



# The elephant in the record: On the multiplicity of data recording work

Health Informatics Journal

1–16

© The Author(s) 2019

Article reuse guidelines:

[sagepub.com/journals-permissions](https://sagepub.com/journals-permissions)

DOI: 10.1177/1460458218824705

[journals.sagepub.com/home/jhi](https://journals.sagepub.com/home/jhi)**Federico Cabitza** 

IRCCS Istituto Ortopedico Galeazzi, Italy; University of Milano-Bicocca, Italy

**Angela Locoro**

University of Milano-Bicocca, Italy

**Camilla Alderighi and Raffaele Rasoini**

IRCCS Fondazione Don Carlo Gnocchi, Italy

**Domenico Compagnone and Pedro Berjano**

IRCCS Istituto Ortopedico Galeazzi, Italy

## Abstract

This article focuses on the production side of clinical data work, or data recording work, and in particular, on its multiplicity in terms of data variability. We report the findings from two case studies aimed at assessing the multiplicity that can be observed when the same medical phenomenon is recorded by multiple competent experts, yet the recorded data enable the knowledgeable management of illness trajectories. Often framed in terms of the latent unreliability of medical data, and then treated as a problem to solve, we argue that practitioners in the health informatics field must gain a greater awareness of the natural variability of data inscribing work, assess it, and design solutions that allow actors on both sides of clinical data work, that is, the production and care, as well as the primary and secondary uses of data to aptly inform each other's practices.

## Keywords

data recording work, data work, inter-rater agreement, inter-rater reliability, observer variability

## Background and motivations

Data work can be defined as any activity that is accomplished on data, such as searching, retrieving, consulting, inscribing, arranging, transcribing, printing, and sending data to accomplish tasks and *have work done*. Much of the medical literature equates data work to

---

### Corresponding author:

Federico Cabitza, Dipartimento di Informatica, Sistemistica e Comunicazione, Università degli Studi di Milano-Bicocca, Viale Sarca 336, 20126 Milano, Italy.

Email: [federico.cabitza@unimib.it](mailto:federico.cabitza@unimib.it)

activities for which medical practitioners use expressions like paperwork, deskwork, and desk-top medicine.<sup>1-3</sup> These expressions are usually associated with a derogatory perspective by comparing the time clinicians spend in clinical care and face-to-face encounters with patients to the time spent on reviewing charts, writing orders, arranging records, producing reports, and signing letters. A recent study<sup>3</sup> analyzed the usage logs of almost one million patients' electronic records by 471 primary care physicians and found that physicians split time evenly between seeing patients and documenting activities. In the mold of the increasing digitization of hospital work, Chen et al.<sup>4</sup> recently coined the phrase "electronic patient record encounter" (EPRE) to denote medical interactions with a *further* interlocutor (i.e. the electronic medical record) other than the patient<sup>5</sup> and quantified that residents have approximately 10 EPREs daily, which last little more than half an hour each, for a total of 5 h in data entry and retrieval daily. This finding confirms other recent studies<sup>1,2,6,7</sup> that found how medical doctors spend slightly more than one-half of their daily work shift "on the record," and have to deal with forms and charts even visiting their patients, when data work accounts for approximately one-third of the time spent in the examination room.

However, it is widely known in the health informatics literature that in a work setting, data not only are produced to document work activities and keep a cumulative record to comply with accountability policies and legal norms, but also crucial for cooperative work, enabling the coordination of actors and the articulation of activities across space, time, and business units. This phenomenon was observed in the domain of hospital care and described by Berg<sup>8</sup> in a seminal and oft-cited paper. Thus, although it is analytically easy to distinguish between care and data work, these two kinds of activities are so intertwined that the former could not unfold without the latter. Even isolating "direct clinical face time"<sup>1</sup> in single encounters would neglect the individual doctors' need to take written notes of encounters so conditions and interpretations can be recalled during future encounters<sup>9</sup> and, in so doing, formulate an understanding of the evolution of the patient's illness and *collaborate with themselves* (at least) over time.<sup>10</sup> Decoupling clinical and data work is even more futile in hospital work, which is cooperative by nature. For instance, physicians' prescriptions cannot be decoupled by order *entry* since the cooperative effort to administer a treatment to a patient needs specific data to be articulated within a *record*, which both mediates and enables asynchronous communication and collaboration between physicians and nurses, so as to act as the *distributed* and *working record* described by Fitzpatrick<sup>11</sup> and Bardram and Bossen,<sup>12</sup> among others, for example, by Greenhalgh et al.<sup>13</sup>

In this article, we will focus on the *inscribing side* of data work in the medical domain, or *data recording work*, and in particular on one aspect of the activities in which physicians and nurses make a textual (or anyway coded) representation of some clinically relevant aspect pertaining to the illness of a patient on either paper or an electronic medium, for any of the reasons mentioned above, usually to trigger further action, reflection, and decision-making. One particular aspect we want to shed light on is the *multiplicity* of data recording work, as this is reflected in terms of *data variability*. By doing so, we complement the research conducted on the variability of the "reading side" of data work. This latter variability, perhaps because of its ties to interpretation and sense-making, has been already acknowledged in the specialist literature and widely discussed in many studies from Garfinkle's work in 2008 to more recent ones:<sup>14,15</sup> different readers of the same record can depict to themselves different narratives<sup>16</sup> of the patient's illness trajectory,<sup>17</sup> also according to their knowledge.<sup>18</sup>

## Multiplicity in data work

This study lies within a research line that has focused on the *multiplicity* inherent in the work processes by which data are inscribed into artifacts. Multiplicity (in data work) occurs when the same

phenomenon observed in the reality of interest is reported more than once in the same textual corpus (such as in a medical record) as either *the same representation* (data redundancy) or *different representations* due to the nature of the artifact hosting the representation itself, the situation in which the representation has been produced, or according to different interpretations of the observers of the phenomenon mentioned above (data variability).

Cabitza et al.<sup>19</sup> reported on some studies focusing on *data redundancy*, which is when multiplicity is reflected in multiple traces left on artifacts by data work. We distinguished “multiplicity by redundancy” according to whether the same or slightly different representations of a clinical condition of interest are either *duplicated* or *replicated* across the many components of the medical record. Data work redundancy is, therefore, a kind of redundancy of effort applied to data production. We observed that it is often *purposely* pursued by health practitioners for a number of reasons, such as to improve cognition in event recall, make data retrieval faster, enrich handing-over conferences across work shifts and professional roles and, although it seems obvious, minimize errors by double-checking. These practices are carried out by practitioners convinced of the value of documenting care from their own perspective at the cost of producing more data than strictly necessary and having to correct the inevitable misalignments and local inconsistencies.<sup>20</sup> For these latter drawbacks (which are perceived as drawbacks according to an engineering-oriented perspective), the digitization of the patient record is advocated by some as a means to eradicate data work redundancy, or to limit this phenomenon to the access of the same data from multiple points in the application and by multiple distributed users at the same time, a position that has been recognized as simplistic at best.<sup>13,21</sup>

