‘Overcoming the Bottleneck’: Knowledge Architectures for Genomic Data Interpretation in Oncology

Alberto Cambrosio, Jonah Campbell, Etienne Vignola-Gagné, Peter Keating, Bertrand R Jordan, Pascale Bourret

‘Overcoming the Bottleneck’: Knowledge Architectures for Genomic Data Interpretation in Oncology

Alberto Cambrosio, Jonah Campbell, Etienne Vignola-Gagné, Peter Keating, Bertrand R. Jordan, and Pascale Bourret

Abstract In recent years, oncology transitioned from its traditional, organ-based approach to ‘precision oncology’ centered on molecular alterations. As a result, it has become to a significant extent a ‘data-centric’ domain. Its practices increasingly rely on a sophisticated techno-scientific infrastructure that generates massive amounts of data in need of consistent, appropriate interpretations. Attempts to overcome the interpretation bottleneck have led to the establishment of a complex landscape of interrelated resources that, while displaying distinct characteristics and design choices, also entertain horizontal and vertical relations. Although there is no denying that the data-centric nature of contemporary oncology raises a number of key issues related to the production and circulation of data, we suggest that the focus on data use and re-use should be complemented by a focus on interpretation. Oncology practitioners refer to data interpretation resources as ‘knowledgebases’, an actor’s category designed to differentiate them from generic, multi-purpose databases. Their major purpose is the definition and identification of *clinically actionable* alterations. A heavy investment in human curation, of a clinical rather than exclusively scientific nature is needed to make them valuable, but each knowledgebase

A. Cambrosio (✉) · J. Campbell
Department of Social Studies of Medicine, McGill University, Montreal, QC, Canada
e-mail: alberto.cambrosio@mcgill.ca; jonah.campbell@mail.mcgill.ca

E. Vignola-Gagné
Department of Social Studies of Medicine, McGill University, Montreal, QC, Canada

Science-Metrix, Montreal, QC, Canada
e-mail: e.vignola-gagne@science-metrix.com

P. Keating
Department of History, University of Quebec at Montreal, Montreal, QC, Canada
e-mail: keating.peter@uqam.ca

B. R. Jordan
ADES, Aix-Marseille Université-EFS-CNRS, Marseille, France
e-mail: bertrand.jordan@univ-amu.fr

P. Bourret
Aix Marseille Univ, INSERM, IRD, SESSTIM, Marseille, France
e-mail: pascale.bourret@univ-amu.fr

S. Leonelli, N. Tempini (eds.), *Data Journeys in the Sciences*,
https://doi.org/10.1007/978-3-030-37177-7_16

appears to have its own peculiar way of connecting clinical and scientific statements. In spite of their common goal, knowledgebases thus adopt very different approaches partly captured by the tension between trust and traceability.

1 Introduction

In March 2018, responding to a request by the US Congress, the National Institutes of Health released a draft version of its "Strategic Plan for Data Science".[1] In its drive to modernize the "Data Repository Ecosystem", the Plan introduced a distinction between *databases* and *knowledgebases*. It defined the first as "data repositories that store, organize, validate, and make accessible the core data related to a particular system or systems", and the second as warehouses that "accumulate, organize, and link growing bodies of information related to core datasets". While admitting to "a grey area … between databases and knowledgebases" and acknowledging that some knowledgebase data "may eventually harden and become core data more appropriate for a database", the document stipulated the NIH's intention to "support each separately". In other words, this was not mere semantics: it entailed organizational and financial consequences.

While most readers are doubtlessly unaware of the database/knowledgebase distinction, it came as no surprise to us. During fieldwork for this paper, numerous respondents invoked it to characterize the computerized resources they had developed to facilitate genomic data interpretation in oncology. Given oncology's pioneering role in the development of 'precision medicine', recourse to the neologism 'knowledgebase' deserves additional investigation. What does it entail and how does it relate to the molecular reconfiguration of oncology practices? More specifically, how and to what extent does the replacement of 'data' with 'knowledge' in the portmanteau word reflect actual differences in the origin, kind, and content of the information in knowledgebases? Does the 'data journey' metaphor (Leonelli this volume; Leonelli 2016; Bates et al. 2016), often used to characterize the dynamics of data repositories, continue to appropriately describe how information elicited from journal articles or databases is incorporated and organized within knowledgebases? To begin to answer these questions we need to examine how knowledgebases are located within the sequence of activities that define genomics-driven oncology, from the initial sequencing of a patient's tumor to treatment decisions. Knowledgebases are specifically geared for *data interpretation* and as such impinge directly on discussions about the actionability and clinical utility of genomic results, i.e. the establishment of predictive relations between molecular profiling results and specific drugs (Nelson et al. 2013). Oncologists perceive them as potential solutions to a major 'bottleneck' that threatens the viability of their endeavor.

[1] https://grants.nih.gov/grants/rfi/NIH-Strategic-Plan-for-Data-Science.pdf

2 The Data Interpretation Bottleneck

In his 2011 address to the American Society of Clinical Oncology, ASCO's president discussed the challenges occasioned by the rapidly decreasing price of genomic sequencing and the ensuing 'tsunami' of genomic data:

> When data are that cheap, every patient's cancer will be informative for tumor biology. And things will get very, very complicated. (George Sledge, cited in Goldberg 2011, 4).

The issue was more than quantitative. Traditionally, tumors have been characterized by organ and/or tissue of origin and stage of development. Following the introduction of genomic platforms that identify a wide range of molecular alterations (mutations, amplifications, etc.), clinical practitioners entertained the possibility of generating an alternative categorization of tumors based on shared alterations, thus "creating a new molecular taxonomy of cancer" (Titus 2014a). Early, simplistic attempts to implement genomic medicine using a 'one cancer gene, one drug' approach, have been replaced by a more detailed understanding of the molecular bases of therapies. Cancer-related genes harbor thousands of variants that require an unprecedented level of granularity in assessing their effects. The problem has thus less to do with the actual *production* of molecular data – the required logistics, their reliability and comparability across instruments – than with their *interpretation* and consequent translation into clinical practices (Jordan 2015). As one oncologist argued, "the fundamental problem is we're generating more information than we can readily interpret as individuals" (Titus 2014b).

While precision medicine has its critics (e.g., Prasad 2016; see Subbiah and Kurzrock 2017 for a rebuttal), all major cancer centers and agencies have jumped on the genomic bandwagon. Publications commonly report on the experience of implementing 'omics' approaches (Schwaederle et al. 2015; Subbiah and Kurzrock 2016; Meric-Bernstam et al. 2013; Johnson et al. 2015). Both descriptive and performative, these publications report on the 'knowledge architecture' (Amin and Cohendet 2004) instituted by leading cancer organizations to operationalize cancer genomics. They simultaneously qualify precision oncology as an endeavor that has escaped the status of mere promissory note. All major cancers have been fragmented into a growing number of rare diseases defined not only by specific genomic variants, but also by their differential reaction to a new generation of 'targeted' and immunotherapy treatments (Vignola-Gagné et al. 2017).

The new approach associates clinical oncologists and pathologists with molecular biologists and bioinformatics specialists, modifying the equilibrium between the traditional components of oncology practices. Following the sequencing of tumor samples and the identification of tumor-specific events, these events must be annotated to establish their functional significance. Potential tumor-driving events must be interpreted, prioritized, and summarized "in the context of published literature, clinical trials, and a multitude of knowledge bases" (Good et al. 2014). Clinicians then evaluate these findings by relating them to clinical data generated from the case history of a particular patient (Van Allen et al. 2013). The increasing use of large-scale approaches, such as whole-exome or whole-genome sequencing (as contrasted with limited gene panels), has made the situation even more fraught. As Ghazani et al. (2017, 787) noted:

> [A]ssigning clinical meaning to each somatic and germ-line variant, including the therapeutic, prognostic, and diagnostic implications for individual patients, poses considerable difficulties in light of the inconsistent state of genome biological annotation.

This issue has recently become known, in the actors' own words, as the 'interpretation bottleneck'.

3 Knowledgebases and Databases

Instancing "the production of dozens to thousands of potential tumor-driving events that must be interpreted by a skilled analyst and synthesized in a report", Good et al. (2014) explained that:

> Each event must be researched in the context of current literature, drug-gene interaction databases, relevant clinical trials and known clinical actionability from knowledgebases. In our opinion, this attempt to infer clinical actionability represents the most severe bottleneck of the process.

