Trust (award RPG-2013-153), the Australian Research Council (award DP160102989), the UK Medical Research Council and Natural Environment Research Council (award MR/K019341/1) and the UK Economic and Social Research Council (award ES/P011489/1). The author is very grateful to Lora E. Fleming and Niccolo Tempini for helpful comments on the manuscript, and Rachel Ankeny, Brian Rappert and Barbara Prainsack for relevant discussions.

The author contracted by the Swiss Science and Innovation Council to produce the present paper bears full responsibility for its contents.

Table of contents

Preface by the SSIC

Investigating the notions of health and disease is an important topic of the Swiss Science and Innovation Council's 2016–2019 working programme. This includes asking under what conditions data-centric research can be a basis for statements about individuals and their background. In her analysis for the Swiss Science and Innovation Council (SSIC) on "Biomedical knowledge production in the age of big data", Prof. Sabina Leonelli looks at the extent to which the emergence of big data is altering biomedical research practices and findings, and highlights the range of challenges confronting researchers. In Leonelli's view, scientific approaches are influenced by ethical, institutional, financial and technical problems.

Building on these considerations, which the SSIC is pleased to present here to a wider readership, the Council decided on 26 June 2017 to discuss certain key issues, focussing on the Swiss context. It will pursue this perspective over the coming two years, consulting with Swiss players and continuing to bring in international expert knowledge. It does so with the view that there are challenges to face as well as opportunities to grasp. Technological developments such as the creation of interoperable databases and machine learning can be of great value, but the main determinant of a successful and sustainable development for biomedicine is ultimately the social environment, in particular institutional, financial and legal factors, and, not least, ethical aspects. The Council will also consider fundamental questions of data access and contextualisation which extend the frame of biomedical and health-related concerns.

Vorwort des SWIR

Für den Schweizerischen Wissenschafts- und Innovationsrat (SWIR) ist die Auseinandersetzung mit dem Verständnis von Gesundheit und Krankheit ein wichtiger Themenbereich des Arbeitsprogramms 2016–2019. Dazu gehört die Frage nach den Voraussetzungen, unter denen es möglich ist, mittels datenzentrierter Forschungsansätze Aussagen über ein Individuum in seinem Lebenszusammenhang zu machen. Die Analyse «Wissensproduktion in der Biomedizin im Zeitalter von Big Data», die Prof. Dr. Sabina Leonelli für den SWIR verfasst hat, bietet einen Überblick darüber, inwiefern das Aufkommen von Big Data biomedizinische Forschungspraktiken und -ergebnisse verändert, und zeigt die Vielfalt der Herausforderungen, mit denen Forschende konfrontiert sind. Dabei sind für Leonelli wissenschaftliche Betrachtungsweisen auch von ethischen, institutionellen, finanziellen und technischen Problemfeldern geprägt.

Ausgehend von diesen Überlegungen, die der SWIR mit der vorliegenden Publikation gerne weiteren interessierten Leserinnen und Lesern zur Verfügung stellt, hat der Rat am 26. Juni 2017 die Diskussion über zentrale Fragen im schweizerischen Kontext begonnen. Er wird seine Arbeit in den kommenden zwei Jahren vertiefen, sich mit Schweizer Akteuren auseinandersetzen und weiterhin auch internationales Wissen einbeziehen. Er tut dies mit der Haltung, dass Herausforderungen zu bewältigen, aber auch Chancen zu ergreifen sind. Technologische Entwicklungen, die die Schaffung interoperabler Datenbanken und maschinelles Lernen beinhalten, können von grossem Nutzen sein, doch entscheidend für eine erfolgreiche und nachhaltige Entwicklung der Biomedizin sind letztlich soziale Umstände, insbesondere institutionelle sowie finanzielle und rechtliche Faktoren und nicht zuletzt ethische Aspekte. Der Rat wird auch grundsätzliche Überlegungen zu Datenzugang und Kontextualisierung einbeziehen, über biomedizinische und gesundheitsbezogene Perspektiven hinaus.

Préface du CSSI

Pour le Conseil suisse de la science et de l'innovation (CSSI), approfondir les notions de santé et de maladie est un thème important du programme de travail 2016–2019. Il se pose notamment la question des conditions permettant de produire des affirmations concernant un individu dans son contexte de vie grâce aux approches de recherche centrées sur les données. L'analyse «La production de connaissances biomédicales à l'ère du Big Data», réalisée par le Prof. Sabina Leonelli pour le CSSI, dresse un aperçu des changements générés par l'émergence du Big Data dans les pratiques et les résultats de recherche biomédicaux et montre la diversité des défis auxquels sont confrontés les chercheurs. Pour le Prof. Leonelli, les perspectives scientifiques sont également marquées par des problèmes éthiques, institutionnels, financiers et techniques.

Sur la base de ces réflexions, que le CSSI tient à disposition d'un cercle élargi avec la présente publication, le Conseil a décidé le 26 juin 2017 d'entamer une discussion sur les questions principales qui se posent dans le contexte suisse. Dans l'optique de relever les défis, mais aussi de saisir les chances qui s'offrent, il approfondira son travail au cours des deux prochaines années, consultera les acteurs suisses et continuera d'intégrer des éclairages internationaux. Les évolutions techniques telles que la création de banques de données interopérables et l'apprentissage automatique peuvent être très utiles, mais le développement durable et réussi de la biomédecine dépendra de facteurs sociaux, institutionnels, financiers, juridiques et, en fin de compte, de questions éthiques. Le Conseil s'intéressera également aux aspects fondamentaux de l'accès aux données et de leur contextualisation, élargissant le cadre de la réflexion au-delà des perspectives biomédicales et liées à la santé.

Prefazione del CSSI

Per il Consiglio svizzero della scienza e dell'innovazione (CSSI) l'approfondimento di nozioni di salute e malattia rappresenta un'importante ambito tematico nel programma di lavoro 2016–2019. Questo comprende anche un'indagine delle condizioni nelle quali metodi di ricerca incentrati su dati possono essere una base atta a formulare ipotesi sull'individuo e sul suo contesto di vita. L'analisi «Produzione di conoscenze biomediche nell'era dei big data», redatta per il CSSI dalla prof.ssa Sabina Leonelli, fornisce un quadro dell'influsso dei big data sulle pratiche e sui risultati di ricerca biomedica, illustrando le diverse sfide che i ricercatori si trovano ad affrontare. Secondo Leonelli, gli approcci scientifici sono condizionati anche da questioni etiche, istituzionali, finanziarie e tecniche.

Sulla base di tali considerazioni, che il CSSI mette a disposizione degli interessati tramite la presente pubblicazione, il 26 giugno 2017 il Consiglio ha avviato una discussione su questioni centrali nel contesto svizzero. Al fine di affrontare le sfide e cogliere le opportunità in quest'ambito, nei prossimi due anni il CSSI intensificherà la propria attività, consultando attori chiave in Svizzera e integrando le conoscenze che maturano a livello internazionale. Gli sviluppi tecnologici, che includono la creazione di banche dati interoperabili e l'apprendimento automatico, possono risultare molto utili, ma in ultima analisi sono le circostanze sociali, in particolare i fattori istituzionali, finanziari, giuridici nonché etici, a essere decisive per uno sviluppo efficace e sostenibile della biomedicina. Il Consiglio terrà conto anche di considerazioni fondamentali sull'accesso ai dati e sulla loro contestualizzazione, che vanno oltre le prospettive biomediche e sanitarie.

Executive summary

Starting from an understanding of *data* as “any product of research activities that is collected, stored, and disseminated in order to be used as evidence for knowledge claims”[1], the author notes in the present analysis that the volume of what constitutes *big data* has grown in parallel to technological possibilities. Although scientists have dealt with large datasets for a long time, a characteristic of today’s situation is that data are now considered a scientific output in their own right, to be made publicly available regardless of whether they have been used as evidence for a particular hypothesis. Several funding agencies are experimenting with rewards systems designed to encourage data publication. In this context, new approaches for data management and governance are needed, based on appropriate infrastructures such as digital repositories.

Different epistemic communities hold divergent criteria on what makes data trustworthy. Within biomedicine, the *evidence-based medicine* movement is grappling with how to integrate disparate types and categories of data while maintaining the tight experimental control which is central to clinical trials. The author identifies four trends developing against this background:

- *Personalized or precision medicine*, which relies on the availability of an extensive and diverse body of evidence in order to target research and therapeutic intervention to specific individuals and groups;

- *Health and environment data integration*, including investigations on climate change and its biomedical implications as well as biomonitoring cohort studies;

- *Self-tracking* and the frequently associated habit of deliberately disseminating personal data through social media;

- *Open health data*, designating a concern with the asymmetry between the quantity of publicly available data and the much larger volume of data owned by private companies.

Biomedical subfields differ in their methods and assumptions, which reflect the complexity of the biological world. Therefore, a central challenge of relating diverse types of data is the preservation of system-specific knowledge tied to particular formats, instruments and methods. These constitute an essential context for data interpretation. The *interoperability* principle describes a delicate balancing act between standardization and flexibility to domains-specific requirements.

Because data are selected and organized by researchers, they remain theory-laden. Thus, decisions on the choice of metadata to report have a profound impact on the epistemic value of a data source. Also of high importance for the biomedical context is the link between data and material samples, as in the case of biobanks.

A key challenge for sustainable data governance will be the identification of mechanisms for allocating responsibilities across the highly interconnected network of data dissemination. To this aim, it is crucial to recognize that ethical and social considerations are part and parcel of extracting biomedical knowledge from big data.

To develop and maintain a reliable biomedical data infrastructure requires a team of individuals with competence in information technology and programming, an understanding of the characteristics of the data stored, an awareness of the needs of prospective users, as well as the legal and ethical implications of their activities. Such tasks cannot be fully assumed by automatic intelligence. Furthermore, performance metrics are called into question for failing to acknowledge this collaborative work and for rewarding the publication of articles over efforts to curate and disseminate data, materials and software.

Another issue of high importance is to design sustainable funding models capable of underpinning data collection and storage. A database that becomes obsolete threatens the reliability of all the data collections to which it is linked. Therefore, database contents need to be updated regularly to reflect both technological and scientific advances.

[1] Leonelli S., 2016, Data-Centric Biology: A Philosophical Study. Chicago, IL: Chicago University Press, p. 77.

Executive Summary

Daten als sämtliche Produkte aus Forschungstätigkeiten, die gesammelt, gespeichert und verbreitet werden, um als Nachweis für Wissensansprüche zu dienen[2] – ausgehend von diesem Verständnis hält die Autorin der vorliegenden Analyse fest, dass das Volumen von *Big Data* parallel zu den technologischen Möglichkeiten gewachsen ist. Wissenschaftlerinnen und Wissenschaftler arbeiten zwar seit Langem mit grossen Datenmengen. Neu ist aber, dass Daten heutzutage als eigenständiger wissenschaftlicher Output betrachtet werden. Und sie sollen, unabhängig davon, ob sie als Evidenz für eine bestimmte Hypothese verwendet wurden, öffentlich zugänglich gemacht werden. Verschiedene Förderagenturen testen derzeit Anreizsysteme, um die Publikation der Daten zu unterstützen. In diesem Zusammenhang braucht es neue Ansätze für das Datenmanagement und die Datengovernance sowie entsprechend geeignete Infrastrukturen wie digitale Repositories.

Welche Kriterien Daten vertrauenswürdig machen, wird von verschiedenen Wissensgemeinschaften unterschiedlich beurteilt. Innerhalb der Biomedizin bemüht sich die *evidenzbasierte Medizin* darum, unterschiedliche Typen und Kategorien von Daten einzubeziehen und gleichzeitig die strenge Kontrolle über die Experimente zu bewahren, die für klinische Studien unabdingbar ist. Die Autorin macht vor diesem Hintergrund vier Trends aus:

- *Personalisierte Medizin oder Präzisionsmedizin*, die darauf beruht, dass eine umfangreiche und vielfältige Menge an Evidenzen verfügbar ist, um die Forschung und therapeutische Interventionen auf bestimmte Personen oder Gruppen auszurichten;

- *Integration von Gesundheits- und Umweltdaten*, die Untersuchungen zum Klimawandel und dessen biomedizinischen Auswirkungen wie auch Biomonitoring-Kohortenstudien einbezieht;

- *Self-Tracking* und die oft damit verbundene Gewohnheit, bewusst persönliche Daten über soziale Medien zu verbreiten;

- *Bewegung für offene Gesundheitsdaten*, die die Bedenken bezüglich der Asymmetrie zwischen der öffentlich verfügbaren Datenmenge und dem deutlich grösseren Volumen an Daten im Besitz privater Unternehmen ausdrückt.

Die einzelnen biomedizinischen Felder unterscheiden sich in den verwendeten Methoden und getroffenen Annahmen, was die Komplexität der biologischen Welt verdeutlicht. Eine zentrale Herausforderung bei der Verbindung verschiedener Datentypen besteht deshalb darin, das systemspezifische Wissen zu bewahren, das an bestimmte Formate, Instrumente und Methoden gebunden ist. Denn diese liefern einen wesentlichen Hintergrund für die Dateninterpretation. Das Prinzip der *Interoperabilität* beschreibt den heiklen Balanceakt zwischen Standardisierung und Flexibilität für bereichsspezifische Anforderungen.

Da die Daten von Forschenden ausgewählt und organisiert werden, bleiben sie theoriegebunden. Die Auswahl der mitzuliefernden Metadaten hat somit einen wesentlichen Einfluss auf den erkenntnistheoretischen Wert einer Datenquelle. Grosse Bedeutung kommt im biomedizinischen Kontext auch der Verknüpfung von Daten und Materialproben zu, beispielsweise bei Biobanken.

Zentral für die nachhaltige Governance von Daten wird sein, Mechanismen zu finden, um in der Datenverbreitung mit ihrem hohen Vernetzungsgrad Verantwortlichkeiten zuzuordnen. Dafür ist zwingend anzuerkennen, dass ethische und soziale Überlegungen ein wesentlicher Teil der Extrahierung von biomedizinischem Wissen aus Big Data sind.

Entwicklung und Unterhalt einer verlässlichen biomedizinischen Dateninfrastruktur erfordern ein Team von kompetenten Personen im Bereich IT und Programmierung, ein Verständnis für die Besonderheiten der gespeicherten Daten sowie ein Bewusstsein für die Bedürfnisse der potenziellen Nutzerinnen und Nutzer wie auch die rechtlichen sowie ethischen Auswirkungen der Tätigkeiten. Solche Aufgaben können nicht vollständig automatisiert werden. Zudem werden Leistungsmesssysteme infrage gestellt, weil sie Zusammenarbeit nicht ausreichend würdigen und die Publikation von Artikeln mehr honorieren als Bemühungen zur Pflege und Verbreitung von Daten, Material und Software.

Ebenfalls ein wichtiges Thema ist die Konzeption nachhaltiger Finanzierungsmodelle, um die Datensammlung und -speicherung zu unterstützen. Eine veraltete Datenbank gefährdet die Verlässlichkeit aller damit zusammenhängenden Datensammlungen. Deshalb müssen Datenbankinhalte regelmässig aktualisiert werden, um technologischen wie auch wissenschaftlichen Fortschritten Rechnung zu tragen.