As hinted above, another kind of data work multiplicity regards the *multiplicity of the originality* of data, that is, information regarding the different actors who could have produced the data and their different situated practices, that is, where and in which conditions they have done it. Also, this characteristic is usually lost in the automation of the means by which health practitioners produce data. In paper-based artifacts, originality is conveyed by penmanship and other subtle signs of individual habits of data recording that may get lost when this activity is transformed into a “robotically checking a bunch of electronic boxes”;<sup>22</sup> unlike a “faceless note,”<sup>22</sup> a handwritten note can indicate to a competent member of a work team who wrote it (from the handwriting) and, in some cases, whether the person was in a frantic and rushed medical situation or completed the “paper work”<sup>29</sup> in a calm and focused manner. Both these elements are important when assessing the reliability of the content of the data (i.e. *Who wrote it? A novice or an expert clinician? Someone I respect or an unreliable colleague?*) and eventually trigger double-checking and other compensating actions.<sup>23</sup> These factors notwithstanding, although originality is information that is easy to conceive as meta-data and store along with the data (as often it is, in usage logs), this indication is seldom conveyed to the users in the digital counterparts of paper-based records. Thus, it is missing, like other aspects of data work,<sup>24,25</sup> the degree of perceived validity or finalization, which is usually related to writings in pencil and side notes in paper-based records.<sup>10,26</sup>

These brief accounts are aimed at the following point. Proponents and designers of the digitized record tend to underestimate the phenomenon of *data work multiplicity*, whether it is related to the ostensible inefficient and sometimes unintended routes along which data flow across the parts of the record (i.e. *data redundancy*) or to the ostensibly interchangeability of the data producers (i.e. *data originality*). In particular, data producers are not abstract actors that translate an objective reality into a standardized language<sup>27</sup> and are not interchangeable: *who* records *what* (and *when*) affects what data are produced and both their validity and reliability.

For this reason, historically, health care facilities have adopted structured forms requiring practitioners to represent cases in terms of well-defined and limited sets of values and codes. These frames and structures notwithstanding, clinical data recording work is not the orderly representation of an objective reality through the structured lens of the medical record; rather, it is a creative process in

which each single physician tells his or her story about the patient by interpreting the same conditions in different ways with respect to colleagues.

In this article, we focus on data work multiplicity not in terms of data redundancy but in terms of data variability. This phenomenon, often termed ‘observer variability’ or ‘inter-rater reliability,’ is well known in the medical literature and does not indicate the extent medical doctors make mistakes in their records; rather, it is an intrinsic and unavoidable uncertainty of clinical conditions and the ambiguity of medical data. As such, this phenomenon is almost completely neglected in health informatics, resulting in information technology (IT) scholars assuming that medical data are reliable for goals other than care, such as accounting, epidemiological research, and decision support design. In what follows, we shed light on this dark side of data work with two field studies undertaken in two purposely different domains: the qualitative interpretation of electrocardiogram (ECG) readings and the more controlled and procedural environment of spine surgery.

### Investigating data recording variability

As hinted above, we investigated the dimension of data recording variability by undertaking two different yet complementary studies. The first one regards a single cardiological case (represented by an ECG) that a large sample of cardiologists belonging to different institutions in Italy was to report on an ad hoc structured form. The second study regards a series of six surgical cases, which nine surgeons belonging to the same surgical unit reported on a standard surgery form that is adopted internationally for epidemiological purposes. As such, these two studies were aimed at getting a comprehensive picture of the phenomenon in two extremely different settings: one regarding a single condition that is typically interpreted with acuity and expertise for its intrinsic ambiguities and nuances (the ECG) and the other one concerning the well-defined sequence of actions performed in a series of representative surgical operations. In the former case, the data producers did not share any reporting conventions: not only did they belong to different clinical facilities across Italy, but also the online form they were asked to fill in was unlike any of the forms they used on a daily basis at work. In the other case, the surgeons were colleagues working closely on a regular basis, were well acquainted with each other, and shared a number of habits, surgical practices, and linguistic and reporting conventions. In particular, the nine participants met regularly before the study (and independent of it) to agree and share the best ways to fill in the standard form we adopted for this study, which had been already used for years in their unit to routinely send surgical data to an international scientific association.

Data recording variability is not a new topic within the medical community. In the specialist literature, it is often referred to with terms such as inter-rater reliability, inter-rater agreement, information bias, and observer variability. This phenomenon has been documented since the beginning of the 20th century, studied since the 1940s (especially in the radiological domain of image interpretation), and attracted the greatest interest in the 1970s.<sup>28</sup> The main studies show that the agreement between the physicians that *codify* (or categorize) the same phenomenon independently (for this reason, they are called raters or coders) is often poor or questionable; less frequently, it is either good or acceptable. In the past 40 years or so, this aspect of medical data work (and practice) has been a minoritarian concern within the medical community. The lack of greater awareness of it, despite its potential impact on any subsequent step in data work and clinical care, has also been traced to the fact that finding suboptimal reliability could “disturb the doctors morale”<sup>29</sup> and perhaps undermine the doctors’ confidence on the objective and scientific side of their discipline.

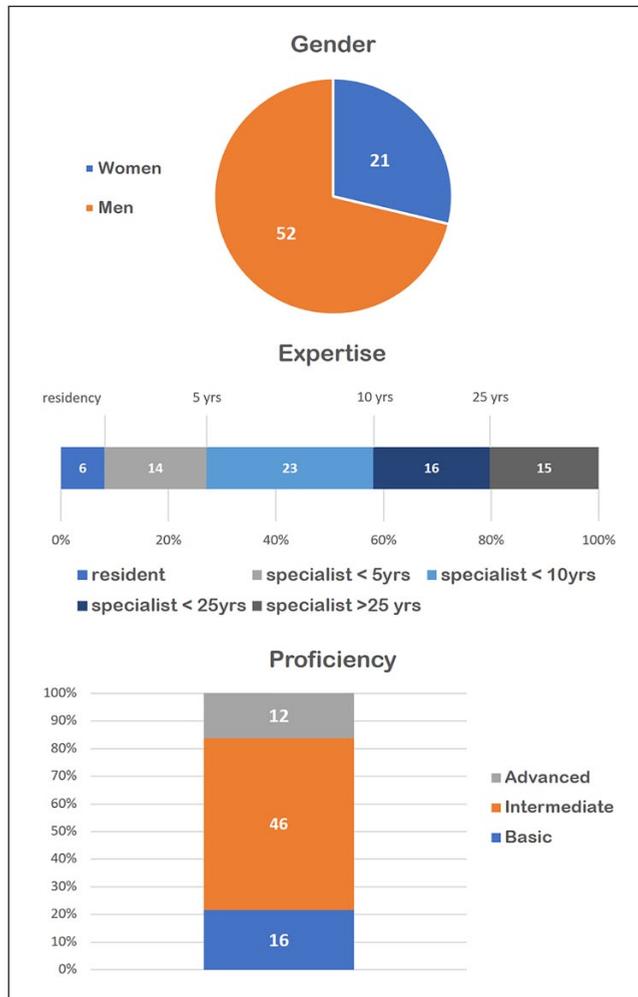
For instance, Jewett et al.<sup>30</sup> reported submitting plain abdominal radiographies to three different radiologists to detect the presence or absence of residual stone fragments; differences among radiologists were found in 52 percent of the reports and 24 percent when the films were reread by the same radiologist. Even more recently, a multicenter, prospective study on 260 digital images of