The Good et al. (2014) paper predates the NIH distinction between databases and knowledgebases by 4 years, which suggests that the distinction has been in use for some time. While the term 'database' needs no further elaboration, having entered common parlance several decades ago, the notion of knowledgebase requires explanation. Although both 'bases' act as repositories for 'data'[2] derived from published papers, conference abstracts, datasets established by large-scale collaborations, and results of tumor profiling analyses of patients enrolled in clinical trials or undergoing routine treatment, it is not clear that we are talking about the 'same' kind of information. It is similarly unsure that both bases treat data in the same way. Are we, in other words, confronted with similar data journeys, and does this metaphor actually apply to knowledgebases?

Both kinds of repositories use equivalent software tools and packages, arguably making one a mere subtype of the other. But as scores of technology studies have shown (e.g. Bijker and Law 1992), it would be simplistic to reduce devices to their technical components. Moreover, the very fact that actors differentiate between them suggests important differences. While acknowledging that many genomic resources incorporate elements from both databases and knowledgebases, Pitel (2017) reiterates the usefulness of the distinction:

> Although data and knowledge are dependent on each other, it is important to understand that data portals contain observations, like those typically seen in the results section of an article. … Knowledgebases, on the other hand, contain critically processed data, contextualized for significance and meaning, much like what you might find in a conclusion section of an article, and are often more appropriate for immediate use in clinical laboratory practice.

At this point, we could be accused of uncritically adopting the actors' categories. Social scientists often contrast native terminology with scholarly notions that enjoy epistemic privilege. A different take on this issue has been proposed by ethnometh-

[2] Adopting an ethnographic stance, we consider as data anything that actors treat as such.

odologists through the notion of 'perspicuous phenomena', i.e. "'things' (and activity settings) that re-tune our sensibilities, so that when we return to the familiar distinctions, concepts, and debates of a social science, we can read them differently" (Lynch 2009, 114). We accordingly eschew the alternative, sometimes referred to as the topic/resource distinction (Lynch 1998, 867n88), that consists in either contenting ourselves with a description of the actors' language or in developing an analytical meta-language divorced from the actors" own meanings and practices. Instead, we seek intersections between the questions that actors ask and the questions we raise, between the practical answers they provide and the theoretical framing we offer. We focus on "topics or themes which preoccupy particular groups, and which resonate with social sciences issues" (M. Lynch, personal communication). As an actor-derived categorization, the database/knowledgebase distinction can be used both by analysts as a language of description, and by concerned groups as a language for action (Lynch 1993; Hatchuel 1996).

Figure 1 depicts the funnel running from the initial sequencing to the bottleneck of interpretation by/for the clinician. It appears that much of what precedes the bottleneck (stage 5) can be categorized as the domain of databases, whereas the bottleneck and its knowledgebases interrupt the data journey. Knowledgebases come into play when oncologists receive a sequencing report listing mutations of possible clinical import. Instead of manually scouring the entire published literature for information about those mutations, they turn to one of several interpretation knowledgebases that offer a synthetic summary and description of a given variant's clinical significance. The 'product' of a knowledgebase is the interpretation itself, an assertion about the clinical actionability of a particular variant. Although it can be traced back to a specific reference (PubMed or otherwise), a given interpretation is likely to differ from those embedded in other knowledgebases for the 'same' variant. What differs is the statement or interpretation itself, the 'level of evidence' associated with a given statement, the suggested therapy or clinical action, and the references supporting the interpretation. In this context, 'the data' no longer enter, leave, or occupy space in the 'base' as immutable entities. The core content of the knowledgebase – the interpretation – arises from the knowledgebase itself wherein the data are recombined and transmogrified into interpretative statements with multiple lineages.

Practitioners contrast databases with knowledgebases in two different (albeit complementary) ways. The first claims that knowledgebases contain *interpretation-laden* and *action-oriented* data, as contrasted with *raw* data.[3] The introduction of the database/knowledgebase distinction thus reifies the content of databases as theory-neutral data unaffected by interpretation. The distinction also elevates the status of the interpretations embedded in knowledgebases as (temporarily) reliable knowledge. The second demarcation refers to the practices and goals that establish those two infrastructures, which we can for now summarize as follows: whereas databases aim at the production of resources that will be available for use by different communities of practice, oncology's knowledgebases are typically the result of initia-

[3] Arguably an oxymoron (Gitelman 2013), the term 'raw data' is easily found in scientific publications and laboratory discussions, where it makes pragmatic sense (Cambrosio and Keating 2000, 263–265).

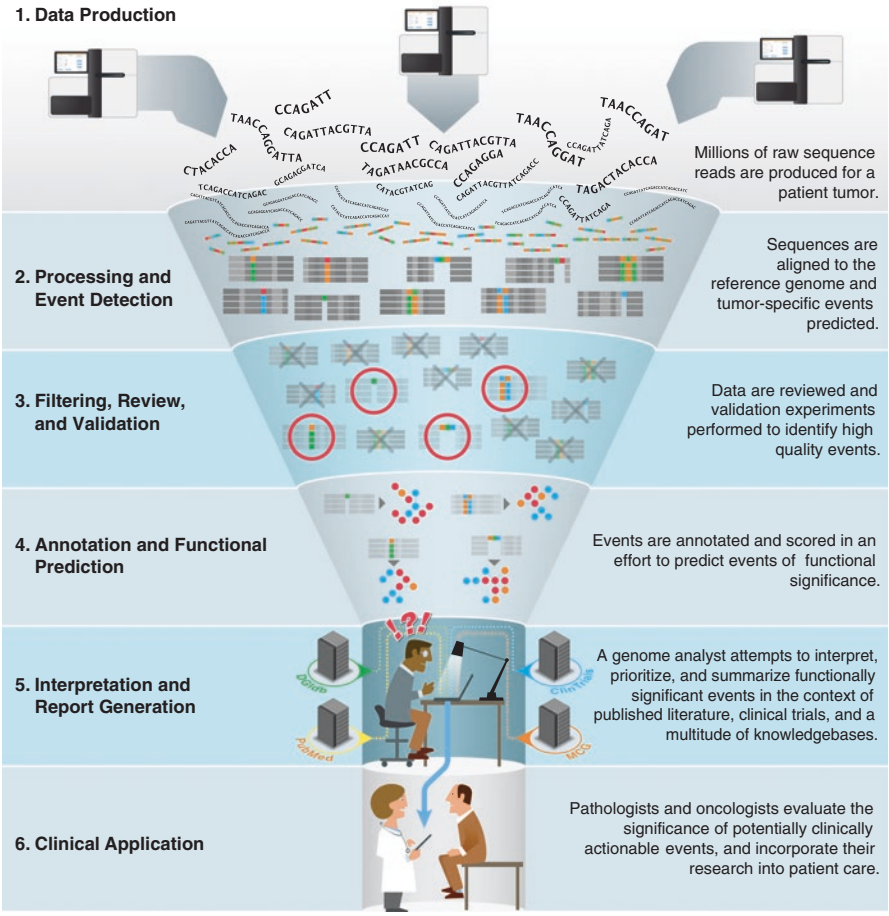

Fig. 1 "The interpretation bottleneck of personalized medicine" (Source: Good et al. 2014; Creative Commons Public Domain image)

tives derived from practical clinical concerns. As compared to much larger databases, knowledgebases address specific audiences. They are characterized by a high degree of 'situatedness' (Suchman 1987), i.e. they act as resources for clinical decision-making that are grounded in a collective understanding of possible therapeutic pathways once the local contingencies of clinical work are considered. For instance, the fact that several knowledgebases consist of an outward-facing website that only reports information with literature support, and an internal component that can exclusively be accessed by members of that institution, is justified as follows:

> In the absence of a community that understands the nuances of the potentially actionable, it's a little harder to relay that [kind of genomic] information. The treating physicians at [our institution] get a report that says: "We think this is potentially actionable because of the following reasons", and they can understand how grey that call is. That is a little more personal personalized therapy, therefore harder to do en masse, so that is indeed not reported currently on our outward-facing website. (Interview with Dr. Funda Meric-Bernstam, July 2017; henceforth FMB).

