Review Article

Bridging the "last mile" gap between AI implementation and operation: "data awareness" that matters

Federico Cabitza[1], **Andrea Campagner**[2], **Clara Balsano**[3,4]

[1]Dipartimento di Informatica, Sistemistica e Comunicazione, Università degli Studi di Milano-Bicocca, Milano, Italy; [2]IRCCS Istituto Ortopedico Galeazzi, Milano, Italy; [3]Dipartimento di Medicina Clinica, Sanità Pubblica, Scienze della Vita e dell'Ambiente, Università degli Studi dell'Aquila, L'Aquila, Italy; [4]Francesco Balsano Foundation, Via Giovanni Battista Martini 6, Rome, Italy

Contributions: (I) Conception and design: F Cabitza, C Balsano; (II) Administrative support: A Campagner; (III) Provision of study materials or patients: None; (IV) Collection and assembly of data: F Cabitza, A Campagner; (V) Data analysis and interpretation: F Cabitza; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

Correspondence to: Federico Cabitza, PhD. Dipartimento di Informatica, Sistemistica e Comunicazione, Università degli Studi di Milano-Bicocca, Milano, Italy. Email: federico.cabitza@unimib.it.

Abstract: Interest in the application of machine learning (ML) techniques to medicine is growing fast and wide because of their ability to endow decision support systems with so-called artificial intelligence, particularly in those medical disciplines that extensively rely on digital imaging. Nonetheless, achieving a pragmatic and ecological validation of medical AI systems in real-world settings is difficult, even when these systems exhibit very high accuracy in laboratory settings. This difficulty has been called the "last mile of implementation." In this review of the concept, we claim that this metaphorical mile presents two chasms: the hiatus of human trust and the hiatus of machine experience. The former hiatus encompasses all that can hinder the concrete use of AI at the point of care, including availability and usability issues, but also the contradictory phenomena of cognitive ergonomics, such as automation bias (overreliance on technology) and prejudice against the machine (clearly the opposite). The latter hiatus, on the other hand, relates to the production and availability of a sufficient amount of reliable and accurate clinical data that is suitable to be the "experience" with which a machine can be trained. In briefly reviewing the existing literature, we focus on this latter hiatus of the last mile, as it has been largely neglected by both ML developers and doctors. In doing so, we argue that efforts to cross this chasm require data governance practices and a focus on data work, including the practices of data awareness and data hygiene. To address the challenge of bridging the chasms in the last mile of medical AI implementation, we discuss the six main socio-technical challenges that must be overcome in order to build robust bridges and deploy potentially effective AI in real-world clinical settings.

Keywords: Artificial intelligence; last mile; DevOps; data hygiene

Submitted Jan 18, 2020. Accepted for publication Feb 27, 2020.
doi: 10.21037/atm.2020.03.63
View this article at: http://dx.doi.org/10.21037/atm.2020.03.63

Introduction

Interest in the applications of machine learning (ML) techniques in medicine is growing fast because of their ability to endow decision support systems with so-called artificial intelligence (AI), a vague but evocative expression to denote, in this context, the capabilities of machines (i.e., algorithms) to classify or stratify clinical cases or predict related conditions with high accuracy—in some cases, even more accurately than human experts (1). If we limit ourselves to counting how many papers indexed on PubMed have the expression "AI" in their title, we can see that, every 2 years from 2012 to 2017, this number has roughly doubled (43, 70, and 169 in, respectively, 2012 to 2013, 2014 to 2015, and 2016 to 2017), but, in the most

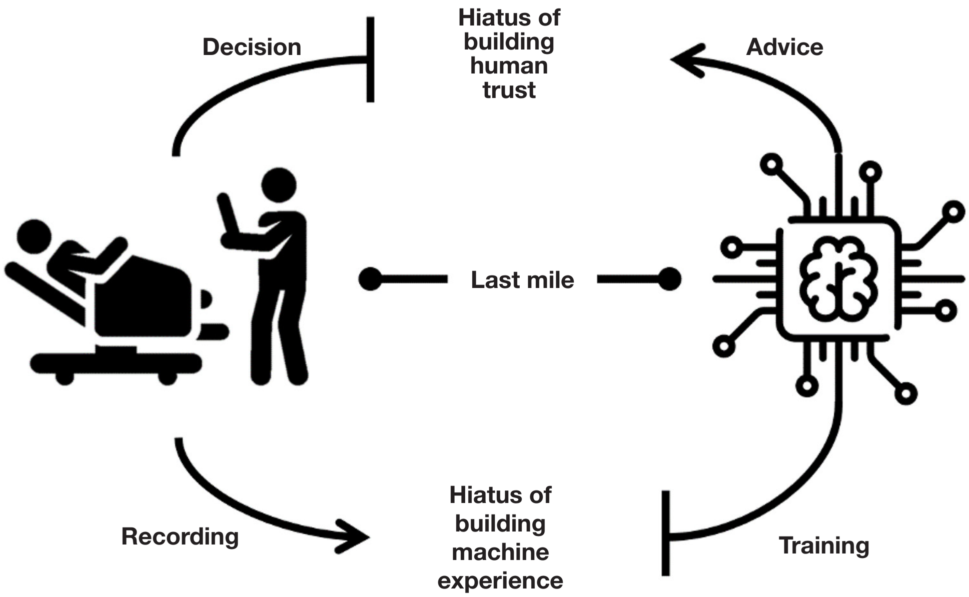

Decision
Hiatus of building human trust
Advice
Last mile
Recording
Hiatus of building machine experience
Training

