Social Science & Medicine 272 (2021) 113760

Contents lists available at ScienceDirect

Social Science & Medicine

journal homepage: http://www.elsevier.com/locate/socscimed

Actionable data for precision oncology: Framing trustworthy evidence for exploratory research and clinical diagnostics

Niccolò Tempini [a,b,c,*], Sabina Leonelli [a,b,c,d]

[a] Egenis, Department of Sociology, Philosophy and Anthropology, University of Exeter, Byrne House, St Germans Road, EX4 6TJ, Exeter, UK
[b] Institute for Data Science and Artificial Intelligence, University of Exeter, UK
[c] Alan Turing Institute, London, UK
[d] School of Humanities, University of Adelaide, Adelaide, 5005, Australia

ARTICLE INFO

Keywords:
Cancer genomics
Actionability
COSMIC
Data infrastructure
Precision medicine
Data trustworthiness
Big data
Data curation

ABSTRACT

Huge amounts of genomic data produced by researchers around the world undermine data-centred discovery and therapeutic development. This paper considers how researchers make decisions about the *actionability* of specific datasets and the conditions that allow such data to be trusted. We discuss the case of COSMIC, a leading cancer genomics database which aggregates a large amount of sources. We research what the actionability of cancer data means in different situations of use, contrasting exploratory and diagnostics research. They highlight different questions and concerns upon genomic data use in medical research. At the same time, strategies and justifications pursued to evaluate and re-use can also share important similarities. To explain differences and similarities, we argue for an understanding of actionability and trust in data that depends on the goals and resources within the situation of inquiry, and the social epistemology of standards.

1. Introduction

The vision of clinical advancement offered by precision medicine implicates a large number of elements (Keating and Cambrosio 2011, 2013; Prainsack 2020). In oncology, where precision medicine has been heavily promoted over the last decade, these include overarching epistemic frameworks, data sources, technologies, organisational processes, strategic partnerships, experts, capital investments, legal and policy frameworks, drugs and algorithms (Cambrosio et al., 2013; Hogle 2016; Levin 2018; Vignola-Gagné et al., 2017). Against this background, this paper highlights the challenges involved in the evaluation and use of sequencing research data in the development of innovative oncological diagnostics and treatments, focusing particularly on the impact of data management procedures and decisions on data interpretation. We argue that data infrastructures shape data interpretation and facilitate users' trust in the actionability of data as evidence for future applications of genomic innovations in clinical settings – an achievement involving a degree of informed speculation about the quality and significance of data.

Genomic data are among the best standardised and most valued data types available to precision medicine. Yet they require complex intermediations to be used as medical evidence (Rheinberger 2010), including curation and visualisation practices that make it possible for these data to move across user communities (Lowe 2018). This considerably complicates the interpretation of genomic sequences (Huang et al., 2016). In what follows, we document the efforts involved in maintaining the Catalogue of Somatic Mutations in Cancer (COSMIC), a database developed to distribute research evidence on associations between genetic sequences and specific cancer types, susceptibility to targeted treatments and other functional implications (Bamford et al., 2004; Tate et al., 2019). COSMIC is widely regarded as a key resource for achieving reliable interpretations of genomic data and related mutations in oncology. An investigation of COSMIC curatorial practices and their reception by COSMIC users provides an excellent window into practices of genomic data management, evaluation and use within oncology (see also Tempini, 2020a).

Finding evidence that suggests that a given mutation is associated with a specific cancer type is a common heuristic strategy for developing mechanistic hypotheses on the drivers of the disease (Bechtel 2019). However, current understandings of the causal role of particular genes and mutations – and more generally the aetiology of cancer – remain limited and contested (Bertolaso 2016; Plutinsky 2018). Consequently,

* Corresponding author. Egenis, Department of Sociology, Philosophy and Anthropology, University of Exeter, Byrne House, St Germans Road, EX4 6TJ, Exeter, UK.
E-mail addresses: n.tempini@exeter.ac.uk (N. Tempini), s.leonelli@exeter.ac.uk (S. Leonelli).

https://doi.org/10.1016/j.socscimed.2021.113760
Received in revised form 22 January 2021; Accepted 4 February 2021
Available online 11 February 2021

it is difficult to decide how to take action and what to target – an issue typically referred to as *actionability* by practitioners. Actionability captures the *"aim to generate predictive relationships between genetic information and drug therapies"* (Nelson et al., 2013:405) in a context where there is no straightforward pairing between observed sequences and definitive assessment of cause and trajectory of the individual cancer. Crucially, making genetic mutation data actionable depends on the constraints posed by related socio-technical systems – and several such considerations are involved when assessing alternative courses of action, *"such as the regulatory status of those mutations and drugs, the availability of testing and treatments within health care systems, and the geographical location and design of clinical trials for drugs still under development"* (Nelson et al., 2013:414). Actionability has a strong organisational connotation, highlighting the implications of data for division of labour (407). Given the current data deluge, researchers have no choice but to resort to databases of genomic evidence and the interpretive decisions that have been embedded within them (Timmermans 2015).

This paper discusses epistemic assessments of data actionability by COSMIC curators and users, with a particular interest in data practices and the organisational interdependencies that shape them (Leonelli, 2016; Tempini, 2017, 2020a). We explore the concept of data actionability in contexts other than clinical settings, thus contributing to studies of genomic evidence in oncology and the associated epistemological, ethical and professional uncertainties (Bergeron et al., 2020; Bergeron and Castel, 2011; Cambrosio et al., 2020; Kerr et al., 2019; Timmermans et al., 2017). Clinical situations are complex (hence the emergence of Molecular Tumour Boards, where different specialists convene – Bourret and Cambrosio 2019) and many steps of data production and manipulation separate the initial clinical encounter from the diagnostic and therapeutic response. Our study focuses on one specific fragment of the genomic medicine mosaic. We analyse the conditions for the actionability of COSMIC data in light of two concepts informed by previous work in the social studies of science: *speculative reasoning* and *trust*. Our analysis aims to show how assessments of actionability are tied to both these concepts.

By speculative reasoning we mean the process whereby the persistent uncertainty related to genomic sequences is harnessed and transformed into possibility, promise and opportunity. This enables the emergence of a research and innovation industry. As Fortun (2008) argues, a genomic industry is sustainable if it is able to promise a 'revolution' in medicine and envision and articulate the key milestones, resources, and methods that will enable it. Our use of the concept follows Fortun in that speculation is not simply a discursive notion – a way of reasoning about epistemic uncertainty by way of hypotheticals and assumptions. Nor is it reducible to a materialist concept grounded in risk-taking, material and financial commitments, and rewards. Instead, we understand speculation as a twofold concept within which these discursive and materialist interpretations are inseparable and sustain each other (Fortun, 2008). When exception rules, the absence of definitive evidence enables actions based upon incomplete assumptions (Fortun, 2008; Kerr et al., 2019). Hence, uncertainty is productive (Timmermans et al., 2017). It also opens the possibility for different initiatives to interface with one another while developing interdependencies and operationalising assumptions. Partaking in a regime of innovation grounded on the promise of a future in the making (Rajan 2006; Fortun 2008; Hilgartner 2017), the production of speculative judgements about the importance of specific mutation data is a key element of data curation in databases like COSMIC. Curators who retrieve and harvest published data must judge which are worth adding to and maintaining in the database. This form of 'speculation' happens in the midst of the deafening 'noise' generated by the observation of a multitude of mutations, which can impede the search for any 'signal'. In this sense, the enterprise of curation will be shown to be about enabling data-based decisions and actions in under-specified situations of subsequent use.