premature infants was performed to investigate intra- and inter-observer reliability among seven experts in the diagnosis of retinopathy of prematurity.<sup>31</sup> Results showed both an inadequate inter-observer agreement among experts with agreement values (calculated in terms of Fleiss' kappa) between 0.24 and 0.41 (according to diagnosis of different features in retinal images) and, as expected, a slightly higher (albeit still inadequate) intra-observer agreement for the same features with values between 0.47 and 0.63. In general, excellent concordance is very rare.<sup>32</sup>

In this study, to assess the agreement with which the clinicians reported the cases, we chose two common measures from the specialist literature. For the sections of the forms that encompass the mutually exclusive type of multiple choice items (rendered as radio buttons on Web forms), we employed a chance-adjusted measure, *Krippendorff's agreement coefficient alpha*. In its most general form, this coefficient is defined as  $\alpha = 1 - (D_o / D_e)$ , where  $D_o$  is a measure of the observed disagreement and  $D_e$  is a measure of the disagreement that can be expected when chance prevails.<sup>33</sup> Intuitively, this coefficient expresses the *degree of reliability of data reported by multiple observers above what can be expected by chance*; for this reason, it is a more conservative and accurate measure of agreement than just percent agreement, which is the proportion of fields on which two observers agree.<sup>34</sup> We chose this coefficient because it is known to be the “most general agreement measure with appropriate reliability interpretations in content analysis”<sup>33</sup> because it is (different from other widely used measures, such as Cohen's kappa or Fleiss' kappa) robust with respect to missing data and can be applied to any kind of data (both nominal and ordinal) reported by any number of coders. In content analysis, as well as much of the medical research, a conventional threshold to consider data reliable is  $\alpha = 0.8$ , meaning that at least 80 percent of the data are reported to a degree better than chance. However, in Krippendorff's<sup>33</sup> words, “[this threshold] is a pretty low standard by comparison to standards used in engineering.” For this reason, the general recommendation is to consider content with reliabilities between  $\alpha = 0.67$  and  $\alpha = 0.80$  only for drawing tentative conclusions and discard as unreliable all data associated with a lower coefficient. To assess the reliability of single items, such as the items in a checkbox item section, we employed a different measure, the *percent observed agreement*, because the above measure is not defined on single units of analysis. In formal terms, this percentage is calculated as follows:  $A = \sum_{k=1}^q (r_k(r_k - 1)) / (r(r - 1))$  where  $q$  is the total number of possible values,  $r$  is the total number of raters that assigned the item to any value, and  $r_k$  is the number of raters that assigned the item to the  $k$ th value. Since this measure is not adjusted for chance, it may yield higher levels of agreement. Thus, considering both measures yields a balanced (yet optimistic) account of the level of agreement reached in our studies.

### The cardiological study

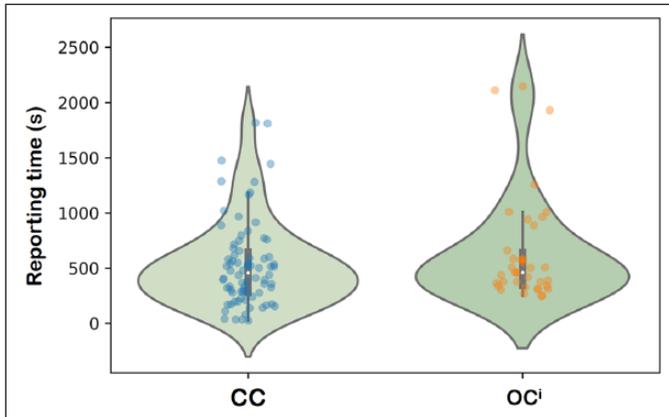
For this study, two authors (cardiologists) conceived a reporting task and questionnaire to be administered to a sample of clinician experts in ECG reading. This task regarded the ECG of a 77-year-old woman, randomly extracted from the *ECG Wave-Maven* (<https://ecg.bidmc.harvard.edu/maven/mavenmain.asp>) database, an online repository of more than 500 ECGs of various difficulty developed for the self-assessment of students and clinicians in ECG reading proficiency. The selection was limited to cases with difficulty levels ranging from 2 to 4 (on a scale of 1–5), so trivial and excessively difficult cases can be avoided. Although the selected case (no. 26, <https://ecg.bidmc.harvard.edu/maven/dispcase.asp?rownum=25&ans=1&caseid=26>) had been assigned a difficulty score of 4, the authors agreed that this was due to the diagnostic condition (i.e. a hard-to-detect renal failure), and the ECG signs were suitable for a standard reading and to the task at hand. In this task, the participants were invited to assess 39 ECG features in terms of presence or absence. This was done so to mimic the standard report forms designed with long lists of checkboxes that have been increasingly adopted in cardiological settings to allow for fast reporting and minimal deviation from



**Figure 1.** The distribution of the participants of the cardiologic study, with respect to gender, expertise, and (self-perceived) proficiency in ECG reading.

structured coding. However, to avoid missing important nuances for each reporting decision, we conceived a four-value scale (certainly present, probably present, probably absent, certainly absent), from which dichotomous data could be derived by subsequent aggregation. The aspects considered in the form were the electrical axis of the QRS vector (4 items), the atrioventricular conduction (10 items), and morphological aspects (25 items).

Participants were personally invited by email to complete an online multi-page questionnaire, implemented on the Limesurvey platform. Invitations were personal and the questionnaires tokenized to avoid multiple compilations and allow for one gentle reminder that was sent after 2 weeks after the initial invitation. When we closed the survey two additional weeks later, 75 cardiologists had completed the form (see Figure 1), completing the task in approximately 9 min on average ( $M=529$ ,  $(SD)=390$  seconds, see Figure 2).