Figure 1 The hiatuses in the “last mile” between medical AI and human agency.

recent 2-year period, this number has increased almost tenfold (1,413 in 2018 to 2019). Moreover, this interest goes beyond the ambit of academic research, as it is mirrored by a doubling of the number of FDA approvals for devices endowed with some form of AI in the last 5 years.

The proof of the pudding in medical AI

Despite this increasing interest, there is oddly still a lack of consensus on how to assess whether the adoption of AI in a healthcare setting is even successful and hence useful. In an editorial in this journal (2), we made the point that we must go beyond statistical validation (which is what it is usually conducted and reported in scientific reports and articles in terms of accuracy measures, such as C-statistics or F-scores) and demand proof that these systems bring clinical benefit when fed with real-world data (what we called pragmatic validation) and when deployed in real clinical settings (ecological validation). Achieving ecological validation of statistically high-performing AI in diagnostic and other medical tasks has been observed to be more complicated than initially expected; in fact, most of the challenges that make technically sound systems perform poorly in real-world settings lie in the “last mile of implementation” (3)—a concept that we equate to the conceptual gap between developing medical ML and the mere application of ML techniques to medical data. Moreover, this “last mile,” although apparently short, is not a flat and regular path, but rather presents two chasms, as shown in *Figure 1*.

In particular, the hiatus of human trust represents the most serious hindrance to the full realization of the potential of AI at the point of care, as even the most accurate systems are affected when they are not trusted by doctors as a result of what we called “prejudice against the machine” (4). This hiatus represents the failure of medical AI to make a positive impact on doctors’ decisions, irrespective of its intrinsic accuracy. This can be because the interface is inadequate, because the good advice comes late or among several—often too many—false alarms (5), or because the decision-maker cannot take advantage of the technological support due to the emergence of either automation bias (6) or automation complacency (7). Efforts to bridge this hiatus are attracting increasing interest from the specialist communities of ML and human-computer interaction (HCI), in which solutions are designed and tested to improve the usability of AI interfaces (8), their causability (9)—that is, the quality of their explanations—and, more generally, their acceptability (10).

This is why we make the point here that, paradoxical as it might seem, the other hiatus, which represents how the clinical experience from which ML models might learn is accumulated and processed, is more neglected than the human trust hiatus. In fact, most data scientists would deny that this step is even a gap: data scientists and ML developers usually assume that the datasets with which their predictive models are trained—what is emphatically called ground truth—are (I) truthful, (II) reliable, and (III) representative of the target population. However, this threefold assumption is seldom tenable and often ill-grounded, at least to some extent. As we have observed (11), ground truth can be considered reliable (note: for our practical aims, we consider a source dataset reliable if it is 95% accurate) only when perfect raters are called to annotate past cases or at least nine averagely accurate raters [that is, at least 85% accurate on average, which is a reasonable estimation (12)] are involved. Such a condition seldom holds true, if ever. Thus, the hiatus of machine experience represents the difference between actual care, as experienced by doctors, and its codified representation in the form of data, which is the only input for any machine, no matter how intelligent it is or might ever become.

A matter of quality

We have thus come to the crux of our position: the primary source for the training of ML models is data produced during the care of patients—with the exception of test results and diagnostic images—by doctors and other clinical practitioners in their daily routines and tasks. However, the quality of data found in medical records is notoriously far from being perfect, with independent studies consistently finding that approximately 5% of records contain errors (13-15). Focusing on diagnostic imaging, radiological reports might be used for automatic annotation (16), though they can have an even higher error rate (17), or, alternatively, images can be deliberately annotated by a pool of radiologists, though they often show a high degree of discordance in their findings (2). Thus, the quality of ML training data is often lower than the level needed to build reliable models for integration into effective decision support systems; what is worse, ML developers and doctors—as end users of the products—usually neglect or underrate this issue.