The other key concept is trust. COSMIC data are used as a standard for the functional interpretation of mutations by several teams and organisations operating in cancer genomic medicine, including some of the most prominent cutting-edge systems. For instance, the Memorial Sloan Kettering Cancer Centre's cBioPortal integrates COSMIC data as one of several layers of data that in-house clinicians juxtapose to individual cancer profiles (Gao et al., 2013). And research and diagnostics services company Personal Genome Diagnostics, a Johns Hopkins' spin-off, performs comparisons of sequenced sample data to COSMIC as part of its routines (see Fig. 1) (Jones et al., 2015). The COSMIC team has published many papers that have been cited thousands or many hundreds of times (e.g., Bamford et al., 2004; Forbes et al., 2011; Forbes et al. 2015; Forbes et al. 2016; Forbes et al. 2017).

The widespread reliance on the contributions of a single team is particularly interesting in light of the epistemic uncertainties involved in working on a complex and elusive disease. COSMIC data seem to function as *terra firma* upon which competing scientific efforts, navigating uncharted waters, can pivot and develop the innovations they lay claim to. *In what ways do attitudes of trust towards COSMIC shape assessments of data actionability? How do researchers speculate on opportunities and risks linked to the use of a community standard such as the COSMIC database?* By addressing these questions, this paper seeks to flesh out the relationship between data actionability, speculative reasoning and trust that characterises data management practices in cancer genomics.

Key STS work has investigated the conditions that allow researchers to trust heterogeneous evidence (e.g., Fortun 2008; Lynch et al., 2010; Porter 1995; Timmermans 2015). Here we focus on the considerations and concerns held by researchers involved in trusting, speculating, and acting with data: 1) that have been made available by curators who, in turn, are concerned with issues of trust, speculation and actionability of data sources; 2) that are routinely produced by an institution – COSMIC – that is devoted to data stewardship and the development of data infrastructure; and 3) whose access and reuse depends substantially on aligning data practices and infrastructures with that institution. We thus move across the gap separating curators and data users, to try and show how consequent experiences of genomic data work of curators and data re-using researchers inform and shape one another. We build on recent STS insights into standardization, data infrastructures and curation, trust and the role of evidence in personalised and precision medicine (Prainsack 2020; Cambrosio et al., 2020) to reveal some of the epistemic, organisational and infrastructural interdependencies that shape the circulation and actionability of cancer genomics data.

2. Methods

Our analysis is based on a qualitative study of COSMIC curation practices carried out between 2015 and 2017. This encompassed twenty-one interviews with team members and data users. With the exception of one, all interviews were carried out during two field visits by Tempini and over a total of six days. Interviews were semi-structured and followed a questionnaire that was customized for each interviewee on the basis of their professional and scientific background, organisational role and work context. These interviews were accompanied by some opportunities for on-site observation, including an exploratory site visit in 2015, participation in an awayday during which the COSMIC team and external collaborators presented their work and reflected on the state of the project, and informal discussions during a relevant workshop on data practices in bioinformatics and biomedicine in January 2017. The case study is also informed by secondary evidence gathered from the COSMIC website, browsing of available data resources, associated documentation of procedures and standards, and the cancer genomics literature related to COSMIC. The aim of the fieldwork was to document perspectives and the roles of COSMIC staff in the development and assemblage of the COSMIC database, as well as the perspectives of users in bringing COSMIC to bear in their own data interpretation processes (cfr. Tempini 2020a). Those interview transcripts which participants have agreed to disclose are accessible from the

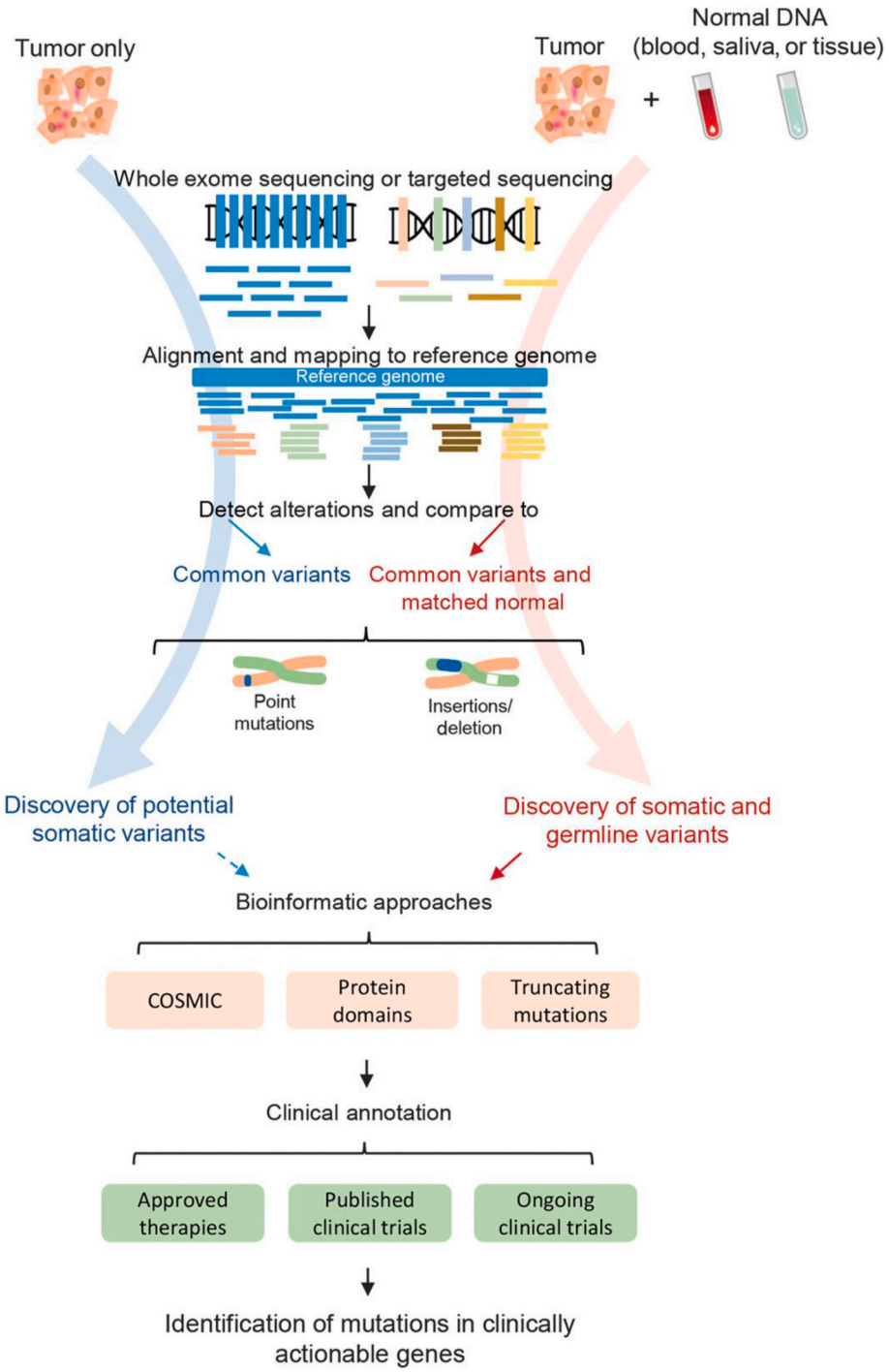

Fig. 1. Schematic description of whole-exome or targeted next-generation sequencing analyses (from: Jones et al., 2015).

Exeter Data Studies data collection hosted by Zenodo (Tempini 2020c; yExeter Data Studies). This case study is one of several carried out within a larger project on the impact of big data methods and infrastructures in the biological and biomedical sciences (www.datastudies.eu).

COSMIC was chosen as an ideal research site for understanding data management practices in a fast-changing field of biomedical research. The opportunity for broader contextualization of research data practices provided by the parent project allowed us to restrict our focus to COSMIC curators and users. Through content analysis of the data we collected at COSMIC, we identified data curators' concerns about the quality of the data stored in the database, the challenges of adhering to the curatorial standards of the project, and the strategies used to meet them. We also gathered user perspectives on the implications of related organisational procedures for the actionability of COSMIC data. Below, we report an analysis of these views and challenges supported by select interview excerpts. The analysis is intended to inform future research on how COSMIC impacts downstream clinical practice, a topic that which is beyond the scope of this paper.