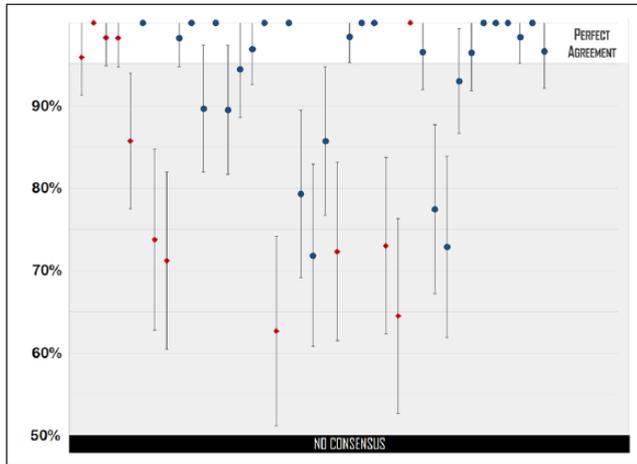


**Figure 2.** Violin plots of the reporting times for the cardiological study (on the left, the *CC* case) and the orthopedic study (on the right, all cases, denoted as *OC'*). Timings are for single cases reported by the coders. Although variances are different because the orthopedic case encompasses all six cases, the two median compilation times (indicated by the white circle within the inner box plot) are similar.

The results surprised the cardiologists; although they considered the ECG of average complexity, the number of doubts expressed and errors made by the coders were beyond expectations. One-third of the respondents chose either one of the two middle values indicating only a likely answer (27%), or the item indicating they were too unsure to make a decision (the ‘don’t know’ (DK)—option chosen by 4%). For 18 out of 39 conditions, diagnostic errors exceeded the 5 percent threshold, which can be considered acceptable. Worse yet, for 10 items, more than 20 percent of the coders were wrong in their reporting decisions. However, the majority decision was always achieved with statistical significance (i.e. the majority group was clearly above 50 percent as indicated by the confidence intervals in Figure 3).

To assess the inter-rater agreement achieved by the involved coders in filling in the form, we considered the DK values as missing values and dichotomized the ordinal answers by aggregating certain and likely options together. By doing so, we could treat the data as nominal variables (i.e. as regular checkboxes). Moreover, to reduce the impact of systematic missing values on the agreement score, we also discarded the answers of 24 cardiologists as “lazy coders,” meaning participants who did not fill in at least 90 percent of the form’s items.

As expected, the agreement was low (52%) (all agreement coefficients are expressed in terms of Krippendorff alphas, computed with IBM SPSS, v. 24, and the KALPHA script developed by Hayes and Krippendorff<sup>34</sup>), which means the coders agreed on only half of the overall content of the form (not considering the fields that could match by mere chance), as shown in Figure 4 (the *CC* data point). To investigate this result, we then distinguished between the items requiring an ECG reading and interpretation (i.e. *ECG items*) and the items requiring the coders to consider the indications provided with the case description (i.e. *review items*) to assess the impact on general reliability of the data that were more prone to interpretation errors, like the ECG features. As expected, agreement on the 12 review items was higher than on the 27 ECG items (57% vs 35%). In any case, we found that by repeating the experiment, we could achieve acceptable agreement in only 1 out of 20 cases on this set of items. Indeed, agreement scores for the ECG case study were well below the threshold suggested for considering ECG data on medical records reliable.<sup>33</sup>



**Figure 3.** For each item of the cardiological form (39 items—on the horizontal axis from left to right), we indicate the proportions (and their confidence intervals) of coders belonging to the majority who reported either the presence or absence of a condition in the ECG considered. ECG items are indicated as red diamonds and review items as blue circles. Majorities were always statistically significant (i.e. the confidence interval never crosses the 50% border on the left, which is when no univocal answer is collectively expressed by the sample of coders). However, in several cases, majorities were smaller than 95 percent.

### *The orthopedic study*

The second study was undertaken at the IRCCS Orthopedic Institute Galeazzi (IOG) of Milan, Italy. This is a large teaching hospital specializing in basic and clinical research on locomotor disorders and associated pathologies where almost 5000 surgeries are performed yearly, mostly arthroplasty (hip and knee prosthetic surgery) and spine-related procedures. From this institute, we involved nine surgeons of the GSpine4 unit, which is the largest spine surgery division at the IOG, including 15 stable members and treating approximately 600 patients annually.

This study adopted an existing standard form designed by an international scientific society, Eurospine, to allow its affiliates (i.e. hospitals, spine units, and single surgeons) to collect data from their practice according to a unified model and send the data to a central registry, the international spine registry ‘Spine Tango.’ The form we considered for our study is called the Spine Tango Surgery (version 2011, available at [http://www.eurospine.org/cm\\_data/SSE\\_PRIM\\_2011\\_ENG.pdf](http://www.eurospine.org/cm_data/SSE_PRIM_2011_ENG.pdf)). This is the international (i.e. English language) version of a two-page form, available in 10 languages, which encompasses only closed-ended options, either multiple-answer or alternative-option items, to minimize recording variability and hence allow for the aggregation of data from multiple sites across Europe.

Two authors (orthopedic surgeons) designed a reporting task where the members of the GSpine4 unit would voluntarily and anonymously consider six cases of varying types and complexity and then independently report them on an electronic version of the paper-based Spine Tango Surgery form for each case. The six cases were real surgical operations that had been performed in the previous 3 months; these cases are considered a random sample of the typical surgical procedures performed at IOG by the GSpine4 team. More precisely, they were the latest operations to occur which were also representative of three main classes of procedures that are typically performed by this spine unit. One case involved myelopathy surgery; this procedure addresses injuries to the

spinal cord due to severe compression and traumas. The other cases involved deformity and degenerative lumbar diseases. Deformity was represented by two cases: one in the idiopathic form in an adolescent and the other in the age-acquired form in adult. Degenerative disease was represented by three cases, including a routine procedure (i.e. a slipped disk) and an intermediate difficult procedure (i.e. spondylolisthesis) to address shifted vertebrae. The third case of this series involved anterior arthrodesis, which is surgery-induced joint ossification between two bones; it is encountered considerably less frequently than the others and considered a more complicated clinical case.

One of the authors then reviewed the medical records and all the available documents regarding each of these six cases and wrote a comprehensive yet concise surgical report (of approximately 700 words) describing the case so that the respondents could recall it (each respondent was likely involved in all of these surgical procedures, although possibly weeks earlier or at different degrees of involvement) and then be supported in the task of completing the surgery form. All the descriptions also included two or three medical high-resolution images and could be consulted by the respondents at all times during the compilation of the form. This form was rendered through a multi-page online questionnaire developed on the Limesurvey platform, adopting a style sheet that would resemble the paper-based form. In addition, no digital support was given to detect and prevent potential inaccuracies in the paper-based format: all the available options were left selectable, even for those fields that were denoted as mutually exclusive choices in the original form.