That said, few concepts are as intuitively comprehensible and yet academically elusive as that of data quality (DQ) in health records. In fact, the concept of DQ can intuitively and concisely be equated to the concept of fitness for use (18). This is in line with the operational definition of DQ given by The Joint Commission, which equates DQ with the adequacy to support a number of medically relevant tasks, like identifying the patient, supporting the diagnosis, justifying care and treatment, documenting the course and results of treatment, and promoting continuity and safety of care (19). Despite this apparent simplicity, a recent review by Juddoo *et al.* (20), which considered 41 high impact papers, extracted a staggering 43 distinct DQ dimensions relevant to health care applications, of which 38 were mentioned more than once. In their words, “This confirmed the impression of a lack of a universal DQ framework and the possible fact that different authors might be using different jargon to express the same idea”. Accuracy, which Cabitza and Batini (21) defined as “health data [that] represent the truth and what actually happened” was the dimension most frequently mentioned (58 times), followed by completeness, consistency, reliability, and timeliness. These are the dimensions that appeared more than ten times and hence were considered by the authors to be “most important in the context of Big Data within the health industry.”

To try to simplify this complex matter, we could consider three main DQ areas. The first is related to accuracy and reliability (where the latter also includes internal consistency and high inter-rater agreement). The second is in regard to completeness and timeliness (in that missing data can be seen as data that is not yet recorded and, conversely, having complete but obsolete data is like not having useful data at all). The third is related to comparability and (external) consistency, which is also a matter of interoperability between information systems and the communities of practitioners who use those systems.

Crossing the chasms between clinical practice and ML

In the following, we will address the issue(s) of DQ in healthcare in order to bridge the “last mile” of the ML hiatus. This is the challenge of bringing a trustworthy (datafied) representation of health conditions and care actions to the opposite side of this chasm or, in other words, the challenge of engineering a workflow to develop an ML-based AI that supports medical decision making—a composite industrial process encompassing various steps, among which are the engineering of the above

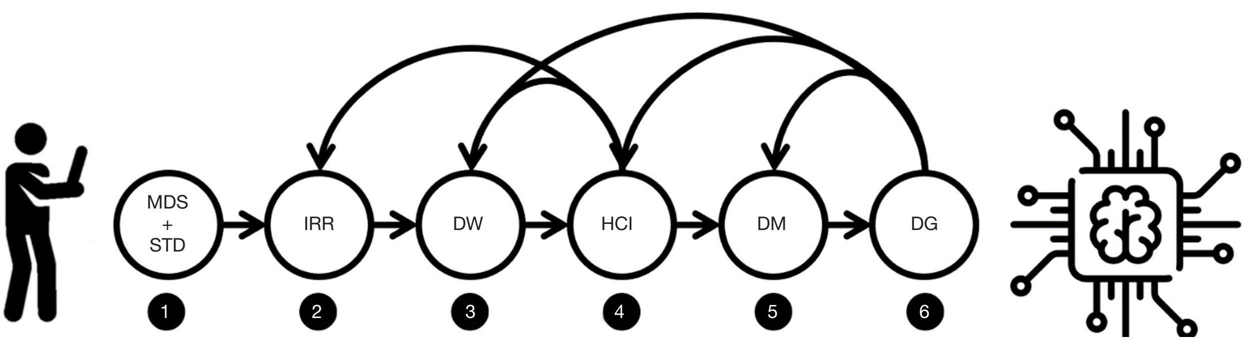


Figure 2 The main factors either bridging or separating clinical work and AI development. Oriented arcs represent strong influence.

representation, the training of predictive models, and their testing and validation.

The main challenges to which we refer relate to the following areas of concern (see *Figure 2*):

(I) A lack of uniformity and consensus on (i) what to record (MDS in *Figure 2*, for minimum data set) and (ii) how to record it (STD, for standards);

(II) The phenomenon of observer variability (IRR, for inter-rater reliability), which is the extent to which multiple raters disagree on how to classify, and hence codify, a given medical phenomenon;

(III) Limitations on doctors' capacity to record and communicate information; this also includes limited education of processing staff (both clinical and administrative), especially with respect to awareness of the consequences of poor DQ (DW, for data work);

(IV) Poorly designed data collection tools—both paper-based and electronic interfaces (HCI, for human-computer interaction);

(V) No single (or central) repositories for vetted and anonymized data for use at scale for secondary purposes, like research and ML training and validation (DM, for data management); and

(VI) Lack of planning (or will to plan) by administrative and managerial staff and higher policy makers with respect to DQ assessment and continuous improvement (DG, for data governance).

As the reader may notice, the problems we have mentioned are all socio-technical in nature, with inter-rater reliability (IRR), data work, and data governance being primarily socio-organizational areas of concern, and HCI and data management being mainly technical. Yet, for the very first concern—achieving consensus on minimum data sets and the related classification schemas—both areas are inextricably intertwined, and distinguishing between them is more futile than in other cases. Let us quickly review each of the concerns.