The paper is structured as follows. First we situate COSMIC in the field of precision oncology. We then report on team members, and users', views; and we identify differences between the views and assumptions held by data users working within exploratory research and within the field of diagnostics. We conclude by discussing, in light of philosophical and social scientific scholarship on trust in evidence and in

organisations (Hawley 2017; Porter 1995; Timmermans 2015), how users' judgement of reliability of the data, speculative value, and actionability shape one another. The procedural standardization of curation processes and the status of COSMIC as a benchmark allow researchers to rely on COSMIC data for their analytic processes, while at the same time ensuring that the evidence remains broadly in line with their own expert assumptions.

3. The Catalogue of Somatic Mutations in Cancer (COSMIC)

The Catalogue of Somatic Mutations in Cancer (COSMIC) is a project of the Sanger Institute based at the Wellcome Genome Campus in Hinxton, UK. It originated from a cancer genetics 'list', a spreadsheet curated by the principal investigators of the Institute's Cancer Genome Project's (CGP). Formalizing their knowledge and state of the art, this list contained several dozen entries on genes that, when mutated, are often implicated in human cancers. Later, it was published as the Cancer Gene Census (Futreal et al., 2004). In 2004, the idea of a gene-centric curation of cancer genetics research data was turned into a full-fledged web project, and COSMIC was created to promote the circulation of research evidence in the community (Bamford et al., 2004). COSMIC launched with an extensive curation of the evidence for the role of four genes in cancer. In the following years, the focus shifted from cancer genetics to cancer genomics (Forbes et al., 2006, 2008). This required adjustments in data aggregation and dissemination strategies, since the then new genomic studies generated much bigger datasets. By 2005, the number of genes curated by COSMIC has grown to 28. Importing new genomic data from CGP's cancer cell lines then added a further 518 mutated genes from over 124,000 samples. The rapid growth continued: by 2008, genes in COSMIC counted over 4800 and sampled 250,000; in 2011, the counts were, respectively, over 18,000 and 542,000; and in 2015, the number of samples exceeded one million (Forbes et al., 2006, 2008, 2011, 2015). *"Originally designed to detail simple coding gene point mutations, COSMIC now describes millions of coding mutations, noncoding mutations, genomic rearrangements, fusion genes, copy number abnormalities and gene expression variants across the human genome"* (Forbes et al., 2015:D805). As the volume and types of data kept multiplying, questions about what data are most important to the community resurfaced. Several integrations of COSMIC with other cancer databases were carried out, the genomic data of the consortia The Cancer Genome Atlas (TCGA) and International Cancer Genome Consortium (ICGC) were imported, and interfaces with other bioinformatics infrastructures (e.g., Ensembl) developed (Forbes et al., 2011).

COSMIC has since become widely used in precision oncology. Research teams and industry routinely produce genomic data from cancer samples and need a trustworthy reference with which to compare their data. COSMIC shows which mutated sequences were observed in other cancer patients and highlights potential associations with other observations and events. The increasingly complex and open-ended picture of the causative processes of cancer (Plutinsky 2018) has increased the need for such a point of reference for brokering research evidence. As COSMIC Director Simon Forbes put it, *"We want to focus on* supporting *precision oncology as much as possible because, while a number of the mutations causing cancer are understood, quite a large number are considered potentially causing cancer, but no one quite knows how or why"* (SF). A collection of cancer-related mutations assembled over fifteen years by a specialised team in a leading research institution, COSMIC provides an authoritative guide to research evidence. It helps the research community and companies to reduce duplication of effort and related costs. As a diagnostics researcher put it: *"Every cancer company would be reinventing the wheel to try and filter out those common mutations. It's better done in one place with one team so that people can benefit"* (D1). By enabling cost-efficiency and procedural consistency while relieving its users of difficult decisions about how to source, aggregate and evaluate the evidence base, COSMIC has become a standard for cancer genomics (see Timmermans 2015).

3.1. Data integration procedures

The rigour of COSMIC's data integration procedures is a consequence of standardization and specialization in the sourcing and curation of cancer genomic data. Members of the curation team focus on either of two processes, which they refer to as in-depth, manual, genic curation and broad, semi-automated, genomic curation, respectively.

In the first kind of curation, post-doctoral researchers with a background in biology aim to curate all key papers that are published on genes they specialize in. A list of key genes is shared and periodically reviewed by the team of curators, who exchange insights into research trends, promising discoveries, and genes with growing evidence of implication in carcinogenesis. Each curator is responsible for curating evidence relating to a subset of genes. They receive a daily automated report of all new publications about any of 'their' genes. They read and extract data from papers and input them into COSMIC by means of a dedicated web interface. New curators are able to keep up with the literature published on just a few genes, while the most experienced may watch over the literature of several dozens. Manual curation is suited to generating deep and high-quality data. Papers deemed to be of high quality or promise are prioritized. Literature that adds evidence on already well-known relationships is of secondary importance and may not be curated. Curators need to carefully evaluate opportunities, and curate only the literature that yields the most value. Blindly curating all the papers published on a particular gene would be wasteful, as not all studies report novel evidence; at the same time, fixating on the same genes risks missing emerging trends.

Coverage is a challenge for this kind of curation. The number of potentially relevant genes is enormous. Most mutations are thought to be a consequence of cancer. These are called 'passenger' as opposed to causative 'driver' mutations, but establishing which mutations are passengers and which may be drivers is laborious. The COSMIC team – nor any other cancer genomics team, to our knowledge – does not have the resources to manually curate all literature about all genes that are suspected to play a role in cancer. At periodic meetings, the collective of curators discusses emerging trends and whether curators' watch lists need updating, taking on genes not hitherto covered. Considerations shaping the selection of genes for curation may include evidence of causal effect as well as broader research trends, and whether new genes are 'neighbours' of already curated, 'old' genes – as publications about the latter often include or cite the former. A list of genes implicated in cancer causation, the Cancer Gene Census, has been maintained by COSMIC since its early years (Sondka et al., 2018). Upon periodic review, genes can be 'promoted' or 'demoted' from the Cancer Gene Census. The list includes more genes than the team can manually curate and an attempt is made to distinguish mutated genes for which there is strong causal evidence from those for which evidence is unclear. The Census is an authoritative 'hotlist' of the most important mutated genes.

COSMIC also incorporates data that are not manually curated. These data are acquired from whole-genome and gene-panel datasets. In this second type of data curation, post-doctoral researchers with programming skills download data from repositories (e.g., The Cancer Genome Atlas) or from publications' supplemental materials. They integrate such data into COSMIC by scripting code. Here, the emphasis is less on selection of the curation sources and targets. Broad genomic data curation is an 'agnostic' process since the team is not integrating these data on the assumption of a relevant and potentially causative association of specific genes in carcinogenesis (as is the case with manual curation procedures), but rather to make comprehensive datasets of mutations observed in cancer samples available in a trusted repository. A greater proportion of this imported data is about genes that are mutated as a result of cancer ('passenger' mutations) rather than as the cause of cancer ('driver' mutations), heightening the need to distinguish between data sources and mutation types. COSMIC updates are released every three months with a new version of the database. Emphasising standardization and data traceability, the team also publishes a summary of its curatorial

processes on COSMIC website.

