Maintaining and Repairing the Cancer Registries' Regime of Knowing in the Turbulent Context of the French National AI Strategy

François Lambotte
LASCO, UCLouvain, Belgium/francois.lambotte@uclouvain.be

Anja Martin-Scholz
CERTOP, Université de Toulouse, France

Abstract

This ethnographic study investigates maintenance and repair practices that underpin the data work of French cancer registries (CR) amid transformations in healthcare data governance driven by France's national AI strategy. CRs' mission is to provide high-quality data to assess public health policies. Through three cases of breakdown, we analyse repair practices in relation to the regime of knowing encompassing practices, value schemes, and authority arrangements. Drawing on this lens, our empirical study extends repair studies by showing how care, expertise and power relations are intertwined within CRs' repair work. When faced with governance transformations, our findings show how CRs resist these shifts and seek to maintain the regime of knowing to sustain their legitimacy within the healthcare data infrastructure. The study highlights how CRs' restoration efforts seem to fail and points to the need for CRs to move beyond repair to preserve domain-specific knowledge and public health values.

Keywords: Data Journey, Data Repair and Maintenance, Public Healthcare, French National AI Strategy, Cancer Registries, Regime of Knowing

Introduction

In 2018, France launched its national AI strategy (NAS) underpinned by three principles: data as a common good, data sharing, and the humanist ethos (Bareis and Katzenbach, 2022; Paltieli, 2022). Firstly, data should be a common good, a resource whose use and governance is defined by the community (Villani, 2018). Secondly, data sharing is conceptualised as a political virtue (Paltieli, 2022), with citizens consciously choosing to share their data for the benefit of the broader community. Thirdly, the NAS is rooted in the humanist ethos according to which AI innovations should be pushed into sectors that enable human flourishing (Bareis and Katzenbach, 2022). The strategy should therefore focus on sectors that serve the general interest, among which healthcare is central. Taken together, these three principles establish data governance, particularly in healthcare, as a central mechanism of the NAS. A concrete consequence of this was the creation of the Health Data Hub (HDH) "charged by law in 2019 with 'gathering, organising and making available the data of the national health data system mentioned in article L. 1461-1 of the public health code'" (Combes and Maquart, 2021: 4). From a strategic point of view, the HDH aims "to foster the development of AI projects and help improve quality of care" (Hogenhout, 2020: 39).

To understand the specific issues of this transformed healthcare data governance, we followed Hoeyer's (2016: 77-78) recommendation to focus on "everyday practices and to scrutinise the infrastructures facilitating data production and flow". Hence, in 2019 we initiated an ethnographic study with the objective of investigating the data work of two organisations responsible for cancer registries (CRs) in France. These organisations are tasked with the collection of nominative data concerning specific diseases (in this case, cancer) within a geographically defined population for research and public health purposes. The objective of the data work is to provide a comprehensive narrative of the care trajectory of individuals diagnosed with cancer. CRs play a pivotal role in the research activities of social epidemiologists who study the social distribution and determinants of health (Fianu et al., 2022), and data quality is a major concern for social epidemiologists. However, health data are rarely constructed for research purposes in the first place (Balka and Star, 2016; Pine and Bossen, 2020); in most cases, they are collected for purposes of diagnosis, treatment, and billing for treatment. CRs are in charge of the whole process (Leonelli, 2016) whereby data initially produced for treatment or economic reasons is recontextualised for scientific and public health purposes. To paraphrase Leonelli (2020: 9), this work occurs during the data journey, defined as "the movement of data from their initial sites of production [to the CR] in which they are processed, mobilised and re-purposed".

During our ethnographic work, we noticed that the data journey implies significant maintenance and repair work that is not only technical or material (Pink et al., 2018; Tanweer et al., 2016) but relates to political, economic, social, or normative elements constitutive of the socio-materiality (Bates et al., 2016) of the CRs. Hence, the journey highlights not only the specificity and complexity of the data work of CRs but also the rather unstable conditions in which the work is done. Taking these circumstances into account, our work seeks to understand the concerns of CRs about the creation of the HDH:; a direct consequence of the French NAS. CRs perceive this transformation of data governance as content-agnostic, treating data as a commodity (Ribes and Jackson, 2013) and wholly disregarding the specific domains of knowledge production (Alaimo and Kallinikos, 2022). Our ethnographic work shows how this transformation of health data governance has exacerbated the work of maintenance and repair. Consequently, we regard the transformation as having further weakened CRs in their ability to fulfil their public health mission.

To explain our empirical findings, we posit that the data work of CRs is specific to a domain of knowledge and thus draws heavily on a specific *regime of knowing* that encompasses "knowing practices through which actors develop and use knowledge; the valuation schemes through which actions, people, and things are evaluated; and the authority arrangements that determine which actors have control over how the work is performed in certain tasks" (Pachidi et al., 2021: 19). Moreover, we emphasise that repair and maintenance practices in healthcare data work extend beyond addressing technical issues to the ongoing preservation and restoration of the regime of knowing that supports the scientific and public health work of CRs at a time when that work is being shaken by the transformation of health data governance. From a theoretical perspective, we intend to show that combining the concepts of repair and maintenance with the regime of knowing helps deepen our understanding of how AI-driven transformation affects domain-specific knowledge practices.

In what follows, we first revisit the literature on data work in healthcare and present our theoretical framework, which combines the concept of regime of knowing with the concepts of repair and maintenance. In the empirical section, we draw on three cases of data maintenance and repair to describe the regime of knowing that sustains the data work of CRs. We then analyse how the transformation of health data governance (the creation of the HDH) has exacerbated the work of maintenance and repair.

Data work in healthcare

Bossen and colleagues define data work as any human activity related to creating, collecting, managing, curating, analysing, interpreting, and communicating data (Bossen et al., 2019: 466). Data work in healthcare has long been among the practices of diverse occupational groups in organisations working with information systems, including physicians, nurses, and administrative workers. The primary focus of these occupational groups is not data work. Scholars have questioned how healthcare practitioners have juggled data work in tension with their care practices (e.g., Mayère, 1990; Mathieu-Fritz and Esterle, 2013). Self-tracking devices have enabled patients, as well as healthcare professionals, to generate data (Kallinikos and Tempini, 2014; Ruckenstein and Schüll, 2017) and thus to participate actively in data work in healthcare. Studies have considered, for example, issues of power and control over patients through the data produced by wearable devices (Ruckenstein and Schüll, 2017), ambiguities in the meaning of the data produced by patients (Lomborg et al., 2020; Marent and Henwood, 2021), and the increasing work that contextualisation of those data implies for practitioners (Haase et al. 2023; Torenholt and Tjørnhøj-Thomsen 2022). Choroszewicz (2022) studied the place of emotional labour in the journey of data from their production to their repurposing for data analytics, highlighting the role played in data repair by care, frustration, and enthusiasm.

With the rise of big data, other data-focused occupations have grown in importance in healthcare organisations. These occupations are centred on data collection, structuration, curation, and validation. For example, Pine and Bossen conducted studies on the work of clinical documentation integrity (CDI) specialists, who monitor clinicians' data work to improve documentation. Their studies highlighted how CDI specialists translate clinicians' work to maintain the quality of coded data (Pine and Bossen, 2020); how CDI programmes coordinate the efforts of health organisations to maintain the quality of coded data for comparison, benchmarking, and quality reports (Pine et al., 2023); and the use of human–AI collaboration to facilitate coding (Bossen and Pine, 2023).

Data work draws on an increasingly complex network of distributed actors, both human and non-human. To map the actors of the data journey, Bossen et al. (2019) considered the different 'orders' present in reused data. For example, data that were produced initially to monitor patient treatment protocols and the healing process can have a second-order purpose in the billing process and then be reused to populate a database on cancer. All these data usages create interdependencies among the actors, entities, and artefacts that progressively create a data infrastructure with a variety of socio-technical issues. To our knowledge, few studies have focused on the maintenance work carried out on the ground (Bossen et al., 2019) by data workers in second- or third-order organisations.

Data work as a practice sustained by a specific regime of knowing

The data of CRs are set in the domain of social epidemiology. This domain of knowledge refers to "specific categories and rules, validation procedures, checks, methods, etc., as well as work profiles and experts" (Alaimo and Kallinikos, 2022: 25). These different elements compose what Pachidi and colleagues (2021) call a regime of knowing, which includes specific knowing practices of data workers, schemes of values, and authority arrangements.

First, *knowing practices* are the actions and methodologies employed by actors to develop and utilise knowledge within a specific domain (Alaimo and Kallinikos, 2022). For data workers, these are the activities through which they acquire the competencies and knowledge essential for performing their tasks. The knowing practices of data workers are constituted through training, interactions with peers, and accumulated experience, but also by the tools and methods that actors use to work on data. Second, these practices draw on certain *schemes of values* that determine which information matters and through which methods it should be acquired. Data processing and movement are value-laden (Fiore-Gartland and Neff, 2015; Leonelli and Tempini, 2020). In a hospital, for example, data workers draw on values such as the financial sustainability of the hospital to justify or legitimise their actions in improving clinical documentation or maximising reimbursements (Pine and Bossen, 2020). As we show in our empirical data, CR data workers draw on specific values such as the representativity of the database, especially in relation to minority groups that are often overlooked in the system. Third, *authority arrangements* are "the sanctioned ways to organise, affording power to actors whose expertise is highly valued, to impact how they and others engage in their work" (Bourgoin et al., 2020, cited by Pachidi et al., 2021: 21). Analysis of authority arrangements facilitates understanding of the distribution of power and resources among the constituents of the data ecosystem, including the question of who has access to data and who has the authority to determine how data should be transferred, decontextualised, and/or recontextualised. For Pachidi and colleagues (2021), authority arrangements are intimately associated with the value scheme employed to evaluate who has the requisite skills and expertise to undertake a task. The *regime of knowing* thus offers a means of valuation to highlight "the deeper challenges arising from the emergence of algorithmic technologies, related not only with how we know, but also with which ways of knowing are more valuable and who determines that" (Pachidi et al., 2021: 39). Moreover, a regime of knowing becomes particularly visible during major transformations, such as technological innovations, that give rise to power struggles among actors seeking to protect or transform elements of the regime of knowing.