[2] Data as “any product of research activities that is collected, stored, and disseminated in order to be used as evidence for knowledge claims”, Leonelli S., 2016, Data-Centric Biology: A Philosophical Study. Chicago, IL: Chicago University Press, p. 77.

Résumé

Se fondant sur une notion de *donnée* comme tout produit d'activités de recherche recueilli, stocké et disséminé dans le but d'être utilisé comme preuve des connaissances[3], l'auteur constate dans la présente analyse que le volume de ce qui constitue le *Big Data* a pris de l'ampleur parallèlement aux possibilités technologiques. Bien que les scientifiques travaillent depuis longtemps avec de grands ensembles de données, l'une des caractéristiques de la situation actuelle est que ces données sont aujourd'hui considérées comme des résultats scientifiques à part entière. Elles devraient donc être rendues publiques, qu'elles aient ou non servi de preuve pour une hypothèse particulière. Différents organismes de financement testent des mesures incitatives visant à encourager la publication de données. Un tel contexte appelle la mise en place de nouvelles approches pour la gestion et la gouvernance des données en se fondant sur des infrastructures appropriées telles que des répertoires numériques.

Les diverses communautés épistémiques utilisent des critères différents pour juger de la fiabilité des données. Au sein de la biomédecine, le mouvement de la *médecine basée sur des preuves* recherche les moyens d'intégrer des catégories et des types de données différents sans perdre la rigueur du contrôle empirique qui assure la qualité des essais cliniques. Face à cette toile de fond, l'auteur identifie quatre tendances qui se développent:

- *la médecine personnalisée ou de précision*, qui s'appuie sur l'accumulation d'un ensemble exhaustif et diversifié d'évidences dans le but de cibler la recherche et les interventions thérapeutiques à l'intention de personnes et de groupes spécifiques;

- *l'intégration des données sur la santé et l'environnement*, y compris les enquêtes sur le changement climatique et ses implications biomédicales, ainsi que les études de cohortes dans le domaine de la biosurveillance;

- *le self-tracking*, fréquemment associé à l'habitude de disséminer délibérément des données personnelles sur les médias sociaux;

- *le mouvement «open health data»*, qui traduit une préoccupation par rapport à l'asymétrie entre la quantité de données disponibles publiquement et le volume beaucoup plus important de données détenues par des entreprises privées.

Les sous-domaines biomédicaux mobilisent des méthodes et présuppositions variées qui reflètent la complexité du monde biologique. Mettre en relation différents types de données soulève donc un défi majeur: celui de préserver les connaissances spécifiques liées à des méthodes, formats et instruments particuliers. Ceux-ci représentent un cadre essentiel pour l'interprétation des données. Le principe d'*interopérabilité* décrit un équilibre délicat entre standardisation et flexibilité par rapport aux exigences spécifiques selon la discipline.

Du fait que les données sont sélectionnées et organisées par les chercheurs, elles restent imprégnées de théorie. Les décisions sur le choix de métadonnées ont donc un impact profond sur la valeur épistémique d'une source de données. Le lien entre données et échantillons de matériaux, comme dans le cas des biobanques, est également très important pour le domaine biomédical.

L'un des principaux défis relatifs à une gouvernance des données durable consistera à trouver des mécanismes d'attribution des responsabilités dans le réseau très interconnecté à travers lequel ces données sont diffusées. A cette fin, il faut reconnaître que les considérations éthiques et sociales font partie intégrante de l'extraction des connaissances biomédicales à travers le Big Data.

Le développement et le maintien d'une infrastructure fiable de données biomédicales exigent une équipe de personnes qui possèdent des compétences en technologie de l'information et en programmation, qui comprennent les caractéristiques des données stockées et qui ont conscience des besoins des utilisateurs potentiels ainsi que des implications légales et éthiques de leurs activités. De telles tâches ne peuvent pas être entièrement prises en charge par l'intelligence automatique. En outre, les systèmes de mesure de la performance sont remis en question parce qu'ils ne reconnaissent pas suffisamment ce travail collaboratif et tendent à récompenser la publication d'articles plutôt que les efforts entrepris pour organiser et diffuser les données, les matériels et les logiciels.

La conception de modèles de financement durables capables d'étayer la collecte et le stockage de données est une autre question primordiale. Une base de données qui devient obsolète menace la fiabilité de toutes les collections de données auxquelles elle est liée. Il faut dès lors que les contenus des bases de données soient mis à jour régulièrement pour pouvoir refléter les avancées technologiques et scientifiques.

[3] Data as "any product of research activities that is collected, stored, and disseminated in order to be used as evidence for knowledge claims", Leonelli S., 2016, Data-Centric Biology: A Philosophical Study. Chicago, IL: Chicago University Press, p. 77.

Riassunto

Considerando un *dato* come un qualsiasi prodotto di attività di ricerca raccolto, archiviato e diffuso ai fini della dimostrazione di ipotesi[4], nella presente analisi l'autrice rileva che il volume di *big data* è cresciuto di pari passo con le possibilità tecnologiche. Sebbene i ricercatori abbiano lavorato per molto tempo con grandi dataset, oggi i dati sono considerati un prodotto scientifico a sé stante, destinato a essere reso disponibile al pubblico a prescindere dal fatto che siano stati usati o meno per verificare determinate ipotesi. Diversi enti di finanziamento sperimentano sistemi volti a incoraggiare la pubblicazione di dati. In questo contesto sono necessari nuovi approcci per la gestione e l'organizzazione di dati, basati su infrastrutture adeguate, come gli archivi digitali.

I criteri per stabilire l'affidabilità dei dati divergono tra le varie comunità epistemiche. Nel settore della biomedicina, la *medicina basata su prove di evidenza* dibatte su come integrare diversi tipi e categorie di dati mantenendo al contempo il rigido controllo sperimentale, essenziale per gli studi clinici. L'autrice individua quattro tendenze in questo contesto:

- *medicina personalizzata o di precisione*, basata sulla disponibilità di un ampio e vario insieme di prove, al fine di indirizzare la ricerca e l'intervento terapeutico su determinati individui e gruppi;
- *integrazione di dati sanitari e ambientali*, inclusi gli studi sul cambiamento climatico e le relative implicazioni biomediche nonché gli studi di biomonitoraggio di coorte;
- *self-tracking*, frequentemente associato alla consuetudine di diffondere deliberatamente dati personali attraverso i social media;
- *dati sanitari accessibili al pubblico*, che evidenziano una disparità rispetto al volume nettamente superiore di dati in possesso di aziende private.

I sottosettori biomedici presentano differenze nei metodi e nei presupposti, che riflettono la complessità del mondo biologico. Pertanto una delle problematiche centrali dell'integrazione di diversi tipi di dati consiste nel preservare le conoscenze specifiche di sistema, connesse a determinati formati, strumenti e metodi. Sono queste conoscenze a costituire il contesto essenziale per l'interpretazione dei dati. Il principio di *interoperabilità* descrive dunque un delicato gioco di equilibrio tra standardizzazione e flessibilità rispetto ai requisiti specifici delle discipline.

Essendo selezionati e organizzati da ricercatori, i dati sono impregnati di teoria. Di conseguenza le decisioni sulla scelta di metadati da riportare hanno un profondo impatto sul valore epistemico della fonte. Nel contesto biomedico assume inoltre particolare importanza il collegamento tra dati e campioni di materiali, come nel caso delle biobanche.

Un aspetto cruciale per la gestione sostenibile di dati è rappresentato dall'individuazione di meccanismi che permettono di attribuire responsabilità all'interno della rete altamente interconnessa di diffusione degli stessi. A tal fine è essenziale riconoscere che le considerazioni etiche e sociali sono parte integrante dell'estrazione di informazioni biomediche dai big data.

Per sviluppare e mantenere un'infrastruttura affidabile di dati biomedici è necessario un team di individui con competenze informatiche e di programmazione, che comprendano le caratteristiche dei dati archiviati e siano consapevoli delle esigenze dei potenziali utenti e delle implicazioni legali ed etiche delle loro attività. Tali compiti non possono essere svolti interamente da sistemi automatici. Vengono inoltre messi in discussione gli indicatori di prestazione, poiché non riconoscono questo lavoro collaborativo e premiano la pubblicazione di articoli anziché gli sforzi volti alla gestione e alla diffusione di dati, materiali e software.

Un altro aspetto fondamentale è costituito dallo sviluppo di modelli di finanziamento sostenibile allo scopo di favorire la raccolta e l'archiviazione di dati. Poiché un database obsoleto mette a repentaglio l'affidabilità di tutte le raccolte di dati alle quali è collegato, il suo contenuto deve essere aggiornato regolarmente sulla base dei progressi tecnologici e scientifici.

[4] Data as "any product of research activities that is collected, stored, and disseminated in order to be used as evidence for knowledge claims", Leonelli S., 2016, Data-Centric Biology: A Philosophical Study. Chicago, IL: Chicago University Press, p. 77.

Introduction

1

Introduction

Big data are widely seen as a game-changer for social relations, communication and governance around the world. The emergence of big data also promises to revolutionise the production of knowledge within and beyond academia, by enabling new and more efficient ways to plan, conduct, institutionalise, disseminate and assess research.[5] The ability to link and cross-reference datasets coming from different sources is expected to increase the accuracy and predictive power of scientific findings, and help researchers to identify future directions of inquiry. The availability of vast amounts of data provides an incentive to search for intelligent procedures and tools to store, organise and analyse these data, so as to improve the reliability and transparency of scientific knowledge creation. There are therefore strong incentives for researchers to find ways to adequately manage big data at every stage of the research process.

This report examines the extent to which the emergence of big data is transforming research practices and outcomes in biomedicine, and the implications of this transformation for researchers in this area. It is divided into three parts. The first part consists of an introduction to the characterisations of big data currently employed in the scientific literature, and the ways in which these definitions fit broader shifts in the status and use of data for research purposes. The second part reviews the opportunities and challenges arising for biomedical researchers in relation to the management and interpretation of big data, focusing particularly on *technical*, *ethical*, *financial* and *institutional* concerns – and the extent to which such concerns overlap in scientific practice. The third part reflects on how big data infrastructures and skills can be organised at the national and international levels to support a data-centric approach to research, and identifies five principles underpinning the effective and sustainable use of big data in biomedicine.

[5] Hey et al. 2009, Mayer-Schönberger and Cukier 2013, Royal Society 2012, Science International 2015.

Relevant background

2

2.1
Big data and the rise of data-centric science

As noted by Shutt and O'Neil (2015, p. 24) in their Introduction to *Doing Data Science*, there are multiple ways to define big data, each of which captures some relevant aspects.[6] Perhaps the most straightforward approach is to characterise big data as large datasets that are produced in a digital form and can be analysed through computational tools. Many commentators focusing on the use of data in research, however, view this definition as overly narrow and misguided in its emphasis on data size and format, especially given the importance of data *provenance* (that is, the conditions under which data were generated and disseminated) to processes of inference and interpretation, the *diversity* of data types used by researchers (which may include data that are not generated in digital formats, or data whose format is not computationally tractable) and the *dependence* of data use on specific queries, skills and research *contexts*. A more popular alternative is thus to define big data not by reference to their physical attributes, but rather by virtue of what can and cannot be done with them. An example of such an approach is provided by Boyd and Crawford, when they identify big data with "the capacity to search, aggregate and cross-reference large datasets".[7]

Yet another popular way to characterise big data is to point to a cluster of features that they need to possess. These typically include various combinations of the following seven "Vs"[8].

Volume

This is the size of datasets being handled. It is worth stressing that what constitutes a "large volume" depends on the technical means available to generate, store, disseminate and visualize the data, which are themselves evolving rapidly. Thus, what constituted a large volume of data in the 1990s may not be considered such today, and contemporary views of big data as "anything that cannot be easily captured in an Excel spreadsheet"[9] are bound to shift rapidly as new analytic software becomes established. This is clearly exemplified by technological developments surrounding the production, storage and dissemination of genomic sequencing data, where the volume being handled has dramatically increased within the last two decades, thus considerably raising the bar for what constitutes a challenging amount of data.

[6] Kitchin and McArdle 2016 identify as many as 26 ways of defining big data within the scientific literature, most of which also focus on their use. See also the various definitions listed by https://datascience.berkeley.edu/what-is-big-data/

[7] Boyd and Crawford 2012, p. 663.

[8] For discussions of the "Vs" used to characterise big data, see: Nordmandeau 2013, Ward and Barker 2013, Mayer-Schönberger and Cukier 2013, Marr 2015, Kitchin 2014, Khoury and Ioannidis 2014, Borgman 2015.

[9] http://www.cmswire.com/cms/information-management/what-is-big-data-anything-that-wont-fit-in-excel-emetrics-020502.php

Velocity

This refers to the speed at which data are generated as well as, where relevant, the regularity with which this happens. Again, what velocity is considered to be high enough for data to qualify as "big" depends on the availability of relevant technology and related processing skills, as exemplified by the innovation brought by high-throughput experimental instruments such as gene expression microarrays or whole genome screening.

Variety

Data come in countless formats and are produced for vastly diverse purposes. This is particularly true of the biomedical realm, where relevant data may include objects as different as samples of animal tissue, physician description of a patient's symptoms, humidity measurements, GPS coordinates, genome sequences and the results of a blood test.[10] Within such a broad landscape, big data use involves the ability to interrogate and interrelate diverse types of data, with the aim to be able to consult them as a single body of evidence.

Veracity

Data with high volume, velocity and variety are at significant risk of containing bias, inaccuracies and noise. In the absence of appropriate validation and quality checks, this could result in a misleading or outright incorrect interpretation of the reality that the data are meant to document. Veracity concerns the extent to which the quality and reliability of big data can be guaranteed, for instance through reports of bias and abnormalities in the data that are being collected and analysed. Upholding the veracity of big data is relevant particularly within research based on secondary data analysis, where the same data are being re-used for a variety of purposes, and the parameters for what counts as reliable and trustworthy data may change accordingly.[11] This is however complicated by two issues: 1. there are no common criteria within biomedicine to sort "good" from "bad" data, since different epistemic communities within biomedicine hold diverging criteria on what makes data trustworthy;[12] and 2. veracity can only be determined in relation to specific research goals and questions, since what constitutes noise in one situation may well count as data in another.[13]

Validity

Following on from the previous characteristic, validity indicates the selection of appropriate data with respect to the intended use, thus ensuring that adequate and explicit background knowledge and rationale underpin the choice of a given dataset as empirical ground for a given investigation.

Volatility

The extent to which data can be relied upon to remain available, accessible and re-usable as time goes by. This is significant given the tendency of data formats and of the instruments that produce and analyse data to become obsolete, and the efforts required to update and maintain data infrastructures so as to guarantee that data are adequately stored and managed – and thus remain accessible and valid – in the long term.

Value

This is perhaps the most difficult characteristic of big data to identify and describe, and yet it is the most significant. While big data are widely agreed to be valuable, the type of value attributed to them may differ widely across sectors of society, with significant implications for determining data access and re-use. Researchers certainly invest data with *scientific* value as potential sources of evidence for knowledge claims. They may also attribute *financial* value to data, as products of previous investments (of time and resources) or as currencies of exchange with peers; and *ethical* and *social* value, particularly in the case of data accumulated on human subjects who may be viewed as tokens of personal identity. Researchers may also view data as valuable in an *affective* sense, for instance as symbols of their reputation, authority and/or expertise.[14] The institutions, funding bodies, governmental agencies, private companies and patient groups involved in research typically have their own ways of valuing data, which may not overlap with the priorities of researchers. Identifying and negotiating different forms of data value is an unavoidable part of managing and interpreting big data, since these valuation practices determine which data is made available to whom, under which conditions and for which purposes.