3.2. Making research data aggregation processes trustworthy

COSMIC curators' reflections in interviews are shaped by mounting pressures for COSMIC to publish reliable data. This is the result of 1) new uses developed for the data by the community on the back of methodological and theoretical innovations, and 2) the ballooning amounts and diversity of cancer research data being published. The staff regularly meet users, organise COSMIC training workshops, and attend international conferences to understand which uses should be better supported, which kinds of data are most valuable and for whom, and how to improve their availability. The current business model, a hybrid of capped charitable funds and private fees, has opened the project to collaborations with private companies. These make the project more sensitive to the specific priorities of its partners but have also created a stable collaborative basis for understanding user needs and to access external expertise on complex translational questions. COSMIC's pivotal position is founded on the maintenance of a core of well-curated datasets. As of 2015, over 30% of the papers examined by COSMIC curators had been rejected on quality grounds (Forbes et al., 2015). Standardised curation practices are also needed to evaluate which types of data do not fall within the current scope of curation at present but might become important in the future: 'new' genes that are not yet part of the set of curation targets, and new kinds of data, e.g., on drug resistance or methylation.

In curating the genomic data deluge, some trade-offs are unavoidable. Human resources are limited, which constrains the choice of curation targets. Also, the research-orientation of the project creates tensions with respect to clinical applications of COSMIC data. The adoption of research-grade databases in the clinic always poses a degree of risk and concern. Curators can be anxious about the consequences of potential curation mistakes: "*I really get worried about making mistakes. […] If you're going to be a good curator, and I think that does apply to other people in the office too, but you've got to … have a certain level of worry, or fear almost, of making a mistake […] I really check everything, and I probably spend a bit too long checking things, because then you don't get quite as much done*" (C1). Worries around quality revolve not only around what is included (errors, false positives and potential mis-categorizations), but also around what may have been omitted that should have been selected for inclusion (neglected discoveries, new data types).

At the same time, the team's concerns are not only related to whether data are ready for widespread publication as to the ways in which methodological caveats and foibles might be ignored, and the diversity of assumptions with which the data might be interpreted in specific situations of re-use. As COSMIC's director puts it: "*It's what people have been wanting to do all along […] It's also terrifying because everyone's frightened of bringing broadly research informatics and analytics to a clinical space, but as long as that's only part of the decision-making process then that's probably sensible*" (SF). The COSMIC team are keenly aware of the complex considerations involved in re-purposing the data for diverse ends and believe that further local quality checks are necessary, a point they also emphasize in their publications. As they note, manual curation is vulnerable to a set of "ascertainment biases" including publication bias (negative results are most often not published), population bias (research surges on populations of sudden interest) and sequencing bias (known cancer mutation hotspots are more likely to be the subject of further research) – all of which may skew COSMIC data in different ways (Forbes et al., 2006, 2008, 2016). Meanwhile, the automated curation of broad datasets can introduce different biases because of its lower reliability in spotting inconsistencies and mistakes, and other limitations associated with sequencing (Forbes et al., 2015). Users are expected to implement a local process of quality control, and COSMIC includes source and sample references to facilitate this (Forbes et al., 2015; Sondka et al., 2018).

COSMIC curators and users operate in different organisational contexts and there is no systematic communication between them. While some users endeavour to provide feedback and specific requests to the COSMIC team, the vast majority of the user base use COSMIC off the shelf. Most decisions taken by curators are of no interest to these users and are not reached via mechanisms of mutual accountability (such as outlined by Jasanoff, 2005). COSMIC curators envision the decision-making process as a distributed chain of scrutiny – but, given the constraints on interaction and fragmented accountability across the research landscape, worries about kinds of data re-use, risk and opportunity persist. To be able to use and benefit from the data, users need to trust COSMIC's operational principles and procedures without having to repeat the data curation process. For the maintenance of a community standard and state-of-the-art database it is crucial that the COSMIC team is able to make judgments concerning curation targets and data sources, so as to relieve users from the burden of data curation.

4. Data actionability between exploratory research and clinical diagnostics

COSMIC releases data in two ways. First, data can be accessed via web browser, where a number of visualizations, search tools and hyperlinks allow the user to explore various relations between clinical observations. Users can select between available datasets, browse cancer genomes by location coordinates or toponymic categories (genes), and follow links to source publications. Second, users can download files of COSMIC data for integration into local systems and proprietary pipelines. The data can then be used to run automated checks and comparisons against what are typically large quantities of data produced locally from cancer samples. Which data formats users end up using, and which suppositions allow them to trust the data and envision their actionability, depends on the specific problem spaces they work in and the operating assumptions and goals of their endeavours. We explore two kinds of situation: exploratory research and diagnostics development.

4.1. The exploratory research space: understanding pathways

Exploratory researchers rely on COSMIC in their endeavour of understanding the complex causal pathways that shape the evolution of cancer. It is now accepted that genetic drivers are most often expressed through complex gene networks, activation and expression triggers and patterns, protein products and environmental interactions. Working in a space saturated with large quantities of different kinds of data, all of which could be key, researchers' concern is to avoid false positive leads. In exploratory research, actionable data are those that warrant positing a causal link between observable features (i.e., mutated sequences, or measurements of protein products) and cancer behaviour. They justify further investigation of new leads and the development of new disease impact hypotheses. Efforts include, among others, 1) testing the behaviour of small molecules on specific cell lines, with a view to selecting candidates for pharmaceutical development; but also 2) testing strategies and candidate drugs for treating the disease (i.e., 'clinical' research).

Researchers in a pharmaceutical company may, for instance, be interested in the known implications of the mutated form of a certain gene. Some such mutations exhibit quite consistent behaviour. Web-based features of COSMIC provide a "high-level picture" of the gene that can be helpful, and the network of links and data made available for navigation is enormous and for many practical purposes inexhaustible. For instance, Rob McEwen, a pharmaceutical developer at Astra Zeneca, uses the web-based COSMIC browser to select patient populations to design drug trials. Web analytics features include distribution statistics of the kinds of cancer that carry a specific mutation. This helps users to select the patients that are most likely to carry cancers susceptible to targeted therapy. These counts are not population samples, i.e., they are not designed to represent proportions in the population of interest, but rather describe the distribution of mutations in the subset of the

published research data that has been analysed and integrated by COSMIC. Not all data about potentially relevant genes are curated, and the data that are available to COSMIC curators are not sampled by users. The lack of comprehensive statistical distributions notwithstanding, COSMIC counts are used as a best approximation. McEwen explains the key issue in designing basket trials as: *"Each of [the compounds] has got a hypothesis around patient selection so that we know which patients to target depending on which mutations they harbour because then they'll get the most benefit from that particular compound. So, we need to decide on the size of the patient segment in that cancer which is going to get the most benefit. Because there's no use going into a cancer where there's very few patients who've got the mutation that will respond to your drug. So that's part of the work I do, looking at the prevalence of mutations in a particular cancer type to see whether there is an opportunity for a compound or a basket trial containing that compound in that particular cancer type"* (P1).

COSMIC may help inform further pharmaceutical development work when unexpected features of cancer emerge during clinical research (Vignola-Gagné et al., 2017), such as when a cancer develops resistance to a test drug. Insights from the clinical trial can then feed back into the pharmaceutical development pipeline. Researchers who download the COSMIC database can integrate it into proprietary sequencing pipelines and juxtapose it to new cancer sample sequences. McEwen illustrates: *"In some cases then you find that a mutation may have gone really high, so there's been a selection for a particular gene, or a mutation which has allowed the patient, then, to become resistant to the drug you're testing. So then that can go back into the discovery, where you then start to look for a compound that inhibits that particular mutation and that gene that's conferring the resistance to the drug that's in the clinic."* (P1).

COSMIC is also used to help determine whether newly observed mutations 'have already been seen', as another pharmaceutical researcher explains: *"When we go in the clinical trials, we know that some patients are responding. Some are not responding. Then we do some kind of genotype of those patients; see what are the genomic differences between them. Some of the genomic changes, we want to see whether they have already been seen. If they have already been seen they must be in the COSMIC database, then we go back into the COSMIC database to check how much already is known and which are normal and which are not there"* (P2). COSMIC is thus used as a census of research observations, the assumption being that if a mutation has been seen, then it must be in COSMIC. Indeed, the warrant is quickly jeopardised in case of discordance with other sources. The informant explains: *"If there is some kind of disagreement between the databases, we might not go with that kind of gene. We think, 'Okay, we don't trust either of the databases.' So, we try to de-prioritise that."* (P2).