We can begin with an expression of optimism: the concern related to the use of standards (STD in *Figure 2*) to establish how to report a medical condition (i.e., how to code it) has affected the healthcare environment for decades and still prevents many centers from exchanging information or documents. The impact of this on the whole sequence of challenges of the ML hiatus should be gradually declining due to the increasing adoption of coding standards—such as ICD-10, SNOMED, and LOINC—that have passed the test of time and reached sufficient maturity (including technological maturity) and their integration into current electronic medical records and hospital information systems. In particular, the latter two coding systems are the most widely used terminology standards to date for health measurements, observations, and documents (22) and their adoption appears to be a growing trend.

The problem of the minimum data set (MDS in *Figure 2*) —what to report—is more complex and relates to the challenge of identifying all the relevant attributes of a clinical condition that, in the case of AI training, are good predictors of the target variable. In this regard, both ML developers and clinical practitioners can help each other. The former can employ the largest number of attributes (or features) available (or even conceivable) in a well-circumscribed experimental setting to train a set of ML models on specific relevant targets (like classifying cases in terms of either the associated diagnosis or a stratum of expected improvements or outcomes) and then perform a quantitative feature ranking [or feature selection (23)] to determine the most useful *N* features for each predictive task at hand. The union set of these features could then be indicated (including in terms of iconic or graphical signs in the user interface of the electronic record) as being the data that it is recommended to report as carefully and accurately as possible for the secondary use of the data (including, but not limited to, AI development).

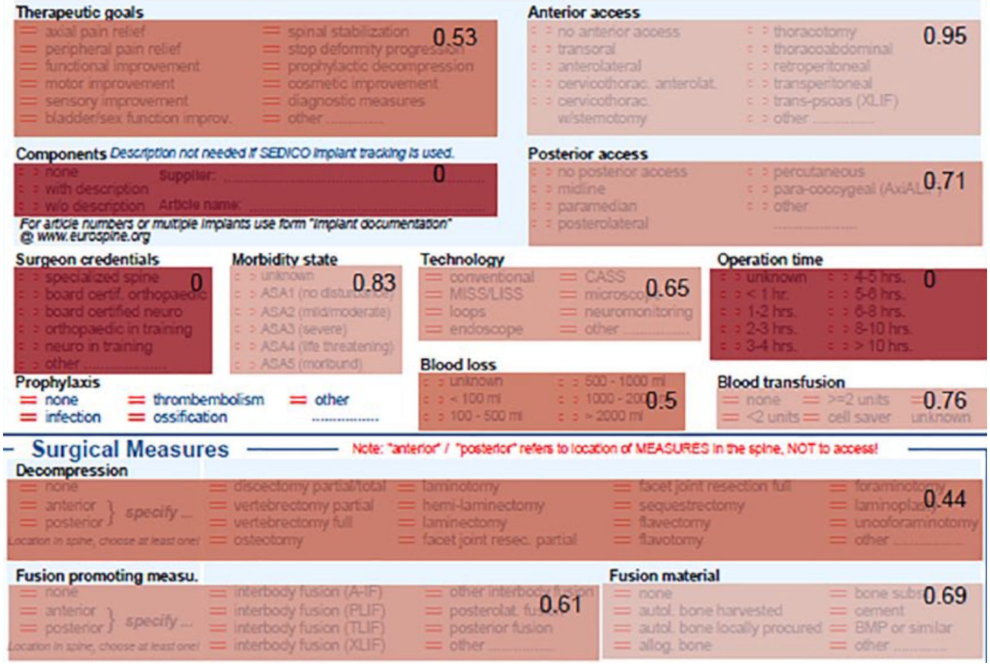
In a similar but complementary fashion, a clinical study could be designed and conducted to review a sufficient number of retrospective cases that are adequately representative of relevant conditions; when multiple raters agree upon what data affected (or would have affected) the right decision at the right time, the study could identify the necessary data without which most of the cases (e.g., 80%) would not be managed appropriately or timely—in other words, the minimum data set with the highest impact on the patients' outcomes. We could call this set the minimum pragmatic data set, which is still lacking for many clinical specialties. In both cases, the common idea is to adopt the motto "less is more" (24), translated into the DQ and ML fields as "less (but good) data is more data." In so doing, we would fully recognize that the doctors' responsibilities cannot be further expanded with new and more intensive reporting tasks (see also problem no. 3) and, perhaps more importantly, that medical data should not be treated as any other type of data and that its quality requirements cannot be borrowed from other domains. As a paradigmatic example of this realization, de Mul and Berg (25) reported a convincing case in which missing data did not necessarily indicate a DQ problem, but rather the occurrence of conditions that practitioners deemed not necessary to document (the "all is well" situation), thus shedding light on the unsuitability of establishing (and enforcing algorithmically) requirements for data completeness that are expressed in terms of some fixed threshold, as is common in other fields, such as administration.