[10] I am defining data as "any product of research activities, ranging from artefacts such as photographs to symbols such as letters or numbers, that is collected, stored, and disseminated in order to be used as evidence for knowledge claims" (Leonelli 2016a, p. 77; see chapter 3 of this book for a discussion of this relational approach).

[11] Cai and Zhu 2015, Floridi and Illari 2014.

[12] Leonelli 2012, Lagoze 2014.

[13] McAllister 2007, Loettgers 2009, Woodward 2015.

[14] Leonelli 2016a, Tempini forthcoming.

These characteristics constitute key challenges for researchers wishing to use big data towards producing new knowledge. As many historians and practitioners have pointed out, these challenges are not new to the history of science.[15] Fields such as astronomy, meteorology and taxonomy have long grappled with how to manage, order and visualise large and complex datasets. Many subfields of biomedical research – such as epidemiology, pharmacology and public health – have also an extensive tradition of tackling data of high volume, velocity, variety and volatility, whose validity, veracity and value are regularly negotiated and contested by patients, governments, funders, pharmaceutical companies, insurances and public institutions. These efforts spurred key developments in the techniques, institutions and instruments used to collect, order and visualise health-related data, including the international standardisation of medical terminology via thesauri, registries and databases, the creation of guidelines and legislation for the management of confidential data, and techniques to integrate and sustain diverse data collected over long periods of time.[16]

What makes the contemporary emergence of big data discourse and practices both novel and revolutionary to scientific knowledge production is, therefore, not only or even primarily the existence of large datasets or the intention to efficiently assemble and analyse data from a wide variety of sources. It is rather the result of two broad features of contemporary research. The first is the *change of status of data* from a mere by-product of the research process to research outputs in their own right. Ever since the creation of scientific journals such as *Philosophical Transactions of the Royal Society* in the 17th century, data have been conceptualised and managed as fundamentally private objects, which are owned by the scientists who produce them. Only a small sample of these objects, claimed by scientists to provide convincing evidence for a given claim, has been made publicly available for validation; and such scrutiny has typically been delegated to a small circle of experts via the process of peer review. Within this approach, the usefulness of data lays in their function as evidence for a specified hypothesis. This perception has shifted in the last decade, with data increasingly portrayed as research components that should be made publicly available regardless of whether or not they have been used as evidence for a particular hypothesis, with no boundaries on who can access, scrutinise and interpret the data. Rather than the birth of a data-driven method, we are witnessing the rise of a *data-centric approach to research*, in which efforts to mobilise, integrate, disseminate and visualise data are viewed as central contributions to discovery.[17]

This has implications for how research is conducted, organised, governed and assessed, and brings us to the second distinctive feature of contemporary big data: *the emergence of new approaches to data management*, including *technologies* for the production and communication of data as well as novel forms of *governance* for data science and related practices. The rise of data-centrism highlights the challenges involved in gathering, classifying and interpreting data, and the concepts, technologies and social structures that surround these processes. On the technological front, tools such as next generation sequencing machines and health apps for smartphones are fast generating large volumes of data in digital formats. In principle, these data are immediately available for dissemination through internet platforms, which can make them accessible to anybody with a broadband connection in a matter of seconds. In practice, however, access to data is fraught with commercial, legal and ethical implications; and that even when access can be granted, it does not guarantee that the data can be fruitfully used to spur further research. This is why the advocates of Open Data are careful to stress that "openness" involves access as well as the ability to re-use freely and effectively.[18] A group of prominent data experts, many of whom working within biomedicine, has recently argued that data put online need to be Findable, Accessible, Interoperable and Reusable (FAIR).[19] The resulting "FAIR principles" have been well received and widely adopted by scientific agencies and governments, including the European Open Science Cloud.[20]

[15] See the special issue "Data-Driven Research in Biology and Biomedicine", *Studies in History and Philosophy of Biological and Biomedical Sciences* (2012), especially Müller-Wille and Charmantier; and the special issue "Big Data in History" in *Osiris* 2018 forthcoming.

[16] See for instance Alexander Broadbent's analysis of the development of inferential techniques in epidemiology (Broadbent 2013) and Rachel Ankeny's discussion of standardisation in case-based reasoning (Ankeny 2014).

[17] Hey et al. 2009, Leonelli 2016a.

[18] Open Knowledge Foundation definition of openness. http://opendefinition.org/od/2.0/en/

[19] Wilkinson et al. 2016.

[20] European Open Science Cloud Report 2017. https://ec.europa.eu/research/openscience/pdf/realising_the_european_open_science_cloud_2016.pdf#view=fit&pagemode=none

Making data usable in the FAIR sense involves specific skills and expertise in data management and curation, as well as the development of appropriate infrastructures such as databases and digital repositories. In response to this demand, fields such as data science, bioinformatics and biocuration are rapidly acquiring prominence alongside traditional scientific disciplines, and large corporations such as Google and IBM are positioning themselves as providers of data analytics and data enrichment tools. Within academia, the expertise of those who produce, curate and analyse data is increasingly acknowledged as indispensable to the effective use of big data. This encourages significant shifts in governance and established hierarchies, subverting the traditional view of laboratory technicians, librarians, administrators and database managers as marginal to knowledge production.

Ideas of research excellence are also being challenged, with some national funders moving away from research evaluations based solely on measurements of the impact of scientific publications. Research agencies in the Netherlands, Finland and Slovenia, for instance, are moving from citation counts and journal impact factors to alternative measures that emphasise the scholarly commitment to public engagement and sustainable data management, including openness indicators.[21] This in turn is forcing scholarly publishers to re-assess their business models and dissemination procedures, and research institutions to adapt their management and administration to this new landscape. These shifts are supported by policy bodies and research funders like the European Commission, the US National Institutes of Health (NIH), the Wellcome Trust and the Gates Foundation. All these major funding institutions view openness as a step towards enhancing scientific excellence and public trust in science, and are leading efforts to foster the dissemination and re-use of data generated through public and private biomedical research.

2.2 Big data in biomedical research

The increasing automation of data mining, visualisation and analysis promises to vastly accelerate the pace of knowledge production, conferring a significant competitive advantage to those who can efficiently handle computational tools. What researchers choose to consider as reliable data (and data sources) is becoming closely intertwined not only with their research goals and interpretive methods, but also with their approach to data production, packaging, storage and sharing. Thus, researchers need to consider what value their data may have for future research by themselves and others, and how to enhance that value – such as through decisions around which data to make public, how, when and in which format; or, whenever dealing with data already in the public domain (such as personal data on social media), decisions around whether the data should be shared and used at all, and how. This is particularly difficult given the ease with which digital data can be moved across locations, disciplines and research environments, making it hard both to retain oversight over how data are used and to predict who may be interested in the data, and for which purposes. The fast pace of change among data infrastructure and analytic tools, together with the rapid and ongoing shifts in Open Science policies and related legal frameworks (particularly around data protection and intellectual property), also complicate matters for researchers, who often struggle to keep up with data management opportunities and requirements.[22]

Within biomedicine these concerns have a particularly strong impact on the conduct and results of research, both because of the sensitivity associated with personal health data and because of the field's history and ongoing intellectual trends. What constitutes an appropriate and reliable source of evidence in biomedicine has long been a matter of heated debate, with substantive implications for which methods of data generation and interpretation are viewed as reliable and which types of expertise should be involved in clinical assessments and interventions – and thus for how clinical research needs to be organised and evaluated. A glaring example is the "hierarchy of evidence" proposed by the evidence-based medicine (EBM) movement (figure 1), with its insistence on considering the results of randomised controlled trials (RCTs) as a "gold standard" over any other type of relevant data.[23]

[21] Next-generation metrics, European Commission https://ec.europa.eu/research/openscience/pdf/report.pdf#view=fit&pagemode=none; see also the forthcoming report by Kim Holmberg, in the context of the ongoing Mutual Learning Exercise of the European Commission on Open Science. https://rio.jrc.ec.europa.eu/sites/default/files/MLE%20Open%20Science_Draft%20%20Modus%20Operandi.pdf

[22] Ossorio 2011, Leonelli 2012, Demir and Murtagh 2013, Hogle 2016.

[23] Guyatt et al. 1992, Ashcroft 2003.

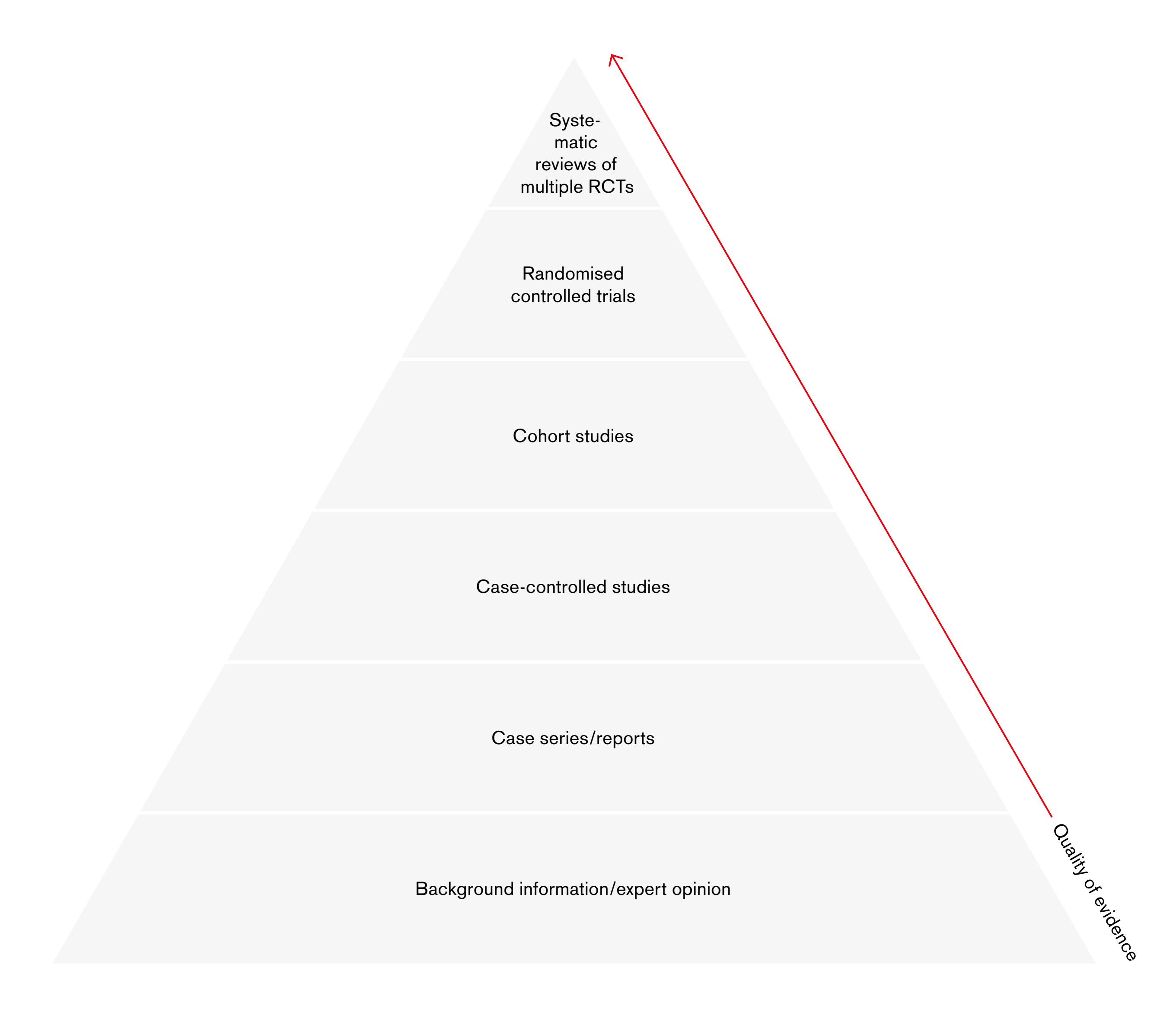

Figure 1. The hierarchy of evidence according to evidence-based medicine; from Griffin et al. (2016)

A key reason for EBM supporters to defend the hierarchy of evidence is the recognition that bringing together disparate types of data from a large variety of data sources means losing the quality assurance that the tight experimental control exercised in clinical trials can provide. Furthermore, data may vary dramatically in their format, target and provenance, making their integration very difficult to achieve from a technical perspective; and the communities producing data can have widely diverse methods, goals and assumptions, including non-overlapping or even conflicting commitments with regards to what constitutes reliable and significant data, under which circumstances, and for what purposes.[24]

Against this background, it is remarkable how successfully big data have been adopted as a rallying platform for researchers interested in a more liberal approach to medical evidence than that championed within EBM.[25] This is evident when considering four trends within biomedicine which have emerged over the last decade as key directions for the future development of research, treatment and care around the globe. All four of these trends, which I shall now proceed to briefly discuss, are strongly tied to the rise of big data discourse and practices, and propose to build on the emerging opportunities to assemble and link diverse types of data through digital technologies.

[24] Edwards et al. 2011, Mauthner and Parry 2013.

[25] It is important to note that many biomedical researchers do not see EBM and personalised medicine as opposing approaches, particularly given that EBM can be viewed as pertaining only to the realm of data interpretation (rather than access and dissemination). Nevertheless, EBM has been widely interpreted (whether rightly or wrongly) as pushing the field towards a specific perception of what counts as "good data" in medical research (Clarke et al. 2013).

1. *Personalised and precision medicine*. These two terms are currently used to denote approaches that target medical research and therapeutic intervention to specific individuals and groups.[26] This goal can only be achieved by consulting a vast and diverse body of evidence, including data collected through lab studies of non-human models, provided by patients through social media, generated through direct-to-consumer genetic services, obtained through clinical and longitudinal studies on various human populations, and produced within clinical sites such as hospitals and individual GP practices.[27] Advances in genomics also underpin the movement towards increasingly fine-grained diagnostic tools, which promise to capture variation among individuals so as to enhance biomedical understanding of how disease may affect any one person, family or community, and how this is best tackled through preventative measures and medical care.