4 A Spectrum of Data Repositories

To further explore the distinction between different kinds of data repositories, consider Leonelli's (2013) analysis of an oncology database that explicitly refrained from selecting and interpreting data, namely the caBIG database, a bioinformatics initiative sponsored by the US National Cancer Institute (NCI). A key component of the 'cyberinfrastructure' destined to "empower a 'third way' in biomedical research" (Buetow 2005), caBIG was launched with great fanfare in 2003. Following recurrent criticism fueled by its overly ambitious plans, it was replaced in 2012 by a new National Cancer Informatics Program (Goldberg 2012; Thomas 2012). According to Leonelli (2013), caBIG was an "all-encompassing" database designed to provide a pluralistic community of clinical and basic researchers in oncology with easy access to a heterogeneous collection of cancer-related data. Interoperability was a key preoccupation, leaving "as much room for selecting and interpreting data as possible to their users". Otherwise put, the motley of data to which caBIG gave access had to be general enough to allow for global circulation and specific enough to fit the needs of local expert communities. A paradigmatic 'boundary object' (Star and Griesemer 1989, 393), its inability to manage this tension between two opposing demands — "fostering the global circulation of data and facilitating their local adoption" — led to caBIG's demise (Leonelli 2013). The relevant issue here is that the database design and structure were not predicated upon a shared understanding among a specific community of practice of its content and possible uses. Rather, it was supposed to "serve as many specialized uses of data as possible", with data reuse enabling collaboration or even integration across communities.

In contrast, the knowledgebases discussed in this paper seek to provide evidence-based, actionable interpretations of genomic data for use by clinical practitioners engaged in the implementation of precision oncology. From this perspective, unlike the metaphorical travelers who maintain their identity in different locations, the constitution and handling of a knowledgebase cannot be reduced to the transfer of free-floating bits of information from publications to knowledgebases through nested database systems. The issue is not simply that each database channels and filters data. Rather, data experience a process of 'extensive' manual curation, whereby, after being extracted from publications, they undergo valuation and ordering by being paired with levels of evidence, levels of actionability, and summary statements that vary from knowledgebase to knowledgebase. As a result, the information provided by knowledgebases qualifies as actionable claims or statements, rather than data, and becomes undistinguishable from the knowledgebases in which it is embedded. This fact also accounts for the difficulties encountered when curators attempt to compare or harmonize knowledgebases.

Prominent oncology knowledgebases include Vanderbilt's My Cancer Genome (MCG), launched in 2011 as the first public somatic variant interpretation resource, MD Anderson's Personalized Cancer Therapy (PCT), Memorial Sloan Kettering's (MSK) OncoKB, and Wash U's Clinical Interpretations of Variants in Cancer (CIViC). These knowledgebases rely on the biomedical literature collected in the PubMed database and in other databases such as the Catalogue Of Somatic Mutations In Cancer (COSMIC). Established in the UK at the Wellcome Trust

Sanger Institute in 2004 with just four genes, COSMIC has now become the "world's largest and most comprehensive resource for exploring the impact of somatic mutations in human cancer".[4] In addition to data manually curated from PubMed, COSMIC contains other datasets such as those produced by multi-center collaborative networks (Forbes et al. 2015). In short, and as the knowledgebase developers admit, theirs and similar resources stand "on the shoulders of these other giants, these other resources that have many more variants, tens of thousands, hundreds of thousands, even millions of observations and variants" (Interview with Drs Obi and Malachi Griffith, December 2016; henceforth MOG1). Figure 2 (Ainscough et al. 2016) illustrates this dependency structure.

Given the existence of multiple knowledgebases, oncologists are confronted with a complex landscape of interrelated resources that, despite recurrent harmonization initiatives, display distinct characteristics and design choices that promote their individuality. An informant spoke, in this respect, of a "very complicated landscape of resources that are pulling multiple different resources together, integrating them in some way, helping things be visualized, or making things more user friendly, and it's a bit Wild West" (MOG1). Rather than standalone devices, these resources maintain both horizontal and vertical relations: some repositories, such as COSMIC, act as de facto quasi-standards on which others explicitly rely, extracting and embedding their content, while simultaneously maintaining an individuality that challenges the seamless interoperability of their data. CIViC, for instance, links its content to COSMIC, perceived as a complementary and yet distinct resource:

> If you have a specific variant and you find it in CIViC, then you know that someone in CIViC believes it is clinically relevant, with some documented evidence, and we link out to

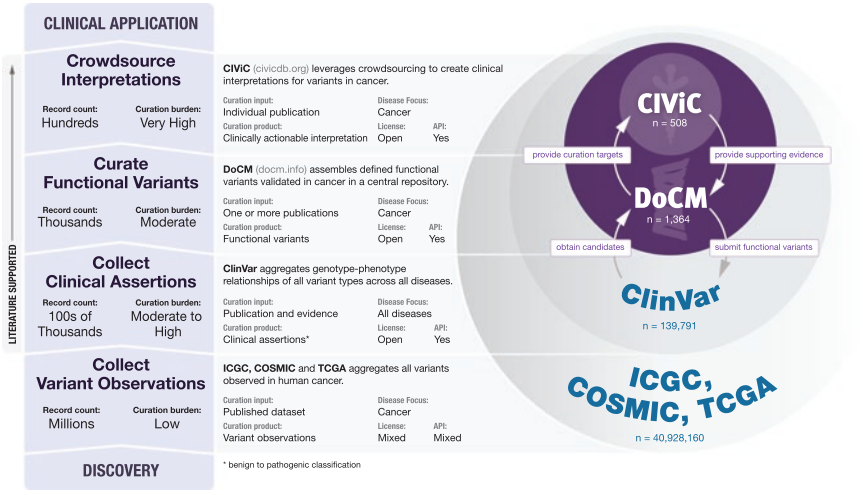

Fig. 2 CIViC in the context of related resources. Reprinted by permission from Springer Nature: *Nature Methods*, DOCM: A database of curated mutations in cancer, B.J. Ainscough et al., Copyright ©2016

[4] https://cancer.sanger.ac.uk/cosmic

> that variant’s record in COSMIC, so you can learn also what COSMIC says about that variant, about how many types of cancers have seen that variant before, and that is useful information. But there are many variants, most variants, that you won’t find in CIViC because they haven’t yet reached this level of documented clinical relevance, but they still exist in COSMIC and that’s still useful information that you could use to design an experiment or understand something about that variant, but it just doesn’t reach that level of clinical relevance for CIViC. (MOG1).

Far from being isolated and self-contained, knowledgebases function within an information ecosystem to whose intricacy and development they contribute. Their curatorial practices, for instance, include the active consulting of other databases and knowledgebases:

> When curators receive a list of gene-variants to curate, they are also given instructions to not limit their search for information to PubMed. They are trained to reference other publicly available knowledgebases such as COSMIC, Jackson Lab’s Jax CKB, and MyCancerGenome for example. Importantly, they are explicitly instructed to not copy or paraphrase the interpretations from these knowledgebases, but to use them as a resource for the primary literature on key gene variants. (Interview with Drs Debyani Chakravarty and J.J. Gao, May 2017; henceforth C/G).

They also openly relate to (or even embed) each other. As part of its data architecture, MSK has developed cBioPortal (now a multi-center endeavor), an advanced data visualization tool that draws on a number of different resources including, most obviously, OncoKB, but also CIViC, MCG, and, as one would expect given its preeminence in the field, COSMIC. When viewing the record for a given variant in cBioPortal, a user can mouse over icons for each of the above resources to bring up a brief summary of the information they contain or click through to proceed to their website. The information excerpted from those resources can thus be accessed directly via the cBioPortal interface, but the kind of information provided in the pop-up windows is different for each resource, so that inclusion of different knowledgebases provides complementary, rather than redundant information, about the ‘same’ molecular entity.

5 Practitioners’ Accounts of the Database/Knowledgebase Distinction

When asked to elaborate on the distinction between databases and knowledgebases, one of the developers of My Cancer Genome offered the following tripartite categorization, borrowed from the ‘data-information-knowledge hierarchy’[5]:

> You take data, say measurements or patient data, then you analyze or aggregate or present those data, and that would be the information, and then if you synthesize information across a bunch of different sources, that would be the knowledge. The point of MCG and CIViC and some of the other resources is really to be a ‘knowledgebase’. (Interview with Dr. Christine Micheel, July 2017; henceforth CM1).