As our ethnographic work will demonstrate in the case of CRs, the regime of knowing sustains the data journey from hospitals, laboratories, and insurance systems to the registries, enabling the production of meaningful social-epidemiological knowledge. Moreover, we show that it is not only the data that require repair and maintenance; when disrupted by the creation of the HDH, the regime of knowing itself is the primary object of these efforts. By disregarding domain-specific logics, the new governance model exacerbates the challenges that CRs face in fulfilling their social-epidemiological mission.

Data work as a practice of repair and maintenance

We adopt the standpoint of authors such as Denis and Pontille (2015: 8), who see "maintenance and repair as deeply inscribed in a logic of care that starts from decay and vulnerability instead of denying them (Tronto, 1993)". In studies of maintenance and repair, it is a matter of considering the order of things, part of the social order, as the ever-vulnerable result of an endless process of correction and repair (Denis et al., 2015; Hoeppe, 2020). Jackson (2014: 221) presents this as "an exercise in broken world thinking" that takes for granted the normality of erosion, breakdown, and decay, rather than of novelty, growth, and progress. Jackson (2014: 221) advocates "an appreciation of the real limits and fragility of the worlds we inhabit—natural, social, and technological". This approach shifts the focus from innovation, often described as a heroic moment that leads to success (Denis and Pontille, 2022), to how innovation is often conceived during repair and maintenance and cannot be consolidated or sustained without them.

In light of these considerations, it is appropriate to consider repair as a process "of accompanying things over time and ensuring that they persist beyond, below, the ruptures" (Denis and Pontille, 2020: 3, our translation). In this respect, things are 'not put back in order' but rather undergo a transformation to a state of 'working order' (Henke and Sims, 2020). As Jackson (2014: 223) states, "the world is always breaking; it's in its nature to break. That breaking is generative and productive […] always being recuperated and reconstituted through repair". Henke and Sims (2020: 4) elucidate the point: "Repair work is not always about directly fixing. [It is also] associated with broader discussions and arguments about what needs to be repaired, how it should be repaired, and even whether it is actually broken in the first place". The concept of repair extends beyond mere technical fixes to encompass the intricate dynamics of infrastructures, organisational systems, and interpersonal relationships, as articulated by Henke (2019).

Repair concerns the "subtle acts of care by which order and meaning in complex socio-technical systems are maintained and transformed, human value is preserved and extended, and the complicated work of fitting to the varied circumstances of organisations, systems, and lives is accomplished" (Jackson, 2014: 222). According to Denis and Pontille (2022), this notion returns to the foreground matters that seemed to be taken for granted, as well as the fragility of the basis on which they rest. Nevertheless, the notion of repair tends to assume that it is only a question of putting things back in order, sometimes neglecting how things are transformed through repair. Denis and Pontille regard repair as reducible to a single moment, namely the saving act. However, maintenance activities tend to focus not on an event but rather on small gestures that are fully part of existence and even vital to the stability of the relationships that humans have with most objects (Denis and Pontille, 2022: 48). Accordingly, these authors propose a distinction between repair and maintenance on the grounds that "they do not refer to exactly the same problems" (Denis and Pontille, 2022: 37, our translation). They thus argue for taking repair into account only as one of the many elements that punctuate maintenance, in that it makes things last. Similarly, Henke and Sims (2020) present a perspective on maintenance work as one end of the continuum of repair work.

In this debate, we take a position that distinguishes between repair and maintenance, although we perceive these activities to be situated on a continuum (Reiss-Sorokin, 2023). Repair is associated with breakdowns and accidents (Denis and Pontille, 2022). It signifies a unity of action and time, often involving "heroic efforts" (Henke and Sims, 2020) and a change of state (Denis and Pontille, 2022), and it is a reaction to an external event (Reiss-Sorokin, 2023). Conversely, maintenance entails prevention, anticipation, planning, and scheduling of actions (Reiss-Sorokin, 2023), regarded as business as usual (Denis and Pontille, 2022), hidden and mundane (Henke and Sims, 2020). This distinction emphasises that data work occasionally entails repair or maintenance: repair when data are absent or when a disruption in the infrastructure necessitates workarounds or improvisation (Schubert, 2019); and maintenance to prevent breakdowns, ensure data quality, and so on.

As we will demonstrate in the following case studies, the identification and prevention of such breakdowns are of equal importance to the repair of data itself. The integration of these two activities into a unified framework facilitates the demonstration of their interconnection and interaction in, for instance, the way maintenance activities can facilitate the identification of failures that require repair, or the manner in which repairs can be executed to enable access to data required for maintenance purposes.

Few studies have investigated data work as maintenance and repair of data. Through their research, Tanweer and colleagues (2016) developed a framework for understanding the breakdown/repair process of broken data. They argued that breakdowns and repairs can be understood as part of a larger process of data assemblage. Data assemblages are collections of data, tools, and practices that are used to produce knowledge. The authors contended that breakdowns and repairs are essential parts of the data assemblage process, as they allow the identification and correction of errors, the improvement of data quality, and the generation of new knowledge. Pink and colleagues (2018) presented the concept of broken data, arguing that data are not always clean and orderly collections of facts, but can be messy, incomplete, or broken.

In studies relating to healthcare data work, Schwennesen (2019) investigated patients' and professionals' efforts to repair broken data in order to make algorithms work. Bossen and Bertelsen (2023) reported that maintenance, cooperation, data quality assurance, and analysis are the most prevalent tasks of data workers. In their inquiry on the repair and correction of data relating to Covid-19, Boisson and Denis (2024) highlighted the work carried out by the 'lower-up' services that is often invisible and far from the heroic figures portrayed in the media.

To sum up, we consider that focusing on the repair and maintenance of the regime of knowing that sustains data work is particularly useful for understanding what people who take care of data through repair and maintenance are attached to and try to sustain in their everyday practices when confronted with major transformations. This approach gives a socio-technical thickness to data and data work by moving away from a standpoint where data are essentialised and their fabrication unquestionable, as if merely assembling and utilising them were enough (Marent and Henwood, 2021). It emphasises the vulnerability and fragility of data, the contingencies that condition their existence, and any factor that can affect their assemblage and processing into meaningful information. Thus, it provides a framework that deepens our understanding of the concrete consequences of national AI strategies on public health systems.

An ethnography of the data work of two French cancer registries

This research draws on ethnographic work carried out in organisations responsible for producing registers in two of the 101 departments (administrative divisions) of France, referred to here as H and T. CRs are typically non-profit organisations headed by a social epidemiologist who oversees the work of investigators responsible for collecting data from various sources and coders who code it in the registry databases. The registries fulfil a dual mission: monitoring cancer incidence and mortality rates within specific geographical areas, and conducting studies and research based on their data to evaluate care trajectories, prevention campaigns, and the influence of social inequalities on incidence and survival rates. In France, registries usually assess cancer incidence at the department level, and the departmental registries are either specialised for a type of organ (e.g., digestive system, thyroid) or general (i.e., like registre-cancers-44-85.fr, they record all cases of cancer in individuals residing in a department, irrespective of age or organ affected). Both the registries we investigated are general registries. The present research focuses on the data journey where data originating from several sources are repurposed in order to achieve the registries' double mission. Hence, we do not cover the processes of interpretation and knowledge creation that follow the creation of a registry.

Our ethnographic work started with researchers specialised in the use of epidemiological data. We conducted three interviews of 3 hours each to better understand the role of data in the domain of public health expertise. The interviews also contributed to our understanding of the current assemblage of organisations producing data in the French public health ecosystem. The epidemiologists who participated had been the first to raise concerns about the creation of the HDH. This exploratory work led us to focus on CRs. Within the two registries investigated, we conducted semi-structured interviews with epidemiological physicians (four interviews of around 2–3 hours each) and supervisors of the regional network of oncology (three interviews of 2–3 hours each). We shadowed investigators and coders in charge of data production and processing in the two registries (11 hours in total) and observed two meetings where the maintenance of the registry was debated (3.5 hours in total). We also collected a set of documents including the manuals used to support coding work and the record sheets used to encode patient data. This fieldwork took place over a period of 3 years from 2019 to 2021.

We analysed the data iteratively, drawing on an inductive approach in a double movement of zooming in and zooming out (Nicolini, 2009). Zooming out shed light on the complex ecosystem in which CRs are inscribed and the political, economic, and social elements at stake in the health data fabric (Martin-Scholz et al. , 2021): What are the interdependencies and power struggles among organisations in the data journey? What are the potential sources of breakdown? Zooming in, we paid particular attention to the activities (Delcambre, 2009) of investigators at the core of the registries and their data work: What are the different tasks needed to find, gather, and assemble the data? What are the breakdowns and in what circumstances do they occur? What are the maintenance and repair activities that follow and how do they relate to tensions and power struggles in the ecosystem?

This analytical approach helped us to reconstitute the data journey from the sites of production to the CR. As Bates and colleagues emphasised, the notion of journey symbolises "the disjointed breaks, pauses, start points, end points and frictions" that the process introduces (Bates et al., 2016: 4). Following Edwards (2013, cited by Bonde et. al 2019: 559), we paid particular attention to data frictions denoting "the costs in time, energy, and attention required simply to collect, check, store, move, receive, and access data".

Through our analysis, four cases of breakdown/repair/maintenance emerged, of which we chose to explore three. (The fourth case, around a treatment protocol, was set aside because of a lack of empirical data.) The case studies highlight that sources of breakdown are not limited to technicalities and that breakdowns occur not only at the level of data but also in the 'pipes' between the registries and the diverse organisations that the data originate from. Presenting our emerging results at conferences and seminars led us to enrich our theoretical framework, reiterate our analysis, and deepen our understanding and interpretation with the concept of the regime of knowing.

The results are presented in two sections. First, we describe the data work undertaken by CRs and unpack the knowing practices, values, and political arrangements that sustain the data journey. Second, we analyse how various events associated with the transformation of health data governance have shaken this regime of knowing and necessitated additional repair work for CRs.