2. *Health and Environment Data Integration*. Building on a growing appreciation of the magnitude of anthropogenic change and its impact on human health, the integration of data from an expanding variety of disciplines and approaches is viewed as critical to addressing the biomedical implications of climate change, as well as the increasing challenges presented by environmental exposure to toxic substances.[28] So-called “data mash-ups” bring together data from clinical practice, climate science, environmental studies, citizen science initiatives and social media (including self-tracking) to identify hot spots where populations and/or ecosystems are particularly vulnerable to environmental shifts. Sophisticated biomonitoring systems are also being implemented and linked with big data generated through social media and governmental interventions, which in turn creates new opportunities to investigate the effects of human exposure to harmful substances. This can inform targeted prevention, interventions, and research, and provide early warning systems to prevent and anticipate environmental impacts on health and wellbeing (as well as, of course, planetary health).[29]

3. *Self-tracking* and its relation to social media. Otherwise referred to as the “quantified self”, self-tracking refers to the emergence of measurement technologies enabling individuals to collect data about their own physiology, behaviour and activities, including information such as blood pressure, heart rate, and intensity and regularity of physical exercise.[30] The collection of such data is often associated with the deliberate dissemination of personal data through social media, either as a form of social engagement with family and friends (e.g. the “sharing activity” option offered by Apple Health) or as a contribution to biomedical research (as in the case of networking sites such as PatientsLikeMe).[31]

4. *Open Health Data*. Depending on the legal framework and user agreements in place in each case, health-related data collected through self-tracking tools may be acquired by data analytics companies and become part of privately owned big data collections.[32] Similarly, many companies that provide direct-to-consumer testing services (for instance in genetics) tend to keep results within their in-house databases.[33] This privatisation of health data raises significant scientific, ethical, legal and political issues, with some physicians and researchers viewing it as an obstacle to research progress and to the swift interpretation of clinical variants. One increasingly popular response is to insist on Open Health Data – that is, the sharing of biomedical data regardless of national boundaries, ownership and, in some cases, confidentiality concerns – as a way to encourage transparency and accountability in medical research and treatment.[34]

[26] Collins 2010, Solomon 2015.

[27] Hood and Friend 2011, Merelli et al. 2014, Lucivero and Prainsack 2015, Green and Vogt 2016, Prainsack and Vayena 2013, Kallinikos and Tempini 2014, Bliss 2018.

[28] E.g. Whitmee et al. 2015, McMichael and Haines 1997, Manrai et al. 2017.

[29] Fleming et al. 2014, Fleming et al. 2017, Nichols et al. 2014.

[30] Lupton and Michael 2017, Beer 2016.

[31] Tempini 2015, Sharon 2017.

[32] Ebeling 2016.

[33] Harris et al. 2016.

[34] For policy support for Open Data including biomedical, see European Union Open Science book 2016 and Strasser and Edwards's report on Open Access for SSIC 2015. For discussions of the scientific value of data sharing, see for instance Royal Society 2012, Science International 2015, McKiernan et al. 2016, Vayena and Gasser 2016. Ethical considerations are discussed in section 2.2 of the report.

Current challenges

3

3.1 Scientific and technical concerns: tackling diversity

One notable result of these emerging trends is an increasing diversification in the types and sources of data considered to be of interest to biomedical researchers. Appropriate infrastructure, algorithms and visualisation tools need to be developed to be able to search, link and integrate these data seamlessly and effectively. This turns out to be extremely complex. Given the different data practices, methods and assumptions characterising biomedical subfields, this may include significant disagreement over:

- preferred experimental methods and laboratory protocols;
- terminology and classification criteria, sometimes resulting in the use of the same term to refer to different phenomena, or vice versa in the use of different definitions for the same term;
- types of data judged to be of relevance to a particular target;
- choice of targets, e.g. populations, individual patients or specific diseases;
- research goals and interests, which may range from understanding biological mechanisms and processes underlying a trait of interest to the testing and validation of therapies, the production of diagnostic tools and the provision of medical care;
- commitments to particular styles of evidential reasoning and evaluation;
- and overarching conceptualisations of disease, pathology and the relationship between medical, biological and environmental knowledge.

These differences are instantiated within the variety of technologies and domain-specific standards used to generate, store, share and analyse biomedical data, each of which reflects and supports a specific perspective on the study of living systems.

Finding ways to tackle this diversity is the main epistemological and technical challenge confronted by data-intensive approaches to biomedicine. Remarkably, the solution to this challenge does *not* lie in the provision of one centralised, unified platform through which all biomedical researchers would be able to conduct their inquiries in the same way. Given the enormous costs involved in managing data and the paramount importance of securing long-term data access, governments and corporations are powerfully drawn by centralised strategies to capture all available evidence on any given topic. However, eliminating diversity in research approaches is not only highly unrealistic given the current landscape, but also undesirable and inefficient as a way to enhance knowledge production. The biological world is so complex and full of variation that no one method or model can capture all of its features. Many of the existing differences between biomedical subfields derive from researchers' effort to adapt their methods, instruments and assumptions to the specific characteristics of their research subjects. This specialised approach is what enables researchers to manipulate living systems and understand their functioning in ever greater detail, and provides essential context for data analysis and interpretation. The challenge of big data integration thus consists in *bringing different traditions and results in dialogue with each other without undercutting system-specific knowledge* tied to particular data formats, instruments and methods. Big data tools and infrastructures must therefore embrace diversity and be widely applicable to different research situations and forms of inquiry. This affects the future role of automated systems and Artificial Intelligence (AI) in biomedicine, which will certainly increase in the near future, and yet will hardly replace the crucial judgement of expert human curators.

3.1.1

Data infrastructures and information technology: interoperability, standardisation and maintenance

The importance of bridging between research traditions without oversimplifying or misinterpreting their use of data is a key reason for data experts to point to *interoperability* as the best approach to data linkage in biology and biomedicine.[35] Making existing datasets interoperable means making them searchable and potentially usable as a single body of evidence, without at the same time losing information about the specificity, granularity and provenance of the data. Interoperable databases effectively become pipelines through which the same datasets can be disseminated across a wide variety of research contexts, clinical situations and interpretation services. This requires a delicate balancing act between standardisation and flexibility to domain-specific requirements. Some degree of standardisation in the formats, algorithms and terminologies used to share data is necessary to guarantee that different databases can be linked and compared with each other. At the same time, standards need to be flexible enough to accommodate the diverse and evolving needs and preferences of database developers and users, as well as the characteristics of the data at hand. The Investigation/Study/Assay (ISA) Commons standards, for instance, propose to classify information about data provenance into three general categories that are common to any research project and can therefore be used to structure data linkage and interpretation: "Investigation", that is the high-level context and types of questions for which data have been collected; "Study", the central unit of analysis containing information on the subject under study and relevant treatments applied; and "Assay", the type of measurement used to generate the data.[36] This system supports interoperability and provides a modular platform for researchers to compare and discuss the motivations and assumptions underlying the differences in their approaches, often to the benefit of all involved. Another two notable examples of interoperable data standards are the bio-ontologies developed within the Open Biomedical Ontology (OBO) Foundry, which attempt to capture and systematise the terminology used to refer to research objects within and across biological and biomedical communities in ways that support biomedical data integration;[37] and the "minimal information" approach to the description of experimental protocols, including tools such as MIBBI (Minimal Information about Biological and Biomedical Investigations) and MIAME (Minimal Information About Microarray Experiments).[38] In 2014, hundreds of similar data standards were brought together under the umbrella of BioSharing, a web portal devoted to assembling community developed standards such as reporting guidelines, data labels and exchange formats, with the aim to "make data along with experimental details available in a standardized manner."[39]

It is important to note that while interoperable data standards are meant to apply to a wide variety of data types, many of them have been developed chiefly with reference to genomic data, which are relatively easy to generate and share in digital formats and continue to receive considerable investment and interest from public and private funding. Indeed, standardisation works best in biomedical fields that make extensive use of genomics, such as for instance oncology.[40] The adoption and implementation of standards that would enable interoperability is far from widespread in areas that are less visible and well-funded, such as research on rare diseases, and that rely on data types that are harder to share online and mine through computational tools, such as imaging data or case reports. In those cases, databases tend to be developed around specific projects, and thus to be highly idiosyncratic, with no links to international initiatives and wider networks of data storage and dissemination. Connecting such project-related databases through interoperable networks then requires high levels of manual labour and case-by-case decision making, for instance to: create ways

[35] Sansone et al. 2012, Weber et al. 2014.

[36] Sansone et al. 2012, pp. 123–124.

[37] Smith, B. et al. 2007.

[38] https://biosharing.org/collection/MIBBI; MIAME, http://fged.org/projects/miame/

[39] https://biosharing.org/

[40] Forbes et al. 2011, Cerami et al. 2012, An et al. 2014, Cambrosio et al. 2017.

to access and contribute to data collections around the world; identify appropriate computational tools to store, retrieve and visualise the data; and build trust among data donors and users as required for effective re-use. To achieve these goals, it will be important for the biomedical community to expand its gaze beyond genomics approaches as a reference model for how data should be curated and shared, and consider other potential sources of insights such as the long-standing traditions and methods used to enhance data interoperability within public health and social epidemiology.[41]

Even more demanding is the labour of *maintaining and updating* data infrastructures and standards once they have been implemented (including the work involved in seeking funding to sustain those activities). Database contents – including data formats, software and knowledge base – need to be updated regularly to reflect cutting edge developments in technology as well as new scientific findings, which may well subvert existing categories and working assumptions underpinning the organisation of the data.[42] Given the unpredictability of new discoveries and systems breakdowns, these updates are impossible to fully automate, and indeed they are so expensive that many data infrastructures cannot afford to carry them out – particularly given the short-term and innovation-driven character of current research funding. In the absence of adequate updates, however, biomedical data infrastructures may stop functioning, disappear or – worse still – stagnate and become obsolete, thus becoming increasingly more unreliable and untrustworthy. The potential for progressive loss of data quality is a particularly big challenge given the nested, inter-dependent nature of interoperable databases, which makes it possible for any unreliable data source to damage the overall reliability of all the data collections to which it is linked.[43] It is here that human judgement and peer review play an essential role, which needs to be supported alongside the increasing automation of computational systems of data collection and analysis. While very promising in terms of providing well-organised grounds for the formulation of human judgement over data ordering, quality and interpretation, Artificial Intelligence is far from being able to replace expert judgement in these areas, and insistence on over-reliance on automated systems for data analysis and diagnosis can have serious repercussions on the quality of medical research and care.

Finding reliable ways to guarantee data quality is particularly relevant in the wake of the “replicability crisis” experienced within psychology and biomedicine.[44] However, the various approaches available for data quality checks, while usefully focusing on aspects such as error detection and countering misinformation, are ultimately tied to domain-specific estimations of what counts as quality and reliability (and for what purposes) that cannot be transferred easily across fields, and sometimes even across specific cases of data use.[45] This does not help towards the development and implementation of mechanisms that can guarantee the quality of the vast amounts of research data stored in large digital repositories for open consultation. Indeed, researchers, publishers and learned societies have not yet managed to establish common guidelines for evaluating data quality.[46] This is illustrated by ongoing debates around what kind of peer review should be used for submissions to data journals such as *F1000*, *GigaScience* and *Data*, as well as what are appropriate incentives to provide for prospective reviewers.[47]

3.1.2
Data re-use: the role of theories, metadata and materials

There has been much debate around the extent to which the advent of big data and data-intensive methods is heralding the “end of theory” in scientific research, and the start of a “data-driven” approach.[48] And yet, a simplistic opposition between inductive and deductive procedures does not help to understand the epistemic characteristics of big data analysis. Data production and interpretation are unavoidably theory-laden, with substantial commitments to particular conceptual frameworks underpinning the processes of selecting, ordering, visualising, retrieving and analysing data. The keywords used to structure databases are a case in point, as attempts to define even basic terms such as “pathogen”, “ecosystem”, “gene” and “regulation” generate controversy and disagreement. The choice and definition of keywords used to order and retrieve data matters enormously to the subsequent interpretation of data by database users. The use of big data in biomedicine may thus be best characterised as theory-*informed* rather than

[41] The research communities working with cancer registries, for instance, have long considered these issues and proposed several technical and organisational solutions to the challenge of data interoperability (e.g. Hiatt et al. 2015).

[42] See for example the laborious ways in which the Gene Ontology responds to shifts in knowledge base (Leonelli et al. 2013).

[43] Tempini and Leonelli 2018.

[44] Open Science Collaboration 2015, Allison et al. 2016.

[45] Floridi and Illari 2014.

[46] Cai and Zhu 2015; for a review of possible approaches to data quality evaluation in open biomedical databases, see also Leonelli 2017.

[47] E.g. Lawrence et al. 2011, Morey et al. 2016.

[48] Anderson 2008.

theory-driven, i.e. as drawing on theories and conceptual assumptions without letting them pre-determine the ultimate outcomes.[49] An awareness of the conceptual choices and background knowledge informing the selection and organisation of data may not always be relevant to researchers looking to mine and interpret big data collections. It is however crucial for big data users to recognise the fact that data (particularly those found within online databases) are never truly "raw", and the theoretical structures that informed their processing may well have a bearing on their future use.[50]

To assess whether and how a given dataset may be used as evidence for new purposes, researchers typically need to be able to compare the situation within which data were originally generated with their own research context. This is achieved by reference to information about the provenance and history of data, which is referred to as *metadata*. This may include the procedures, equipment and materials used to generate them, the relevant environmental conditions, and the ways in which data have been subsequently handled and manipulated. Decisions around which metadata to report, why, and how can significantly affect how data available online are interpreted, as well as how researchers plan, describe and execute their studies – thus having a profound impact not only on experimental practice, but also on scientific reasoning and methods. This again highlights the importance of monitoring and updating curatorial efforts to standardise and automate data dissemination, to make sure that they reflect the latest advances and diversity of approaches characterising each biomedical subfield. The same goes for the use of artificial intelligence, whose implementation can significantly accelerate data linkage and the identification of correlations, but whose underlying assumptions and constraints need to be regularly reviewed and subjected to wide-ranging feedback in order to lower the risks of data misuse, overinterpretation or misinterpretation.

The link between metadata and the availability and management of *material samples* deserves a specific mention here. Most data of relevance to biomedical research are extracted from organic specimens, tissues, cells and microbial cultures, as well as of course from human subjects. Continuous access to original samples is well known to enhance data reproducibility and re-use, by providing researchers with better opportunities to replicate experiments and re-contextualise data. Links between data and samples, the virtual and material, the analogue and the digital are therefore very important for data re-use. However, continuous access to materials is often impossible in the case of human subjects, who may not be traceable or willing to continue their participation in a study, or may simply die. Even when it is physically possible, as in the case of tissue samples stored in biobanks, access is hard and laborious to organise. Organic resources are fragile and depletable, so access needs to be monitored and appropriately limited.

The management of participation by both donors and users of samples is also fraught with controversy, with serious questions posed around the moral obligation, technical challenges and scientific opportunities involved in engaging donors not only at the point of sample collection, but throughout the life and re-use of the sample (especially if the research goals and prospects change considerably from context in which donors provided their consent, due to shifts in science and technology).[51] The fact that sample collections need stewardship and curation as much as data infrastructures, and should ideally be intertwined and inter-dependent with their digital counter-parts, is rarely recognised outside professional biomedical circles, and the associated costs make it prohibitive to support for most public funding agencies. Biobanks are therefore struggling with a lack of funding and visibility, and are typically not well-coordinated with each other.[52] The extent to which materials can be traced and linked to data collections has a significant impact on which metadata are reported in digital databases, and how.

[49] Waters 2007

[50] Gitelman 2013, Leonelli 2016a.

[51] Global research commons must be managed to facilitate not only use, but also re-contributions from users (Schofield et al. 2009; see also Wyatt et al. 2013).

[52] There are exceptions of course, such as the UK Biobank, yet these exceptions further underscore the large amount of resources required to ensure systematic links between sample collections and digital infrastructure, and the development, use and maintenance of adequate information technologies.