Speculative reasoning about the epistemic value of COSMIC data is not something that actors do in isolation. It is something that researchers imagine to be assessed collectively by the broader community of cancer genomics. The database is envisioned as a space where the observations of the research community are shared and curatorial judgments made by COSMIC staff are regarded as a reliable assessment of the relevance of genes and mutations. In the words of McEwen *"Quite often we have a long table of genes and mutations […] it's useful then to […] remove anything that's not in the cancer gene census* [the COSMIC Cancer Gene Census is a list of mutated genes for which evidence is strongest of their implication in cancer – see Sondzka et al. 2018; Tempini, 2020a]. *So, you're only looking at cancer-relevant genes or ones where there's a bit more of a layer of evidence that they might be cancer-related"* (P1). Delegating evidence assessment to a trusted curatorial process enables cost-efficiency and adherence to community standards at once.

4.2. The diagnostics development space: Locating reliable markers

Diagnostics researchers seek to identify observable biological features or events that can serve as reliable biomarkers. This often involves forgoing a full understanding of the causal pathways through which cancer evolves in favour of the identification of strong predictive signals. These must be robustly correlated with outcomes, and as early and discriminating as possible, to diagnose and treat patients while cancer is easier to kill and has had fewer chances to spread. A biomarker which predicts type, location and further evolution of cancer will facilitate the formulation of a tailored treatment regime. Here, actionable data are those that warrant a predictive association between measured feature and target outcome, where the latter can include cancer type, stage of advancement and/or preferred treatment. Diagnostics researchers we interviewed are working on a cutting-edge technique, liquid biopsy, that is considered highly promising. Small quantities of DNA released from cancer cells start circulating in the blood stream from the earliest stages of the disease. Liquid biopsy developers aim to intercept and characterise this DNA to generate predictions.

In this context, COSMIC data are considered actionable if they enable the identification of a genomic variant as reliable predictor of cancer trajectories. Using COSMIC data as a comparator helps to find clues about mutations of interest: "*Are those previously reported mutations? Are they likely to be deleterious? And this is where COSMIC comes into play. It tells us whether others have found it before"* (D1). This socialisation and intersubjectivity of observations strengthens the warrant that data confer to an interpretation. Users can trust their own observations because they are compatible with the observations of others, and because they can assume that to be trusted in turn. We saw in the previous section how exploratory research users emphasised the same point. These practices echo those reported in the sociology of witnessing (Shapin and Schaffer 1985): being able to use the data is a matter of being able to witness and to trust others' witnessing. A diagnostics informant explains: *"So you've seen something; somebody else has seen something, so that adds weight to the fact. You say, 'Okay, it's more likely to be real than just something random that I've seen, something random in the sequencing.'"*

Yet, diagnostics developers, too, are concerned about both false positives and false negatives (unnoticed sequences with high predictive power). They are particularly concerned that the data may not be comprehensive or statistically representative since they are trying to identify genotype–phenotype correlations. As an informant puts it: *"The data were not representative of the true diversity of mutations in cancer. There was a bias towards mutations being tested a lot."* (D1). These users need to carefully examine COSMIC data, assessing their consistency with respect to their own knowledge of the state of the art and working hypotheses and assumptions. The presence of experts able to do this is key. Possible corrections can only be identified on the basis of personal experience: *"It was useful but it, in a lot respects, wasn't trustworthy, so you still had to do your own due diligence. […] we relied on the knowledge of people who'd spent a lot of time in the field of cancer. They had a good feel for what was real, what was not real. They knew what they knew. I did trust them that they had their finger on the pulse of the literature."* (D2).

Given that a source of statistically representative cancer genomic data does not exist, users have to speculatively weigh the risks and rewards of using and scrutinizing COSMIC data. More complex questions about the very conceptualization of cancer and the causal assumptions underpinning the identification of specific mutations as cancer 'drivers' or 'passengers' are likely less relevant. Improving diagnostic tests' by exploiting the 'low-hanging fruits' (i.e., efficiently interpreting the most common mutations) is seen as enough progress for the time being: *"We're young companies. So, you take the low hanging fruit first. You can see its obvious application, you get your pipeline set up, use it, and then yeah, maybe we'll use* [COSMIC] *for further investigation"* (D1). The extent to which the data will be relied upon is dependent on situated evaluations of opportunity, where the relative accessibility of available evidence, and the risks and advantages of adopting it, are weighed against one another. However, the presence of data in COSMIC that cannot be used does not undermine the high status accorded to the database and the curatorial work undertaken by its team. Crucially for our analysis, users draw a distinction between the trust they have in the fidelity with which COSMIC staff maintain their curatorial processes and standards, and the

degree of reliance that they are willing to invest in some data over another: *"There's plenty of expertise […] The bulk of the data is trustworthy. There are a few odd things I don't trust myself either, but it's uncommon"* (D1).

5. Actionability, trustworthiness and speculation

An inviting contrast is emerging between notions of curation, actionability, trust and speculative reasoning in different kinds of cancer genomics research. As we have seen, in line with an emphasis on putative causal links between molecular makeup and cancer behaviour, exploratory researchers value a 'high level picture' of genes and mutations. They highlight an interest in 'derivative' data (Tempini, 2020c) that consolidate and interpret large amounts of sequence data. They are also interested in 'new' mutations as they try to explain cancer behaviour (such as drug resistance) in the context of experiments with new targeted drugs, and resort to COSMIC as a database of 'what has already been seen' in their search for potentially responsible novelties. In contrast, diagnostics researchers emphasize mutation counts and problematise the relationship between COSMIC counts and real-world distributions. Their interest is in weighing the importance of mutations that have already been identified, and the clinical outcomes of those cases. This concerns not only the frequency of mutations but also the associated harms. As interesting as fleshing out this contrast would be, we leave this task to future research. In the remainder of this paper, we instead focus on explaining why exploratory and diagnostics users are nevertheless drawn to use COSMIC. Both kinds of research user are interested in the dynamics of collective seeing, so we turn our attention to explaining why researchers would use COSMIC as a reference to establish what has and what has not been seen.

Practices of research evidence making, sharing and interpretation in cancer genomics have converged around a small number of standard databases. COSMIC is one of them and acts as a widely accepted point of reference, institutional memory of the research community, repository of witnessing, and off-the-shelf resource of computable data. A standard database helps to reduce local uncertainty in the many judgements involved in ascertaining whether cancer mutation data are relevant. At the same time, the routinization of a standard database carries risk: it delegates some interpretive choices that would otherwise be made by users and can make scrutiny of the data more difficult. Our interviewees often ignored the details and choices inherent in COSMIC's curatorial processes (with respect to curation targets, prioritization of specific research trends and literature sources, and quality standards). They assumed that COSMIC provides a reliable census of evidence. Further, some users have invested in deep integration with COSMIC, developing capabilities and infrastructure in parallel with COSMIC in order to eschew the high running costs and uncertain returns involved in directly curating a database of research evidence. As one informant observed, *"I'm not going to put someone to reanalyse TCGA for six months; that's just not a good investment of time, in my opinion."* (D1). The incentive to rely so deeply on COSMIC is not limited to convenience and cost-efficiency gains in the local setting. The trustworthiness of COSMIC has been broadly sanctioned by the community. Familiarity to researchers, deep integration in technological platforms and pipelines, and community endorsement create a robust warrant for relying on the database. In the words of the diagnostics researcher who noted its longevity, *"a lot of people are collecting data. COSMIC is the biggest, but it's also one of the oldest, as in it's been around, and people understand it"* (D2).