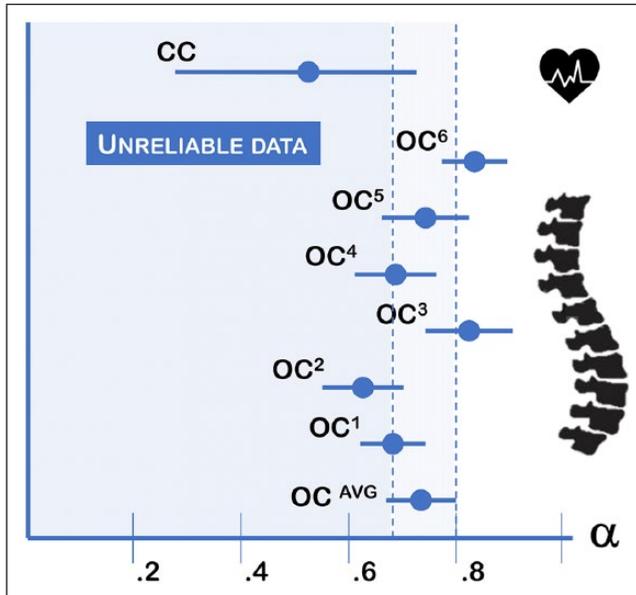
The team members who were enrolled as participants of the study were told that the study was aimed at evaluating the reporting time and assessing its impact on the team workload. Thus, they were instructed to fill in the forms as carefully as possible but in real-life conditions during their regular work shift, to minimize bias in the conclusions that we could draw from the analysis of the responses.

Nine surgeons filled in the six forms, completing each case in approximately 9 min on average ( $M=521$ ,  $SD=254$  seconds), although most of the cases were reported in approximately 6 minutes (see Figure 2). Agreement levels varied, also according to the relative (and varying) complexity of the cases. The case-wise average agreement scores are reported in Figure 4, while Figures 5 and 6 show the average reliability of each section of the standard form.

## Discussion and implications for design

In this article, we presented two studies aimed at giving a complementary view of the phenomenon of data work multiplicity. In the orthopedic setting, the data work under consideration was the standardized reporting of surgical procedures by those who had performed them (although after a wash-out period and on the basis of a written account). However, in the cardiological setting, data work focused on how proficiently clinicians could document what they read in some sample ECG tracings. These two studies yielded different results (see Figure 4). The agreement among the orthopedic coders was found to be higher than in the cardiological study but not significantly. This was partly expected due to the nature of the task, the acquaintance of the coders with the reporting tool, and the common background of the coders involved.<sup>18</sup> However, in both cases, the average agreement did not achieve the threshold recommended by Krippendorff<sup>33</sup> for medical data. However, with this contribution, we do not aim to yield “yet further” evidence of the phenomenon of observer (or coder) variability in medicine, which is well known and documented in the literature. Rather, we make a point about the importance of multiplicity in data work (especially in data recording work), which is generally not widely considered in medical informatics and in the design of electronic medical records. This multiplicity is the “elephant” we refer to in the title of this article.

The elephant in the room is an idiomatic expression indicating a matter of concern that is not considered or discussed despite its lumbering and placid existence. We use the phrase *elephant in*



**Figure 4.** The agreement coefficients ( $\alpha$ ) with their 95 percent confidence intervals for each clinical case considered. The cardiological case is labeled as CC. For the orthopedic cases, we show the alphas for each case ( $OC^1 - OC^6$ ) and the average agreement achieved considering all the six cases ( $OC^{AVG}$ ). Data left of the 80 percent threshold should be considered unreliable for epidemiological purposes; data lower than 67 percent are unsuitable for any purpose; the closer to the vertical axis (i.e.  $\alpha = 0$ ) the data points, the less reliable the recorded data.

*the record* to describe the phenomenon of variability in data recording work and how this data multiplicity affects the consequent reliability of the record's data for other aims than the ones for which the data have been produced (mostly for medical care). This phenomenon is reminiscent of the so-called first law of medical informatics: "data shall be used only for the purpose for which they are collected"<sup>35</sup> and its corollary, which states, "if no purpose was defined prior to the collection of the data, then the data should not be used." In this multi-site study, we showed that even if data producers know the purposes for which they record data related to their practice (accountability and research), feel genuinely committed to these purposes, and use forms specifically designed to minimize variability and increase the objectivity of judgment, secondary uses of their data (such as statistical analysis for either monitoring or epidemiological research and the statistical optimization of predictive models) can be undermined by the *variability* among their reports. In this article, we specifically spoke of the *multiplicity* of their reports instead of considering this variability in terms of *unreliability* or poor agreement among the coders as it is usually done in the specialist literature. In fact, our point is that the observed multiplicity (i.e. multiple descriptions of the same events and procedures), such as in the orthopedic study, or multiple interpretations of the same health condition, such as in the cardiological study, should not be traced (entirely) to incompetence or reporting errors. While we cannot rule out the existence of reporting and interpretation errors in our studies, in discussing these results with some specialists (including the four medical authors of this article), we concurred that errors had a low impact. Instead, data multiplicity can be traced back to the intrinsic uncertainty, ambiguity, and manifoldness of the medical phenomena being reported by Simpkin and Schwartzstein<sup>36</sup> and, therefore, it is created where *data work* and

# SPINE TANGO



# SURGERY 2011

Internal Use Only / Not read by scanner

Last name	First name	Gender
Street		M.R.N.
Country code	Zip code	City
Social security number		Birthdate (DD.MM.YYYY)

**Directions**  
 • Use a #2 soft pencil for marking.  
 • Text answers must be entered with the web interface.  
 • All questions must be answered unless otherwise indicated.  
 • Completely fill in boxes to record answers.

**Question types**  
 only 1 answer allowed  
 multiple answers allowed  
 mandatory questions  
 ..... please specify

<b>Format</b>	<input type="checkbox"/> minimal	<input type="checkbox"/> complete
---------------	----------------------------------	-----------------------------------

**Level of intervention**  
 upper cervical     cervicothoracic     thoracic     thoraco-lumbo-sacral     lumbo-sacral     coccyx  
 mid lower cervical     cervico-thoraco-lumbar     thoracolumbar     lumbar     sacral

### Admission / Pathology

Day:  1-30    Month:  1-12    Year:  19-20

<b>Main pathology</b>	<b>0.83</b>
<input type="checkbox"/> degenerative disease <input type="checkbox"/> fracture/trauma <input type="checkbox"/> spondylolisthesis (non-degen.) <input type="checkbox"/> infection <input type="checkbox"/> repeat surgery <input type="checkbox"/> non-degen. deformity <input type="checkbox"/> pathological fracture <input type="checkbox"/> inflammation <input type="checkbox"/> tumor <input type="checkbox"/> other, specify .....	

**Specification of Main Pathology**    Only answer questions related to Main Pathology (Main Pathology "other" requires no specification).