[5] See https://en.wikipedia.org/wiki/DIKW_pyramid

Asked to clarify her statement by comparing, for instance, COSMIC and MCG, she answered that "COSMIC catalogues the alterations that have been observed in cancer, and MCG explains how that may impact therapeutic decisions" (Interview with Dr. Christine Micheel, August 2017). For his part, when asked a similar question the COSMIC director replied:

> We focused on the database angle until fairly recently. We wanted to collect as much information in the one place to empower others to investigate it to look for new genes, new targets. And we kept feeding the database. The increasing breadth and depth of that database just gives other scientists more power for their investigations. … Are we a database or a knowledgebase? We're probably focused more on the database angle of this than the knowledgebase angle. (interview with Dr. Simon Forbes, May 2017; henceforth SF2).

Rather than attempts to build robust ontological categories, these definitions of the database/knowledgebase distinction qualify as pragmatic categorizations within a rapidly evolving context. They situate each kind of 'base' in relation to the aforementioned spectrum that ranges from large-scale repositories, such as the now defunct caBIG, to single-purpose knowledgebases, via intermediate entities such as COSMIC that qualify as 'information bases' insofar as they systematically arrange information. The case of COSMIC, given its liminal position, is a useful starting point for clarifying this issue.

Compared to other endeavors COSMIC qualifies as a 'giant' because of the millions of data it contains in contrasted to the thousands typically found in a knowledgebase. Because of its 'database-ish' nature, and its comprehensive reach, COSMIC "is different things for different people … in some sense, it is just a large bucket of information that you can sift through with different perspectives in mind" (interview with Dr. Simon Forbes, February 2017; henceforth SF1). COSMIC, however, is not an undifferentiated 'bucket', but a bucket of baskets: it includes data subsets targeted to specific users. For instance, the Cancer Gene Census subset that catalogues genes causally implicated in cancer has been recently upgraded by adding annotations related to the traits that govern carcinogenesis, known as the 'hallmarks' of cancer (Hanahan and Weinberg 2000, 2011). As noted by the director of COSMIC, the CGC "the way it looks at the moment is more 'database-ish' as well, but with the new hallmarks annotations we're aiming more toward knowledge, we can describe the functional impact of each gene in cancer rather than just that it causes cancer" (SF2). This is part of a broader plan to transition from an exclusive focus on data acquisition, to the inclusion of annotations about the value of the information, leading, for instance, to the design of a "targeted, specific subset of the database toward clinicians and diagnostics". As acknowledged, however, by the same informant:

> If you're a clinician you might want to get in [COSMIC] for some clues around the impact of mutations, but it' not going to tell you that information because it wasn't really built with that in mind. We built it to gather large quantities of information. (SF1).

Is "looking for some clues around the impact of mutations" then the primary motivation for creating knowledgebases?

6 Why Knowledgebases?

Given COSMIC's pre-eminent position in the field, why did oncology practitioners feel the need to develop knowledgebases? Part of the answer lies in the need for dedicated clinical information to guide therapy. As noted by a cancer genomics researcher:

> COSMIC is just cataloguing pure genetics data online, so in the end we don't know much about clinical outcomes of these cases. It's very limited in scope. It still tells us whether a mutation has been observed more frequently than expected, which tells us something about whether it is likely to be a driver or not, but it still needs much more. We need much more data in these databases. (Interview with Dr. Marco Gerlinger, January 2016).

The missing data are bio-clinical, i.e. data that re-specify genomic entities by tying them to clinical insights; "what we're really interested in, is the clinical data that will be useful for interpretation of the molecular data, and to integrate that" (C/G). According to the same respondents, "in the development of OncoKB one thing became very clear: without clinician insight, OncoKB will be useless for clinical decision support". The information embedded in OncoKB links biological, clinical, and therapeutic information from multiple sources, which include not only the medical literature, but also FDA labeling, clinical guidelines, and abstracts from major conference proceedings, such as the American Society of Clinical Oncology (ASCO), the European Society of Medical Oncology (ESMO), and the American Association for Cancer Research (AACR).

Most importantly, in OncoKB annotations derived from these sources are not merely selected and organized by curators but vetted by a Clinical Genomics Annotation Committee (CGAC) consisting of MSK clinicians who represent leaders in their respective disease-specific fields:

> MSK has some of the best clinical and research expertise in the country. For OncoKB it was not sufficient to simply curate the available literature, our loftier goal was to capture, in a database readable format, the interpretation of these data through the lens of MSK in-house clinical expertise. (C/G).

The following example illustrates the nature of the clinicians' vetting:

> Our initial OncoKB curation efforts cast a wide net, allowing inclusion of information with any possible opportunity for clinical intervention based on the presence of a genetic variant. However, it became very clear very quickly that MSK is conservative in its definition of precision oncology. Thus, for example, we had initially included TP53 as potentially clinically actionable, based on an open phase I clinical trial testing a specific chemotherapy in TP53 mutant patients. However, the clinical committee made us immediately remove TP53 based on their real-world experience, i.e. TP53 alterations are present in 40% of patient tumors, [but] to-date there have been no therapies that have been able to effectively utilize TP53 as a predictive biomarker of activity for a targeted therapeutic. (C/G).

CIViC also focuses on data interpretation: "the meat of what we're trying to create, the data or content that we're creating, is actually the interpretation" (MOG1). The presence or absence of a clinical input is used to draw a line not only between CIviC and COSMIC but also between knowledgebases:

> Knowledgebases such as CIViC are great tools for use in the research space. They comprehensively capture the scientific literature and present this data in a research intuitive way. The development of knowledgebases such as MD Anderson's PCT and our OncoKB have been, from their inception, guided with the clinician in mind as the end-user. For OncoKB, there was an institutional mandate that physician scientists who represent disease experts were to guide the curation by specifying which information would be useful for clinical treatment decisions, and which information was considered extraneous. (C/G).

CIViC developers counter that:

> Resources like OncoKB and PCT talk a lot about their clinician review, but I haven't seen much of a structured representation of what that is, like which clinician reviewed which elements in what ways. You're just told: "You look at something in OncoKB or PCT, you should feel more confident in it because we have had it reviewed by clinicians." But that fact doesn't seem to be represented in the data model in any sophisticated way. (Interview with Drs Obi and Malachi Griffith, June 2017; henceforth MOG2).

The kind of curation, rather than the mere presence of curation, broadly defined, is thus at the very heart of the valuation processes that underlie the database/knowledgebase distinction.

7 Modes of Curation

During the Obama administration, when confronted with the challenges raised by precision medicine, the US FDA began considering a scenario according to which test developers might use information derived from a 'regulatory quality database' to support their claims. To qualify as 'regulatory quality', a database would be curated, have standards, and preferably provide levels of evidence, all of which differentiates it from a data repository (interview with an FDA official, March 2015). So, here is a first distinction: a non-curated repository and a curated database. But things are not so simple, because when asked for an example of a repository, our respondent mentioned a database that maintained in fact a relatively large team of curators. It thus looks as if it is not curation *per se* that is at stake here, but the *kind* of curation, namely research-oriented vs. clinically oriented curation. For instance, having attended a meeting of the International Society of Biocuration, one of the developers of MCG explained:

> Those are the folks that really started and maintained those research-oriented resources … I think the primary difference is the intended audience. When [Drs. Pao and Levy] conceived of MCG they were really focused on the clinician audience … both were practicing oncologists, intimately familiar with the workflows of a clinician, the way a clinician thinks, and the amount of time they have to look at a resource. The research-focused resources are really not what a clinician needs. (CM1).

The issue is not simply to avoid wasting a clinician's precious time, but, more importantly, to protect clinicians from being fed inaccurate or potentially damaging information derived from inappropriate contexts:

> For example, a patient with early stage disease is annotated to have this alteration and therefore they should get this therapy, without recognition that really in that context it is not within clinical guidelines to make that actionable. … There are a lot of manuscripts written

> without a clinical implication in mind, and saying this biomarker is associated with this drug sensitivity while the drug doses being used are clinically not relevant at all, or that biomarker association was really not a very strong one. (FMB).