Moving to the next concern: even if a community of specialists agree upon what and how to report, data could still be unreliable whenever more than one clinician is involved in its production (as should be the case for data for ML training) and these practitioners do not agree on how to report the same clinical phenomenon. This is a well-known situation in the medical literature (26,27), which has been studied for almost a century, variously referred to as observer variability, inter-rater agreement, or IRR (see *Figure 2*). In Cabitza *et al.* (28), we contributed to raising renewed awareness of the potential distortion that IRR, which is intrinsic to and probably ineradicable from the interpretation of medical conditions, can induce in any medical dataset, especially those used to train ML models.

To address this factor, we proposed further investigating the viability and efficacy of some socio-technical solutions, which we can also relate to the HCI concern and to the solution we proposed above for raising awareness of the importance of careful completion of selected fields of the record. In particular, we proposed highlighting the fields that presented low IRR scores during adoption, in a way that is not too different from that shown in *Figure 3* (28). In this solution, IRR scores can be computed on the basis of a small user study—or even at regular intervals—by asking two or more clinicians to fill in the same data on a random basis and computing this score on the fly.

Data work (DW in *Figure 2*) is a recent expression (29) that was introduced to cover all the tasks that doctors and nurses perform to document care and coordinate with each other (30) and that produce (and consume) medical data. The concerns with this kind of work (ontologically different from care) are related to excessive paperwork, with the consequent frustration and alienation of health practitioners, possibly leading to potentially serious consequences for the quality of care and health of their patients (31). Thus, if we just assume the limitations of doctors with respect to DQ (and avoid treating it as either their fault or an organizational failure), we can support data work in several ways—partly organizationally and partly technologically. As an example of a radical solution of the former type, we suggested relieving doctors from directly using data collection tools and flanking them with medical scribes (29). These would be "non-licensed health care team members that document patient history and physical examination contemporaneously with the encounter" (32) and who are trained in transcribing the doctors' orders and notes as well as in describing medical cases in standardized and more consistent and comparable ways. Another example is the technological counterpart of this organizational solution, the so-called virtual scribes. This term can denote either the outsourcing of the medical scribing service (33) or, less frequently (and together with the alternative term digital scribes), the full automation of this service through AI systems that perform speaker diarization (understanding who spoke when), speech recognition, named entity (or knowledge) recognition, and the processing of structured data (32). These would be, in short, a sort of specialized AI that fills in the electronic medical record autonomously and with reduced effort on the part of the medical staff (who are involved in vetting the AI output).

We acknowledge that both the above solutions require additional resources (both human and economic), but we can here paraphrase the famous quotation often



Therapeutic goals 0.53
☐ axial pain relief
☐ peripheral pain relief
☐ functional improvement
☐ motor improvement
☐ sensory improvement
☐ bladder/sex function improv.
☐ spinal stabilization
☐ stop deformity progression
☐ prophylactic decompression
☐ cosmetic improvement
☐ diagnostic measures
☐ other ....

Anterior access 0.95
☐ no anterior access
☐ transoral
☐ anterolateral
☐ cervicothorac. anterolat.
☐ cervicothorac. w/sternotomy
☐ thoracotomy
☐ thoracoabdominal
☐ retroperitoneal
☐ transperitoneal
☐ trans-psoas (XLIF)
☐ other ....

Components *Description not needed if SEDICO implant tracking is used.* 0
☐ none
☐ with description
☐ w/o description
Supplier: ....
Article name: ....
For article numbers or multiple implants use form "Implant documentation" @ www.eurospine.org

Posterior access 0.71
☐ no posterior access
☐ midline
☐ paramedian
☐ posterolateral
☐ percutaneous
☐ para-coccygeal (AxiAL
☐ other ....

Surgeon credentials 0
☐ specialized spine
☐ board certif. orthopaedic
☐ board certified neuro
☐ orthopaedic in training
☐ neuro in training
☐ other ....

Morbidity state 0.83
☐ unknown
☐ ASA1 (no disturbance)
☐ ASA2 (mild/moderate)
☐ ASA3 (severe)
☐ ASA4 (life threatening)
☐ ASA5 (moribund)

Technology 0.65
☐ conventional
☐ MISS/LISS
☐ loops
☐ endoscope
☐ CASS
☐ microscop
☐ neuromonitoring
☐ other ....

Operation time 0
☐ unknown
☐ < 1 hr.
☐ 1-2 hrs.
☐ 2-3 hrs.
☐ 3-4 hrs.
☐ 4-5 hrs.
☐ 5-6 hrs.
☐ 6-8 hrs.
☐ 8-10 hrs.
☐ > 10 hrs.

Prophylaxis
☐ none
☐ infection
☐ thrombembolism
☐ ossification
☐ other ....

Blood loss 0.5
☐ unknown
☐ < 100 ml
☐ 100 - 500 ml
☐ 500 - 1000 ml
☐ 1000 - 200
☐ > 2000 ml

Blood transfusion 0.76
☐ none
☐ <2 units
☐ >=2 units
☐ cell saver
☐ unknown

Surgical Measures
Note: "anterior" / "posterior" refers to location of MEASURES in the spine, NOT to access!