3.2
Ethical concerns: making knowledge production sustainable

The interconnected and international nature of big data dissemination makes it impossible for any one individual to retain oversight over the quality, uses, import and potential social impact of the knowledge being produced. Many individuals, groups and institutions end up sharing responsibility for the social outcomes of specific data uses. A key challenge for data governance is to find mechanisms for allocating responsibilities across this complex network, so that any fraudulent, unethical, abusive, discriminatory or misguided actions can be singled out, corrected and appropriately sanctioned.[53] To this aim, it is crucial for policy-makers and researchers to recognise that there is *no simple technological fix* for monitoring the social impact of data re-use; and that *ethical and social considerations are part and parcel of extracting biomedical knowledge from big data*, helping to foster the reliability and long-term sustainability of results, especially when they are made openly accessible.

3.2.1
Ownership and the value of data

Data are defined by the *scientific value* that they are given as evidence for knowledge claims.[54] At the same time, they can be viewed as valuable in many other ways, which has a strong impact on their scientific uses. Data have *political* and *financial* value, for instance as the result of costly investments, as tools to legitimise or oppose governmental policies, or as trade currency among national governments, lobby groups, social movements and industries. They have *cultural* and *social* value through their link to the identity, histories, norms, sensitivities and behaviours of the communities and individuals which they are taken to document. And they can have *affective* value for the researchers that invest time and effort in their production and interpretation, as well as (particularly in the case of the personal data used in biomedicine) for the human subjects from whom data are extracted.[55]

To make it at all feasible for data to travel across contexts and thus possibly increase their scientific value, market structures, research institutions and policy bodies need to acknowledge and negotiate their value as political, financial and social objects. This can generate frictions among different actors involved in biomedical data handling and interpretation, including tensions around who owns the data and what constitutes acceptable re-use. For instance, some researchers feel that individuals who have not been involved in the production of certain kinds of data – particularly those which are highly susceptible to shifts in environmental conditions – may not be able to evaluate their significance appropriately as evidence, and thus that sharing such data could harm scientific progress by encouraging misleading interpretations.[56] This interpretation of the affective and scientific value of data contrasts strongly with the encouragement of data sharing promoted by the Open Health Data movement. Other examples are the clash between financial valuations of data, which provide an incentive for commercial entities such as data analytics companies and pharmaceutical industries to retain competitive advantage by keeping their research data in-house, and political and scientific valuations which favour Open Data to enhance the transparency of the research process and its outputs; and the financial valuation of the economic worth of personal data extracted from social media, which may vary considerably between the providers of the communication platform at hand, researchers who use that platform as a data source and individual contributors to that platform (who may not even be aware of contributing personal data for research or commercial purposes).

How such clashes are negotiated shapes the travel of biomedical big data and the effectiveness with which they are used to produce new knowledge. What has propelled data into becoming protagonists of contemporary biomedicine is precisely their multifaceted perception as at once local and global, free commodities and strategic investments, common goods and grounds for competition, potential evidence and meaningless information. If they are to travel across academic labs, industrial development departments, policy discussions and social media, data need to be of interest to all actors involved, but it is no wonder that motivations and incentives behind efforts to collect, share and re-use data should diverge widely. This is another demonstration that flexibility to multiple uses and future scenarios, as well as to the diverging interests of potential users, is crucial to the success of databases in enabling big data analysis, and thus to the future of knowledge-production in biology. Researchers and research-facing institutions need to find ways to deal with possible conflicts among data values in their work, thus balancing the constraints and decisions internal to scientific reasoning, and the broader landscape of opportunities, demands and limitations within which researchers operate.

[53] Aicardi et al. 2016.

[54] The term "value" captures the modes and intensity of the attention and care devoted by individuals, groups or institutions to given objects or processes.

[55] Leonelli 2016a, Ebeling 2016, Rajan 2017, Murphy 2017.

[56] Leonelli et al. 2013, Fecher et al. 2015, Borgman 2015.

3.2.2 Security, privacy and confidentiality

Another set of frictions concerns the status of personal data used for biomedical research, and particularly the use of "private" data – that is, data that should not be shared with others or that should only be shared selectively for specific purposes, with the consent of the data originator. Despite the various legal frameworks protecting individuals from undesired dissemination and misuse of their personal data (e.g. data protection law and privacy rights), there is at present no internationally recognised framework specifically targeted to health data, and the application of such laws in the biomedical domain is fraught with ambiguity.[57] While it is widely agreed that personal data are not always private and private data are not always confidential, who is ultimately responsible for determining what constitutes private and confidential data, and with which rationale, remains unclear.[58]

One option is to regard the individuals from whom data are being extracted as the ultimate arbiters of whether and how their data should be disseminated and re-used, and for which purposes. The Global Alliance for Genomics and Health (GA4GH), an international coalition of academia, industry, and patient groups, has strongly advocated this interpretation of responsibility towards data sharing, arguing that it is each individual's right to donate their data to science if they so wish, and thus to contribute to the advancement of biomedical knowledge (an idea sometimes linked to the adoption of the "right to science" as a fundamental human right).[59] The GA4GH has drawn up an international Framework for Responsible Sharing of Genomic and Health-Related Data that balances individual privacy, recognition for researchers, and the right of citizens to benefit from the progress of science.[60] In parallel to this view, solidarity is increasingly regarded as an important motivation for the sharing of personal data for research purposes.[61] It is argued that there are strong ethical grounds for individuals to donate their personal data to research activities that serve the public interest and could benefit them or their communities in the future.[62]

At the same time, given the complexity of data dissemination pathways and the unpredictability of data re-use, it is arguably impossible for individual data donors to assess what implications data sharing may have for themselves and others in the future. Indeed, traditional notions of informed consent and informational control are widely recognised as out of step with interoperable data infrastructures, where new opportunities for data linkage foster extensive data re-purposing in ways that cannot be foreseen at the moment of data collection.[63] This may be taken to imply that the individuals from which data are extracted should not hold responsibility for data sharing strategies.[64]

Furthermore, there are several conflicting ways of interpreting what may constitute "public interest" and "common good" in the case of big data sharing and re-use for medical purposes.[65] It may well be that data disseminated to foster biomedical research end up causing harm to individuals or groups, or that such data are misappropriated and misused for purposes other than research. In such cases, governments and research institutions need to take responsibility for protecting individuals (and particularly patients) from harm, whether or not those individuals have decided to exercise their right to science. This is particularly important in light of the increasing commercialisation of personal data described above, and the vast potential for fraud and manipulation. Given the ease with which individual data points can be aggregated and de-anonymisation procedures can be reversed, confidentiality and patient protection remain paramount in medical research practice.[66] It is also argued that privacy frameworks need to be extended to groups and local communities, for example in the case of biomedical research mining personal and geolocation data from social media like Twitter and Facebook.[67] As recognised by the Organisation for Economic Co-operation and Development (OECD) in its latest report on the governance of health data, in the age of big data patients' privacy needs legal and social protection like never before.[68]

[57] O'Brien et al. 2017, OECD Recommendations on Health Data Governance, http://www.oecd.org/els/health-systems/health-data-governance.htm

[58] Nuffield Report 2015, https://www.nuffieldhealth.com/local/d2/5b/f79006c9445ca00ba0c188eb7b04/annual-report-2015-full-report.pdf

[59] Shaver 2010, Vayena and Tasioulas 2015. See also the "your genome belongs to you" movement, http://www.yourgenome.org/

[60] https://www.ga4gh.org/

[61] Prainsack and Buyx 2017.

[62] Although there are several conflicting ways of interpreting what may constitute "public interest" and "common good" in the case of big data sharing and re-use for medical purposes (Floridi 2014, Nuffield Report 2015, p. 55, Burton et al. 2015, Prainsack and Buyx 2017).

[63] Kaye et al. 2015, Vayena et al. 2013.

[64] E.g. Nuffield Report 2015, p. 75: "Where a person providing data about themselves cannot foresee or comprehend the possible consequences when data are to be available for linkage or re-use, consent at the time of data collection cannot, on its own, be relied upon to protect their interests."

[65] Floridi 2014, Nuffield Report 2015, p. 55, Burton et al. 2015, Prainsack and Buyx 2017.

[66] Hogle 2016.

[67] E.g. Floridi 2014.

[68] See OECD latest report on Governance of Health Data, https://www.oecd.org/health/health-systems/Recommendation-of-OECD-Council-on-Health-Data-Governance-Booklet.pdf

Concerns around potential harm deriving from the misuse and misinterpretation of the data also put pressure on standards of information security, understood as the requirement to preserve the integrity of data from external attacks, degradation and misuse.[69] These issues are compounded by the shift towards the digital, which has increased the ease with which data can be copied, circulated, corrupted and leaked. As a result, information security requirements need to be put in place which exert a strong influence on the trajectories and outcomes of data sharing efforts, such as for instance masking and anonymisation efforts (i.e. techniques used to prevent data from being associated with specific individuals or groups).

An important question emerging from consideration of the complexities and vulnerabilities involved in handling personal data is whether data sharing is the best way to facilitate data re-use for biomedical purposes. An alternative route is to keep data collections in separate silos, while fostering the development of mechanisms and procedures through which researchers can search through those collections and acquire only the information that is of potential relevance to their ongoing research.[70] These mechanisms can include software that automatically mines existing data collections for specific parameters and correlations,[71] and governance structures for each data collection that can help to mediate access and point researchers to relevant content (such as those implemented by institutions like the Secure Anonymized Information Linkage databank in Wales[72]). Whether or not these mechanisms facilitate an efficient, fast and thorough exploration and mining of big datasets remains a controversial question, yet they do offer substantial support towards research that is ethically sound, socially responsible and sustainable in the longer term.

Big data integration exemplifies the extent to which social and ethical concerns around the potential impact of biomedical data sharing on individuals and communities are inextricably linked with scientific concerns around the quality, validity, and security of data. The ways in which privacy, security and confidentiality concerns are handled are critical to the study and treatment of human subjects; affect how data are integrated and interpreted; and are therefore an integral part of the research process.[73] This can arguably have negative effects on the pace and scope of research, such as making it insensitive to detecting target relationships (because granularity is lost through anonymisation processes, for instance), vulnerable to technological and political developments (such as hacking, viruses, etc.), or subject to cumbersome bureaucracy which may obstruct creative and exploratory uses of data by favouring projects with well-defined hypotheses. Privacy protecting procedures such as information governance panels can be particularly unwieldy when the research is multi-sited, raising questions around how consistency in judgement can be maintained while avoiding duplication of efforts.[74] At the same time, ethical and security measures can enhance the overarching quality of the research effort, for instance by providing opportunities for devising mechanisms for the long-term sustainability of data infrastructures; and by re-configuring relationships between research partners in ways that enable better and more inclusive reflection on potential downstream implications of big data linkage and re-use.

3.2.3

Bias, inequality and the digital divide

The use of big data in biomedical research can give rise to two forms of discrimination, both of which can have pernicious effects on the credibility and veracity of the knowledge being produced. One is the extent to which the data available for analysis contains *bias* – in other words, whether big data available online constitute a good sample for the research questions at hand.[75] We already discussed how computational tools for data tracking and monitoring unavoidably rely on human judgement about what counts as data in the first place and how data should be ordered, labelled and visualised. These judgements are particularly significant given that not all data are equally easy to digitally collect, disseminate and link through existing algorithms, resulting in a highly selective data pool that does not accurately reflect reality (and in some cases actively distorts it). At the same time, the existing distribution of resources, infrastructure and skills determines high levels of *inequality* in public participation to the production, dissemination and use of data. In government as much as in academic research and industry, big players with large financial and technical resources are leading the development and uptake of data analytics tools, leaving the rest of society at the receiving end of innovation in this area. Indeed, contrary to popular depictions of the data revolution as harbinger of transparency, democracy and social equality, the *digital divide* between those who can access and use data technologies, and those who cannot, continues to widen. The vast majority of the population is thus encouraged to provide more and more personal data for access to digital services, but does not have the means to consider the multiplicity of uses to which such data can be put and the potential for negative repercussions on themselves and their communities. This results in potentially unfair modes of participation in data collection and analysis, with some social groups being represented more heavily than others, and little protection from their resulting visibility (or lack thereof) as research subjects and the claims derived from the analysis of such data.

[69] Gold 2010, Tempini 2016.

[70] This is what the Royal Society recommended in their report "Science as an Open Enterprise" 2012.

[71] An example of this is DataShield. www.datashield.ac.uk

[72] https://saildatabank.com/

[73] Dove et al. 2015, Mittelstadt and Floridi 2016, Leonelli 2016b.

[74] Dove et al. 2016.

[75] Boyd and Crawford 2012.

Another type of discrimination concerns the extent to which big biomedical data are accessible and re-usable to researchers and other communities around the world, and what differences in access mean in terms of the kinds of approaches, disciplines and locations conducting big data analysis. Databases mostly display the outputs of rich, English speaking labs within visible and highly reputed research traditions, which deal with "tractable" data formats (such as "omics").[76] The involvement of poor/unfashionable labs and researchers working in low-income countries is low and almost always at the receiving end (meaning that they are not involved in developing resources, just consulting them). The issue of big data access compounds existing digital divides locally and internationally with a new divide in access to data as well as appropriate technology, resources, and trained personnel to be able to re-use such data. Furthermore, the resources required to collect, store, and analyse big data are increasingly being appropriated and developed by a few multi-national corporations and governments, with little opportunity left for less powerful and internationally recognised players to participate in shaping the relevant technologies and strategies. In the private sector, it is also unclear what the status of data from clinical trials is, and which data are being shared with whom. This divide in who has and has not the capacity to become involved in big data usage has severe implications for researchers based in low-resource environments, with inequalities in visibility, power and location being reinforced, rather than mitigated, by big data dissemination.[77] The divide also results in the scarcity of biomedical data relating to certain subgroups and geographical locations, and thus to questions around whether individuals that remain excluded from big data collections benefit from advances such as personalised medicine. This limits the comprehensiveness of available data resources, restricts the potential for big data use to tackle global health challenges, and constitutes an additional source of potential bias.

3.3 Institutional and management issues

3.3.1 Division of labour and shifts in roles

The need for the reliable development and oversight of data infrastructures calls for a skillset not typically found in traditional medical, biological, statistical or computational training. Database developers need to acquire competence in information technology and programming; an understanding of the characteristics of the data stored in the infrastructure and of how they may inform research; and an awareness of the needs and interests of prospective users – as well as of the legal and ethical implications of their activities, the diverse commitments and methods characterising different research communities, and the different forms of value that data can acquire across social contexts. It is becoming increasingly clear that one individual cannot possibly cover the variety of skills required to develop and maintain a reliable biomedical data infrastructure, and thus that big data management requires teams of individuals with relevant complementary skills and the ability to communicate and interact efficiently.