This also accounts for the decision by more clinically oriented knowledgebases to include information from oncology conferences. While results presented at conferences are “generally not held to the same standard of quality or validation that a publication will be” (MOG2), they do contain relevant clinical information that is not otherwise available:

> When annotating the clinical implications of gene variants, our clinicians frequently referred us to interim clinical trial data from the proceedings of disease-specific and general clinical oncology conferences. Importantly, tumor-type specific negative data and information as to whether a drug is being discontinued from further development due to poor efficacy data is only available through conference proceedings. (C/G).

Several knowledgebases are deeply embedded in the clinical infrastructure of their parent organizations, thus providing further evidence of their situatedness. MD Anderson’s PCT, for example, acts as the external window of its Precision Oncology Decision Support (PODS) service (Meric-Bernstam et al. 2015; Kurnit et al. 2017, Dumbrava and Meric-Bernstam 2018). PODS is a prime internal resource for MD Anderson’s physicians who need assistance with the interpretation of genomic reports. It provides a rapid assessment of the quality of the testing platform, of the alterations seen in actionable genes, and of variant interpretation. In order to make it available for in-house physicians with similar patients, the information goes into a back-end database behind the institution’s firewall, whereas the information included in the external PCT knowledgebase concerns only those variants that have literature support.

A similar situation prevails at MSK, where thousands of patients are sequenced and subsequently matched with a large trial portfolio via a sophisticated IT infrastructure (Eubank et al. 2016). OncoKB annotation is included in the sequencing report that provides summaries of relevant information about alterations for which there are FDA-approved biomarkers and drugs, or compelling clinical data justifying enrolment in a specific clinical trial (C/G). The treating oncologist (who makes the final therapeutic decision) can then interact with the OncoKB team and other colleagues to further discuss the recommendations. As with MD Anderson, the public version of OncoKB does not include all internal information.

8 Trust and Transparency

Knowledgebases deploy different curatorial strategies that define how each positions itself vis-à-vis the others in a climate defined by both competition and collaboration within the oncology community. Rather than clinical expert knowledge, CIViC resorts to crowdsourcing, arguing that the sheer amount of potentially relevant references available in PubMed makes such an approach inescapable, a claim supported by the fact that the overlap between the publications curated by different knowledgebases is extremely low. CIViC’s Wikipedia-like crowdsourcing nonethe-

less involves, in addition to external curators (any user can in principle be a curator), internal curators, site editors, and domain experts in charge of ensuring quality. Crowdsourcing offers the advantage of introducing a measure of transparency:

> CIViC is the only database that actually allows a user to log in and comment and say: "Hey I disagree with this", or "You're missing this important paper", or "I would like to modify this to make it clearer". The other resources generally have behind the scenes a team of experts and they work as a sort of editorial board, almost like writing mini reviews about each variant and each gene, and they may have a collaborative process, but it's hidden and it's not occurring inside the interface and there's not the same degree of provenance about who exactly said what, and how did the knowledge evolve from its initial state to the current state, and so on. (MOG1).

To which other practitioners counter:

> Crowdsourcing as a theoretical concept is amazing. However, it comes with the assumption that clinicians, who have very limited time and bandwidth, will buy into that concept. I think one of the key factors contributing to the success of OncoKB is that MSK clinicians were mandated to guide OncoKB development since it was slated to be an institutional clinical decision support system. Additionally, we had carefully trained medical fellows and translational cancer biologists as curators who were well versed in the quality control of information that we would allow into OncoKB. (C/G).

The emergence of knowledgebases devoted to the same purpose is less an expression of redundancy than of the existence of different curatorial approaches that embed and enact each knowledgebase's strategy held together by a tension between trust (in expert judgment) and transparency (of the curatorial process). When asked what motivates the proliferation of knowledgebases, a practitioner explained:

> If you are a center or a company and you are interpreting a variant for an actual patient, a real patient, and you're acting on that information, what information do you trust? [What information] gives you confidence that you could actually act on that mutation to do something for that patient? … So, what's ended up happening is that every center just says: "We don't know who we're going to trust, so let's just recreate the whole thing over again and we control it." … There's kind of this tension between openness and trust. (Interview with Dr. Ethan Cerami, April 2017).

This tension is reflected in the different solutions adopted by CIViC and OncoKB. Both knowledgebases originated in an attempt to streamline interpretation work. Their development, however, diverged as CIViC adopted traceability and transparency as its trademark, whereas OncoKB is vetted by, and therefore representative of, MSK clinical expertise.

9 Curation, Interpretation, and Levels of Evidence

Thanks to its transparent curatorial system, CIViC offers a more granular view of those practices. The debates between curators and editors are available on the CIViC interface, and although a vast majority of them are relatively short and 'technical', some involve choices that escalate to concerns about underlying principles and the meaning of curation and data interpretation. Here is an example:

> Curator A posts evidence concerning the EML4-ALK E20 variant on the webpage.
>
> Editor B deletes part of the evidence summary arguing that it amounts to speculations. She also reduces the evidence trust rating from 5 to 3 stars.
>
> Curator A replies that he recognizes the speculative nature of his summary, but that this is part of his philosophy of evidence-statement production and his interpretation of CIViC’s mission, namely, to add context and speculate on possible connections and significance.
>
> In the ensuing discussion, Editor B asks Editors C, D, and E to weigh in on the discussion of the group’s philosophy of interpretation and evidence-statement production.
>
> She also attempts to clarify the meaning of a 5-star trust rating that should refer to highest-quality, standard-of-care studies, and be based on how well the evidence supports a given predictive statement, not the overall quality of the original paper.
>
> Concerning the deleted passages, Editor B suggests that “the additional text would be well suited to a comment at the time of submission, but I believe it to be tangential to the main point.”
>
> Editor D steps in, noting that information extracted from case reports warrants by definition a lower star rating, because of its anecdotal nature. He agrees with Editor B, and this ends the discussion.

This vignette shows how curation debates can be framed by the essential tension between the clinical purpose and utility of the knowledgebase (see the reference to standards of care), and the scientific validity and the future of evidence statements. Reminiscent of the work of guideline developers (Knaapen et al. 2010), it also highlights the textual dimension of curatorial practices, whereby data are polished into statements. A further example of this dynamic is provided by the following example:

> Following the posting of a new evidence-summary statement, the discussion focuses on whether certain kinds of lower-evidence statements, in this case about mutation co-occurrence, belong in CIViC because they could subsequently turn out to be useful.
>
> – Editor A questions the clinical utility of the evidence, whether the information actually fits into the evidence schema offered by CIViC, whether it qualifies as diagnostic, and whether it has been given the appropriate evidence-quality grade. He nonetheless acknowledges the importance and potential usefulness of the study behind the evidence statement.
> – Editor A ultimately rejects the submission, but with an encouragement to produce a new evidence statement that more clearly articulates its relevance.

It thus appears that ‘data’ excerpted from publications or databases are transformed through interpretation because they are turned into different kinds of evidence, or evidence for different things. Again, the issue is not about data or evidence *per se*, but about the *textual framing of evidence statements* and their relation to clinical utility. A key device, in this respect, is the attribution of Levels of Evidence (LoE) to statements, which act as markers of the degree of uncertainty characterizing the actionability of that statement. All the knowledgebases we investigated include LoE, and this, once again, reminds us of the centrality of this device in relation to clinical utility:

> I think the levels of evidence is instrumental, because for a clinical decision support tool to have any sort of utility a clinician needs to know: “What am I doing? Is it backed by consensus?” (G/C).

Knowledgebases have adopted different approaches to LoE. For instance, OncoKB's LoE are tied to the sum of evidentiary support that a specific mutational event is predictive of response to a targeted therapy, whereas CIViC's LoE reflect the source of the evidentiary support that comes with the statement. CIViC items are additionally accompanied by a 'Trust Rating' that indicates how compelling that evidence is judged to be. There are, moreover, differences in how knowledgebases advertise their LoE component. For instance, CIViC is described as a "community knowledgebase for expert crowdsourcing" (Griffith et al. 2017), whereas OncoKB is presented as a "a precision oncology knowledge base" that includes a distinctive system of Levels of Resistance (LoR) predictive of resistance to a specific targeted therapy (Chakravarty et al. 2017).