Yet, as previous noted by its director, COSMIC is built on the idea that in order to achieve reliable interpretations of cancer sequences, a degree of local judgement over the fitness for purpose of the data is required (see also Timmermans 2015). More generally, it is understood that a universal standard can thrive in different local contexts as long as it is under-specified so as to enable local adaptation through expert judgement (Berg and Timmermans 2000). However, local judgements about evidence concerning specific mutations do not resolve the uncertainty entailed in the curation and use of COSMIC data. Speculative reasoning is unavoidable, but this is not scientific hypothesizing in the abstract, rather it is embedded in the organisational reality of procedures, routines, priorities, resource constraints, and business strategies among other factors. While the choices of the COSMIC team aim to pre-empt user questions about the interpretation of the data, local adoption of the database makes users ask new questions about the context of COSMIC's interpretations.

There is thus a crucial interplay between speculative reasoning and reliance on COSMIC data, which shapes perceptions of data actionability and underpins the practice of cancer genomics. Whether users are able to regard COSMIC data as actionable depends on 1) the choice to trust the curatorial process that determines which data may be most valuable, which in turn requires speculative reasoning and risk-taking (by both curators and users) and 2) the judgment that specific COSMIC data are fit for purpose in each situation of genomic data interpretation.

There is a difference between the two conditions. Hawley (2017) argues that individuals can be trusted, while organisations and objects can only be relied upon. This is because trust implies the possibility of betrayal or deception. For a COSMIC user it is thus reasonable to ask whether the curators can be trusted. As there is no reason to suspect the curators of deception, users trust that the curators are doing, give or take, as fine a job as they themselves would. The procedural principles guiding curation represent best practice and, most importantly, are promulgated on dedicated pages on the COSMIC website and in the team's publications. The professional experience of curators supports the belief that their curation will be accurate and reliable. These are some of the reasons why a user might believe that COSMIC curators are trustworthy. Beyond that, the expectation is that any curatorial oddities will be errors, but not deceptions.

Instead, COSMIC data, as objects, can only be relied upon, but not trusted (Hawley 2017). Data can be relied upon as epistemically significant if the organisation generating them has made a commitment to purposefully designed standardised procedures (Hawley 2017), although reliance will depend also on assumptions about their suitability for the problem at hand. Indeed, the commitment of the COSMIC team to explicit curatorial principles and goals warrants other researchers and organisations' reliance on the data. Users we interviewed cited episodes when the data were not relied upon because they did not appear accurate or sound. And yet, the fact that data may on occasion be faulty did not invalidate the main reasons for adopting COSMIC as a routine step in the sequence interpretation process. The separation between, on the one hand, placing trust in the curator who is handling the data and, on the other hand, reliance on curated data allows one to conceive of unreliable data and a trustworthy curator at once. We elaborate on these conditions of actionability in what follows, starting from the practices of speculative reasoning and trust.

5.1. Speculative reasoning in data interpretation

COSMIC curators process only a fraction of the data generated by the field. They make decisions regarding curation targets (e.g., genes most worthy of curation) and emerging opportunities (such as important research trends as well as new methods and kinds of data that curation could cover). They compile authoritative evidence summaries. Their curatorial decisions, which are by necessity speculative in their assumptions about what is relevant, cannot fit everybody's needs. Yet these decisions are key to the trustworthiness of the team and the COSMIC project as a whole. Users delegate to COSMIC important decisions about data processing, and value its focused specialization and steady commitment to curation. COSMIC is valued not only for what it includes but also for what it excludes. In every user organisation, reasoning with COSMIC data involves evaluating how much of the data to rely upon and how much to deprecate through local validation (e.g., second checks with other resources, manual review). Using "imperfect" databases of genetic observations (Timmermans 2015) makes a degree

of risk unavoidable. Local expertise is required to lead the re-contextualization of data – Timmermans (2015) calls this reflexive standardization. Evaluating the actionability of data is, accordingly, a speculative judgement, but neither mere risk-taking nor discursive reasoning. Users concerned with the data's lack of statistical representativeness speculate as they estimate the 'true' statistical ratios. Other users take COSMIC as a 'census' of observations (which it is not) to see if something 'has already been seen' and how often. The fact that users may ignore the details of COSMIC's curation procedures highlights how constrained their deliberation is. Some users seem not to give much thought to the specific commitments to standardised procedures that, according to Hawley, warrants the epistemic significance of COSMIC's assessments. A standard database that initially emerged in the community thanks to its leadership in setting curatorial standards has thus become an obligatory passage point that is relied upon without detailed scrutiny of its assumptions and methodological choices.

5.2. Trust in the database and reliance on data

That the COSMIC database project is considered trustworthy even if the data are not always considered reliable highlights how speculative reasoning shapes attitudes of trust and evaluations of actionability. It is in this purview that the topic of speculative reasoning should be framed: not only as characterizing the epistemic practices of individual research organisations, but as the basic underpinning for a whole field that performs data interpretation under conditions of high uncertainty. Similarly to what Porter notoriously showed for numerical evidence (1995), and Timmermans argues in respect to genomic data (2015), a database such as COSMIC is used by the cancer research community as common ground upon which evidence, interpretations and discoveries can be assessed. If every research team compiled its own evidence base of cancer mutations, interpretations would be more difficult to scrutinize. As Porter observes (Porter 1995; Timmermans 2015), trusted standards emerge when the objectivity of the evidence shared within a research community is systematically questioned. The inability to rely on the data is therefore a pre-condition, not a limitation, for adopting specific resources, methods and procedures as trusted standards for the production of objective evidence. In Porter's seminal study, the favourable evaluation of quantitative methodologies (typically consisting of repeatable processes through which numbers are produced) was a key step in the rise of numbers and quantitative evidence as a gold standard in several scientific and administrative fields. Trust was not invested in numbers per se but in the processes that reliably produced them. As Hawley similarly observes from a philosophical perspective (2017), organisational processes can be considered reliable and epistemically significant if an organisation is committed to operational principles designed to ensure reliability and earn trustworthiness. With respect to COSMIC, procedural standardization emerges as the key factor behind its success in attracting user trust, despite the sporadic lack of reliance on the data. Collecting and distributing publicly available data by following well-documented procedures and public commitments is a key step in the construction of objective sources of evidence. COSMIC's circulation of standardised data across the community enables the use of data as a common point of reference in the generation, comparison and evaluation of cancer sequence interpretations. Procedural standardization makes it possible for users to trust the database even when individual datasets are not seen as trustworthy in and of themselves.

5.3. Speculation and trust

Hawley (2019) defines trust as a combination of the trustor's practical reliance on the trustee and the trustee's commitment to fulfilling a specific task. Commitments can come into being, she notes, because: the trustee makes them; the trustor thinks that the trustee has made them; or the trustee allows the trustor to continue to rely upon her (2019:21). Because of the high level of uncertainty about the role of genetic sequences in the expression of phenotypical outcomes (Timmermans et al., 2017), users accept that error is possible in COSMIC's curation processes. Uncertainty puts limits on expectations about what constitutes fulfilment of COSMIC's commitment to high quality curation. Procedural standardization gives weight to this commitment, securing trust in COSMIC. Trust in COSMIC is not based on the intrinsic qualities of data but rather on the alignment between the organisation and the user's material and epistemic commitments. As Timmermans (2015) argues, despite the concerns that users have about the reliability of specific data, they remain dependent on sequence databases, and have limited ability to amend data, especially when their use is routinized. The process constrains the scrutiny that users can exercise. Operating procedures such as those performed by automated filters, algorithms and database schemas further embed standardised decisions into cancer genomics practice.