How such teams should be composed and trained continues to be a contested matter, whose resolution bears significant implications for the future development of big data biomedicine. Universities are rapidly assembling new training programmes (typically in bioinformatics, computational medicine and/or data science) which can help to form data managers and curators, thus filling the current gap in expertise.[78] Private providers of computational systems, such as Google, IBM, Microsoft and Apple, are also interested in expanding their research support towards data enrichment and interpretation services, with the idea that biomedical researchers and clinicians could delegate data management to external providers, and rely on them to develop trustworthy data infrastructures and related tools for data visualisation and analysis (a dependence which is of course already manifest in the use of corporate websites, portals and search engines).[79] The extent to which knowledge extraction from big data requires the negotiation of multiple forms of data value is not always acknowledged when developing this training. There is also variation in the ways in which different programmes interpret the division of labour between biomedical, data science, data management and clinical work, as well as the respective value of the contributions from each of these types of expertise.

[76] On the linguistic divides, see Amano et al. 2013.

[77] Bezuidenhout et al. 2017. This is particularly striking given the volume of research-relevant data being collected, shared, and consulted by a widening portion of the population around the world.

[78] All Ivy League universities in the United States are offering courses in data science and big data analysis. In the UK alone, there are currently over 126 degrees and MSc level courses on offer in data science (http://www.mastersportal.eu/study-options/268927258/data-science-big-data-united-kingdom.html)

[79] E.g. IBM Watson Healthcare https://www.ibm.com/watson/health/; Deepmind Health https://deepmind.com/applied/deepmind-health/; Apple Healthcare https://www.apple.com/healthcare/

Biomedical researchers, clinicians and data scientists all need to play a role in ensuring the soundness of decisions made around how data and related metadata are described and disseminated. The necessity to provide input on data sharing procedures is however likely to affect the overall workload and goals of all three groups. This is likely to have particularly significant effect on the workload and responsibilities of clinicians, whose position in relation to patients and care settings provides them with unique insight on the potential implications of specific ways of handling data, and yet limits their ability to respond to new sources of information and methods of analysis, given the severe time pressures and legal responsibilities already involved in their work. Particularly in the case of personalised treatment regimes, clinical decision-making is increasingly being flanked by interdisciplinary panels and analytic services specifically geared to help piece together and interpret the clinical situation.[80] These help clinicians to address the organisational and methodological challenges generated by the ongoing changes and increasing fragmentation in the medical knowledge base, and interpret lab results in view of broader data resources to address individual situations. At the same time, these interventions affect the loci of clinical decision-making, and clinicians need guidance on what this involves for their overarching role and responsibilities, especially in case of conflicts around data interpretation.[81]

The ongoing and well-documented difficulties in accommodating genetic counselling and testing alongside other medical procedures (including the need for the introduction of new experts and related services as part of clinical care) are useful reminders of the challenges involved in bringing big data analysis to patients, and the potential burdens that this may place on patients' role in medical decision-making.[82] Indeed, another key group whose role and expertise is shifting due to the emergence of big data is that of patients and their families – particularly considering the variety of cultures of medical care, conceptions of medical expertise and the role assigned to patients within different national healthcare systems. This needs to be taken into account when devising forms of patient engagement in big data collection and analysis. More broadly, governing data use requires a participatory approach to the production and oversight of tools for data management and analysis, in which technicians work alongside people who may not have technical skills in data science, but do have the experience and expertise to make informed and considered decisions around data use and its social implications. Data processing strategies and tools should never be developed separately from the situations of data use where ethical and social concerns emerge. Interesting prototypes for such an exchange can be found in biomedical projects and medical schools who systematically engage groups of interested citizens as members of steering groups, advisors on the ethical and social implications of ongoing procedures, or sounding boards over emerging ideas.[83]

3.3.2

Reward system and open data enforcement

There is a growing recognition among research institutions that the current credit system supporting academic hires and promotions hinders the flourishing of a research culture of openness and care for data. Performance metrics such as impact factors and citation counts are coming under fire for failing to reward collaborative work and rewarding the publication of articles in prestigious venues over efforts to curate and disseminate data, metadata and related materials and software. Researchers cite the lack of appropriate rewards as having a major effect on the resources and time that they are willing and able to allocate to data sharing activities, not least since they are under pressure to secure not only their own track records, but also those of their collaborators and students. Hence many scientists, and particularly those subject to severe financial and social pressures, perceive data sharing and data curation activities as an inexcusable waste of time, despite being aware of their scientific importance.[84] This leads to the vast majority of data produced through publicly funded research not being made Open, even in countries where Open Data are strongly favoured by funding bodies.

[80] E.g. the Molecular Tumor Boards, Cambrosio et al. forthcoming.

[81] Wang and Krishnan 2014.

[82] Kelly 2008, Markens 2013, Paul 2017.

[83] Such arrangements have been successfully implemented, for instance, within various departments of the University of Exeter (http://www.exeter.ac.uk/cbma/getinvolved/magpies/; http://www.exeter.ac.uk/research/centre/hepe/)

[84] Fecher et al. 2015, Levin et al. 2016, Treadway et al. 2016.

Major funders (such as the European Research Council, the Wellcome Trust and the Gates Foundation) have reacted to this detrimental situation by overhauling their evaluation system for grant applications, placing increasing value on researchers' open science practices and on the strategic planning of data management. The Leiden manifesto published in *Nature* in 2015 offers a set of ten principles to underpin alternative research metrics, highlighting particularly the need for qualitative assessments that consider a wide range of ways in which researchers contribute knowledge and expertise to their peers and society at large.[85] Some prominent universities, such as the members of the League of European Research Universities (LERU), have committed to implementing the Leiden principles.[86] However, this shift in evaluative cultures remains challenging for most research institutions and particularly biomedical research environments and medical schools, where quantitative measures of the impact of publications often constitute the primary evaluative criterion. There is also scarce evidence of uptake of alternative evaluative measures among peer reviewers of academic journals and grant applications – and more generally, little consensus as yet on how evaluative measures supporting open, ethical and sustainable data handling should be enforced, with research funders and governments unclear on whether and how non-compliant researchers, peer reviewers, companies and institutions should be penalised, and concerns around balancing requirements to foster data re-use with a commitment to support research excellence.

This situation urgently needs to change for big data collections and analytics to become reliable sources for knowledge production. Both the quality and long-term survival of data infrastructures, and of related tools for the critical scrutiny and interpretation of data, depend on biomedical researchers having institutional support and incentives to openly and rigorously document experimental practices, evaluate the risks and ethical implications of sharing data, and appropriately curating the storage and dissemination of data, materials and metadata. These requirements only come into conflict with research cultures in situations where competitiveness around discovery claims is privileged over collaboration and attention to the robustness and external validation of research results. By contrast, the development of adequate data packaging requires the support and co-operation of the broader community and institutional structures within which researchers operate, whether this is academic community or industry. As mentioned above, the most substantial engagement at the moment tends to come from researchers working in prestigious institutions which are more likely to capitalise on their visibility and international links, thus valorising and advertising data curation activities as providing new platforms for research collaboration and exchanges.[87]

3.4 Financial concerns: which business models for data infrastructures?

Funding for database curation (and associated training) remains relatively scarce, especially when compared to the investments made by public and private institutions in other research activities. This is true even in areas as successful as model organism biology, where the best stocked and curated databases are those concerned with few, highly standardised data types (such as sequencing, transcriptomics and, to some extent, proteomics), while curators are still struggling to incorporate more labour-intensive data such as those used in metabolomics, cell biology, physiology, morphology, pathology and environmental science – not to speak of the complex datasets used in public health and epidemiology.[88] Many research projects in big data and human health are typically set up at most for up to five years, with no possibility to extend funding further so as to maintain and update the datasets and related infrastructures that have been produced. When the funding ends, access to data deteriorates and is sometimes lost entirely, leading to a loss of knowledge resources. Public and private funders also tend to focus on innovation associated with the production of new research, thus implicitly favouring the development of new resources over the maintenance of existing ones, and typically lack the resources and willingness to support ongoing infrastructures or attempts to link/enrich/curate data within them. Even the funding streams that are explicitly dedicated to long-term data infrastructures, such as the UK BBSRC infrastructure funds, operate on a short-term basis and require applicants to look for ways to become financially self-sufficient.

There is a lack of clarity over what kind of business models could sustainably underpin data infrastructures and related efforts of data collection. It is not clear how to make the associated costs viable in the long term, and how to support related expertise. Various models have been proposed, including a variety of subscription models and the involvement of private companies whose activities benefit from the free and well-curated dissemination of research data. However, the current landscape still sees most well-functioning data infrastructures in biomedicine being funded by government. On the one hand, this may be seen as sensible given that nation states tend to be more resilient and financially robust entities than any single research institution or initiative. On the other hand, given the current tendency for public services to become privatised, shrinking research budgets and intense competition over public spending, the reliance on governmental funds puts data-centric research at the mercy of political trends and short-term priorities.

[85] Hicks et al. 2015.

[86] Ayris et al. 2013.

[87] Leonelli 2016a, Bezuidenhout et al. 2017.

[88] Bastow and Leonelli 2010.

Conclusions

4

4.1 Two key challenges: managing infrastructures and engaging stakeholders

The emergence of big data and related digital technologies and skills has immense potential to transform biomedical research, which however needs to be unlocked through the development of responsible, sustainable and effective ways to generate, disseminate, analyse and re-use relevant big data. Two sets of challenges stand out as crucial to the future impact of big data on biomedical knowledge production:

- the first revolves around the *management of biomedical data infrastructures*, and includes questions around whether and how to mediate access to available information, and who should be in charge of overseeing this process; which types of data are available, and in which format; who should be responsible for deciding which potential uses the data can be put towards; who evaluates the quality of the data, standards and algorithms within data infrastructures, and how; and what business models can sustainably underpin the development and maintenance of data infrastructures.

- the second concerns *participation and engagement in data-centric biomedicine*, and involves questions around which types of personal data should be included in big data collections, under which conditions and with what kinds of accountability; how should data selection and inclusion in databases be conducted and prioritised, given the considerable resources required to appropriately curate data for re-use, and the resulting constraints around research directions; what are the consequences of inclusion or exclusion from the production and re-use of big biomedical data collections for prospective and current patients, and what kinds of exclusions are being created by reliance on digital data infrastructures as sources of biomedical information – both in terms of the kinds of expertise and skills being considered as central to data analysis, and in terms of potential bias around the types of subjects and topics being represented and studied.

Technologies such as interoperable databases, data mining, AI and interpretive tools can be of great use in addressing these challenges, but the deciding factors are human: social and institutional. Research-facing organisations need to set up platforms to foster sustainable and responsible big data collection and analysis, including relevant training programmes, adequate reward systems and regulation, and long-term financial support. National governments and funding bodies should endorse these activities and provide cultural and economic incentives to their realisation, while also supporting the development of legal frameworks and assessment procedures for ethical, non-discriminatory and sustainable data sharing and re-use – including a clear way to allocate responsibility when things go wrong, and provide compensation against harm deriving from the inappropriate use of personal data. The involvement of patient groups and private providers of data-related services (such as data analytics companies, providers of direct-to-consumer tests and social media) in addressing these issues is paramount to *building public trust* and an *awareness of the scope, limits and opportunities linked to the use of big data for biomedical research*. The nature and quality of such engagement needs to be critically discussed on a case-by-case basis, as what constitutes "good" and "useful" involvement (and for whom) is likely to change depending on the specific projects, goals and data types at hand.

International co-ordination is also crucial, given that biomedical research in both the public and the private sectors operates through collaborative networks that transcend national borders and increasingly relies on the use of common standards, technologies and institutional agreements – such as those underpinning the interoperability of data infrastructures. The emergence of international consortia coordinating the management of research data across national borders, such as the European Open Science Cloud, GO-FAIR and Elixir, is thus a welcome move, which should be supported by national governments and research institutions as a necessary step towards the harmonisation of approaches to biomedical knowledge production, innovation and sustainability.

4.2 Towards a federated approach

At the same time, and particularly with respect to biomedical research, excessive centralisation of data standards and management can be counterproductive. There are important lessons to learn from history about the dangers of setting up large, top-down data infrastructures – particularly the ease with which centralised management can lose contact with and thus the trust of data producers and users.[89] The diversity of biomedical knowledge, methods, sources is precious and needs to be exploited, otherwise there is a strong risk of encouraging conservatism in research and thus to stifle innovation and creativity. There is a tangible danger that reliance on data-centric analysis will improve the visibility of research approaches that are already well-established, while less well-known or popular traditions get side-lined, no matter their innovative potential. By regulating the extent to which data-centric methods incorporate or exclude dissent, diversity and creative insights, the ways in which data are managed can determine the degree of conceptual conservatism characterising future biomedicine, and the extent to which it can support radically new insights.

Thus, data handling initiatives need a federated implementation, with coordinating bodies tasked with fostering dialogue among local initiatives and new and diverse approaches to data management and interpretation. The scale of international data management efforts makes it exceptionally hard to create standards that may accommodate the vast diversity of epistemic cultures involved, and still remain accessible for scrutiny and feedback to such a wide variety of expertise. Diversity in intellectual property regimes, ethical regulations and expectations of patients concerning personal data and medical interventions is also tremendous, as is the variation in the kinds of commercial interests surrounding big data analytics in the medical domain – including many kinds of companies, instruments, sequencing, algorithms, counselling, software, testing, genome editing. Consortia have been flagged as an excellent social structure to manage big data and open science initiatives across all domains, while taking account of the scale and implications of diversity in their constituents and the potential audience of this research.[90] Whenever they select standards, labels and data sources, and strategise over future priorities and funding bids, national and international initiatives for biomedical data management and analysis need to acknowledge that their decisions determine who is included and who is excluded from contributing to data-centric biomedicine.

[89] A useful example is the failure of the Cancer Biomedical Informatics Grid, a multi-million investment by the National Institute of Health in the United States with the ambitious aim to provide data sharing, curation and analysis tools for the whole of biomedicine. The project folded not because technically flawed, but because of its lack of engagement with users and target communities (Leonelli 2013; https://deainfo.nci.nih.gov/advisory/bsa/archive/bsa0311/caBIGfinalReport.pdf)

[90] Leonelli 2009, Cutcher-Gershenfeld et al. 2017.

4.3
Using big data to produce biomedical knowledge: five principles

4.3.1
Principle 1: ethics and security concerns are an integral part of data science

Data producers, curators, users make key choices about what constitutes data, for which purpose, and at each stage of their dissemination and use. These choices are affected by the materials from which data are extracted, data formats and vehicles, the ethos of relevant research communities, existing standards for what counts as reliable data, conditions for data access and use, understandings of data ownership and value, the availability of other research components (such as software, instruments, protocols, models), and institutional and infrastructural support for big data analysis and re-use (or lack thereof). There is thus no clear separation of concerns around ethics and information security from epistemic concerns. Building safeguards for social and ethical concerns with data re-use can help to make the resulting science methodologically sound, accountable to and engaged with diverse stakeholders and robust to continuously changing requirements and challenges. This general principle underpins many of the norms currently being proposed for responsible big data research, with a recent formulation including the following rules: “acknowledge that data are people and can do harm”; “practice ethical data sharing”; “debate the tough, ethical choices”; and “develop a code of conduct for your organization, research community, or industry”.[91] The implementation of such rules calls for careful planning around the scheduling and impact of big data projects, taking into account interdependencies and potential delays[92] and assessing the potential ethical and legal implications of data re-use at regular intervals throughout any research project.[93]

4.3.2
Principle 2: public engagement and trust are crucial to the successful dissemination and use of big data in biomedicine

The emergence of the citizen science movement and “quantified self” technologies call for increasingly engaged relationships between biomedical researchers, data curators and analysis, and the wider public – particularly patients and their families. Public engagement should include activities aimed to discuss research plans and modalities of data sharing, raise awareness of how data are being used in research, and develop new research directions and hypotheses. It is important for researchers to clarify what is meant by “public” whenever planning such collaborations, as well as what is meant by “public good” in relation to scientific activities.[94] Scholarly work on the information commons, non-rivalrous uses of knowledge and alternative forms of consent is growing, and could provide a crucial reference point for setting up such engagement practices.[95] Such reflection provides a needed counterbalance to the current tendency towards data appropriation and privatisation. In a world where data, as the “new oil”, is a financially valuable commodity, it is all too easy to transform acts of data sharing and open data regimes into mechanisms for the infringement of privacy and theft of intellectual property. This is facilitated by the variety of national legislations surrounding the use of personal data and the transfer of materials (such as samples), and the legitimate worries raised by bioprospecting practices.