Decompression 0.44
☐ none
☐ anterior
☐ posterior
} specify ...
Location in spine, choose at least one!
☐ discectomy partial/total
☐ vertebrectomy partial
☐ vertebrectomy full
☐ osteotomy
☐ laminotomy
☐ hemi-laminectomy
☐ laminectomy
☐ facet joint resec. partial
☐ facet joint resection full
☐ sequestrectomy
☐ flavectomy
☐ flavotomy
☐ foraminotomy
☐ laminoplasty
☐ uncoforaminotomy
☐ other ....

Fusion promoting measu. 0.61
☐ none
☐ anterior
☐ posterior
} specify ...
Location in spine, choose at least one!
☐ interbody fusion (A-IF)
☐ interbody fusion (PLIF)
☐ interbody fusion (TLIF)
☐ interbody fusion (XLIF)
☐ other interbody fusion
☐ posterolat. fu
☐ posterior fusion
☐ other ....

Fusion material 0.69
☐ none
☐ autol. bone harvested
☐ autol. bone locally procured
☐ allog. bone
☐ bone subs
☐ cement
☐ BMP or similar
☐ other ....

Figure 3 A standard form to collect surgery data, with indications of IRR for each field (the darker the red, the lower the agreement among raters). Adapted from Ref. (28).

misattributed to Derek Bok: “if you think ensuring high DQ is expensive, try low DQ.” In any case, we recall the success of some cost-effective initiatives to improve doctors’ hand hygiene (34) and can envision that similar cognitive-behavioral solutions [“nudges” (35)] could also be applied to data work practices in order to spread good practices *of data hygiene*. These solutions would likely be more effective than mere economic incentives or disciplinary sanctions, although their effectiveness in the long term may be uncertain and make them of limited sustainability.

The concerns regarding the quality of the electronic data collection tools—their low usability and poor HCI (see *Figure 2*)—have several implications, including for medical errors, patient safety, and clinician burnout (36). However, the design of structured and orderly graphical interfaces has also been found to have a positive role in improving DQ for ML training; for instance, Pinto Dos Santos *et al.* (37) provided proof of the concept that data extracted from structured reports written during clinical routines can be used to successfully train deep learning algorithms. While a plea to improve the usability of the interfaces of electronic medical records would hardly be considered inappropriate, we recognize the difficulties inherent in its realization. Nevertheless, we believe it is important that scholars talk about these problems, that research is done into the role of human factors in the performance of doctors (38), and that healthcare stakeholders become more aware of the opportunities to improve medical AI not only in terms of its accuracy, but also in terms of the usability of the systems through which we interact with it and, ultimately, in terms of the satisfaction of its users.

The concern about data management (DM in *Figure 2*) is the most technical one of those mentioned so far; in this regard, we can observe the wider diffusion and stronger reliability of third-party cloud storage solutions in which health facilities can store the data they produce in a secure and safe environment, often maintained by dedicated staff using state-of-the art equipment. This is justified not only by cost savings and economies of scale, but also by the more robust infrastructure against malicious threats like data poisoning (39), in which an adversary injects bad data into a model’s training dataset to get it to learn something that could make it vulnerable (attack its integrity) or inaccurate for a particular input (attack its availability or usefulness), or against adversarial attack, in which an adversary changes the input (e.g., by adding random pixels to a diagnostic digital image) to prevent the system from classifying the resulting input (without the knowledge of the healthcare provider). Recent events have also raised the awareness of stakeholders, managers, and policy makers of data management risks and made it clear how the greater dependence on technology that AI induces—precisely because of its quality and potential—is also mirrored by greater vulnerability and fragility of the health system as a whole.

Finally, and related to this latter point, the greatest organizational concern is the last one mentioned: data governance (DG in *Figure 2*). We emphasize the difference between data management, which is a set of practices around the good operation of an information system and the adequate quality of its data flows, and data governance. This is a term denoting a strategic attitude toward the information assets "under the control of a hospital or health system [which encompasses] all policies and procedures to guide, manage, protect, and govern the electronic information" (40).

We consider this factor at the very end of the ML hiatus, as we regard it as the "last yard" of the path from the point of care to the entrance to the ML development pipeline, although its influence, as can be seen in *Figure 2*, can be easily traced back to almost all of the steps preceding it. In fact, all of the previous elements can be set and aligned to bring high quality data to the ML development pipeline, but if healthcare facilities do not exert full data governance over this flow and process—including governance of the processes by which predictive models are created, validated, updated, and applied to new cases on the basis of daily needs and routine—the ML hiatus depicted in *Figure 1* might be closed for a while, but it will open up again, sooner or later, under the attacks of cyber-hackers or just the erosion caused by the drift of practices and the passage of time.

