Editorial Commentary

The proof of the pudding: in praise of a culture of real-world validation for medical artificial intelligence

Federico Cabitza[1,2], **Jean-David Zeitoun**[3,4]

[1]IRCCS Istituto Ortopedico Galeazzi, Milano, Italy; [2]University of Milano-Bicocca, Milano, Italy; [3]Centre d'Epidémiologie Clinique, Hôtel Dieu Hospital, Assistance Publique-Hôpitaux de Paris, Paris, France; [4]Gastroenterology and Nutrition, Saint-Antoine Hospital, Assistance Publique-Hôpitaux de Paris, Paris, France

Correspondence to: Federico Cabitza. University of Milano-Bicocca, Viale Sarca 336, Milano, Italy. Email: federico.cabitza@unimib.it.

Submitted Feb 06, 2019. Accepted for publication Feb 20, 2019.
doi: 10.21037/atm.2019.04.07
View this article at: http://dx.doi.org/10.21037/atm.2019.04.07

Artificial Intelligence (AI) in healthcare has become a quasi-normal subject (1). In the last few years, there has been an impressive increase in the number of publications concerning the application of machine learning (ML), a set of techniques and models for building data-driven AI systems, to medical tasks, such as the diagnosis, prognosis and anticipation of treatment effects and complications (2). Several studies and clinical trials have been conducted in virtually all medical specialties, including oncology (3), cardiology (4) and ophthalmology (5), often reporting machine performance on par with or even superior to human performance in both diagnostic and prognostic tasks (2). Due to its allegedly superhuman accuracy and black-box nature (6), we refer to this new generation of AI as oracular to distinguish it from traditional rule-based expert systems and decision support systems. The US Food and Drug Administration (FDA) has recently approved quasi-oracular AI products for detecting atrial fibrillation (AliveCor, Apple Watch), diabetic retinopathy (IDx), cancer (Arterys) and wrist fractures (Imagen). However, the barrier for entry of these algorithms has been low, and the FDA's fast-track approval plan published in 2018 (2) is seen by some as a sign of the inadequacy of current regulatory standards (7). Two years after AI systems (based on ML) were first recognized as riding "atop the peak of inflated expectations" (8), we can now behold from a distance what lies beyond this peak, what still separates us from the "plateau of productivity"[1] and what will attract major interest from multiple stakeholders: validation. Validation is a commonly used term that, unfortunately, is associated with the same meaning in different scientific communities (e.g., medicine and AI). In its broadest terms, we view AI validation as *proof that these systems can and do deliver value consistently*[2] and, more prosaically, "live up to their (vendors') claims" (9). What is usually referred to as the "clinical validation" or "prospective validation" of AI systems is urgent in that it has been "high on promise and relatively low on data and proof" (2). Indeed, a recent paper (10) analysed 516 published studies (in the first six months of 2018) reporting the performance of AI algorithms for the diagnostic analysis of medical images and found that only six percent (31 studies) externally validated their algorithms.

From this perspective, as authors belonging to the two different and yet increasingly closer communities mentioned above, we aim to characterize this concept of validity further and increase awareness of the complexities

[1] Here the reference is to Gartner's Hype Cycle for Emerging Technologies, a diagram in which technology applications are interpreted in terms of how expectations regarding their impact change over time, passing through distinct phases, from a steady increase that saturates at a peak and then decreases with similar steepness through a shallow trough of disillusion, finally settling on a more constant and reasonable plateau of productivity.

[2] This general definition is compatible with the definition of validation used in the software engineering field, which is an evaluation of a system based on the satisfaction of specified requirements (cf. IEEE Std 610.12-1990 R2002), if we agree that the main requirement of such a system is to provide value and be helpful for doctors and patients.