4.3.3
Principle 3: effective public engagement depends on big data literacy across society as a whole

Biomedical researchers should participate in programmes aimed at communicating basic data science skills and an understanding of individual rights relative to big data access, use and interpretation. This is crucial both within the research world, where different disciplines have different levels of understanding of the implications of sharing and integrating large datasets, and within society as a whole, where a minimum level of literacy concerning the nature and implications of big data sharing/use is becoming indispensable to be able to engage cogently with public and private services, technologies and social media.

[91] Zook et al. 2017.

[92] Tempini 2016.

[93] Leonelli 2016b.

[94] Wood 2015.

[95] Ostrom 2005, Eschenfelder and Johnson 2014, Eschenfelder and Shankar 2016.

4.3.4

Principle 4: biomedical research needs to build on multiple sources of evidence, taking account of the novel types of discrimination created by big data dissemination and re-use

Many sources and types of data can contribute to the efficient development of medical diagnostics and treatments, and their efficient targeting to individuals and groups. Efforts to facilitate the mining, integration and re-use of various types of data need to be supported by individuals with widely different expertise, so as to maximise the evidential value that can be extracted from the data. This does not necessarily require making all data of biomedical relevance open, as tools can be developed to tailor data access to research need and goals. While open access to biomedical data can stimulate creative research and novel interpretations, it can also increase the risk of misuse and harm to human subjects. Every instance of big data sharing needs to be evaluated by weighing its research value against the potential harm that it may yield, particularly when integrated with other data.

4.3.5

Principle 5: data infrastructures and related data management skills are essential to extracting biomedical knowledge from big data

Integrating big biomedical data requires labour-intensive conceptual and material scaffolding, in the form of appropriate IT infrastructures, technologies, institutional support, regulatory frameworks and skills. Long-term funding is crucial to the maintenance and appropriate updating of data infrastructures. It is also essential for biomedical researchers to interact with existing databases from the very start of their research activities (as required by data management plans), both as users and as data providers. This requires cultivating an awareness of which databases are being developed in their field, and a commitment to provide feedback to the work done within these databases whenever possible. Clinicians' training and incentive structures need to be updated to reflect these changes. At the same time, the new requirements and expertise relating to big data methods and tools cannot all be shouldered by existing biomedical researchers. Expertise in data science and data management needs to be recognised as an emerging professional role that complements and supports biomedical research and interventions, and requires adequate institutional support and specific career pathways.

References

Aicardi, C., Del Savio, L., Dove, E. S., Lucivero, F., Tempini, N. and Prainsack, B. 2016. Emerging ethical issues regarding digital health data. On the World Medical Association Draft Declaration on Ethical Considerations Regarding Health Databases and Biobanks. *Croatian Medical Journal* 57 (2), pp. 207–213. doi:10.3325/cmj.2016.57.207.

Allison, D. B., Brown, A. W., George, B. J. and Kaiser, K. A. 2016. A tragedy of errors. *Nature* 530 (7588), pp. 27–30. doi:10.1038/530027.

Amano, T., González-Varo, J. P., Sutherland, W. J., Montgomery, S. L., Meneghini, R., Packer, A. L., Guha, B. et al. 2013. Languages Are Still a Major Barrier to Global Science. *PLOS Biology* 14 (12), e2000933. doi:10.1371/JOURNAL.PBIO.2000933.

An, O., Pendino, V., D'Antonio, M., Ratti, E., Gentilini, M. and Ciccarelli, F. D. 2014. NCG 4.0: The Network of Cancer Genes in the Era of Massive Mutational Screenings of Cancer Genomes. *Database: The Journal of Biological Databases and Curation*, bau015. doi:10.1093/database/bau015.

Anderson, C. 2008. The end of theory: The data deluge makes the scientific method obsolete. Wired. See https://www.wired.com/2008/06/pb-theory/.

Ankeny, R. A. 2014. The overlooked role of cases in casual attribution in medicine. *Philosophy of Science* 81 (5), pp. 999–1011.

Ashcroft, R. E. 2003. Current Epistemological Problems with Evidence-Based Medicine. *J Med Ethics* 30, pp. 131–135. doi:10.1136/jme.2003.007039.

Ayris, P. et al. 2013. LERU Roadmap for Research Data. See https://www.fosteropenscience.eu/sites/default/files/pdf/598.pdf.

Bastow, R. and Leonelli, S. 2010. Sustainable digital infrastructure. *EMBO Reports* 11 (10), pp. 730–735. doi:10.1038/embor.2010.145.

Beer, D. 2016. *Metric Power.* Basingstoke: Palgrave Macmillan.

Bezuidenhout, L., Leonelli, S., Kelly, A. and Rappert, B. (in press, 2017). Beyond the Digital Divide: Towards a Situated Approach to Open Data. *Science and Public Policy.*

Bliss, C. 2018. *Social by Nature: How Sociogenomics is Redefining What it Means to be Human*. Stanford University Press.

Borgman, C. 2015. *Big Data, Little Data, No Data*. MIT Press.

Boyd, D. and Crawford, K. 2012. Critical questions for big data. *Information, Communication, & Society* 15 (5), pp. 662–679.

Broadbent, A. 2013. *Philosophy of Epidemiology*. Basingstoke: Palgrave Macmillan.

Burton, P. R., Murtagh, M. J., Boyd, A., Williams, J. B., Dove, E. S., Wallace, S. E. et al. 2015. Data Safe Havens in health research and healthcare. *Bioinformatics* 31 (20), pp. 3241–3248. doi:10.1093/bioinformatics/btv279.

Cai, L. and Zhu, Y. 2015. The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. *Data Science Journal* 14, 2. doi:10.5334/dsj-2015-002.

Cambrosio, A., Bourret, P., Keating, P., Nelson, N. C. 2017. Opening the regulatory black box of clinical cancer research: transnational expertise networks and disruptive technologies. *Minerva* 55, 161. doi:10.1007/s11024-017-9324-2.

Cerami, E., Gao, J., Dogrusoz, U., Gross, B. E., Sumer, S. O., Aksoy, B. A., Jacobsen, A., Byrne, C. J., Heuer, M. L., Larsson, E., Antipin, Y., Reva, B., Goldberg, A. P., Sander, C. and Schultz, N. 2012. The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data. *Cancer Discovery* 2 (5), pp. 401–404. doi:10.1158/2159-8290.CD-12-0095.

Clarke B., Gillies D., Illari P., Russo F. and Williamson J. 2013. The evidence that evidence-based medicine omits. *Preventative Medicine* 57, pp. 745–747. doi:10.1016/j.ypmed.2012.10.020.

Collins, F. S. 2010. *The Language of Life: DNA and the Revolution in Personalized Medicine*. 1st Edition. HarperCollins e-books. Retrieved from http://www.amazon.co.uk/The-Language-Life-Revolution-Personalized-ebook/dp/B003100UQU.

Cutcher-Gershenfeld, J. et al. 2017. Five ways consortia can catalyze Open Science. *Nature* 543 (7647), pp. 615–617. See http://www.nature.com/news/five-ways-consortia-can-catalyse-open-science-1.21706.

Demir, I. and Murtagh, M. J. 2013. Data Sharing across Biobanks: Epistemic Values, Data Mutability and Data Incommensurability. *New Genetics and Society* 32 (4), pp. 350–365. doi:10.1080/14636778.2013.846582.

Dove, E. S., Barlas, I.O., Birch, K. et al. 2015. An Appeal to the Global Health Community for a Tripartite Innovation: An "Essential Diagnostics List," "Health in All Policies," and "See-Through 21(st) Century Science and Ethics". *Omics* 19 (8), pp. 435–442. doi:10.1089/omi.2015.0075.

Dove, E. S., David T., Meslin, E. M., Bobrow, M., Littler, K., Nicol, D., de Vries, J. et al. 2016. Ethics Review for International Data-Intensive Research. *Science* 351 (6280), pp. 1399–1400. doi:10.1126/science.aad5269.

Ebeling, M. F. E. 2016. *Healthcare and Big Data: Digital Specters and Phantom Objects*. New York: Palgrave Macmillan US.

Edwards, P. N., Mayernik, M. S., Batcheller, A. L., Bowker, G. C. and Borgman, C. L. 2011. Science friction: Data, metadata, and collaboration. *Social Studies of Science* 41 (5), pp. 667–690. doi:10.1177/0306312711413314.

Eschenfelder, K. R. and Johnson, A. 2014. Managing the Data Commons: Controlled Sharing of Scholarly Data. *Journal for the Association for Information Science and Technology* 65 (9), pp. 1757–1774. doi:10.1002/asi.23086.

Eschenfelder, K. and Shankar, K. 2016. Designing Sustainable Data Archives: Comparing Sustainability Frameworks. *iConference 2016 Proceedings*, pp. 1–7. doi:10.9776/16243.

European Commission (EC) 2016. *Open innovation, open science, open to the world – A vision for the future*. Directorate-General for Research and Innovation. See http://bookshop.europa.eu/en/open-innovation-open-science-open-to-the-world-pbKI0416263/.

European Data Protection Supervisor 2015 Towards a new digital ethics. Data, dignity and technology. Opinion 4/2015. See https://secure.edps.europa.eu/EDPSWEB/webdav/site/mySite/shared/Documents/Consultation/Opinions/2015/15-09-11_Data_Ethics_EN.pdf.

Fecher B., Friesike S. and Hebing M. 2015. What Drives Academic Data Sharing?. *PLOS ONE* 10 (2): e0118053. doi:10.1371/journal.pone.0118053.

Fleming, L. E., Haines, A., Golding, B., Kessel, A., Cichowska, A., Sabel, C. E. et al. 2014. Data mashups: Potential contribution to decision support on climate change and health. I*nternational Journal of Environmental Research and Public Health* 11, pp. 1725–1746.

Fleming L. E., Tempini N., Gordon-Brown H., Nichols G., Sarran C., Vineis P., Leonardi G., Golding B., Haines A., Kessel A., Murray V., Depledge M. and Leonelli S. 2017. Big Data in Environment and Human Health: Challenges and Opportunities. *Oxford Encyclopaedia for Environment and Human Health*. Oxford University Press.

Floridi, L. 2014. *The fourth revolution: how the infosphere is reshaping human reality*. Oxford, UK: Oxford University Press.

Floridi, L. and Illari, P. (eds) 2014. *The Philosophy of Information Quality*. Synthese Library 358. Cham, Switzerland: Springer. doi:10.1007/978-3-319-07121-3.

Forbes, S. A., Bindal, N., Bamford, S., Cole, C., Kok, C. Y., Beare, D., Jia, M., Shepherd, R., Leung, K., Menzies, A., Teague, J. W., Campbell, P. J., Stratton, M. R. and Futreal, P. A. 2011. COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucl. Acids Res*. 39, D945–D950. doi:10.1093/nar/gkq929.

Gitelman, L. 2013. *"Raw data" is an Oximoron*. Cambridge: MIT Press.

Global Alliance for Genomics and Health (GA4GH, 2016). Framework for responsible sharing of genomic and health-related data. See https://www.ga4gh.org/.

Gold, S. 2010. Securing the National Health Service. *Computer Fraud and Security* 2010 (5), pp. 11–14. doi:10.1016/S1361-3723(10)70053-1.

Green, S. and Vogt, H. 2016. Personalizing medicine in silico and in socio. *Humana.Mente Journal of Philosophical Studies* 30, pp. 105–145.

Griffin J., Jordan D. J. and El Gawad A. 2016. Teaching Evidence-Based Medicine in Surgical Education. Open Medicine Journal 3, pp. 337–345. doi:10.2174/18742203016030 10337.

Guyatt G., Cairns J., Churchill D., Cook D., Haynes B., Hirsh J., Irvine J., Levine M., Nishikawa J., Sackett D., Brill-Edwards P., Gerstein H., Gibson J., Jaeschke R., Kerigan A., Neville A., Panju A., Detsky A., Enkin M., Frid P., Gerrity M., Laupacis A., Lawrence V., Menard J., Moyer V., Mulrow C., Links P., Oxman A., Sinclair J. and Tugwell P. 1992. Evidence-Based Medicine: A New Approach to Teaching the Practice of Medicine. *JAMA* 268 (17), pp. 2420–2425. doi:10.1001/jama.1992.03490170092032.

Harris, A., Kelly, S. and Wyatt, S. 2016. *CyberGenetics: Health Genetics and New Media*. (*Genetics and Society*). London: Routledge/Taylor and Francis Group.

Hey, T., Tansley, S. and Tolle, K. 2009. *The fourth paradigm: Data-intensive scientific discovery*. Redmond, WA: Microsoft Research.

Hiatt, R. A. et al. 2015. Leveraging State Cancer Registries to Measure and Improve the Quality of Cancer Care: A Potential Strategy for California and Beyond. *J Natl Cancer Inst* (2015) 107 (5): djv047. doi:10.1093/jnci/djv047.

Hicks, D., Wouters, P., Waltman, L., de Rijcke, S. and Rafols, I. 2015. The Leiden Manifesto for Research Metrics. *Nature* 520, pp. 429–431. doi:10.1038/520429a.

Hogle, Linda F. 2016. Data-Intensive Resourcing in Healthcare. *BioSocieties* 11 (3), pp. 1–22. doi:10.1057/s41292-016-0004-5.

Hood, L. and Friend, S. H. 2011. Predictive, personalized, preventive, participatory (P4) cancer medicine. *Nat Rev Clin Oncol* 8, pp. 184–187. doi:10.1038/nrclinonc.2010.227.

Kallinikos, J. and Tempini, N. 2014. Patient data as medical facts: Social media practices as a foundation for medical knowledge creation. *Information Systems Research* 25 (4), pp. 817–833. doi:10.1287/isre.2014.0544.

Kaye, J. et al. 2015. Dynamic Consent: a patient interface for 21st century research networks. *European Journal of Human Genetics* 23, pp. 141–146. doi:10.1038/ejhg.2014.71.

Kelly, S. 2008. Choosing not to choose: reproductive responses of parents of children with genetic conditions or impairments. *Sociology of Health and Illness* 31 (1), pp. 81–97.

Khoury, M. J. and Ioannidis, J. P. A. 2014. Big data meets public health: Human well-being could benefit from large-scale data if large-scale noise is minimized. *Science* 346 (6213), pp. 1054–1055. doi:10.1126/science.aaa2709.