These differences can be compounded with the fact that establishing LoE is notoriously contentious as it involves a large degree of interpretative flexibility and because of the conflicting sources that can be used to perform that task:

> The interpretation of the genomic variants is subjective – I mean a fifty percent response rate for you is responsive? What about five percent? … For that individual patient, one of twenty that responded, this gene-drug-disease match was perfect. Just one out of twenty. Five percent. So, is this responsive if I consider a broader population? … We have one interpretation that is different from OncoKB: they have their own strengths because they have internal data, but [our source] is published, we have the connection. (Interview with an oncology data scientist, October 2016).

This brings us back to the tension between trust in the clinical expertise available at leading cancer centers and the traceability of statements to published sources. The process at MSK illustrates how the clinical consensus of an institute is captured by knowledgebase annotation:

> Several MSK physician-scientists, who represent a broad spectrum of opinion have provided insight into what a given OncoKB annotation should or should not include. One key role of OncoKB is to generate a consensus of opinion from these varied voices. Discussions and compromise have taken place through this process, no one voice has dominated, and the OncoKB annotation represents the middle ground. (C/G).

The excerpt highlights the role of local context and shared understandings in the valuation processes underlying the trustworthiness of specific statements, and thus the worth of individual knowledgebases.

10 Heterogeneity

Knowledgebases differ in terms of the kind and amount of information they carry and the assessment and interpretation of the evidence they include. In fact, they overlap very little in terms of the specific variants included and the literature they reference. When they do overlap, they may actually interpret variants differently, either because their curation relies on different publications, or because they interpret those publications differently (Patel et al. 2016).

Knowledgebases contain interpretations rather than 'data' as such (Pitel 2017). These interpretations consist of statements about associations, i.e. claims about the

evidence that a given mutation plays a particular role in cancer, and the evidence that a drug or intervention may be associated with that variant and have clinical relevance. Even in a database such as COSMIC the 'data' is not the variant itself, but the pairing of a set of genomic coordinates that represent the variant with a given biopathological process. In the case of knowledgebases, the unit of analysis consists less of 'data' than evidence records, which amount to sets of locations, cross references, and literature citations leading to an interpretation. The interpretation defines which variants are clinically relevant and the description of that clinical relevance varies from one knowledgebase to another. Factors that account for this variation include the sheer number of available publications, so that the overlap of the literature covered by a given knowledgebase can be quite small. Moreover, as noted by the developer of the PathOS decision support system (Doig et al. 2017), "a PubMed article is a pretty large body of data, and actually finding the sentence that confirms that the action is positive or negative or related to something is actually a very hard job" (Interview with Dr. Ken Doig, June 2017).

Other sources of heterogeneity include temporality and granularity. Temporality refers to the rapidly evolving knowledge in oncology, so that information presented at a conference, or even published, can be quickly disproved or replaced:

> We get a lot of requests to add [information from conference abstracts] because there are clinicians who want the most amazing cutting-edge stuff, and then you have other clinicians where we have the feedback that this published NEJM paper from three years ago [is] not good enough because it was debunked by a subsequent JAMA paper two years later, with a much larger clinical trial that was better statistically powered. (MOG2).

As for granularity, while the knowledge at the level of a gene expressed in guidelines and regulatory documents might be relatively stable, the same does not apply to gene variants:

> The FDA-labeling of approved targeted agents in a specific indication can be vague. For example, the FDA-approval of erlotinib in patients with EGFR-mutant non-small cell lung cancer was irrespective of EGFR mutation status. This is because in these cases, the drug's approval predates much of the sequencing data that determined the specific patient populations that benefit from the targeted agent. (C/G).

Similar considerations apply to guidelines that include mutations for which there are established data:

> But what does a clinician do when faced with a sequencing result that includes a known actionable gene but a lesser known variant? … That kind of information is critical in supporting clinical care, and that's where the levels of evidence represent a practical and immediate way to communicate this information. (C/G).

Knowledgebase developers are well aware of the issue of heterogeneity which is viewed as both problematic and unsurprising given the extent of the field and the complexity of interpretation. They have recently established the Variant Interpretation for Cancer Consortium (VICC), to "harmonize global efforts for clinical interpretation of cancer variants".[6] Rather than building yet another knowledgebase (a 'meta-

[6] https://genomicsandhealth.org/working-groups/our-work/variant-interpretation-cancer-consortium

knowledgebase'), the idea is to construct a portal giving access to the content of multiple knowledgebases. Thus, the field may move toward addressing the problem of heterogeneity without having to sacrifice either the latent mistrust embedded in or the pragmatic role fulfilled by locally maintained knowledgebases. This suggests that rather than a solution to the 'data interpretation bottleneck', knowledgebases and their claims and statements are still part of that same bottleneck, requiring additional bioinformatic and expert clinical human work.

11 Conclusion

Oncology has recently transitioned from its traditional, organ-based approach to a 'precision oncology' of molecular alterations. As a result, it has become 'data-centric' (Leonelli 2016). Its practices increasingly rely on a sophisticated techno-scientific infrastructure that generates large amounts of data that demand consistent, appropriate interpretations. In turn, attempts to overcome the interpretation bottleneck have led to the establishment of a complex landscape of interrelated resources that, while displaying distinct characteristics and design choices, also entertain horizontal and vertical relations. Although there is no denying that the data-centric nature of contemporary oncology raises a number of key issues related to the production and circulation of data — issues that can be explored using the 'data journeys' metaphor — we suggest in this paper that the focus on data use and re-use should be complemented by a focus on interpretation. Interpretation here refers to both the 'interpreting' activities performed by bio-clinical collectives, and to the outcomes of those activities under the guise of *actionability claims or statements*, rather than 'data'.

Oncology practitioners refer to data interpretation resources as 'knowledgebases', an actor's category designed to differentiate them from generic, multipurpose databases. While in most cases publicly accessible, albeit in a pared-down format compared to their in-house version, knowledgebases are deeply embedded in the clinical pathways of their home institutions. Their major purpose is the definition and identification of *clinically actionable* alterations, i.e. those that drive tumors and can be matched to treatments. This is no easy task, as shown by the existence of several knowledgebases that, in spite of their common purpose, adopt very different approaches partly captured by the tension between trust and traceability. To investigate what makes different knowledgebases 'valuable' to genomic practitioners confronted with a rapidly evolving domain, we have examined their structure and dynamics. The nature, amount, and quality of curation underwriting each knowledgebase appear to be major contributors to these valuation processes. A heavy investment in human curation, of a clinical rather than exclusively scientific nature is needed to make them valuable, but each knowledgebase appears to have its own way of connecting clinical and scientific statements elicited from publications, conference abstracts, clinical trials, genomic datasets, and even in-house expert statements.

The main goal of the NIH "Strategic Plan for Datascience" mentioned at the beginning of this paper is to facilitate "the modernizing [of] the NIH-funded biomedical data-resource ecosystem". The Plan refers to the development of core data

repositories to be used across different scientific domains, but also marks out a special place and a distinct role for knowledgebases within the data ecosystem. Knowledgebases that, as just mentioned, involve large amounts of human curation have been developed by "targeted communities for the benefit of scientists in that community", and they are here to stay, as they will "still serve the functions of their own communities the way they always have, [as] distinct entities with their own priorities, their own goals and objectives" (Interview with Dr. Susan Gregurick, May 2018). While, according to the same respondent, part of the information they contain could be 'hardened', by for instance being made compliant with the FAIR principles for data management (Wilkinson et al. 2016), and thus transferred at some future point to a data repository, the situated and ever-changing nature of the information collected in knowledgebases make such a prospect somewhat difficult to entertain, especially in clinical domains characterized by the ongoing realignment of the normal and the pathological.