<p><b>Degen. disease</b></p> <p>Type of degeneration    Specify grade of spondyl. → <b>0.5</b></p> <p> <input type="checkbox"/> disc herniat./protrusion    <input type="checkbox"/> degen. spondylolisthesis  <input type="checkbox"/> central stenosis    <input type="checkbox"/> other instability  <input type="checkbox"/> lateral stenosis    <input type="checkbox"/> myelopathy  <input type="checkbox"/> foraminal stenosis    <input type="checkbox"/> facet joint arthrosis  <input type="checkbox"/> degen. disc disease    <input type="checkbox"/> other .....  <input type="checkbox"/> degen. deformity    Specify type of deformity below         </p> <p><b>Type of deformity</b>    Also specify type of degenerative deformity <b>0.7</b></p> <p> <input type="checkbox"/> scoliosis    <input type="checkbox"/> scoliosis  <input type="checkbox"/> kyphosis    <input type="checkbox"/> other .....            Type of scoliosis  <input type="checkbox"/> single curve    <input type="checkbox"/> double curve         </p> <p><b>Predominant etiology</b> <b>1</b></p> <p> <input type="checkbox"/> idiopathic    <input type="checkbox"/> posttraumatic  <input type="checkbox"/> congenital    <input type="checkbox"/> M. Scheuermann  <input type="checkbox"/> neuromuscular    <input type="checkbox"/> other .....         </p>	<p><b>Spondylolisthesis</b> <b>0.67</b></p> <p>Type of spondylolisthesis    Grade of spondylolisthesis</p> <p> <input type="checkbox"/> Type I (congenital, dysplastic)    <input type="checkbox"/> Grade 0  <input type="checkbox"/> Type II (isthmic)    <input type="checkbox"/> Grade I  <input type="checkbox"/> Type III see type of degeneration    <input type="checkbox"/> Grade II  <input type="checkbox"/> Type IV (traumatic)    <input type="checkbox"/> Grade III  <input type="checkbox"/> Type V (pathologic)    <input type="checkbox"/> Grade IV  <input type="checkbox"/> Type VI (postsurgical)    <input type="checkbox"/> Spondyloptosis (V)         </p> <p><b>Inflammation</b></p> <p>Type of inflammation</p> <p> <input type="checkbox"/> inflammatory arthritis (seropos)  <input type="checkbox"/> seronegative arthritis  <input type="checkbox"/> ankylosing spondylitis (M. Bechterew)  <input type="checkbox"/> other .....         </p> <p><b>Infection</b></p> <p>Infection specification    Affected structure(s)</p> <p> <input type="checkbox"/> pyogenic    <input type="checkbox"/> fungal    <input type="checkbox"/> spondylitis    <input type="checkbox"/> paravertebral infection  <input type="checkbox"/> parasitic    <input type="checkbox"/> other    <input type="checkbox"/> discitis  <input type="checkbox"/> tuberculous    <input type="checkbox"/> other    <input type="checkbox"/> epidural space    <input type="checkbox"/> other .....         </p> <p><b>Tumor</b> <b>1</b></p> <p>Type of tumor    Localization</p> <p> <input type="checkbox"/> primary malignant    <input type="checkbox"/> extraosseous soft tissues  <input type="checkbox"/> primary benign    <input type="checkbox"/> intraosseous (superficial)  <input type="checkbox"/> secondary malignant    <input type="checkbox"/> intraosseous (deep)  <input type="checkbox"/> tumor like lesion    <input type="checkbox"/> extrasosseous (extradural)  <input type="checkbox"/> other .....    <input type="checkbox"/> extrasosseous (intradural)  <input type="checkbox"/> other .....    <input type="checkbox"/> other .....         </p> <p>Specify type of tumor .....</p>
---	--

**Additional fractures w/different treatments require separate forms**

<b>Type of (pathological) fracture/trauma</b>	<b>1</b>
<input type="checkbox"/> bony/frag (C0) <input type="checkbox"/> C2 other fracture <input type="checkbox"/> C0/1 dislocation <input type="checkbox"/> soft tissue injury neck <input type="checkbox"/> C1 fracture <input type="checkbox"/> fracture C3-L5/S1 <input type="checkbox"/> C1/2 instability <input type="checkbox"/> sacrum fracture <input type="checkbox"/> C2 dens fracture <input type="checkbox"/> other .....	
<b>Dens fracture type</b> C3-L5/S1 AO fracture type <input type="checkbox"/> I    Type <input type="checkbox"/> A <input type="checkbox"/> B <input type="checkbox"/> C <input type="checkbox"/> II    Group <input type="checkbox"/> 1 <input type="checkbox"/> 2 <input type="checkbox"/> 3 <input type="checkbox"/> III    Subgroup <input type="checkbox"/> 1 <input type="checkbox"/> 2 <input type="checkbox"/> 3	
<b>Pathological fracture due to ...</b> Fracture age <input type="checkbox"/> osteoporosis <input type="checkbox"/> fresh fracture <input type="checkbox"/> tumor <input type="checkbox"/> old fracture <input type="checkbox"/> other .....    In case of tumor, answer questions "Type of tumor" and "Localization" in section "TUMOR"	

**Repeat surg.**    Type or reason of repeat surgery **0.72**

hardware removal     neurocompression     implant failure  
 non-union     postop. infection     sagittal imbalance  
 instability     superficial     adiac. segment  
 failure to reach     postop. infect. deep     pathology  
 therapeutic goals     implant malposition     other .....

**Comments regarding main pathology:**

In segments, mark cranial VB    SA = sacrum (S3-5) / CO = coccyx

<b>Most severely affected</b> <input type="checkbox"/> segment <input type="checkbox"/> vertebral body <b>0.17</b>	<b>Extent of lesion (segments/vertebral bodies)</b> <b>0.33</b>
--	---

**Additional pathology (Answer to question "Main pathology" is excluded.)** **0.67**

none     non-degen. deformity     pathological fracture     inflammation     tumor     other, specify .....  
 degen. disease     fracture/trauma     spondylolisthesis (non-degen.)     infection     repeat surgery

**Number of previous spine surgeries** **0.82**    Previous surgeries at same level    Previous treatment for main pathology (by specialist) **0.5**

Knows "0" excludes both "Previous surgery questions (at same level" and "at same hospital)"

**Risk factors**    **Current smoker**    **Presence of flags - low back pain**

BMI:  <20     20-25     >25     26-30     unknown  
 Current smoker:  yes     no     unknown  
 Presence of flags - low back pain:  none     orange     not assessable/applicable  
 red     blue     yellow     black

Red: Biomedical Factors; serious spinal pathology  
 Yellow: Psychosocial or behavioral factors  
 Orange: Abnormal psychological processes indicating psychiatric disorders  
 Blue: Socioeconomic/work factors  
 Black: Occupational and societal factors

SA = sacrum / CO = coccyx    Copyright ©EMMO, 2011 All rights reserved 31.12.2011 / Version v1

**Figure 5.** The front side of the Spine Tango Surgery form. Agreement coefficients ( $\alpha$  and  $A$ ) are indicated for the main sections of the form in terms of saturation (the darker the color, the lower the agreement) and figures (the closer to 1.0 the values, the higher the agreement).