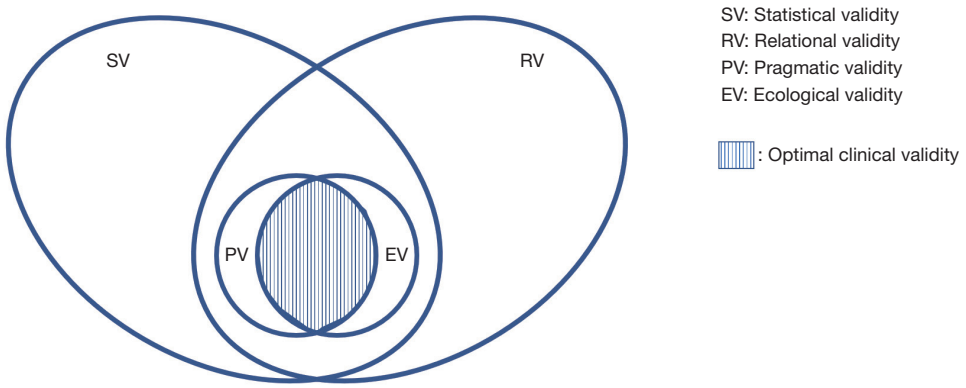

Figure 1 Types of validity.

intrinsic in its evaluation. As stated earlier, the requirement for further "clinical validation" of AI is now legitimate and widespread, but we propose viewing it as an emerging property of nested validation tasks, accomplished from different yet complementary perspectives (*Figure 1*), aimed at defining different kinds of validity. Among these, we distinguish between statistical validity, relational validity, pragmatic validity and ecological validity. To better understand the characteristics of these four kinds of validity, we must first take a step back.

Different kinds of validity

Two main approaches can be considered to validate a computational system. The first is objectivistic, which is focused on technology as an active agent, and the other is consequentialist, which is focused on the effects of technology on a given setting. The former attaches to AI systems the nature of objects that are characterized by properties (e.g., accuracy and reliability), and these are in turn considered intrinsic and essential to the systems and therefore susceptible to quantitative and objective evaluation. The latter approach considers AI systems as inseparable from—and not really distinct from—their actual use, that is, as computational processes performed within human practices of collaborative care and decision-making. In turn, these latter practices can be assessed either in terms of their unfolding (or performance) or their consequences (e.g., the impact they have on the user's work) and, even more importantly, on patients' health. The objectivistic approach is by far the most common in the computer science and engineering disciplines, as it is grounded in the cognitivist model of computation (11), and it is employed in most (if not all) of the latest studies aimed at validating AI systems in medicine. According to this approach, the validity of AI systems is based on accuracy, sensitivity or other similar error-based measures [e.g., the positive predictive value or the area under the receiver operating characteristic (ROC) curve]. This is what we call statistical validity. This kind of validity entails being proved valid with respect to a numerical, statistical threshold (usually the best state-of-the-art algorithm, posed as a benchmark) under well-specified experimental conditions. Unfortunately, the AI research is pervaded by terminological confusion that does not spare the concept of validation (12), making it difficult to know whether or not what has been reported by a research group is in fact sound (statistical) validation. As odd as it might seem, such validation is not accomplished during the phase called cross-validation. Cross-validation is often performed on a subset of the training data (called, not without irony, a validation set) for a twofold purpose: to obtain an average estimate of the system prediction performance (or skill, as it is called within the AI community) in different cases and to tune some parameters of the system's model (e.g., the number of hidden units in an artificial neural network) before training it on all the available data. Although performing cross-validation on mutually exclusive partitions of the validation dataset (instead of on a single holdout set) and reporting the related metrics is a recommended—and often neglected—practice in medical AI papers (2,13), the final proof of the system's validity should be given in terms of an unbiased estimate of the system's skill, which is obtained by testing the system on a "test dataset" (i.e., a set of cases that were originally held back from the available data and were used neither to train nor to tune the system's model).