Final remarks

In this contribution, we have focused on the socio-technical elements that we recognize must all line up to allow for the deployment of potentially effective AI in real-world clinical settings. Rather than focusing on the theoretical performance and accuracy of medical AI, which is a rather new and surprising concern that computer scientists seem to have passed on to doctors, we have shed light here on a still relatively neglected and underrated set of concerns regarding the quality of the data that is used to train and adjust the AI algorithms to fit the situated needs of a community of health practitioners.

At this point, the famous adage comes to mind, which appears whenever ML and DQ are near each other: "garbage in, garbage out". This expression refers to the fact that, no algorithm, no matter how smart or intelligent it is, can produce value if its input lacks value in the first place. However, the situation in healthcare is unfortunately worse than this common engineering phrase might suggest in other, less critical, domains. In fact, if inadequate input is used to optimize the performance of a decision support system, yet remains undetected, and thus the input is not appropriately improved nor the "support" discarded, but instead erroneously considered truthful and fit for purpose, the resulting garbage output risks being viewed as the proper advice of an accurate tool. A unreliable indication may then be made more "objective" and indisputable thanks to the armor of algorithmic legitimacy that we tend to ascribe to this class of machines, and we eventually risk allowing experts to be misled in more complex decisions by this garbage-in-disguise output and risk novices being deskilled in what should be easy decisions (41). Thus, before AI can unleash its full potential to help practitioners deliver better—and more human—care, we must, when the implementation chasm is closed and human and machine intelligence converge (42), build robust bridges that close both the sides of the machine and human hiatuses. This requires a full range of interventions, both organizational and technical, which alone would be either over ambitious or useless, together with the awareness that the accuracy of any technological support is nothing in medicine without the power of doctors to use it to the best of their knowledge and judgment: in a word, responsibly.

Acknowledgments

The authors are grateful to the Scientific Department of the IRCCS Galeazzi, and to its head, prof. Giuseppe Banfi, for their continuous and unconditional support.
Funding: None.

Footnote

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at http://dx.doi.org/10.21037/atm.2020.03.63). FC serves as an unpaid editorial board member of *Annals of Translational Medicine* from Apr 2018 to Mar 2020. The other authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

References

1. Faes L, Liu X, Kale A, et al. Deep Learning Under Scrutiny: Performance Against Health Care Professionals in Detecting Diseases from Medical Imaging-Systematic Review and Meta-Analysis. Available online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3384923
2. Cabitza F, Zeitoun JD. The proof of the pudding: in praise of a culture of real-world validation for medical artificial intelligence. Ann Transl Med 2019;7:161.
3. Coiera E. The Last Mile: Where Artificial Intelligence Meets Reality. J Med Internet Res 2019;21:e16323.
4. Cabitza F. Biases Affecting Human Decision Making in AI-Supported Second Opinion Settings. In: Torra V, Narukawa Y, Pasi G, et al. editors. Modeling Decisions for Artificial Intelligence. Springer, Cham: Lecture Notes in Computer Science, 2019.
5. Bezemer T, de Groot MC, Blasse E, et al. A Human (e) Factor in Clinical Decision Support Systems. J Med Internet Res 2019;21:e11732.
6. Lyell D, Coiera E. Automation bias and verification complexity: a systematic review. J Am Med Inform Assoc 2017;24:423-31.
7. Merritt SM, Ako-Brew A, Bryant WJ, et al. Automation-Induced Complacency Potential: Development and Validation of a New Scale. Front Psychol 2019;10:225.
8. Cai CJ, Winter S, Steiner D, et al. "Hello AI": Uncovering the Onboarding Needs of Medical Practitioners for Human-AI Collaborative Decision-Making. Proc ACM Hum-Comput Interact 2019;3:article 104.
9. Holzinger A, Langs G, Denk H, et al. Causability and explainabilty of artificial intelligence in medicine. Wiley Interdiscip Rev Data Min Knowl Discov 2019;9:e1312.
10. Fan W, Liu J, Zhu S, et al. Investigating the impacting factors for the healthcare professionals to adopt artificial intelligence-based medical diagnosis support system (AIMDSS). Ann Oper Res 2018. Available online: https://doi.org/10.1007/s10479-018-2818-y
11. Cabitza F, Ciucci D, 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, Magni M. editors. Organizing for the Digital World. Springer, Cham: Lecture Notes in Information Systems and Organization, 2018.
12. Graber ML. The incidence of diagnostic error in medicine. BMJ Qual Saf 2013;22:ii21-ii27.
13. Gillies A. Assessing and improving the quality of information for health evaluation and promotion. Methods Inf Med 2000;39:208-12.
14. Huaman MA, Araujo-Castillo RV, Soto G, et al. Impact of two interventions on timeliness and data quality of an electronic disease surveillance system in a resource limited setting (peru): a prospective evaluation. BMC Med Inform Decis Mak 2009;9:16.
15. Liaw ST, Rahimi A, Ray P, et al. Towards an ontology for data quality in integrated chronic disease management: a realist review of the literature. Int J Med Inform 2013;82:10-24.
16. Zech J, Pain M, Titano J, et al. Natural language-based machine learning models for the annotation of clinical radiology reports. Radiology 2018;287:570-80.
17. Bruno MA, Walker EA, Abujudeh HH. Understanding and confronting our mistakes: the epidemiology of error in radiology and strategies for error reduction. Radiographics 2015;35:1668-76.
18. Wang RY, Strong DM. Beyond accuracy: what data quality means to data consumers. J Manag Inf Syst 1996;12:5-33.
19. World Health Organization. Regional Office for the Western Pacific. (2003). Improving data quality: a guide for developing countries. Manila: WHO Regional Office for the Western Pacific.
20. Juddoo S, George C, Duquenoy P, et al. Data governance in the health industry: investigating data quality dimensions within a big data context. Appl Syst Innov 2018;1:43.
21. Cabitza F, Batini C. Information quality in healthcare. In: Batini C, Scannapieco M (eds). Data and Information Quality Dimensions, Principles and Techniques, Chapter: 13. Springer, 2016.
22. McKnight J, Wilson ML, Banning P, et al. Effective coding is key to the development and use of the WHO Essential Diagnostics List. Lancet Digital Health 2019;1:e387-8.