*interpretation work* overlap, as they often do in medicine. Indeed, it is noteworthy here to recall the words of Monteiro about interpretation work. He said,

[this] is embodied on two levels: one, when the digital objects [in our case, reports and records] are produced; two, when they are handled in weekly meetings. Meaning about natural phenomena [in our case, medical phenomena] is thus being continually constructed in these various processes.<sup>37</sup>

The data work we described in this study is clearly on the former level: the embodiment of interpretation in some written form. However, our observations on data redundancy and the related practices,<sup>38</sup> and the similar work on *intertextuality* occurring among medical records<sup>39</sup> suggest that data work is also crucial to interpretation work on the second level mentioned above.

Interpretation work makes sense of seemingly incoherent data (by interpolating between them or even transcending them) and reduces the potentially negative impact of data variability on diagnostic accuracy and therapeutic efficacy, which are constantly (yet slowly) improving over time.<sup>40</sup> However, this kind of work usually eludes the abstract models and algorithmic requirements of data quality as these are usually defined for (or imposed by) health information systems.<sup>41,42</sup> When automation, rather than human interpretation, creates output from input, the old saying holds true (i.e. *if garbage is put into a machine, then garbage goes out*<sup>43</sup>). This simple consideration cannot be overrated, especially in this data-driven age, in which increasingly larger amounts of clinical data are harvested and used for epidemiological research and to create the so-called *ground truth* for the development of predictive models with machine learning techniques. To these latter aims, medical informaticians usually focus on problems of data quality such as completeness and timeliness<sup>42</sup> and take the reliability of record data for granted, leaving the multiplicity of data recording work and its consequences on automation almost completely unaddressed.<sup>44</sup>

However, sweeping data multiplicity under the rug can just contribute to fiction and hence undermine both human interpretation work (by offering data that are flattened into hard-wired codes) and the automation support of real-life care. A more viable alternative could be using automation to detect data multiplicity, assess it, and hence *bounce* it to the awareness of the practitioners (rather than trying to get rid of it). This is the first and necessary step to make this phenomenon duly recognized and an object of discussion among clinicians and health informaticians. A second step could involve undertaking further research on whether *representing data multiplicity* could improve the collaborative process of making sense of data (which is another form of data work) and contribute to the achievement of *shared decision-making*<sup>45</sup> that could be grounded not only on firm and established certainties but also on the unavoidable ambiguities and gray areas of many medical cases (where multiple views and interpretation coexist). *Representing multiplicity* (i.e. to detect, assess, and display it to the users of the medical record) is a challenging task. For instance, observer variability could be automatically assessed by a system that, for a random selection of cases, would ask two or three physicians (unaware of the others' assignments) to fill in the *same* patient record. In so doing, the system would have multiple data for the same patient and compute a reliability measure associated with these attributes. This information could be conveyed at the interface level in terms of  $\alpha$ ,  $A$ , or any other suitable measure or even just rendered in terms of color saturation of form backgrounds as depicted in Figures 5 and 6. This would make subsequent consumers of those data aware of the intrinsic variability of the data in the record and possibly promote more discussion about those specific cases by the members of the team.

Moreover, the system could extract the *mode value* (i.e. the value reported in the same field of the same form by most of the respondents, if such a value exists) and determine whether this majority value had been chosen by a significant majority of coders (i.e. higher than 50%). This approach has been found useful to improve the quality of medical records and make their data reliable

enough for *ground-truthing*,<sup>46</sup> at the expense of involving a sufficient number of coders (e.g. 12) on a random basis, if not on a regular one. In our field cases, this kind of consensus data could be extracted for all of the items in the cardiological case (not surprisingly, given the high number of coders involved), but less frequently in the orthopedic study, when mode values extracted from the forms were found to be representative of a team consensus for just two-thirds of the form items (65%).

In conclusion, we believe that the following conjecture is worth further investigation: does a system that *exposes* data recording work variability help the involved stakeholders discuss this issue and understand whether there is a will (and the necessary resources) to improve concordance in data work? This aim could be addressed through a number of socio-technical interventions, which include gently inviting colleagues to a stricter compliance with compilation rules,<sup>47</sup> as well as by suggesting or promoting the adoption of check-lists;<sup>48</sup> binding incentives to the continuous improvement of average concordance rates; adopting (or requiring) an electronic medical record that implements data checks which prevent further use if anomalies and divergences among multiple coders are detected; planning additional training sessions to improve the coders' acquaintance with classification schemas, coding standards, and compilation conventions; and, finally, hiring (or asking for) medical assistants, or scribes,<sup>49</sup> who are specialized in data recording work and committed to high-quality record-keeping. Any of the above interventions could have a positive impact on data work and care, as well as unintended consequences, which must be partly envisioned *ex ante* and duly assessed in post hoc analyses.

With the support of digital IT, the elephant of the variability of recording data work would not (or better yet, should not) be removed from the rooms of medicine and hospital care. Rather, data workers could be made more aware of it and could be provided with some means to coexist with, and possibly leverage, it for the sake of better data, smoother data work, and the respectful observance of the laws of medical informatics.

## Authors Note

Angela Locoro is now affiliated with Università Carlo Cattaneo - LIUC, Italy.

## Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Grant SPINEREG - CC-2015-2365325 funded by the Italian Ministry of Health. Grant CO-2016-02364645 - The influence of baseline clinical status, comorbidity, and surgical techniques on the risks and benefits of spine surgery: use of the registry to supplement the evidence from the cohort study funded by the Italian Ministry of Health.

## ORCID iD

Federico Cabitza  <https://orcid.org/0000-0002-4065-3415>

## References

1. Sinsky C, Colligan L, Li L, et al. Allocation of physician time in ambulatory practice: a time and motion study in 4 specialties. *Ann Intern Med* 2016; 165(11): 753–760.
2. Arndt BG, Beasley JW, Watkinson MD, et al. Tethered to the EHR: primary care physician workload assessment using EHR event log data and time-motion observations. *Ann Fam Med* 2017; 15(5): 419–426.
3. Tai-Seale M, Olson CW, Li J, et al. Electronic health record logs indicate that physicians split time evenly between seeing patients and desktop medicine. *Health Affairs* 2017; 36(4): 655–662.