The regime of knowing of French CRs

Figure 1 shows how the data journey is presented in official documents. The process consists of four

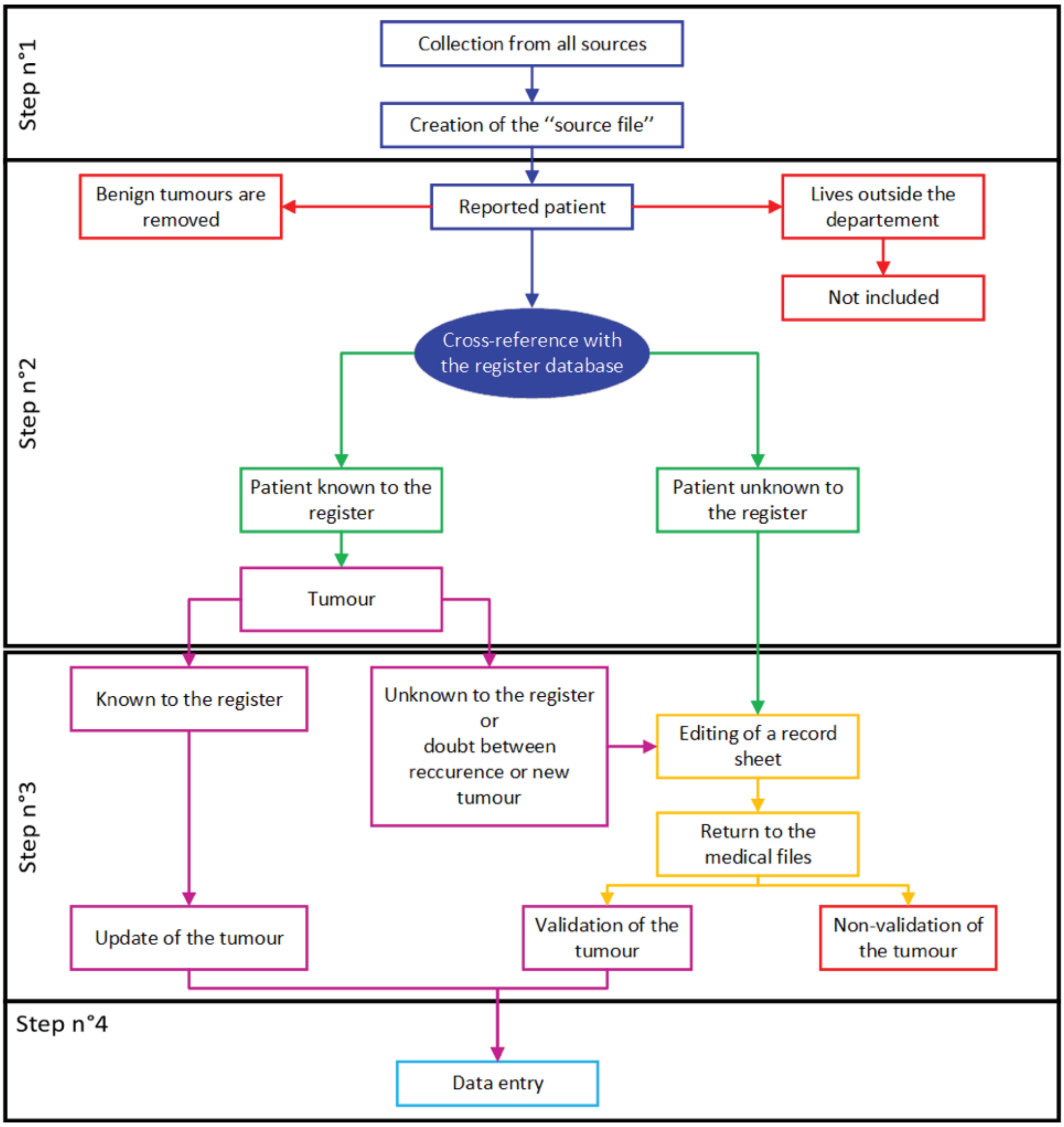


Figure 1. Case validation process (source: file for the evaluation of the H registry).

steps. In Step 1, the CR receives batches of files from source organisations (mostly hospitals and labs). In Step 2, they check the reported patient file. If the tumour is benign or if the person is from another geographical department, the file is removed. In Step 3, the remaining files are cross-checked with the CR database. If the patient is known, the CR has to verify whether the tumour is new or has already been encoded. If the patient is unknown or the tumour is new, they create a record sheet. Finally, in Step 4, data are encoded or updated.

The representation of the data journey in Figure 1 gives a false impression of linearity and automation in the CRs. In fact, our ethnographic work showed that this scheme encapsulated a much more complex mapping of the actors and data that participate in the CR data journey (see Figure 2). We will use this figure in the analysis of the three cases of maintenance and repair work within CRs.

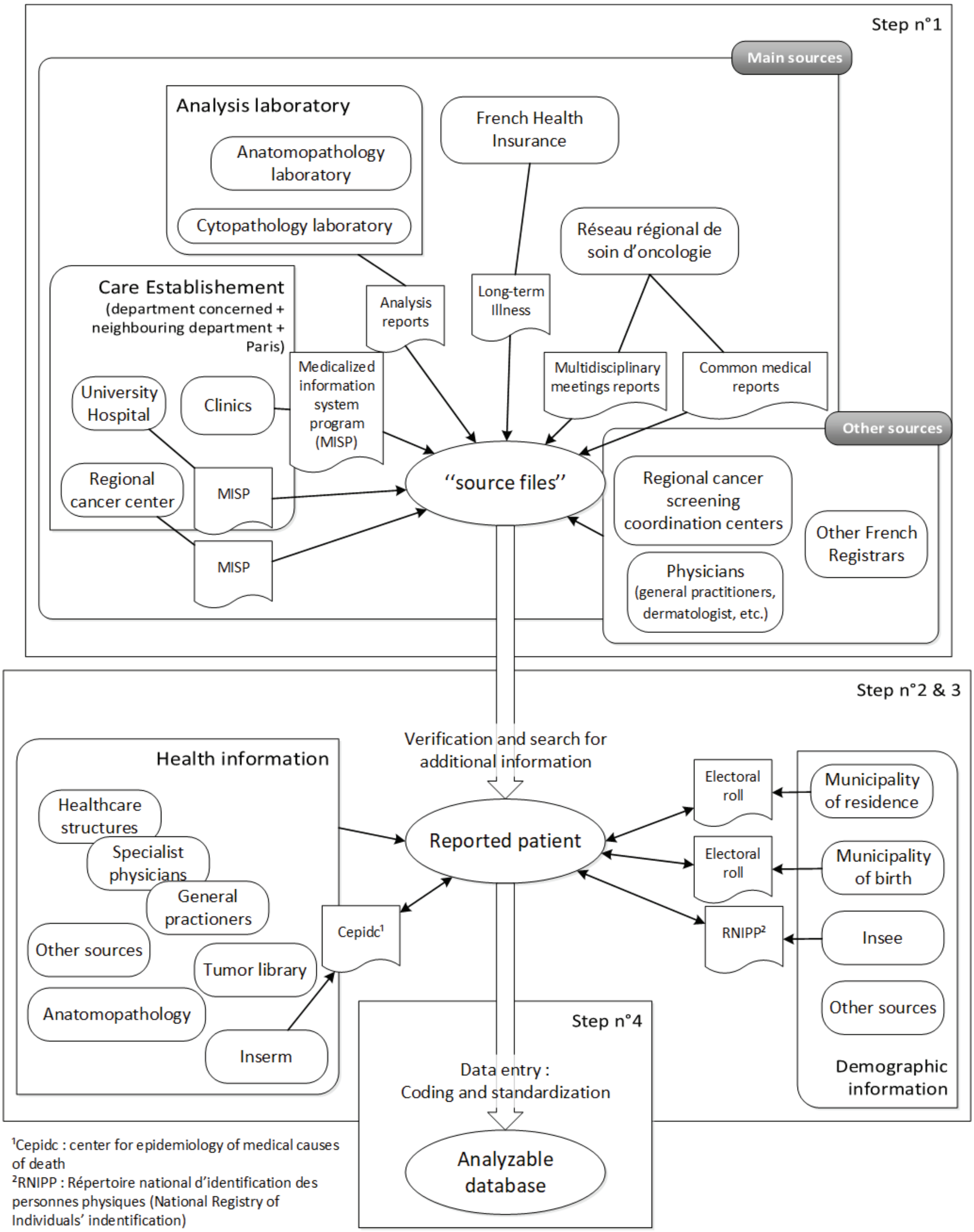
The definition of data quality in the CRs is based on the accuracy of the data collected, their completeness for a sample of recorded cases, and the exhaustiveness of cases recorded for a given geographical area. Although the CRs attempt to register all cases of cancer for a given department, insufficient resources mean they are not able to record all the information about each patient. Hence, they focus their efforts on entering complete information on the stage of cancer development and the treatment protocols for a sample of 10 per cent of cases per year. A central value in public health studies, namely that in incidence analyses for a given type of cancer the sample should be representative of the broader population, implies making considerable efforts to reach marginalised populations for whom data are difficult to access. Consequently, the CRs prioritise not only the completeness of the data but also the completeness and uniqueness of the cases registered. This is achieved by dedicating ample work and attention to repair and maintenance of data.

Case 1: Repairing identity and location data

To ensure quality and consistency, for each new case encoders pay particular attention to the person's identity to avoid duplicate or erroneous encoding in the registry's database. Many countries rely on social security numbers to link health data across systems and to centralise health data (e.g., Denmark; Hoeyer, 2016). However, in France social security numbers cannot be used as identification keys to connect different databases (Lang, 2018). The different traces left by individuals in their various interactions with state-controlled structures or public services are split among databases with different identifiers and no reliable common denominator. Investigators use the patient's given name, surname at birth, date of birth, and address for identification purposes, but any of these elements can be broken: the date of birth may have been incorrectly recorded; the Insee (Institut national de la statistique et des études économiques) code identifying a city can be wrongly encoded (e.g., when two villages merge, the resulting municipality has a new Insee code); the patient may legally change their name (e.g., after marriage or naturalisation).