23. Cheng TH, Wei CP, Tseng VS. Feature selection for medical data mining: Comparisons of expert judgment and automatic approaches 19th Proc IEEE Int Symp Comput Based Med Syst (CBMS'06), 2006.
24. Grady D, Redberg RF. Less is more: how less health care can result in better health. Arch Intern Med 2010;170:749-50.
25. de Mul M, Berg M. Completeness of medical records in emergency trauma care and an IT-based strategy for improvement. Med Inform Internet Med 2007;32:157-67.
26. Khan L, Mitera G, Probyn L, et al. Inter-rater reliability between musculoskeletal radiologists and orthopedic surgeons on computed tomography imaging features of spinal metastases. Curr Oncol 2011;18:e282-7.
27. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb) 2012;22:276-82.
28. Cabitza F, Locoro A, Alderighi C, et al. The elephant in the record: on the multiplicity of data recording work. Health Informatics J 2019;25:475-90.
29. Bossen C, Chen Y, Pine K. The emergence of new data work occupations in healthcare: The case of medical scribes. Int J Med Inform 2019;123:76-83.
30. Berg M. Accumulating and coordinating: occasions for information technologies in medical work. Comput Suppport Coop Work 1999;8:373-401.
31. 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 Aff (Millwood) 2017;36:655-62.
32. Lorenzetti DL, Quan H, Lucyk K, et al. Strategies for improving physician documentation in the emergency department: a systematic review. BMC Emerg Med 2018;18:36.
33. Benko S, Idarraga A, Bohl DD, et al. Virtual Scribe Services Decrease Documentation Burden Without Affecting Patient Satisfaction: A Randomized Controlled Trial. Foot & Ankle Orthopaedics 2019. doi: 10.1177/2473011419S00105.
34. Caris MG, Labuschagne HA, Dekker M, et al. Nudging to improve hand hygiene. J Hosp Infect 2018;98:352-8.
35. Devisch I. Progress in medicine: autonomy, oughtonomy and nudging. J Eval Clin Pract 2011;17:857-61.
36. Khairat S, Coleman C, Newlin T, et al. A mixed-methods evaluation framework for electronic health records usability studies. J Biomed Inform 2019;94:103175.
37. Pinto Dos Santos D, Brodehl S, Baeßler B, et al. Structured report data can be used to develop deep learning algorithms: a proof of concept in ankle radiographs. Insights Imaging 2019;10:93.
38. Heponiemi T, Kujala S, Vainiomäki S, et al. Usability Factors Associated With Physicians' Distress and Information System-Related Stress: Cross-Sectional Survey. JMIR Med Inform 2019;7:e13466.
39. Jagielski M, Oprea A, Biggio B, et al. Manipulating machine learning: Poisoning attacks and countermeasures for regression learning 2018 IEEE Symp Secur Priv. Available online: https://arxiv.org/pdf/1804.00308.pdf
40. Reeves MG, Bowen R. Developing a data governance model in health care: although the term may be unfamiliar, data governance is a longstanding obligation of the healthcare industry. Healthc Financ Manage 2013;67:82-6.
41. Cabitza F, Rasoini R, Gensini GF. Unintended consequences of machine learning in medicine. JAMA 2017;318:517-8.
42. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019;25:44-56.

Cite this article as: Cabitza F, Campagner A, Balsano C. Bridging the "last mile" gap between AI implementation and operation: "data awareness" that matters. Ann Transl Med 2020;8(7):501. doi: 10.21037/atm.2020.03.63