Whatever way it is statistically proven, this kind of validity is not sufficient for the clinician to responsibly rely

upon or for the health manager to procure and the payer to reimburse, as it completely lacks proof of efficacy, safety and other clinically relevant concepts. The question becomes, to effectively support medical decision-making, what is an acceptable level of accuracy that a technology should exhibit? Some doctors might appreciate receiving correct advice from a computerized decision aid seven times out of ten (with the conviction that they will be able to recognize the wrong three cases and find reassurance and confirmation in the other seven); however, others may demand better performance. Paradoxically, a more accurate AI system is not necessarily better. This is because the AI's "tremendous predictive power" (7) might lead doctors to over-rely on its advice, even when it is wrong. Hence, the doctors might not be fully vigilant, and they could fail to consider all the available evidence or to seek out alternative evidence. In this way, "oracularity" could affect decision-making and lead to automation bias (14). To go beyond single-figure metrics (e.g., the F1 score and the area under the ROC curve, which are commonly used to express statistical validity), some doctors might wish to distinguish between specificity (the capability of the machine to not induce doctors to commit a "false positive" error) and sensitivity (the machine's capacity to contribute to avoiding "false negative" errors) and "choose" the operating point at which the machine should work to exhibit the safest (sensitive) behaviour, without inducing unspecific overuse (15). For instance, in (16) Corey and colleagues describe an oracular AI (appropriately called Pythia), which, based on the patient's age, race, sex, medication and comorbidity history, is able to determine the risk of morbimortality after surgery. At a sensitivity level of 0.75, one out of three patients flagged by Pythia experiences a postsurgical complication within 30 days and hence could benefit from targeted enhanced assessment and management. However, the other two patients would receive unnecessary, and potentially harmful, treatments.

This example, among many others, suggests that, even if we defined a sound benchmark for statistical validity for a class of systems, it would likely not be a sufficient condition for assessing the usability of AI systems. In this case we speak of relational validity. We choose a purposely novel term in order to increase the awareness of aspects that are usually neglected and which we consider to have equal importance to let clinical validity emerge (*Figure 1*). With the term relational validity, we refer to the extent to which physicians *can relate to the AI*, attach some clinical meaning to its advice and integrate its use in their daily workflows and routines. Relational validity is more difficult to assess than statistical validity due to the difficulty in setting objective requirements. Relational validity requires the oracular AI to be explainable and interpretable (17). It must be open to the doctor's scrutiny so as to address multiple questions: in which cases is it more likely to fail? The most complex ones (whatever this means)? The less frequent and unusual ones? Those that involve patient groups that are not adequately represented in the training data of the system? In this latter case, will the AI reproduce or even corroborate the (sampling) bias possibly hidden in the training data? (18). Or rather, will the AI be unable to provide an accurate diagnosis in any case where some data are missing, incomplete or less than totally accurate or where the observer variability is high (19)?

The above reference to usability is no coincidence, as addressing these questions requires that doctors are able to interact with the decision support, undertake counterfactual analysis (20), compare local surrogate models and inspect the relative importance of specific features (predictors) and patients for the prediction produced; in other words, their ability to get (and remain) "in the loop" (21). Evaluating the usability of AI systems requires going beyond the concept of accuracy that statistical validity refers to and also assessing these systems in terms of security, efficiency and satisfaction. Security, broadly speaking, involves protection against deliberate incidents. Efficiency is related to the resources an AI system consumes to deliver its main function and hence to the extent to which its adoption in a real setting can actually yield higher throughputs (with the resources employed for service provision being equal), relevant savings (with the service level guaranteed by the healthcare provider being equal) or less administrative paperwork for the clinicians (which is unproductive). This is a condition that is usually (and optimistically) associated with improved opportunities for clinicians to spend more time with patients (22). Perhaps more realistically, it will contribute to reaching the higher throughput mentioned above. Last but not least, satisfaction can regard either the direct users of a medical AI (i.e., physicians, nurses) or the indirect ones (e.g., patients, who can be involved in providing useful data for the training of prognostic models in terms of how they feel, as with the Patient-Reported Outcome) (23). The satisfaction of different users is usually intertwined (24) and is closely related to both effectiveness and efficiency. However, it is much more neglected than the other dimensions explicitly mentioned in standard glossaries of usability (cf. ISO 9241-11:2018). In fact, the patient experience has been

found to correlate with outcome in complex ways (25,26), and, when less than optimal, it has been found to have a negative effect on treatment adherence and patient anxiety. In turn, a negative physician experience is associated with error-prone behaviours, such as alert fatigue (27) and burnout (24), which several recent studies relate to the increased paperwork and information overload often associated with health information technologies (28-31). Statistical and relational validity are often disjointed ideals, but we are interested in systems that can exhibit both these kinds and provide the right conditions for users to exploit their accurate performance.