Investigators are allowed to cross-check identification data with Insee files and electoral lists (see Figure 2) that contain administrative details such as birth name, birth date, and current address. In some cases, the patient records are incomplete, with the postcode missing or not recorded by the CR investigator, but this information is required for purposes of deduplication and specific matching, such as associating cancer cases with a geolocalised socio-economic index. A breakdown due to the absence of this data affects the value of the CR data, as the cancer cases are then incomplete. The coders repair this by consulting other databases to find the missing information; they undertake the repair work knowing which data may be missing and where and how to find them.

For the CRs, exhaustiveness is different from completeness. Whether a registry is general or specialised, one of the main concerns of its researchers is to ensure the exhaustiveness of the cases recorded for a given area, in this case at the department level. The definition of the geographical area where cancers are registered is very important, both for ensuring that the area remains constant over time and for comparison with the population of that area. CR investigators pay particular attention to the address associated with the tumour at the time of diagnosis because



[1]Cepidc : center for epidemiology of medical causes of death
[2]RNIPP : Répertoire national d'identification des personnes physiques (National Registry of Individuals' indentification)

Figure 2. The complex ecosystem constituting the data journey of French CRs.

it enables socio-environmental impact studies to be carried out: Are patients in the area being exposed to a particular industrial pollutant? Does geographical remoteness influence compliance with a care protocol?

Mobility can also be a source of breakdown. For incidence studies, the CRs always encode the address linked to the diagnosis, and this address is linked to the cancer. Because CRs try to follow the evolution of the disease for each patient, they also need to keep a record of the patient's address. However, if a patient moves to another address, it becomes very difficult to follow them, and CR agents are aware of the biases and imperfections of the data they register.

> [Medical epidemiologist]: But knowing if they move is very complicated. That's why, in fact, […] we act when someone tells us that the address has changed.

This quote highlights a tension between the pursuit of completeness and accuracy in data collection and the limited resources available to investigators. This tension leads to the unfortunate circumstance where incomplete data, deemed less significant, are often completed on an opportunistic basis. This is not perceived as a breach by CR workers, as it does not hinder the functioning of the registers. Investigators are encouraged to exercise particular vigilance and to adopt an opportunistic approach to updating the addresses of previous cases as and when they are discovered fortuitously during the investigation of the year's cases in the patient files.

Case 1 illustrates that the process of repairing patient identity and location data is situated within a specific and constraining regime of knowing. CR agents conduct these repairs on the basis of their understanding of established practices, which enables them to identify what missing data to seek, where to locate them, and how to access them. They are compelled to balance the value system that defines the quality of the CR database against a set of constraints that includes the fragmentation of patient data across multiple sources and their own limited resources.

Case 2: Ensuring completeness and avoiding duplicate records

If a patient has been identified as having a cancer of a certain type in a certain year, their name is likely to reappear some years later if the cancer recurs. In such a case, the investigator must go back to the patient file to make sure that the tumour is a new cancer and not linked to the cancer recorded earlier.

> [Medical epidemiologist]: [When] a new cancer [gets reported] in the defined location, typically, a lady who has had cervical cancer. All went well in '90, then she comes back, we see her, and she's reported to us as carrying breast cancer in 2017. There was no previous history of breast cancer. On the other hand, if she had breast cancer in 1990 and now has a new breast cancer in 2017, so now, we're in the same topography, the same location. Our rules that are enacted at the international level are to look at the histological type, because there are groups. Tumours are classified. We know that they're the same tumours, the same group, so we're going to consider that if the 2017 tumour corresponds to the same histological type as the 1997 tumour, it's a recurrence of the 1990 cancer. On the other hand, if it's a different histological group, we'll consider it a new cancer diagnosis.

In order to avoid the occurrence of duplicates, CR workers check whether a new cancer case is a recurrence or a metastasis: they verify the date of diagnosis, the topography and location of the tumour, and the type of cancer cell. These data are not easily found and require a search of the patient's records and translation of medical information into cancer data that comply with international classification standards. In fact, medical information is produced in hospitals and laboratories to observe patients and their diseases from a perspective tailored to each domain. To align with the objectives of the CRs, the data must undergo processes of decontextualisation and recontextualisation carried out by the investigators who complete the files and by the encoders who read and interpret topographical and histological reports. These processes enable the repurposing of medical data and ensure the completeness and accuracy of the data produced for the CR.

Practically, the CR coders print record sheets (Figure 3) that are dispatched according to the care facilities and cancer topographies (Figure 4). These files are then distributed among the CR investigators, who collect and pre-code certain information by looking at patient files in the various facilities where patients are treated or where certain types of tumours are treated.

> [CR investigator]: For the pancreas, I've summarised everything on one page. And then there's all the details. In fact, it's because of what's behind it that I made myself a summary sheet to make it go faster.
> [Researcher]: You made your own summary sheets.

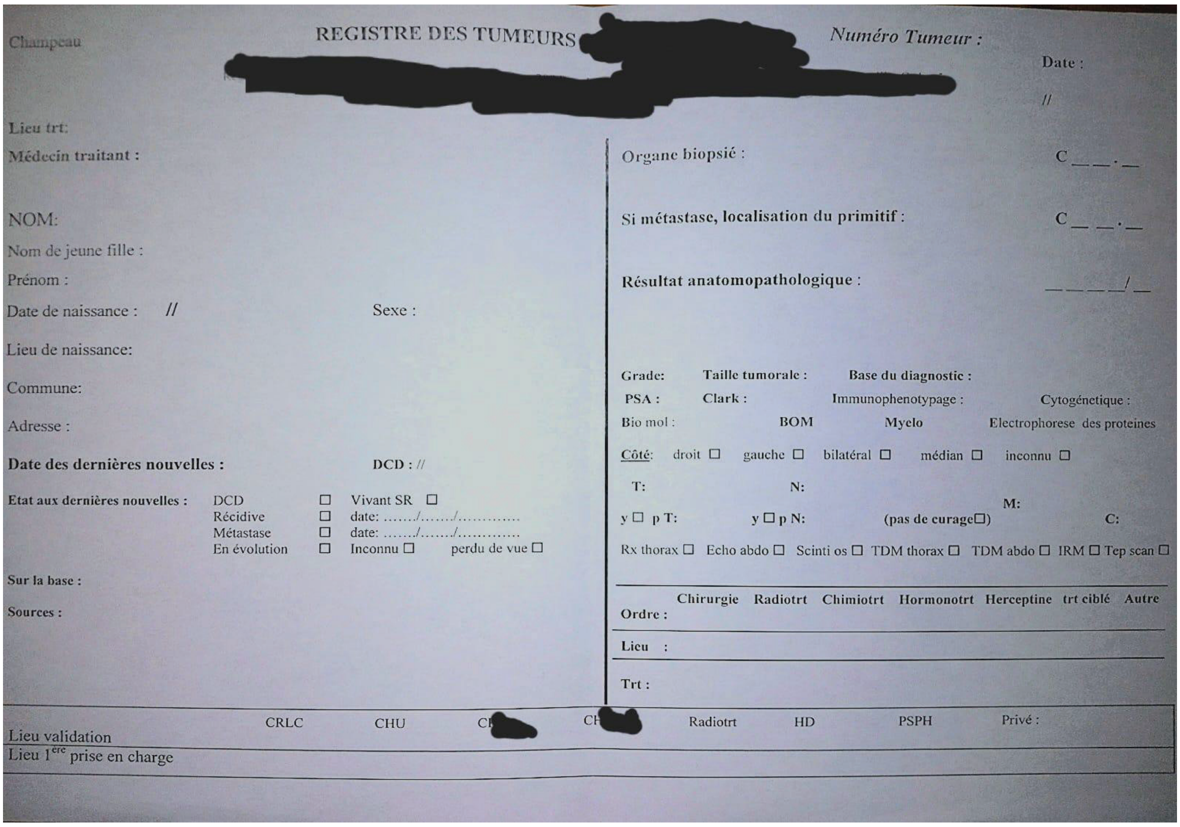
> [CR investigator]: Yes. So, when I go to the hospital, because we have a lot of them, I take these [the coding guide and its summary sheet] and if I'm doing pancreas, I get out my pancreas sheet and I do all the pancreas [cancer cases] at once. And I only need one thing. And if the next time I go for the liver, I take that out.

The investigators who collect data in the field have to know what the coders need. Similarly, the coders have to know what difficulties the investigators face in the field. Additionally, both the investigators and the coders have developed competencies for detecting inconsistencies in order to prevent data entry errors, as this interaction illustrates.

> [CR coder]: The diagnosis date doesn't match.
> [CR investigator]: Multidisciplinary meeting, University Hospital […] it's [other investigator], the University Hospital, I think … And what's wrong?
> [CR coder]: The diagnosis date doesn't correspond to the operation date. I think there's an error in the diagnosis date.
> [CR investigator]: Oh yes, that's okay … It's the 16th. Up there … [the patient] had the direct surgery … There was no surgery, so it's the biopsy.

As we can see in the extract above, the coder's daily work routine involves the maintenance of the CR database, with particular attention paid to the identification of inconsistencies as cancer cases are recorded, and the investigator must know where and how to look in the file to give thickness to the data to resolve discrepancies. In this particular instance, the coder identifies an apparent error and engages in a discussion with the investigator to ascertain whether it constitutes a breakdown. The basis of the discussion is the coder's and the investigator's shared knowledge of the practice.