6. Conclusion

Our case study highlights a deep yet indirect interdependence between curation and data use in cancer genomics. Cancer genomics research depends on the collective actions of curators and users – actions that are imperfect, constrained by conditions of high uncertainty about the wider ecology of practices of the field and yet reflexive (albeit indirectly) about one another. Paying attention to organisational procedures, commitments and standards at both the curation and user sites, we emphasised the level of delegation that adoption of a database standard involves as an element that needs more consideration in the social studies of data curation. COSMIC is entrusted by researchers to provide an epistemic resource that can be relied upon to adhere to widely accepted standard procedures. In so doing, it provides a whole research domain with a centralised function of evidence aggregation, operationalising the assumption that empirical studies generate consistent and comparable signals.

When evaluating whether or not to rely on data, users need to consider the speculative judgments of COSMIC curators, together with methodological foibles associated with the data and the possibility of curatorial mistakes. Nevertheless, trusting COSMIC holds advantages for user organisations which go well beyond immediate convenience. A standard database such as COSMIC creates the opportunity for the cancer genomics field to share a collection of facts about cancer sequences (Cambrosio et al. 2020; Tempini 2020a). The existence of such a space, a database that aggregates findings in ways that make them concordant and comparable, makes it possible to assemble a coherent audience from a fragmented, distributed network of researchers communicating asynchronously and inconsistently through publications shared over the Internet. Echoing Anderson's seminal analysis of the birth of national identities in imagined communities Anderson (2006), the database here contributes to the imagination of a community, which converges on specific means of communication and is involved in practices of collective witnessing – of 'seeing' and agreeing on what was seen. In contrast to 18th Century audiences of empirical witnesses, cancer genomics observations are convened by standardised evidence curation, aggregation and representation procedures. Although COSMIC data may occasionally contribute to a mistaken interpretation (be it due to uncertainty, curation error, or user error), this does not immediately undermine users' trust in COSMIC. Instead, the *routine consultation* of a record of 'what others have already seen', compiled and dynamically maintained through trusted standardised procedures, becomes a key step in the transformation of user organisations' local practices of sequence interpretation into standardised procedures. This is a strategic move by which user organisations assert their own trustworthiness (O'Neill, 2002). In the same vein as COSMIC retains trust despite caveats and imperfections, user organisations aim to demonstrate their own commitment to high standards of evidence to their own users by using COSMIC as a key component of their own sequence interpretation routines.

The actionability of cancer genomics data thus rests on a nested architecture of dynamic, yet standardised, procedures of data management (Cambrosio et al., 2020; Tempini and Leonelli 2018a), in which multiple organisations committed to demonstrating trustworthiness intervene in a sequence of operations of data management and interpretation. Beyond the trust in standard procedures, little else is firm in this chain of custody (Wylie 2020; Tempini and Leonelli 2018b). These trustworthiness-seeking data management practices sit squarely at the heart of cancer genomics, making it possible to bear an increasing amount and diversity of data upon each other, while the chance for individual researchers to directly scrutinize sources is decreased.

Credit statement

Niccolò Tempini: Conceptualization, Methodology, Investigation, Resources, Data Curation, Writing – Original Draft, Writing – Review & Editing. Sabina Leonelli: Conceptualisation, Methodology, Resources, Writing - Review & Editing, Supervision, Project administration, Funding acquisition.

Acknowledgments

We are thankful to our informants for the time and effort they donated to us to make this research possible, to the editor and the three anonymous reviewers for their valuable comments and support. We are also thankful to colleagues who shared comments on early drafts or presentations of this paper: in the Exeter Centre for the Study of the Life Sciences, and including among others Brian Rappert, Adrian Currie, Sam Wilkinson and Benjamin Smart; in Edinburgh's Department of Science, Technology and Innovation Studies, and including among others Miguel Garcia-Sancho, James Lowe, Robin Williams, Lukas Engelmann, Jane Calvert and Steve Sturdy; the organisers and participants to the November 2018 workshop "Organisational and epistemic innovation in precision cancer medicine" at the Centre for the Sociology of Organisations in Paris, and including among others Alberto Cambrosio, Pascale Bourret, Henri Bergeron, Patrick Castel, Anne Kerr, Sarah Cunningham-Burley, Stefano Crabu and Carsten Timmermans; the organisers and participants to the track on Precision Medicine at EASST 2018 in Lancaster, including Nadav Even Chorev, Ilaria Galasso and Christopher Goldsworthy; the organisers and participants to the April 2018 workshop "La Cura del Cancro e l'Innovazione Tecnologica: Aspetti Etico-Sociali" at the Accademia dei Concordi in Rovigo, including Giovanni Boniolo and Federico Neresini. We are thankful to Adam Bostanci for further feedback on a late draft. This research was funded by ERC grant award 335925 (DATA_SCIENCE), and by the Alan Turing Institute under the EPSRC grant EP/N510129/1.