Kitchin, R. 2014. *The data revolution: Big data, open data, data infrastructures and their consequences*. London: SAGE.

Kitchin, R. and McArdle G. 2016. What Makes Big Data, Big Data? Exploring the Ontological Characteristics of 26 Datasets. *Big Data & Society* 3 (1), pp. 1–10. doi:10.1177/2053951716631130.

Lagoze, C. 2014. Big data, data integrity and the fracturing of the control zone. *Big Data & Society* 1 (2), pp. 1–11. doi:10.1177/2053951714558281.

Lawrence, B., Jones, C., Matthews, B., Pepler, S. and Callaghan, S. 2011. Citation and Peer Review of Data: Moving Towards Formal Data Publication. *International Journal of Digital Curation* 6 (2), pp. 4–37. doi:10.2218/ijdc.v6i2.205.

Leonelli, S. 2009. Centralising Labels to Distribute Data: The Regulatory Role of Genomic Consortia. In Atkinson, P., Glasner, P. and Lock, M. (eds) *The Handbook for Genetics and Society: Mapping the New Genomic Era*. London: Routledge, pp. 469–485.

Leonelli, S. 2012. When Humans Are the Exception: Cross-Species Databases at the Interface of Clinical and Biological Research. *Social Studies of Science* 42 (2), pp. 214–236. doi:10.1177/0306312711436265.

Leonelli, S. 2013. Global Data for Local Science: Assessing the Scale of Data Infrastructures in Biological and Biomedical Research. *BioSocieties* 8 (4), pp. 449–465. doi:10.1057/biosoc.2013.23.

Leonelli, S. 2016a. *Data-Centric Biology: A Philosophical Study*. Chicago, IL: Chicago University Press.

Leonelli, S. 2016b. Locating ethics in data science: responsibility and accountability in global and distributed knowledge production. *Philosophical Transactions of the Royal Society: Part A*, 374 (2083), 20160122. doi:10.1098/rsta.2016.0122.

Leonelli, S. 2017. Global Data Quality Assessment and the Situated Nature of "Best" Research Practices in Biology. *Data Science* 16(32): pp. 1-11. doi:10.5334/dsj-2017-030.

Leonelli, S., Smirnoff, N., Moore, J., Cook, C. and Bastow, R. 2013. Making Open Data Work in Plant Science. *Journal for Experimental Botany* 64 (14): pp. 4109–4117. doi:10.1093/jxb/ert273.

Levin, N., Leonelli, S., Weckowska, D., Castle, D. and Dupré, J. 2016. How Do Scientists Understand Openness? Exploring the Relationship between Open Science Policies and Research Practice. *Bulletin for Science and Technology Studies* 36 (2), pp. 128–141. doi:10.1177/0270467616668760.

Loettgers, A. 2009. Synthetic Biology and the Emergence of a Dual Meaning of Noise. *Biological Theory* 4 (4), pp. 340–355. doi:10.1162/BIOT_a_00009-.

Lucivero, F. and Prainsack, B. 2015. The lifestylisation of healthcare? "Consumer genomics" and mobile health as technologies for healthy lifestyle. *Applied & Translational Genomics* 4, pp. 44–49. doi:10.1016/j.atg.2015.02.001.

Lupton, D. and Michael, M. 2017. "For Me, the Biggest Benefit Is Being Ahead of the Game": The Use of Social Media in Health Work. *Social Media + Society* 3 (2). doi:10.1177/2056305117702541.

Manrai A. K., Cui Y., Bushel P. R. et al. 2017. Informatics and data analytics to support exposome-based discovery for public health. *Annu Rev Public Health* 38, pp. 279–294. doi:10.1146/annurev-publhealth-082516-012737.

Markens, S. 2013. "It just becomes much more complicated": Genetic counselors' views on genetics and prenatal testing. *New Genetics and Society* 32, pp. 302–321.

Marr, B. 2015. *Big Data: Using smart big data, analytics and metrics to take better decisions and improve performance*. John Wiley & Sons.

Mauthner, N. S. and Parry, O. 2013. Open Access Digital Data Sharing: Principles, Policies and Practices. *Social Epistemology* 27 (1), pp. 47–67. doi:10.1080/02691728.2012.7606 63.

Mayer-Schönberger, V. and Cukier, K. 2013. *Big data: A revolution that will transform how we live, work, and think*. New York: Eamon Dolan/Houghton Mifflin Harcourt.

McAllister, J. W. 2007. Model Selection and the Multiplicity of Patterns in Empirical Data. P*hilosophy of Science* 74 (5), pp. 884–894. doi:10.1086/525630.

McKiernan, E. C., Bourne, P. E., Brown, C. T., Buck, S., Kenall, A., Lin, J., McDougall, D. et al. 2016. How Open Science Helps Researchers Succeed. *eLife* 5 (July), pp. 1–19. doi:10.7554/eLife.16800.

McMichael, A. J. and Haines, A. 1997. Global climate change: The potential effects on health. *BMJ* 315 (7111), pp. 805–809.

Merelli, I., Perez-Sanchez, H., Gesing, S. and D'Agostino, D. 2014. Managing, Analysing, and Integrating Big Data in Medical Bioinformatics: Open Problems and Future Perspectives. *Biomed Research International*, vol. 2014. doi:10.1155/2014/134023.

Mittelstadt, B. D. and Floridi, L. 2016. The ethics of biomedical big data: Current and Foreseeable Issues in Biomedical Contexts. *Sci Eng Ethics* 22 (2), pp. 303–341. doi:10.1007/s11948-015-9652-2.

Morey, R. D. et al. 2016. The Peer Reviewers' Openness Initiative: incentivizing open research practices through peer review. *R. Soc. opensci.* 3, 150547. doi:10.1098/rsos.150547.

Müller-Wille, S. and Charmantier, I. 2012. Natural history and information overload: The case of Linnaeus. *Studies in the History and the Philosophy of the Biological and Biomedical Sciences* 43, pp. 4–15. doi:10.1016/j.shpsc.2011.10.021.

Murphy, M. 2017. The *Economization of Life*. Durham: Duke University Press.

Nichols, G. L., Andersson, Y., Lindgren, E., Devaux, I. and Semenza, J. C. 2014. European monitoring systems and data for assessing environmental and climate impacts on human infectious diseases. *International Journal of Environmental Research & Public Health* 11 (4), pp. 3894–3936. doi:10.3390/ijerph110403894.

Normandeau, K. 2013. Beyond volume, variety and velocity is the issue of big data veracity. *Inside big data*. See http://insidebigdata.com/2013/09/12/beyond-volume-variety-velocity-issue-big-data-veracity/.

Nuffield Council on Bioethics 2015. *The collection, linking and use of data in biomedical research and health care: ethical issues*. London, UK: Nuffield Council on Bioethics. See http://nuffieldbioethics.org/wp-content/uploads/Biological_and_health_data_web.pdf.

O'Brien D., Ullman J., Altman M., Gasser U., Bar-Sinai M., Nissim K., Vadhan S. and Wojcik M. J. 2017. OECD Recommendations on Health Data Governance. See http://www.oecd.org/els/health-systems/health-data-governance.htm.

Open Science Collaboration 2015. Estimating the reproducibility of psychological science. *Science* 349 (6251), pp. 943–952. doi:10.1126/science.aac4716.

Ossorio, P. 2011. Bodies of data: Genomic data and bioscience data sharing. *Social Research* 78 (3), pp. 907–932.

Ostrom, E. 2005. *Understanding institutional diversity*. Princeton: Princeton University Press.

Paul, D. 2017. Norm change in genetic services: How the discourse of choice replaced the discourse of prevention. *Varia Historia*, Belo Horizonte, 33, pp. 21–47.

Prainsack, B. and Buyx, A. 2017. *Solidarity in Biomedicine and Beyond*. Cambridge: Cambridge University Press.

Prainsack, B. and Vayena, E. 2013. Beyond the clinic: "direct-to-consumer" genomic profiling services and pharmacogenomics. *Pharmacogenomics* 14 (4), pp. 403–412. doi:10.2217/pgs.13.10.

Rajan, K. S. 2017. *Pharmocracy: Value, Politics, and Knowledge in Global Medicine*. Durham: Duke University Press.

Royal Society 2012. Science as an open enterprise. See https://royalsociety.org/~/media/Royal_Society_Content/policy/projects/sape/2012-06-20-SAOE.pdf.

Sansone, S. A., Rocca-Serra, P., Field, D., Maguire, E., Taylor, C., Hofmann, O., Fang, H. et al. 2012. Toward Interoperable Bioscience Data. *Nature Genetics* 44 (2), pp. 121–126. doi:10.1038/ng.1054.

Schofield, P. N., Bubela, T. et al. 2009. Post-publication sharing of data and tools. *Nature* 461, pp. 171–173. doi:10.1038/461171a.

Science International 2015. Open data in a big data world. See https://www.icsu.org/cms/2017/04/open-data-in-big-data-world_short.pdf.

Sharon, T. 2017. Self-Tracking for Health and the Quantified Self: Re-Articulating Autonomy, Solidarity, and Authenticity in an Age of Personalized Healthcare. *Philosophy & Technology* 30 (1), pp. 93–121. doi:10.1007/s13347-016-0215-5.

Shaver, L. 2010. The right to science and culture. *Wisconsin Law Rev.* 1, pp. 121–184. See https://www.aaas.org/sites/default/files/Shaver_ScienceandCulture.pdf.

Shutt R. and O'Neill, C. 2015. *Doing Data Science: Straight talk from the front line*. Cambridge: O'Reilly.

Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L. J., Eilbeck, K., Ireland, A., Mungall, C. J. et al. 2007. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotechnol* 25 (1), pp. 1251–1255. doi:10.1038/nbt1346.

Solomon, M. 2015. *Making Medical Knowledge*. Oxford: Oxford University Press.

Tempini, N. 2015. Governing PatientsLikeMe: Information production and research through an open, distributed and data-based social media network. *The Information Society* 31 (2), pp. 193–211. doi:10.1080/01972243.2015.998108.

Tempini, N. 2016. Science Through the "Golden Security Triangle": Information Security and Data Journeys in Data-Intensive Biomedicine. In: *Proceedings of the 37th International Conference on Information Systems* (*ICIS 2016*). Dublin.

Tempini, N. (forthcoming). "Till Data Do Us Part: Understanding Data-based Value Creation in Data-Intensive Infrastructures". *Information & Organization*.

Tempini, N. and Leonelli, S. (2018). Genomics and Big Data. In: Gibbons, S. et al. (eds). *Routledge Handbook for Genomics, Health and Society*. Routledge, UK.

Treadway, J., Hahnel, M., Leonelli, S., Penny, D., Groenewegen, D., Miyairi, N., Hayashi, K., O'Donnell, D. and Hook, D. 2016. The State of Open Data Report. figshare. doi:10.6084/m9.figshare.4036398.v1.

van Dijck, J. and Poell, T. 2016. Understanding the Promises and Premises of Online Health Platforms. *Big Data and Society*, no. June, pp. 1–11. doi:10.1177/2053951716654173.

Vayena E. and Gasser, U. 2016. Between openness and privacy in genomics. *PLoS Med* 13, e1001937. doi:10.1371/journal.pmed.1001937.

Vayena E. and Tasioulas J. 2015 "We the scientists": a human right to citizen science. *Philos.Technol.* 28, pp. 479–485. doi:10.1007/s13347-015-0204-0.

Vayena E., Mastroianni A. and Kahn J. 2013. Caught in the web: informed consent for online health research. *Sci. Transl. Med.* 5, 173fs6. doi:10.1126/scitranslmed.3004798.

Wang, W. and Krishnan, E. 2014. Big data and clinicians: a review on the state of the science. *JMIR Medical Informatics* 2 (1), e1. doi:10.2196/medinform.2913.

Ward, J. S. and Barker, A. 2013. Undefined by data: A survey of big data definitions. School of Computer Science University of St Andrews, United Kingdom. See http://www.adambarker.org/papers/bigdata_definition.pdf.

Waters, C. K. 2007. The nature and context of exploratory experimentation: An introduction to three case studies. *His. Phil. Life Sci.* 29 (3), pp. 275–284.

Weber, G. M., Mandl, K. D. and Kohane, I. S. 2014. Finding the missing link for big biomedical data. *JAMA* 311 (24), pp. 2479–2480. doi:10.1001/jama.2014.4228.

Whitmee, S., Haines, A., Beyrer, C., Boltz, F., Capon, A. G., Ferreira de Souza Dias, B. et al. 2015. Safeguarding human health in the Anthropocene epoch: Report of The Rockefeller Foundation–Lancet Commission on planetary health. *The Lancet* 386, pp. 1973–2028.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A. et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 3, 160018. doi:10.1038/sdata.2016.18.

Wood, A. 2015. Integrating approaches to privacy across the research lifecycle: when is information purely public? *Berkman Center Research Publication* no. 2015-7. doi:10.2139/ssrn.2586158.

Woodward, J. 2015. Data, Phenomena, Signal, and Noise. *Philosophy of Science* 77 (5), pp. 792–803. doi:10.1086/656554.

Wyatt, S., Harris, A., Adams, S. and Kelly, S. 2013. Illness Online: Self-reported Data and Questions of Trust in Medical and Social Research. *Theory, Culture & Society* 30 (4), pp. 131–150. doi:10.1177/0263276413485900.

Zook, M., Barocas, S., Boyd, D., Crawford, K., Keller, E., Gangadharan, S. P., Goodman, A. et al. 2017. Ten Simple Rules for Responsible Big Data Research. *PLOS Computational Biology* 13 (3), e1005399. doi:10.1371/journal.pcbi.1005399.

Abbreviations

AI	Artificial intelligence
BBSRC	Biotechnology and Biological Sciences Research Council
CSSI	Conseil suisse de la science et de l’innovation / Consiglio svizzero della scienza e dell’innovazione
EBM	Evidence-based medicine
EU	European Union
FAIR	Findable, Accessible, Interoperable and Reusable
GA4GH	Global Alliance for Genomics and Health
GP	General practitioner
GPS	Global Positioning System
ISA	Investigation/Study/Assay
IT	Information Technology
LERU	League of European Research Universities
MIAME	Minimal Information About Microarray Experiments
MSc	Master of Science
NIH	National Institutes of Health
OBO	Open Biomedical Ontology
OECD	Organisation for Economic Co-operation and Development
SSIC	Swiss Science and Innovation Council
SWIR	Schweizerischer Wissenschafts- und Innovationsrat
UK	United Kingdom
US	United States

Imprint

Swiss Science and Innovation Council SSIC
Einsteinstrasse 2
CH-3003 Bern
T +41 (0)58 463 00 48
F +41 (0)58 463 95 47
swir@swir.admin.ch
www.swir.ch

ISBN 978-3-906113-52-4
Bern 2017

Copy-editing: Doris Tranter, Stéphane Gillioz
Concept and Design: Modulator, Branding + Design
Cover image: amiak

Swiss Science and Innovation Council SSIC
Einsteinstrasse 2
CH-3003 Bern

T +41 (0)58 463 00 48
F +41 (0)58 463 95 47
swir@swir.admin.ch
www.swir.ch