Completing the record sheet also requires good knowledge of the organisation of hospitals and the ways data are acquired in their information systems, as well as good relationships with data owners. Hence, the CRs draw on a multiplicity of authority arrangements that connect them to hospitals, biopsy laboratories, and medical doctors, among others (see Figure 2), that produce data related to cancer patients and their socio-economic environment. These arrangements occur at different levels, from the national agreement between the registries network (Francim) and other national institutions, to indi-



Champeau

REGISTRE DES TUMEURS [redacted]

Numéro Tumeur :

Date :
//

Lieu trt:
Médecin traitant :

NOM:
Nom de jeune fille :
Prénom :
Date de naissance : // Sexe :
Lieu de naissance:
Commune:
Adresse :

Date des dernières nouvelles : **DCD :** //

Etat aux dernières nouvelles :
DCD ☐ Vivant SR ☐
Récidive ☐ date: ……./……./…………
Métastase ☐ date: ……./……./…………
En évolution ☐ Inconnu ☐ perdu de vue ☐

Sur la base :
Sources :

Organe biopsié : C__ __.__

Si métastase, localisation du primitif : C__ __.__

Résultat anatomopathologique : __ __ __ __/__

Grade: **Taille tumorale :** **Base du diagnostic :**
PSA : **Clark :** Immunophenotypage : Cytogénetique :
Bio mol : BOM Myelo Electrophorese des proteines

Côté: droit ☐ gauche ☐ bilatéral ☐ médian ☐ inconnu ☐

T: **N:** **M:**
y ☐ p **T:** y ☐ p **N:** **(pas de curage☐)** **C:**

Rx thorax ☐ Echo abdo ☐ Scinti os ☐ TDM thorax ☐ TDM abdo ☐ IRM ☐ Tep scan ☐

	Chirurgie	Radiotrt	Chimiotrt	Hormonotrt	Herceptine	trt ciblé	Autre
Ordre :							
Lieu :							
Trt :							

	CRLC	CHU	CH[redacted]	CH[redacted]	Radiotrt	HD	PSPH	Privé :
Lieu validation								
Lieu $1^{\text{ère}}$ prise en charge								

Figure 3. Record sheet for the H registry.

Figure 4. Cabinet containing the files to be investigated and patient files collected.

vidual arrangements designed to bypass the absence of a formal agreement. These arrangements are neither automated nor given but have to be maintained over time.

Case 3: Ascertaining and translating the stage of cancer for comparison purposes

To compare incidence and mortality rates in order to determine whether the number of cancers is increasing or decreasing from one year to the next, researchers must ensure that they are measuring the same things. The definitions on which the registration of a case of cancer is based are therefore aligned with internationally and nationally shared criteria.

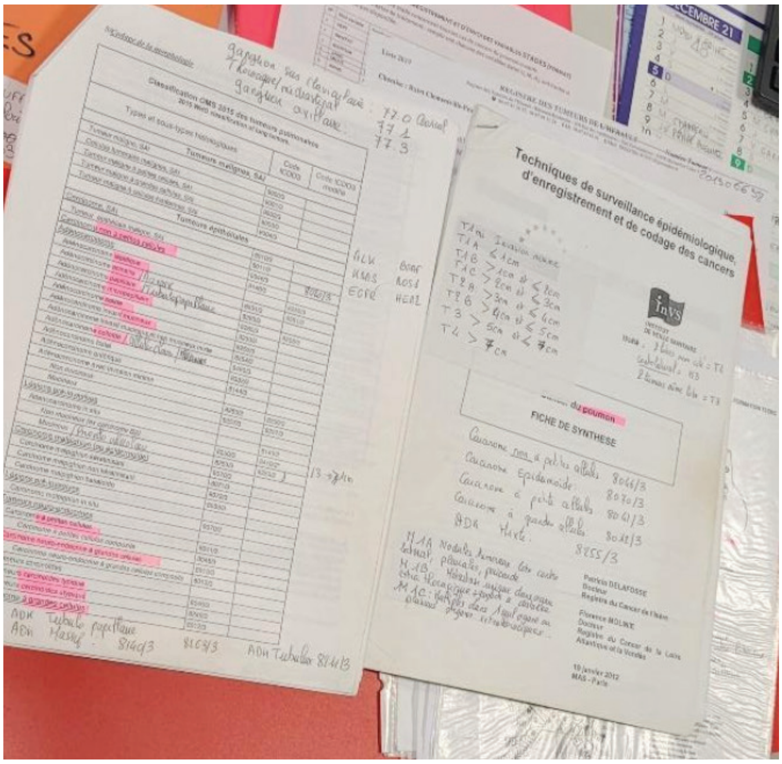
> [Medical epidemiologist]: In fact, the [hospital] considers that it's a cancer for its own management, but we can't take it according to our own criteria, because we follow standard criteria, international standards … We must register a certain type of cancer, we have to register a certain benign tumour of the central nervous system, of the bladder, but we don't register everything.

To allow comparison, nationally or internationally, with other registries, the case registration must ascertain the diagnosis of cancer, the tumour's location, and the stage of the cancer, and code it according to the correct classification. However, hospital codes used by physicians draw on a nomenclature that is different to the oncology nomenclature used by the CRs, and the differences can create messiness in the codification of the cancer's stage.

The activity of coding and translation therefore requires a good knowledge of cancers and their specificities, as well as the ability to find information in patients' files and translate it using international coding standards. The following excerpt illustrates the knowing practices of an investigator who is explaining the use of annotated coding manuals to code the cancer stage using international nomenclature.

> [CR investigator]: With this, I have my coding [guide] for each topography and when I do the lungs, I take out my lungs folder and my coding [guides] [see Figure 5 and Figure 6] because the coding is totally different from one organ to another. So, there I have all my record sheets …
> [Researcher]: You annotate the guide progressively according to …?
> [CR investigator]: As things evolve. I've put some notes on [the front of the coding guide] because it's changed.

In this example, the CR investigator has translated the evolving guidelines into notes and used his know-how to determine what to look for and where to look to translate medical data into cancer stage codes. This illustrates how investigators' knowing practices evolve over time to meet the CRs' standards of accuracy as well as international norms ensuring comparability.



Classification OMS 2015 des tumeurs pulmonaires

ganglion sus claviculaire : 77.0
thoracique/médiastinal 77.1
ganglion axillaire 77.3

ALK BRAF
KRAS ROS1
EGFR HER2

Techniques de surveillance épidémiologique, d'enregistrement et de codage des cancers

T1mi Invasion minime
T1A ≤ 1cm
T1B > 1cm et ≤ 2cm
T1C > 2cm et ≤ 3cm
T2A > 3cm et ≤ 4cm
T2B > 4cm et ≤ 5cm
T3 > 5cm et ≤ 7cm
T4 > 7cm

du poumon
FICHE DE SYNTHESE

Carcinome non à petites cellules 8046/3
Carcinome Epidermoïde 8070/3
Carcinome à petites cellules 8041/3
Carcinome à grandes cellules 8012/3
ADK Mixte 8255/3

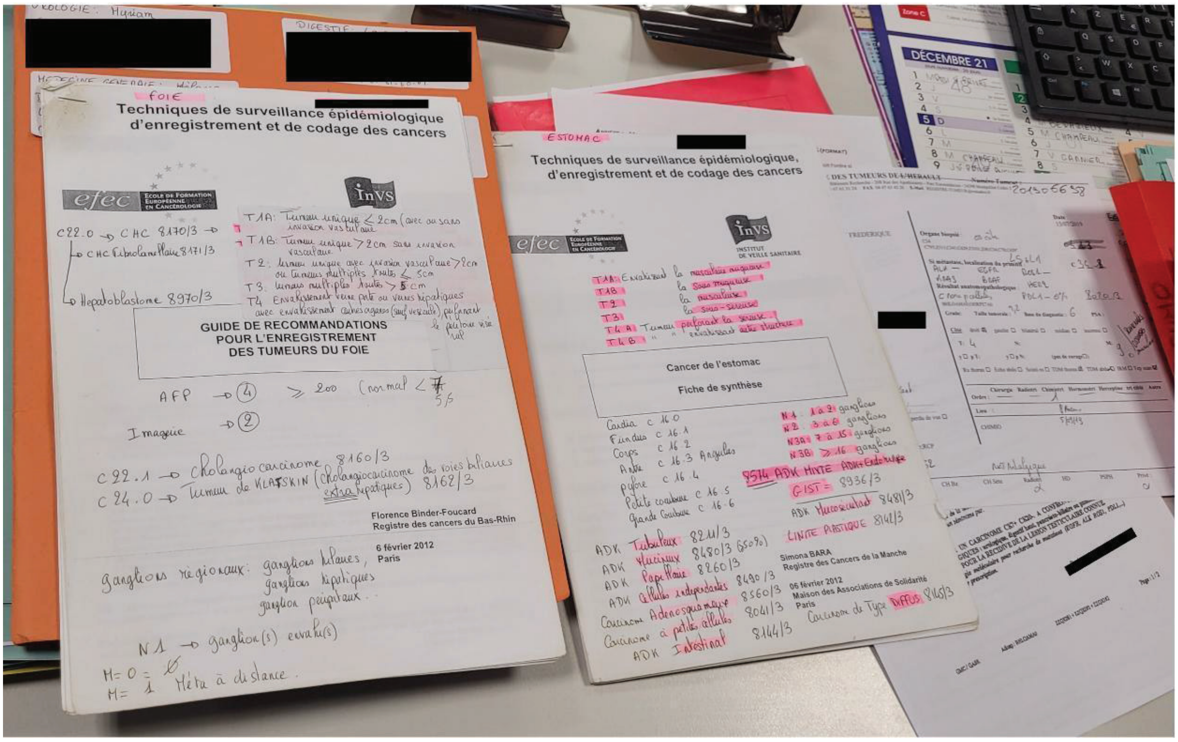
Patricia DELAFOSSE
Docteur
Registre du Cancer de l'Isère

Florence MOLINIÉ
Docteur
Registre du Cancer de la Loire Atlantique et la Vendée

19 janvier 2012
MAS - Paris

Figure 5. Coding guide 1 (lungs).

Overall, this data journey is taken care of by investigators and coders who draw on skills, scientific knowledge, methods, and tools to record all the cases of cancer in a given territory over a given period. The values of exhaustiveness, complete-



UROLOGIE : Myriam
[redacted]
DIGESTIF
[redacted]

FOIE
[redacted]
Techniques de surveillance épidémiologique d'enregistrement et de codage des cancers

efec École de Formation Européenne en Cancérologie
InVS

C22.0 → CHC 8170/3 →
→ CHC Fibrolamellaire 8171/3
→ Hepatoblastome 8970/3

T1A: Tumeur unique ≤ 2cm (avec ou sans invasion vasculaire
T1B: Tumeur unique > 2cm sans invasion vasculaire
T2: tumeur unique avec invasion vasculaire > 2cm ou tumeurs multiples toutes ≤ 5cm
T3: tumeurs multiples toutes > 5 cm
T4: Envahissement veine porte ou veines hépatiques avec envahissement autres organes (sauf vésicule), perforant le péritoine viscéral

GUIDE DE RECOMMANDATIONS
POUR L'ENREGISTREMENT
DES TUMEURS DU FOIE

AFP → ④ ≥ 200 (normal < ~~7~~ 5,5)
Imagerie → ②

C22.1 → Cholangiocarcinome 8160/3
C24.0 → Tumeur de KLATSKIN (Cholangiocarcinome des voies biliaires extra hépatiques) 8162/3

Florence Binder-Foucard
Registre des cancers du Bas-Rhin

6 février 2012
Paris

ganglions régionaux: ganglions hilaires, ganglions hépatiques, ganglion périportaux.