Admittedly, the database/knowledgebase distinction is ideal-typical, given that COSMIC, for instance, is shifting from its initial exclusive focus on data acquisition to highlighting the value of its data (SF2). Oncologists consult COSMIC for research purposes but also to gather information about alterations detected in their patients, although they might do so via local resources that embed COSMIC. While there is an overlap, in terms of use, between COSMIC and the more specialized knowledgebases, the latter lie at one end of a wide spectrum of resources that range from large databases to smaller interpretative resources. In the case of a database such as COSMIC that sits in the middle of this spectrum, the data journey metaphor may be used to describe how curators survey the literature, extract and refashion bits of information, assess their evidentiary strength, and decide whether and how to include them in the database. The addition of the PubMed reference number to those data in principle should allow users to travel back to the original source although, as already mentioned, this is not a straightforward task given the amount of curatorial work needed to locate specific statements. Knowledgebases, however, are less a data repository than a tool for (clinical) action, and the data journey metaphor misses this key aspect. Within knowledgebases bits of information are triangulated with other evidence, associated with levels of evidence and actionability, and embedded in carefully crafted statements that re-specify their meaning. This explains, in part, the major differences between knowledgebases, whereby the 'same' genomic variant is transmogrified into different entities connected to different actions.

In a domain where genomic information is becoming increasingly important for clinical decision-making, but drastically outpacing the genomic literacy of the average oncologist/clinician, knowledgebases are an attempt to fill a translational gap and provide clinicians with information about the actionability of molecular alterations, and the kind and strength of the evidence that underpins it. Knowledgebases, in this context, are designed to act, in a sense, as a virtual, *in-silico* ersatz for the multi-disciplinary gathering of oncology practitioners, molecular biologists, and bioinformaticians who come together to reach a consensus about actionable suggestions (Bourret and Cambrosio 2019). In the case of institutions such as MSK, the sheer number of sequenced patients (Zehir et al. 2017; Eubank et al. 2016) makes such a solution impossible. Instead, a tumor profiling report associated with a clinical decision support tool, OncoKB, is sent electronically to

the treating physician who can trust the provided clinical annotations because they are clinically vetted. "OncoKB", in this context, refers not merely to the knowledgebase, narrowly defined, but to the entire *dispositif*, that includes, for instance, the Clinical Genomics Annotation Committee staffed with leading clinicians.

Knowledgebases, rather than a mere data repository, embed and perform interpretations that deploy a distinctive form of bio-clinical *expertise*. Conversely, in data-centric oncology human expertise can only be enacted *via* bio-clinical collectives properly equipped with tools and devices such as those provided by knowledgebases. This apparently vicious circle becomes virtuous when those tools and devices are constituted and utilized at different places and different times by different collectives. Hence the temporal and relational nature of oncology databases and knowledgebases, which evolve in response to a number of other initiatives, for instance the introduction of new data-sharing projects sponsored by leading cancer centers. Last but not least, we should not forget the strictures that oncology, as a *clinical* domain, imposes upon knowledge production and knowledge flows, and which largely account for the difference between clinical-grade knowledgebases and the kind of databases deployed in other scientific domains.

Acknowledgements Research for this paper was made possible by the following grants: Canadian Institutes of Health Research MOP-133687 and French National Cancer Institute (INCa) SHSESP14-002. We would like to thank all the knowledgebase developers who agreed to be interviewed, in some cases repeatedly. Special thanks to Sabina Leonelli who singlehandedly coerced us into writing this paper.

References

Ainscough, Benjamin J., Malachi Griffith, Adam C. Coffman, et al. 2016. DoCM: A Database of Curated Mutations in Cancer. *Nature Methods* 13: 806–807.

Amin, Ash, and Patrick Cohendet. 2004. *Architectures of Knowledge: Firms, Capabilities, and Communities*. Oxford: Oxford University Press.

Bates, Jo, Yu-Wei Lin, and Paula Goodale. 2016. Data Journeys: Capturing the Socio-Material Constitution of Data Objects and Flows. *Big Data & Society* 2016 (July–December): 1–12. https://doi.org/10.1177/2053951716654502.

Bijker, Wiebe E., and John Law, eds. 1992. *Shaping Technology/Building Society: Studies in Sociotechnical Change*. Cambridge, MA: MIT Press.

Bourret, Pascale, and Alberto Cambrosio. 2019. Genomic Expertise in Action: Molecular Tumour Boards and Decision-Making in Precision Oncology. *Sociology of Health & Illness* 41: 1568–1584.

Buetow, Kenneth H. 2005. Cyberinfrastructure: Empowering a "Third Way" in Biomedical Research. *Science* 308: 821–824.

Cambrosio, Alberto, and Peter Keating. 2000. Of lymphocytes and Pixels: The Techno-Visual Production of Cell Populations. *Studies in History and Philosophy of Biological and Biomedical Sciences* 31: 233–270.

Chakravarty, Debyani, Jianjiong Gao, Sarah Phillips, et al. 2017. OncoKB: A Precision Oncology Knowledge Base. *JCO Precision Oncology* 1: 1–16. https://doi.org/10.1200/PO.17.00011.

Doig, Kenneth D., Anthony Fellowes, Andrew H. Bell, et al. 2017. PathOS: A Decision Support System for Reporting High Throughput Sequencing of Cancers in Clinical Diagnostic Laboratories. *Genome Medicine* 9 (1): 38.

Dumbrava, Ecaterina I., and Funda Meric-Bernstam. 2018. Personalized Cancer Therapy. Leveraging a Knowledge Base for Clinical Decision-Making. *Cold Spring Harbor Molecular Case Studies* 4: a001578.

Eubank, Michael H., David M. Hyman, Amritha D. Kanakamedala, et al. 2016. Automated Eligibility Screening and Monitoring for Genotype-Driven Precision Oncology Trials. *Journal of the American Medical Informatics Association* 23: 777–781.

Forbes, Simon A., David Beare, Prasad Gunasekaran, et al. 2015. COSMIC: Exploring the World’s Knowledge of Somatic Mutations in Human Cancer. *Nucleic Acids Research* 43: D805–D811.

Ghazani, Arezou A., Nelly M. Oliver, Joseph P. St Pierre, et al. 2017. Assigning Clinical Meaning to Somatic and Germ-Line Whole-Exome Sequencing Data in a Prospective Cancer Precision Medicine Study. *Genetics in Medicine* 19: 787–795.

Gitelman, Lisa, ed. 2013. *“Raw Data” is an Oxymoron*. Cambridge, MA: The MIT Press.

Goldberg, Paul. 2011. Prepare for “Tsunami” of Genomic Information, Sledge Urges in ASCO Presidential Address. *The Cancer Letter* 37 (23): 1–7.

———. 2012. NCI Bioinformatics After Kenneth Buetow: Varmus Launches Fundamental Redesign. *The Cancer Letter* 38 (1): 1–6.

Good, Benjamin M., Benjamin J. Ainscough, Josh F. McMichael, et al. 2014. Organizing Knowledge to Enable Personalization of Medicine in Cancer. *Genome Biology* 15: 438.

Griffith, Malachi, Nicholas C. Spies, Kilannin Krysiak, et al. 2017. CIViC is a Community Knowledgebase for Expert Crowdsourcing the Clinical Interpretation of Variants in Cancer. *Nature Genetics* 9: 170–174.

Hanahan, Douglas, and Robert A. Weinberg. 2000. The Hallmarks of Cancer. *Cell* 100: 57–70.

———. 2011. Hallmarks of Cancer: The Next Generation. *Cell* 144: 646–674.

Hatchuel, Armand. 1996. Les axiomatiques de la production: éléments pour comprendre les mutations industrielles. In *La performance économique en entreprise*, ed. Jacques-Henri Jacot and Jean-Pierre Micaëlli, 35–53. Paris: Hermes.

Johnson, Amber, Jia Zeng, Ann M. Bailey, et al. 2015. The Right Drugs at the Right Time for the Right Patient: The MD Anderson Precision Oncology Decision Support Platform. *Drug Discovery Today* 20: 1433–1438.

Jordan, Bertrand. 2015. Recherche cibles, désespérément. *Médecine/Science* 31: 214–217.

Knaapen, Loes, Hervé Cazeneuve, Alberto Cambrosio, et al. 2010. Pragmatic Evidence and Textual Arrangements: A Case Study of French Clinical Cancer Guidelines. *Social Science & Medicine* 71: 685–692.

Kurnit, Katherine C., Ann M. Bailey, Jia Zeng, et al. 2017. ‘Personalized Cancer Therapy’: A Publicly Available Precision Oncology Resource. *Cancer Research* 77: e123–e126.

Leonelli, Sabina. 2013. Global Data for Local Science: Assessing the Scale of Data Infrastructures in Biological and Biomedical Research. *BioSocieties* 8: 449–465.