4. Chen L, Guo U, Illipparambil LC, et al. Racing against the clock: internal medicine residents' time spent on electronic health records. *Journal of Grad Med Educ* 2016; 8(1): 39–44.
5. Swinglehurst D, Roberts C and Greenhalgh T. Opening up the “black box” of the electronic patient record: a linguistic ethnographic study in general practice. *Commun Med* 2011; 8(1): 3–15.
6. Read-Brown S, Hribar MR, Reznick LG, et al. Time requirements for electronic health record use in an academic ophthalmology center. *JAMA Ophthalmol* 2017; 135(11): 1250–1257.
7. Young R, Burge S, Kumar K, et al. A time-motion study of primary care physicians work in the electronic health record era. *Fam Med* 2018; 50(2): 91–99.
8. Berg M. Accumulating and coordinating: occasions for information technologies in medical work. *CSCW* 1999; 8(4): 373–401.
9. Gregory J, Mattison J and Linde C. Naming notes: transitions from free text to structured entry. *Method Inform Med* 1995; 34(1–2): 57–67.
10. Bricon-Souf N, Bringay S, Hamek S, et al. Informal notes to support the asynchronous collaborative activities. *Int J Med Inform* 2007; 76: S342–S348.
11. Fitzpatrick G. Integrated care and the working record. *Health Inform J* 2004; 10(4): 291–302.
12. Bardram JE and Bossen C. A web of coordinative artifacts: collaborative work at a hospital ward. In: *GROUP '05*, Sanibel, FL, 6–9 November 2005, pp. 168–176. New York: ACM.
13. Greenhalgh T, Potts HW, Wong G, et al. Tensions and paradoxes in electronic patient record research: a systematic literature review using the meta-narrative method. *Milbank Q* 2009; 87(4): 729–788.
14. Garfinkel H. “Good” organizational reasons for “bad” clinic records. In: Krippendorff K and Bock MA (eds) *The content analysis reader*. Thousand Oaks, CA: SAGE, 2008, pp. 45–53.
15. Bansler JP, Havn EC, Schmidt K, et al. Cooperative epistemic work in medical practice: an analysis of physicians clinical notes. *CSCW* 2016; 25(6): 503–546.
16. Hunter KM. *Doctors' stories: the narrative structure of medical knowledge*. Princeton, NJ: Princeton University Press, 1991.
17. Reddy MC, Dourish P and Pratt W. Temporality in medical work: time also matters. *CSCW* 2006; 15(1): 29–53.
18. Hobbs P. Islands in a string: the use of background knowledge in an obstetrical resident's notes. *J Sociolinguistics* 2002; 6(2): 267–274.
19. Cabitza F, Locoro A, Ellingsen G, et al. Repetita iuvant: exploring and supporting redundancy in hospital practices. *CSCW* 2018; 27(3–6): 1051–1084.
20. Heath C and Luff P. Documents and professional practice: “bad” organisational reasons for “good” clinical records. In: *CSCW '96*, Boston, MA, 16–20 November 1996, pp. 354–363. New York: ACM.
21. Fitzpatrick G and Ellingsen G. A review of 25 years of CSCW research in healthcare: contributions, challenges and future agendas. *CSCW* 2013; 22(4–6): 609–665.
22. Wachter R. *The digital doctor, hope, hype and at the dawn of medicines computer age*. New York: McGraw-Hill Education, 2015, pp. 78, 80.
23. Berg M. Practices of reading and writing: the constitutive role of the patient record in medical work. *Sociol Health Ill* 1996; 18(4): 499–524.
24. Harper RH, O'Hara KP, Sellen AJ, et al. Toward the paperless hospital? *Brit J Anaesth* 1997; 78(6): 762–767.
25. Zhou X, Ackerman MS and Zheng K. I just don't know why it's gone: maintaining informal information use in inpatient care. In: *CHI '09*, Boston, MA, 04–09 April 2009, pp. 2061–2070. New York: ACM.
26. Cabitza F, Simone C and Sarini M. Leveraging coordinative conventions to promote collaboration awareness. *CSCW* 2009; 18(4): 301–330.
27. Winthereik BR and Vikkelsø S. ICT and integrated care: some dilemmas of standardising inter-organisational communication. *CSCW* 2005; 14(1): 43–67.
28. Koran LM. The reliability of clinical methods, data and judgments. *New Engl J Med* 1975; 293(14): 695–701.
29. Reiser SJ. *Medicine and the reign of technology*. Cambridge: Cambridge University Press, 1981, p. 193.
30. Jewett M, Bombardier C, Caron D, et al. Potential for inter-observer and intra-observer variability in x-ray review to establish stone-free rates after lithotripsy. *J Urology* 1992; 147(3): 559–562.

31. Gschließer A, Stifter E, Neumayer T, et al. Inter-expert and intra-expert agreement on the diagnosis and treatment of retinopathy of prematurity. *Am J Ophthalmol* 2015; 160(3): 553–560.
32. Kraemer HC. The reliability of clinical diagnoses: state of the art. *Annu Rev Clin Psycho* 2014; 10: 111–130.
33. Krippendorff K. *Content analysis: an introduction to its methodology*. Thousand Oaks, CA: SAGE, 2004, pp. 242, 278.
34. Hayes AF and Krippendorff K. Answering the call for a standard reliability measure for coding data. *Commun Methods Meas* 2007; 1(1): 77–89.
35. van der Lei J. Use and abuse of computer-stored medical records. *Methods Inf Med* 1991; 30: 79–80.
36. Simpkin AL and Schwartzstein RM. Tolerating uncertainty the next medical revolution? *New Engl J Med* 2016; 375(18): 1713–1715.
37. Monteiro M. Reconfiguring evidence: interacting with digital objects in scientific practice. *CSCW* 2010; 19(3–4): 335–354.
38. Cabitza F and Simone C. Supporting practices of positive redundancy for seamless care. In: *CBMS'08*, Jyvaskyla, 17–19 June 2008, pp. 470–475. New York: IEEE.
39. Christensen LR. On intertext in chemotherapy: an ethnography of text in medical practice. *CSCW* 2016; 25(1): 1–38.
40. Ma J, Ward EM, Siegel RL, et al. Temporal trends in mortality in the united states, 1969–2013. *JAMA* 2015; 314(16): 1731–1739.
41. Cabitza F and Simone C. Whatever works: making sense of information quality (chapter 6). In: Viscusi G, Campagnolo GM and Curzi Y (eds) *Phenomenology, organizational politics, and IT design*. Hershey, PA: IGI Global, 2012, pp. 79–110.
42. Cabitza F and Batini C. Information quality in healthcare. In: Batini C and Scannapieco, M (eds) *Data and information quality*. Cham: Springer, 2016, pp. 403–419.
43. Kim Y, Huang J and Emery S. Garbage in, garbage out: data collection, quality assessment and reporting standards for social media data use in health research, infodemiology and digital disease detection. *J Med Internet Res* 2016; 18(2): e41.
44. Cabitza F, Ciucci D and Rasoini R. A giant with feet of clay: On the validity of the data that feed machine learning in medicine. In: Cabitza F, Batini C and Magni M (eds.) *Organizing for the Digital World* (2019). Springer, Cham, 2019. pp. 121–136.
45. Lehman R. Sharing as the future of medicine. *JAMA Inter Med* 2017; 177(9): 1237–1238.
46. Svensson CM, Hubler R and Figge MT. Automated classification of circulating tumor cells and the impact of interobserver variability on classifier training and performance. *J Immunol Res* 2015; 2015: 573165.
47. Thaler RH and Sunstein CR. *Nudge: improving decisions about health, wealth, and happiness*. New York: Penguin Books, 1975.
48. Gawande A. *The checklist manifesto: how to get things right*. Gurgaon, India: Penguin Books, 2011.
49. Shultz CG and Holmstrom HL. The use of medical scribes in health care settings a systematic review and future directions. *J Am Board Fam Med* 2015; 28(3): 371–381.