References

Anderson, B., 2006. Imagined Communities: Reflections on the Origin and Spread of Nationalism. Verso, London.
Bamford, S., Dawson, E., Forbes, S., Clements, J., Pettett, R., Dogan, A., Flanagan, A., Teague, J., Futreal, P.A., Stratton, M.R., Wooster, R., 2004. The COSMIC (Catalogue of somatic mutations in cancer) database and website. Br. J. Canc. 91 (2), 355–358.
Bechtel, W., 2019. From parts to mechanisms: research heuristics for addressing heterogeneity in cancer genetics. History and Philosophy of the Life Sciences.
Berg, M., Timmermans, S., 2000. Orders and their others: on the constitution of universalities in medical work. Configurations 8, 31–61.
Bergeron, H., Castel, P., 2011. Continuity, capture, network: the professional logics of the organization of care. Sociol. Travail 53, e1–e18.
Bergeron, H., Castel, P., Vézian, A., 2020. Beyond full jurisdiction: pathology and inter-professional relations in precision medicine. New Genetics and Society 0, 1–16. https://doi.org/10.1080/14636778.2020.1861543.
Bertolaso, M., 2016. Philosophy of Cancer: A Dynamic and Relational View. Springer.
Bourret, P., Cambrosio, A., 2019. Genomic expertise in action: molecular tumour boards and decision-making in precision oncology. Sociol. Health Illness 41 (8), 1568–1584.
Cambrosio, A., Keating, P., Mogoutov, A., 2013. What's in a pill? On the informational enrichment of anti-cancer drugs. In: Gaudillière, J.-P., Hess, V. (Eds.), Ways of Regulating Drugs in the 19th and 20th Centuries, Science, Technology and Medicine in Modern History. Palgrave Macmillan UK, London, pp. 181–205.
Cambrosio, A., Campbell, J., Vignola-Gagné, E., Keating, P., Jordan, B.R., Bourret, P., 2020. 'Overcoming the bottleneck': knowledge architectures for genomic data interpretation in oncology. In: Leonelli, S., Tempini, N. (Eds.), Data Journeys in the Sciences. Springer International Publishing, Cham, pp. 305–327.
Forbes, S., Clements, J., Dawson, E., Bamford, S., Webb, T., Dogan, A., Flanagan, A., Teague, J., Wooster, R., Futreal, P.A., Stratton, M.R., 2006. COSMIC 2005. Br. J. Canc. 94, 318–322.
Forbes, S.A., Bhamra, G., Bamford, S., Dawson, E., Kok, C., Clements, J., Menzies, A., Teague, J.W., Futreal, P.A., Stratton, M.R., 2008. The Catalogue of Somatic Mutations in Cancer (COSMIC). Curr Protoc Hum Genet CHAPTER, Unit-10.11.
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., Futreal, P.A., 2011. COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. Nucleic Acids Res. 39, D945–D950.
Forbes, S.A., Beare, D., Gunasekaran, P., Leung, K., Bindal, N., Boutselakis, H., Ding, M., Bamford, S., Cole, C., Ward, S., Kok, C.Y., Jia, M., De, T., Teague, J.W., Stratton, M. R., McDermott, U., Campbell, P.J., 2015. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. Nucleic Acids Res. 43, D805–D811.
Forbes, S.A., Beare, D., Bindal, N., Bamford, S., Ward, S., Cole, C.G., Jia, M., Kok, C., Boutselakis, H., De, T., Sondka, Z., Ponting, L., Stefancsik, R., Harsha, B., Tate, J., Dawson, E., Thompson, S., Jubb, H., Campbell, P.J., 2016. COSMIC: high-resolution cancer genetics using the Catalogue of somatic mutations in cancer. Current Protocols in Human Genetics 91, 10.11.1-10.11.37.
Forbes, S.A., Beare, D., Boutselakis, H., Bamford, S., Bindal, N., Tate, J., Cole, C.G., Ward, S., Dawson, E., Ponting, L., Stefancsik, R., Harsha, B., Kok, C.Y., Jia, M., Jubb, H., Sondka, Z., Thompson, S., De, T., Campbell, P.J., 2017. COSMIC: somatic cancer genetics at high-resolution. Nucleic Acids Res. 45, D777–D783.
Fortun, M., 2008. Promising Genomics. University of California Press, London.
Futreal, P.A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N., Stratton, M.R., 2004. A census of human cancer genes. Nat. Rev. Canc. 4, 177.
Gao, J., Aksoy, B.A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S.O., Sun, Y., Jacobsen, A., Sinha, R., Larsson, E., Cerami, E., Sander, C., Schultz, N., 2013. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci. Signal. 6, pl1.
Hawley, K., 2017. Trustworthy groups and organizations. In: Faulkner, P., Simpson, T. (Eds.), The Philosophy of Trust. Oxford University Press, pp. 230–250.
Hawley, K., 2019. How to Be Trustworthy. Oxford University Press.
Hilgartner, S., 2017. Reordering Life: Knowledge and Control in the Genomics Revolution. MIT Press.
Hogle, L.F., 2016. Data-intensive resourcing in healthcare. BioSocieties 11, 372–393.
Huang, et al., 2016. The path from big data to precision medicine. Expert Review of Precision Medicine and Drug Development 1 (2), 129–143.
Jasanoff, S., 2005. Designs on Nature: Science and Democracy in Europe and the United States. Princeton University Press.
Jones, S., Anagnostou, V., Lytle, K., Parpart-Li, S., Nesselbush, M., Riley, D.R., Shukla, M., Chesnick, B., Kadan, M., Papp, E., Galens, K.G., Murphy, D., Zhang, T., Kann, L., Sausen, M., Angiuoli, S.V., Diaz, L.A., Velculescu, V.E., 2015. Personalized genomic analyses for cancer mutation discovery and interpretation. Sci. Transl. Med. 7, 283ra53-283ra53.
Keating, P., Cambrosio, A., 2011. Cancer on Trial: Oncology as a New Style of Practice. University of Chicago Press, Chicago, IL.
Keating, P., Cambrosio, A., 2013. 21st-century oncology: a tangled web. Lancet 382, e45–e46.
Kerr, A., Swallow, J., Chekar, C.K., Cunningham-Burley, S., 2019. Genomic research and the cancer clinic: uncertainty and expectations in professional accounts. New Genet. Soc. 38, 222–239.
Leonelli, S., 2016. Data-centric biology: a philosophical study. University of Chicago Press, Chicago.
Levin, N., 2018. Big data and biomedicine. In: Meloni, M., Cromby, J., Fitzgerald, D., Lloyd, S. (Eds.), The Palgrave Handbook of Biology and Society. Palgrave Macmillan UK, London, pp. 663–681.
Lowe, J.W.E., 2018. Sequencing through thick and thin: historiographical and philosophical implications. Stud. Hist. Philos. Sci. C 72, 10–27.
Lynch, M., Cole, S.A., McNally, R., Jordan, K., 2010. Truth Machine: the Contentious History of DNA Fingerprinting. University of Chicago Press.
Nelson, N.C., Keating, P., Cambrosio, A., 2013. On being "actionable": clinical sequencing and the emerging contours of a regime of genomic medicine in oncology. New Genet. Soc. 32, 405–428.
O'Neill, O., 2002. A Question of Trust: the BBC Reith Lectures 2002. Cambridge University Press.
Plutinsky, A., 2018. Explaining Cancer. Oxford University Press, Oxford.
Porter, T.M., 1995. Trust in Numbers: the Pursuit of Objectivity in Science and Public Life. Princeton University Press, Princeton.
Prainsack, B., 2020. The meaning and enactment of openness in Personalised and Precision Medicine. Sci. Publ. Pol., scaa013
Rajan, K.S., 2006. Biocapital. Duke University Press, London.
Rheinberger, H.-J., 2010. An Epistemology of the Concrete: Twentieth-Century Histories of Life. Duke University Press.
Shapin, S., Schaffer, S., 1985. Leviathan and the Air-Pump: Hobbes, Boyle, and the Experimental Life. Princeton University Press.
Sondka, Zbyslaw, Bamford, Sally, Cole, Charlotte G., Ward, Sari A., Dunham, Ian, Forbes, Simon A., 2018. The COSMIC cancer gene census: describing genetic dysfunction across all human cancers. Nat. Rev. Canc. 18 (11), 696–705.
Tate, J.G., Bamford, S., Jubb, H.C., Sondka, Z., Beare, D.M., Bindal, N., Boutselakis, H., Cole, C.G., Creatore, C., Dawson, E., Fish, P., Harsha, B., Hathaway, C., Jupe, S.C., Kok, C.Y., Noble, K., Ponting, L., Ramshaw, C.C., Rye, C.E., Speedy, H.E.,

Stefancsik, R., Thompson, S.L., Wang, S., Ward, S., Campbell, P.J., Forbes, S.A., 2019. COSMIC: the Catalogue of somatic mutations in cancer. Nucleic Acids Res. 47, D941–D947.
Tempini, N., 2017. Till data do us part: Understanding data-based value creation in data-intensive infrastructures. Inf. Organ. 27 (4), 191–210.
Tempini, N., 2020a. The Reuse of Digital Computer Data: Transformation, Recombination and Generation of Data Mixes in Big Data Science. In: Leonelli, S., Tempini, N. (Eds.), Data Journeys in the Sciences. Springer International Publishing, Cham, pp. 239–263.
Tempini, Niccolò, 2020b. Interviews: The Catalogue of Somatic Mutations in Cancer (COSMIC) Zenodo. http://doi.org/10.5281/zenodo.3944466.
Timmermans, S., 2020c. Data curation-research: practices of data standardization and exploration in a precision medicine database. New Genet. Soc. 1–22. https://doi.org/10.1080/14636778.2020.1853513 (online first).
Tempini, N., Leonelli, S., 2018a. Genomics and Big Data in Biomedicine. In: Gibbon, S., Prainsack, B., Hilgartner, S., Lamoreaux, J. (Eds.), Routledge Handbook of Genomics. Health & Society, Routledge, pp. 24–31.
Tempini, N., Leonelli, S., 2018b. Concealment and discovery: The role of information security in biomedical data re-use. Soc. Stud. Sci. 48 (5), 663–690.
Timmermans, S., 2015. Trust in standards: transitioning clinical exome sequencing from bench to bedside. Soc. Stud. Sci. 45 (1), 77–99.
Timmermans, S., Tietbohl, C., Skaperdas, E., 2017. Narrating uncertainty: variants of uncertain significance (VUS) in clinical exome sequencing. BioSocieties 12, 439–458.
Vignola-Gagné, E., Keating, P., Cambrosio, A., 2017. Informing materials: drugs as tools for exploring cancer mechanisms and pathways. HPLS 39, 10.
Wylie, A., 2020. Radiocarbon dating in archaeology: triangulation and traceability. In: Leonelli, S., Tempini, N. (Eds.), Data Journeys in the Sciences. Springer International Publishing, Cham, pp. 285–301.