———. 2016. *Data-Centric Biology. A Philosophical Study*. Chicago: University of Chicago Press.

———. this volume. Learning from Data Journeys. In *Data Journeys in the Sciences*, ed. Sabina Leonelli and Niccolò Tempini. Cham: Springer.

Lynch, Michael. 1993. *Scientific Practice and Ordinary Action: Ethnomethodology and Social Studies of Science*. Cambridge, UK: Cambridge University Press.

———. 1998. The Discursive Production of Uncertainty. The OJ Simpson ‘Dream Team’ and the Sociology of Knowledge Machine. *Social Studies of Science* 28: 829–868.

———. 2009. Working Out What Garfinkel Could Possibly Be Doing with “Durkheim’s Aphorism”. In *Sociological Objects*, ed. Geoff Cooper, Andrew King, and Ruth Rettie, 101–118. London: Routledge.

Meric-Bernstam, Funda, Carol Farhangfar, John Mendelsohn, et al. 2013. Building a Personalized Medicine Infrastructure at a Major Cancer Center. *Journal of Clinical Oncology* 31: 1849–1857.

Meric-Bernstam, Funda, Amber Johnson, Vijaykumar Holla, et al. 2015. A Decision Support Framework for Genomically Informed Investigational Cancer Therapy. *Journal of the National Cancer Institute* 107 (7): djv098.

Nelson, Nicole, Peter Keating, and Alberto Cambrosio. 2013. On Being ‘Actionable’: Clinical Sequencing and the Emerging Contours of a Regime of Genomic Medicine in Oncology. *New Genetics & Society* 32: 405–428.

Patel, Jaymin M., Joshua Knopf, Eric Reiner, et al. 2016. Mutation Based Treatment Recommendations from Next Generation Sequencing Data: A Comparison of Web Tools. *Oncotarget* 7: 22064–22076.

Pitel, Beth. 2017. *Introduction to Publically Available Knowledgebases to Aid Interpretation of Genomic Findings in Oncology*. https://www.youtube.com/watch?v=4dBh1Qkp8os. Accessed 21 Aug 2019.

Prasad, Vinay. 2016. The Precision-Oncology Illusion. *Nature* 537: S63.

Schwaederle, Mzria, Gregory A. Daniels, David E. Piccioni, et al. 2015. On the Road to Precision Cancer Medicine: Analysis of Genomic Biomarker Actionability in 439 Patients. *Molecular Cancer Therapeutics* 14: 1488–1494.

Star, Susan L., and James R. Griesemer. 1989. Institutional Ecology, 'Translations' and Boundary Objects: Amateurs and Professionals in Berkeley's Museum of Vertebrate Zoology, 1907–39. *Social Studies of Science* 19: 387–420.

Subbiah, Vivek, and Razelle Kurzrock. 2016. Universal Genomic Testing Needed to Win the War Against Cancer: Genomics IS the Diagnosis. *JAMA Oncology* 2: 719–720.

———. 2017. Debunking the Delusion that Precision Oncology Is an Illusion. *The Oncologist* 22: 881–882.

Suchman, Lucy A. 1987. *Plans and Situated Actions: The Problem of Human-Machine Communication*. Cambridge, UK: Cambridge University Press.

Thomas, Uduak G. 2012. NCI reorganizes cancer informatics efforts; cuts some caBIG programs, moves others to NCIP. *GenomeWeb.* https://www.genomeweb.com/informatics/nci-reorganizes-cancer-informatics-efforts-cuts-some-cabig-programs-moves-others. Accessed 21 Aug 2019.

Titus, Karen. 2014a. Molecular tumor boards: Fixture or fad? *CAP Today,* October 14. http://www.captodayonline.com/molecular-tumor-boards-fixture-fad. Accessed 21 Aug 2019.

———. 2014b. From tumor board, an integrated diagnostic report. *CAP Today*, December 15. http://www.captodayonline.com/tumor-board-integrated-diagnostic-report. Accessed 21 Aug 2019.

Van Allen, Eliezer M., Nikhil Wagle, and Mia A. Levy. 2013. Clinical Analysis and Interpretation of Cancer Genome Data. *Journal of Clinical Oncology* 31: 1825–1833.

Vignola-Gagné, Etienne, Peter Keating, and Alberto Cambrosio. 2017. Informing Materials: Drugs as Tools for Exploring Cancer Mechanisms and Pathways. *History and Philosophy of the Life Sciences* 39: 10.

Wilkinson, Mark D., Michel Dumontier, IJsbrand J. Aalbersberg, et al. 2016. The FAIR Guiding Principles for Scientific Data Management and Stewardship. *Scientific Data* 3: 160018.

Zehir, Ahmed, Ryma Benayed, Ronak H. Shah, et al. 2017. Mutational Landscape of Metastatic Cancer Revealed from Prospective Clinical Sequencing of 10,000 Patients. *Nature Medicine* 23: 703–713.

Alberto Cambrosio A Professor at McGill University's Department of Social Studies of Medicine, Alberto Cambrosio's recent work examines "genomics in action", i.e. as applied to concrete instances of medical work, by investigating public, academic and commercial programs that capitalize on the therapeutic insights offered by the new molecular genetics of cancer. His most recent book (*Cancer on Trial: Oncology as a New Style of Practice*, University of Chicago Press, 2012, coauthored with Peter Keating) argues that, contrary to common assumptions, clinical trials do not boil down to a mere "technology" or a few methodological principles: rather, they are an institution that corresponds to a profound transformation of biomedical activities and rise to the level of a "new style of practice". This work builds on a previous book (*Biomedical Platforms: Realigning the Normal and the Pathological in Late-Twentieth-Century Medicine*, MIT Press, 2003, also coauthored with Peter Keating) that analysed the transformation of medicine into biomedicine.

Jonah Campbell has an MA degree in Medical Sociology and is presently a Regular Research Assistant in the Department of Social Studies of Medicine. He has over 7 years’ work experience on a variety of research projects in the sociology and history of biomedicine, with a particular focus on precision medicine, genomic oncology and “Big Data”.

Etienne Vignola-Gagné is an analyst at Science-Metrix, where he conducts projects on science and innovation policies and research management. At McGill University, and the University of Vienna before (where he obtained his doctoral degree), he combined policy analysis and science and technology studies to track the history of “translational research” programs and to follow the introduction of genomics sequencing technologies in clinical oncology. He has authored or coauthored scientific contributions published in venues such as *History and Philosophy of the Life Sciences*, *Science and Public Policy* and *Scientometrics*.

Peter Keating is an Associated Professor at the University of Quebec at Montreal. Currently semiretired, he worked for many years in the history of immunology and oncology and, most recently, clinical cancer trials. He has coauthored several books with Alberto Cambrosio on these topics including *Cancer on Trial: Oncology as a New Style of Practice* (University of Chicago Press, 2012).

Bertrand R. Jordan (b. 1939) obtained his PhD in Particle Physics (CERN, 1965), then moved to molecular biology and spanned many topics during his career as CNRS Research Director (mostly at the Marseille-Luminy Immunology Institute). He notably isolated the first HLA gene in 1982 before turning to genomics, expression profiling and medical and cancer genetics. He has edited two multi-author treatises on gene expression and microarrays in diagnostics and also published 12 books on genetics aimed at the general public. He is a Consultant for several biotech companies in the field of cancer diagnostics and therapy and publishes a monthly *Chronique Génomique* (Genomic Chronicle) in the French journal *Médecine/Sciences*. He is now retired from CNRS but remains Associate Researcher at the ADÈS laboratory.

Pascale Bourret is Associate Professor at Aix-Marseille Université where she teaches sociology. She is also a Researcher at SESSTIM (Economy and Social Sciences, Health Care Systems and Societies), an INSERM-IRD-Aix-Marseille Université UMR. At the crossroads of science studies and sociology of medicine, her work focuses on the transformation of biomedical practices in connection to the development of genomic tools, with a focus on biology/clinic interface, the transformation of clinical work and the production of clinical judgement and clinical decision-making. She has published articles on bio-clinical collectives in the domain of BRCA testing, on regulation issues linked to new genomics tools and on the emergence of genomic-driven clinical trials. Her present projects investigate the implementation of precision medicine in oncology and explore the conditions surrounding the development of targeted therapies in the context of clinical and translational research.