N1 → ganglion(s) envahi(s)
M= 0 = Ø
M= 1 Méta à distance.

ESTOMAC
[redacted]
Techniques de surveillance épidémiologique, d'enregistrement et de codage des cancers

efec École de Formation Européenne en Cancérologie
InVS INSTITUT DE VEILLE SANITAIRE

T1A: Envahissant la musculaire muqueuse
T1B la sous muqueuse
T2 la musculeuse
T3 la sous-séreuse
T4A Tumeur perforant la séreuse
T4B " " envahissant autre structure

Cancer de l'estomac
Fiche de synthèse

Cardia C 16.0
Fundus C 16.1
Corps C 16.2
Antre C 16.3 Angulus
pylore C 16.4
Petite courbure C 16.5
Grande Courbure C 16.6

N1: 1 à 2 ganglions
N2: 3 à 6 ganglions
N3A: 7 à 15 ganglions
N3B: ≥ 16 ganglions

8574 ADK MIXTE ADK+Endocrine
GIST = 8936/3
ADK Mucosécrétant 8481/3
LINITE PLASTIQUE 8142/3

ADK Tubuleux 8211/3
ADK Mucineux 8480/3 (≥50%)
ADK Papillaire 8260/3
ADK Cellules indépendantes 8490/3
Carcinome Adenosquameux 8560/3
Carcinome à petites cellules 8041/3
ADK Intestinal 8144/3
Carcinome de Type Diffus 8145/3

Simona BARA
Registre des Cancers de la Manche

06 février 2012
Maison des Associations de Solidarité
Paris

[redacted]

DÉCEMBRE 21

Figure 6. Coding guides 2 (liver, left) and 3 (stomach, right)

ness, accuracy, and comparability of recorded data drive the work of these investigators in achieving their mission. The CRs have built their authority and legitimacy on their methods of cross-checking and verification using multiple sources of data, which entails multiple agreements with source organisations. Their ability to preserve these agreements is dependent on their legitimacy as high-quality data providers who enable the production of scientific knowledge regarding public health issues, and on their belonging to an internationally recognised network of CRs.

How the transformation of healthcare data governance is shaking the CRs' regime of knowing

Access to sources of information (see the main sources in Figure 2) requires negotiations between the different organisations involved and is maintained over time through an ecosystem of fragile socio-technical relationships that are subject to constant threat and questioning according to socio-economic contingencies. Our empirical data reveal the fragility of the agreements that enable registries to collect and exploit data for public health purposes. Here, we analyse the fragilisation of the CRs' regime of knowing due to the transformation of health data governance that follows from the French NAS.

The turbulence started with the implementation of the General Data Protection Regulation (GDPR) in 2018, which triggered work to maintain the agreements with different sources of notification.

> [Medical epidemiologist]: So, we're negotiating now, like everyone else. Before, it was: "You'll come when the secretary isn't in, and you'll take her code to access the file". GDPR isn't really compatible with that. So, we're in the process of negotiating specific codes for the registry's investigators.

The creation of the HDH in 2019 as a central player in health data governance has made things even more complex for the CRs. By law, the HDH assumes the role of secretary of, the ethical and scientific committee for health research, studies and evaluations (CESREES) and facilitates procedures with the French data protection authority in charge of GDPR compliance (CNIL). The HDH has thus become an "obligatory point of passage" (Callon, 1984) for all data producers, including the CRs, thus complicating the negotiation of bilateral agreements.

Since 2021, the HDH is also responsible for the 'expanded' national health data system (SNDS), which no longer comprises only medico-administrative data but also data from registries, research cohorts, hospital data warehouses, and so on. These data are not centralised in a single file; instead, data producers are invited to share their data in a single 'HDH catalogue'. This expansion has created numerous tensions among data producers. In order to circumvent potential political difficulties with the HDH, the CRs have been transmitting restricted data sets to the HDH catalogue while retaining more precise and complete data for themselves. By this means, the CRs seek to safeguard the value of the data under their control, thereby supposedly sustaining the significance of their work and their authority over the data they produce.

> [Medical epidemiologist]: They [the HDH] had also approached the hospitals … Everyone has been tapping away. We're among those who give the least impression of having tapped on the sidelines, because nobody pointed out that those were the basic data that we already sent elsewhere [to the World Health Organization]. So, we didn't say anything. Santé publique France [the administrative body charged with promoting health at the population level] is well aware of this. The INCa [Institut national du cancer], too … We've kept some variables so we can say, "Oh well, no, if you need one more variable in your study, so it can't be the data of the Health Data Hub, it's ours you can use, but that's all".

However, the legitimacy and authority of the CRs as quality data providers is being contested. Recently, the registries' association lodged a formal objection regarding a misrepresentation of data by researchers of the SNDS, which is controlled by the HDH, in an article in a scientific journal concerning the quality of histological confirmation data:

> [Medical epidemiologist]: We [the registries] always give the histological confirmation rate of

> our patients. All the registries do this. The other European data were registry data, so the others had given their histological confirmation rates. The [author from the SNDS for the French data set] didn't bother, he just looked at it and said, "Ah 89, that won't hurt". He put in 89% [as the histological confirmation rate of our patients for the French data set]. This information does not exist in the SNDS.

The CR representatives were able to voice their concerns through a letter published in the journal. The purpose of this action was to defend the legitimacy and the quality of the data they produce, as explained in Case 2. However, they expressed scepticism about the letter's potential impact on the reputation and dissemination of the scientific article.

Repeated transformations in health data governance have thus generated additional maintenance and repair work that extends beyond the data itself. Increasingly, this work takes on a political-economic dimension, as the CRs seek to preserve access agreements with data sources and defend their legitimacy by invoking the domain-specific, high-quality nature of the data they produce. Nevertheless, they acknowledge that their capacity to safeguard the regime of knowing is constrained by their relatively weak position within the broader network of data governance actors. As we explain in the following example, the gradual transformation of the values that govern the production of health data has only accentuated the fragility of the registers.

Medical biology laboratories, despite their private status, are able to survive financially thanks to the financing provided by the health insurance industry, which largely reimburses the costs of biological examinations. The fees for each type of examination are determined by the health insurance industry and the Ministry of Health, enabling the latter to exert control over biological spending (Bienvault, 2019). In 2020, health insurance companies imposed savings of €170 million on spending in medical biology. Consequently, laboratories started a soft strike that resulted in a breakdown of access to their data by other institutions, including the registries. Following this strike, laboratories started to reorganise into larger groups, and this entailed a renegotiation of access to data.

> [Medical epidemiologist]: The problem is: our relationships … every time, we have to establish relationships with new structures. And when they join forces, we're faced with a new structure. For example, two labs in the T region are in the process of teaming up with the bad guys, so we don't really know how to proceed, because these people have always refused to transfer data to the registry. Whereas the two labs did so without any problem. So, we've been trying to get in touch with them for 6 months now.

To repair this breach caused by the reorganisation of labs (see Figure 2, Anatomopathological labs), the investigators have developed a tactical workaround: retrieving the medical biology reports by searching in the appendices of the patient files that are accessed in each hospital.

> [CR investigator]: I collect the anapaths from the [hospital centre] … I take all the binders; they're big binders and I take, I read, I look at the ADICAP[1]code, but I mostly read the minutes and when it's clear, I scan. I have a small computer with a portable scanner. And it gives results. Then, here, we have all the reports, just like that. Because they don't know how to make queries to give me all the names that match.

The ADICAP code does not directly indicate the stage of the cancer, but it is very important in determining the stage. To establish the stage, the CR investigator has to look closely at the anapath reports. This task may involve some discussion, especially with the medical epidemiologist, and the investigator must anticipate this and collect all the results to support it. As the access to anapath results has been compromised, the CR investigator must find another way to obtain this access. As it was not the hospital's job to provide access to anapath reports, there are no query tools to make it easier to find and collect them, making this repair work a tedious and time-consuming task.

The strike revealed that the data intensification resourcing movement (Hoeyer, 2019) has precipitated a shift in the prevailing value scheme, whereby the economic value of data as a bargaining chip prevailed over its value for therapeutic or scientific purposes. This illustration elucidates how disparities in data valences (Fiore-Gartland and Neff, 2015) engender concrete and unpleasant outcomes for the data work undertaken by CRs.

In this section, we have highlighted the additional repair work by the CRs on the prevailing regime of knowing that the transformation in the governance of healthcare data has triggered. We illustrate how this repair work concerns not only the data but also the political arrangements guaranteeing access to the data, the legitimacy of investigators as knowledgeable data workers, and the values of the data produced primarily for scientific and public health purposes.

Discussion

Our contribution to repair studies

Drawing on the regime of knowing, our empirical study extends repair studies by showing how care, expertise and power relations are intertwined within repair work. Although existing studies emphasise how each of these elements - care (Denis and Pontille, 2018; Jackson, 2014), expertise (Schubert, 2019; Henke, 2019), and power relations (Henke and Sims, 2020) - matter to understand repair, these studies do not question how their articulation advances our understanding of maintenance and repair. While Sims and Henke (2012) analyse repair articulating the material, institutional, and discursive as levers to negotiate infrastructure order, the concept of a regime of knowing provides us with an analytical angle that foregrounds what CRs care about and what it takes as technical, organisational skills and political arrangements to maintain or repair it.

The three cases presented in the first part of the analysis exhibit how the elements composing the regime of knowing sustain CRs data maintenance. Through their data work, CR members promote a specific conception of public healthcare and *the value scheme* (the representativity of the database, especially in relation to minority groups that are often forgotten in the system) that supports it. They are committed to their dual mission of providing high-quality data to assess current public health policies and to produce scientific research on prevention through testing, disparities in care trajectories, and health inequalities more generally. Thus, their practices draw heavily on the logic of care proposed by Jackson (2014: 232): “Care reconnects the necessary work of maintenance with the forms of attachment that so often (but invisibly, at least to analysts) sustain it. We care because we care”. In order to achieve this mission, investigators have developed a *know-how*, a ‘professional seeing’ (Goodwin, 1995) that is highly specialised in the cancer research domain for collecting appropriate data and coding it according to international scientific standards. This work also requires the development of a socio-technical perspective (Neves et al., 2024) to access sources of notification in the field, to understand the practices of data work at the primary sources, and to detect potential breakages. Finally, it entails *agreements* with various sources of notifications and secondary sources for the validation of socio-demographic data. However, CR representatives do not perceive these agreements as unified and stable. Instead, they regard them as an “unsettled [and fragile] assemblage of partly unknown elements, constantly subject to external and internal disruptions” (Denis and Pontille, 2022: 288). This fragility gives rise to a further type of maintenance work of a political nature (Henke, 2019) that is required to preserve access to the various sources of data needed to constitute a registry’s database while meeting its quality standards.

In the subsequent analysis, we highlight how the implementation of the French National AI Strategy (NAS) —via policy and resource allocation — has reshaped healthcare data governance, thereby destabilising the CRs’ established regime of knowing. This transformation produces new uncertainties and attendant loss of control, leading to increased workloads and required improvisation (Schubert, 2019). Our findings contrast with the work of Jackson (2014) and Henke and Sims (2020) which tend to emphasise the transformative nature of repair. Instead, it reveals that CR experts resist these shifts and seek to maintain their increasingly fragile regime of knowing, aiming “to stabilize existing social and material relations with respect to a given infrastructure” (Henke, 2019: 272). In line with Sims and Henke (2012: 328), who argue that credibility is “a cultural and institutional process where everything from documents, methodologies, and scientific reputations come into play,” we contend that

CR legitimacy is anchored in a regime of knowing. Thus, when registries repair their knowledge regime, they are simultaneously attempting to repair their legitimacy, existence, identity, and boundaries (Sims and Henke, 2012: 324).

Since the beginning of the implementation of the French NAS, CRs' regime of knowing has been challenged from all sides. In preparation for the GDPR, a number of bilateral contacts were initiated to prevent any disruption in the logic of maintenance. However, the establishment of the HDH as an 'obligatory passage point' (Callon, 1984) has substantially disrupted the pre-existing, often interpersonal ad hoc *arrangements* that existed between CRs and other institutions. It has led to unexpected political work involving the creation of a new authority agreement with the HDH. Arrangements to access data are also being reconsidered by the changing perception of data as a source of wealth. The French NAS-supported data intensification resourcing, which posits that data-driven innovation can catalyse economic growth, has precipitated a paradigm shift, underscoring the economic value of data in contemporary value schemes. This shift was revealed by the lab strike, which led to a renegotiation of the data-sharing agreements between private labs and other organisations. The resulting frictions (Bates et al., 2016) constrain the movement of data to and from CRs. When the arrangement with private labs breaks down, it is the access to data that is broken from a political point of view. This disruption in the flow of data necessitated a repair process on the part of the CR investigators, who devised a bypass for accessing data from hospital files. Such repairs rely on investigators' capabilities to locate the required data and their legitimacy in the eyes of nurses and hospitals facilitating access to the relevant files. However, their legitimacy is being eroded as other institutional players, such as the SNDS (controlled by the HDH), doubt their uniqueness in producing high-quality data. This leads them to question the need for CRs to provide data on cancer, given that the SNDS claims that its database already contains this information. CRs therefore face a dilemma: by withholding the full dataset through HDH, they seek to protect their assets and bargaining power; however, doing so limits their visibility and opportunities to demonstrate the distinctiveness of their data compared to other datasets.

Hence, restoring their regime of knowing to a "normal order" in the sense of Ureta (2014) entails an extra burden that seems useless, since the transformation of the French healthcare data infrastructure appears irreversible. Therefore, should the CRs recognise that their regime of knowing is irreparable and move beyond repair (Ureta, 2014 : 388)? In other words, should the CRs take this failure as an opportunity to move from a logic of maintenance to a more transformative one (Henke, 2019)? Future research is needed to determine whether CRs can adapt from one mode to another—thereby preserving their legitimacy in cases where exclusive data access is lost or where material and expertise are fundamentally altered by digitisation.

Our contribution to the field of healthcare data work studies

The present study responds to the call made by Bertelsen et al. (2024) for new theories to be introduced to the field of healthcare data work studies. The introduction of a new framework enables a thicker understanding of data work in healthcare, which cannot be reduced to production/collection/interpretation, and makes it possible to elaborate a description of the knowledge, values, and arrangements required to prevent data in healthcare from being broken, through maintenance or repair.

The study also contributes to the existing body of knowledge by offering an ethnography of data work in the French healthcare system. Research in this field is dominated by studies in countries considered leaders in the digitalisation of health, such as the Netherlands, Denmark, and Finland. The French ecosystem of healthcare data is fragmented and subject to numerous power struggles, which makes the sharing and centralisation of data difficult to achieve. The use of paper files in the CRs to retrieve data from hospitals exemplifies this fragmentation and its repercussions for CR data work. This may suggest that France is less advanced in the digitalisation of healthcare, or it may be perceived as a manifestation of democratic vitality, where the discourse surrounding data sharing and the inherent value of AI (Hoff,

2023) remains a subject of debate. Future research should address this debate.

Finally, we shed light on the activities on the ground (Bossen et al., 2019) of data workers in second- or third-order organisations that are less often investigated. In particular, we demonstrate the maintenance and repair efforts necessary to recontextualise data for cancer research purposes. Recontextualisation, as defined by Leonelli (2016: 194), is the process of rendering data comprehensible to researchers who are unfamiliar with them, making it possible to assess the evidential value of the data and utilise them for specific research endeavours. Haase et al. (2023: 526) delineated the evidential value of data as the extent to which they can be employed to substantiate a specific claim. It is clear that recontextualisation is an inherent aspect of the process by which data attain evidential value. Our findings are thus consistent with the conclusion of Torenholt and Tjørnhøj-Thomsen (2022) that recontextualisation necessitates professional competencies and experience, as well as collaboration with colleagues, to substantiate specific interpretations of data. Notably, however, our study builds on previous research by emphasising that recontextualisation constitutes a pivotal aspect of the data work conducted by CR investigators. Consequently, it cannot be disregarded, as is the case for general practitioners concerning data from patients' wearable devices (Haase et al., 2023). Instead, recontextualisation is a key element in the effort to repair and maintain data so as to secure their evidential value.

Conclusion

Throughout this ethnographic investigation, we have demonstrated that the work of CRs cannot be reduced to mere technological data management but must be understood as the ongoing repair and maintenance of a regime of knowing. We show that the maintenance of this regime entails painstaking work that includes political and institutional contingencies. All these efforts are made to ensure data quality and to preserve the legitimacy of CRs in the healthcare public health system. It is evident that the French ecosystem of healthcare data remains fragmented, and this context is likely to be a contributing factor to the additional repair work that we identified. Further research on registries in different countries, in less fragmented contexts, can offer a more comprehensive perspective on the findings.

Our study also provides a critical reflection on AI policies that frequently construe data production as content-agnostic (Alaimo and Kallinikos, 2022), thereby disregarding the labor that domain-specific knowledge production entails. As public health data governance evolves rapidly in the name of AI innovation, it is imperative to render visible the often-overlooked work of maintenance and repair that underpins data quality as defined within fields such as cancer research and social epidemiology. Health is not the only sector experiencing tensions between AI-driven innovation and the production of domain-specific knowledge. Future research might investigate these tensions in other contexts where human labor is central to the construction of data—for instance, in open-source knowledge repositories such as Wikipedia or Stack Exchange.

Acknowledgements

We would like to thank the members of the Cancer Registries who agreed to participate in this research. Thanks to the reviewers and guest editors of this special issue for their very constructive comments, which substantially helped improve the paper. We dedicate this article to Professor Anne Mayère (1957–2023), who initiated this project and collaboration.

References

Alaimo C and Kallinikos J (2022) Organizations Decentered: Data Objects, Technology and Knowledge. *Organization Science* 33(1): 19–37.

Balka E and Leigh Star S (2016) Mapping the body across diverse information systems: Shadow Bodies and How They Make Us Human. In: Bowker G C, Timmermans S, Clarke A E and Balka E (eds) *Boundary Objects and Beyond: Working with Leigh Star*. Cambridge: MIT Press, pp. 417–434.

Bates J, Lin Y-W and Goodale P (2016) Data journeys: Capturing the socio-material constitution of data objects and flows. *Big Data & Society* 3(2): 2053951716654502.

Bareis J and Katzenbach C (2022) Talking AI into Being: The Narratives and Imaginaries of National AI Strategies and Their Performative Politics. *Science, Technology, & Human Values 47*(5): 855-881. https://doi.org/10.1177/01622439211030007.

Bertelsen P S, Bossen C, Knudsen C and Pedersen A M (2024) Data work and practices in healthcare: A scoping review. *International Journal of Medical Informatics 184*: 105348. https://doi.org/10.1016/j.ijmedinf.2024.105348.

Bienvault P (2019) En grève, les laboratoires de biologie protestent contre la baisse de leurs tarifs. *La Croix*, 23 September. Available at: https://www.la-croix.com/Sciences-et-ethique/Sciences/En-greve-laboratoires-biologie-protestent-contre-baisse-tarifs-2019-09-23-1201049445 (accessed 3 November 2024).

Boisson M and Denis J (2024) Faire remonter les chiffres du Covid. *Revue d'anthropologie des connaissances 18*(4): 4.

Bonde M, Bossen C and Danholt P (2019) Data-work and friction: Investigating the practices of repurposing healthcare data. *Health Informatics Journal 25*(3): 558-566. https://doi.org/10.1177/1460458219856462.

Bossen C and Bertelsen PS (2023) Digital health care and data work: Who are the data professionals? *Health Information Management Journal*: 18333583231183083.

Bossen C and Pine KH (2023) Batman and Robin in Healthcare Knowledge Work: Human-AI Collaboration by Clinical Documentation Integrity Specialists. *ACM Transactions on Computer-Human Interaction* 30(2): 26:1-29. https://doi.org/10.1145/3569892.

Bossen C, Pine KH, Cabitza F et al. (2019) Data work in healthcare: An Introduction. *Health Informatics Journal* 25(3): 465–474.

Bourgoin A, Bencherki N and Faraj S (2020) "And Who Are You?" : A Performative Perspective on Authority in Organizations. *Academy of Management Journal*, *63*(4), 11341165. https://doi.org/10.5465/amj.2017.1335.

Callon M (1984) Some Elements of a Sociology of Translation: Domestication of the Scallops and the Fishermen of St Brieuc Bay. *The Sociological Review* 32(1_suppl): 196-233. https://doi.org/10.1111/j.1467-954X.1984.tb00113.x.

Combes S and Maquart B (2021) *Charte responsable de données : Les engagements de la plateforme des données vis-à-vis de ses partenaires responsables des données*. Available at: https://health-data-hub.fr/sites/default/files/2021-04/Charte%20Responsable%20de%20Donn%C3%A9es-compressed.pdf (accessed 3 November 2024).

Choroszewicz M (2022) Emotional labour in the collaborative data practices of repurposing healthcare data and building data technologies. *Big Data & Society* 9(1): 20539517221098413.

Delcambre P (2009) L'activité d'évaluation et les systèmes d'information. L'évaluation est aussi un travail : langagier, assisté, organisé. Études de communication. langages, information, médiations (33): 79–100.

Denis J and Pontille D (2015) Material Ordering and the Care of Things. *Science, Technology, & Human Values* 40(3): 338–367.

Denis J and Pontille D (2020) Maintenance epistemology and public order: Removing graffiti in Paris. *Social Studies of Science* 51(2): 233–258.

Denis J and Pontille D (2022) *Le soin des choses: Politiques de la maintenance*. Paris: La Découverte.

Fianu A, Favier F, Naty N, et al. (2022) Analyse des déterminants sociaux pour concevoir une intervention de santé publique locale. *Revue d'Épidémiologie et de Santé Publique* 70: S166.

Fiore-Gartland B and Neff G (2015) Communication, mediation, and the expectations of data: data valences across health and wellness communities. *International Journal of Communication* 9: 1466-1484.

Goodwin C (1995) Seeing in Depth. *Social Studies of Science* 25(2): 237–274.

Haase C B, Ajjawi R, Bearman M, Brodersen J B, Risor T and Hoeyer K (2023) Data as symptom : Doctors' responses to patient-provided data in general practice. *Social Studies of Science 53*(4): 522-544. https://doi.org/10.1177/03063127231164345.

Henke C R (2019) Negotiating Repair : The Infrastructural Contexts of Practice and Power. In: Strebel I, Bovet A and Sormani P (eds) *Repair Work Ethnographies : Revisiting Breakdown, Relocating Materiality*. Singapore: Springer Nature Singapore, pp. 255-282. https://doi.org/10.1007/978-981-13-2110-8.

Henke C R and Sims B (2020) *Repairing Infrastructures : The Maintenance of Materiality and Power*. Cambridge: The MIT Press. https://doi.org/10.7551/mitpress/11771.001.0001.

Hoeppe G (2020) Sharing Data, Repairing Practices: On the Reflexivity of Astronomical Data Journeys. In: Leonelli S and Tempini N (eds) *Data Journeys in the Sciences*. Heidelberg: Springer Nature, pp. 171–190.

Hoeyer K (2016) Denmark at a Crossroad? Intensified Data Sourcing in a Research Radical Country. In: Mittelstadt BD and Floridi L (eds) *The Ethics of Biomedical Big Data*. Law, Governance and Technology Series. Cham: Springer International Publishing, pp. 73–93.

Hoeyer K (2019) Data as promise: Reconfiguring Danish public health through personalized medicine. *Social Studies of Science* 49(4): 531–555.

Hoff J-L (2023) Unavoidable futures? How governments articulate sociotechnical imaginaries of AI and healthcare services. *Futures 148*: 103131. https://doi.org/10.1016/j.futures.2023.103131.

Hogenhout E (2020) *Artificial Intelligence in Health*. Report commissioned by the Netherlands Enterprise Agency, December 2020.

Jackson SJ (2014) *Rethinking Repair. Media technologies: Essays on communication, materiality, and society*. Cambridge: MIT Press, pp. 221–239.

Kallinikos J and Tempini N (2014) Patient Data as Medical Facts: Social Media Practices as a Foundation for Medical Knowledge Creation. *Information Systems Research* 25(4): 817–833.

Lang T (2018) Histoire : Éléments pour une histoire du « Numéro de sécurité sociale ». *Statistique et Société* 6(1): 37–45.

Leonelli S (2016) The philosophy of data. In: Floridi L (ed) *The Routledge Handbook of Philosophy of Information*. London: Routledge, pp. 191-202.

Leonelli S (2020) Learning from Data Journeys. In: Leonelli S and Tempini N (eds) *Data Journeys in the Sciences*. Heidelberg: Springer Nature, pp. 1–26.

Leonelli S and Tempini N (2020) *Data Journeys in the Sciences*. Heidelberg: Springer Nature.

Lomborg S, Langstrup H and Andersen TO (2020) Interpretation as luxury: Heart patients living with data doubt, hope, and anxiety. *Big Data & Society* 7(1): 2053951720924436.

Marent B and Henwood F (2021) Digital health. In: Chamberlain K and Lyons A (eds) *Routledge International Handbook of Critical Issues in Health and Illness*. London: Routledge, pp. 261-275.

Martin-Scholz A, Mayère A, and Lambotte F (2021) Rendre compte du travail d'instauration et d'assemblage des données en santé : Le cas de Registres des cancers en France. *Communication et organisation* 59: 215-230. https://doi.org/10.4000/communicationorganisation.10195.

Mathieu-Fritz A and Esterle L (2013) Doctors and the Single Electronic Health Record. *Reseaux* 178179(2): 223–255.

Mayère A (1990) *Pour une économie de l'information*. Lyon: Éditions du CNRS.

Neves BB, Omori M, Petersen A, et al. (2024) Navigating Artificial Intelligence in Care Homes: Competing Stakeholder Views of Trust and Logics of Care. *Social Science & Medicine*: 117187.

Nicolini D (2009) Zooming In and Out: Studying Practices by Switching Theoretical Lenses and Trailing Connections. *Organization Studies* 30(12): 1391–1418.

Pachidi S, Berends H, Faraj S, et al. (2021) Make Way for the Algorithms: Symbolic Actions and Change in a Regime of Knowing. *Organization Science* 32(1): 18–41.

Paltieli G (2022) The political imaginary of National AI Strategies. *AI & SOCIETY 37*(4): 1613-1624. https://doi.org/10.1007/s00146-021-01258-1.

Pine KH and Bossen C (2020) Good organizational reasons for better medical records: The data work of clinical documentation integrity specialists. *Big Data & Society* 7(2): 2053951720965616.

Pine KH, Landon LA, Bossen C, et al. (2023) Innovations in clinical documentation integrity practice: Continual adaptation in a data-intensive healthcare organisation. *Health Information Management Journal* 52(2): 119–124.

Pink S, Ruckenstein M, Willim R et al. (2018) Broken data: Conceptualising data in an emerging world. *Big Data & Society* 5(1): 2053951717753228.

Ribes D and Jackson SJ (2013) Data Bite Man: The Work of Sustaining a Long-Term Study. In: Gitelman L (ed.) *Raw Data Is an Oxymoron*. Cambridge: MIT Press, pp. 147–166.

Reiss-Sorokin A (2023) Stewards of the Organization: The Management of Repair and Maintenance Work in Large Organizations. *Engaging Science, Technology, and Society* 9(3): 100–120. https://doi.org/10.17351/ests2023.1347.

Ruckenstein M and Schüll ND (2017) The Datafication of Health. *Annual Review of Anthropology* 46(1): 261–278.

Schubert C (2019). Repair Work as Inquiry and Improvisation : The Curious Case of Medical Practice. In: Strebel I, Bovet A and Sormani P (eds) *Repair Work Ethnographies : Revisiting Breakdown, Relocating Materiality*. Singapore: Springer Nature Singapore, pp. 31-60. https://doi.org/10.1007/978-981-13-2110-8.

Schwennesen N (2019) Algorithmic assemblages of care: imaginaries, epistemologies and repair work. *Sociology of Health & Illness* 41(S1): 176–192.

Sims B and Henke C R (2012) Repairing credibility : Repositioning nuclear weapons knowledge after the Cold War. *Social Studies of Science 42*(3): 324347. https://doi.org/10.1177/0306312712437778.

Tanweer A, Fiore-Gartland B and Aragon C (2016) Impediment to insight to innovation: understanding data assemblages through the breakdown–repair process. *Information, Communication & Society* 19(6): 736–752.

Torenholt R and Tjørnhøj-Thomsen T (2022) 'Is this something I should be worried about?' : A study of nurses' recontextualisation work when making clinical decisions based on patient reported outcome data. *Social Science & Medicine 294*: 114645. https://doi.org/10.1016/j.socscimed.2021.114645.

Ureta S (2014) Normalizing Transantiago : On the challenges (and limits) of repairing infrastructures. *Social Studies of Science 44*(3) : 368392. https://doi.org/10.1177/0306312714523855.

Villani C (2018) *Rapport de Cédric Villani : Donner un sens à l'intelligence artificielle (IA)*. Available at: https://www.enseignementsup-recherche.gouv.fr/fr/rapport-de-cedric-villani-donner-un-sens-l-intelligence-artificielle-ia-49194 (accessed 4 March 2025).

Notes

1 Reference terminology for the field of anatomo-pathology in France, produced by the ADICAP association (Association for the Development of Information Technology in Cytology and Pathological Anatomy). This terminology is used to code an anatomo-cytopathology analysis (sample origin, sampling mode, analysis technique, and result).