However, it is worth noting that relational validity, unlike statistical validity, is not an intrinsic feature of a technological system. Rather, it emerges from the interaction between the system and its users within a situated context: it is "quality in use" (32) and "fit to purpose" (33). Thus, we move from the intrinsic characteristics of an object, which are not different in different settings (provided that a clear, reproducible and immutable standard is also given), to those that characterize a system in a specific context, utilizing a common and shared definition of success (i.e., optimal outcome) that is agreed upon by all the stakeholders involved (34) but would not necessarily hold outside that context.

Can AI work? Does AI work?

The local scope is the main common element between statistical and relational validity. However, while it allows for sound and accurate measurement, it is also their main shortcoming. In fact, both statistical and relational validity address the first question that Haynes once considered (35) for the testing of healthcare interventions: "can it work?" Statistical validity regards a more specific question, "can it work effectively?" while relational validity regards questions like "can it work efficiently, or in a fair way?" and the like. A positive answer to this question means that validity has been demonstrated in at least one controlled experimental setting or in one single real-world setting, where pipelines are deployed that ingest requests, pre-process and cure electronic medical record (EMR) data and allow inference to be run at scale. Proving the good statistical performance of a system by training it on clinical data from multiple settings and gathering real-world data, which are engineered and cured in a feasible and sustainable manner in routine operations, is an important step toward guaranteeing external (statistical) validity, or what we call pragmatic validity, mirroring the idea of pragmatic, real-world trials. Commentators often equate pragmatic validity to clinical validity, while we see it as a kind of external validity (34), which must still be complemented by an external relational validity (i.e., relational validity observed in multiple and heterogeneous settings), or what we call ecological validity (*Figure 1*). Ecological validity regards the impact of a technology not only on strictly clinical (e.g., outcome and care) or workflow- and productivity-related aspects but also on the overall social context, such as career prospects, occupational hazards and salaries for those working with the technology. In this respect, the requirements (against which to match a system) are more difficult to pinpoint (e.g., in the case of deskilling and fairness). Nevertheless, recognizing the importance of ecological validity for medical AI requires considering the impact of AI on workers' skills (avoiding the simplistic idea of augmentation) and defining an "Algorithmic Impact Assessment" model. Such a model should take into account related equality and human rights laws, particularly with regard to discrimination, and also assess the AI in terms of compliance with third-party audit certifications (e.g., ORCAA[3]), guidelines and recommendations (e.g., the WLinAI network[4]). Regarding augmentation, for instance, this seems plausible in terms of the cognitive skills that are related to sign interpretation and to rational reasoning, as pointed out by Obermeyer *et al.* (36). However, this sort of "rational" augmentation could exacerbate the reliance on imaging and laboratory tests, thus raising concerns about its impact on patient safety (37), clinical skills (38) and costs related to overuse. Moreover, emotional, interpersonal and linguistic skills are also universally recognized as important qualities for a caring and effective practitioner and part of the "caring" ecology. Several studies have shown how communication with patients—and even colleagues—deteriorates due to information technology (39,40). We cannot predict how AI will impact group dynamics, communication patterns and decision-making processes or whether AI, being heavily grounded in big data, will reinforce the ideal of the "quantified patient" (41), according to which the

[3] http://www.oneilrisk.com/

[4] http://womenleadinginai.org/category/wlinai-network

measurable aspects of an illness are more important than the context-dependent and existential ones. Promises of AI to relieve doctors from administrative and documental tasks (e.g., through virtual scribes) (42) have to be balanced with darker predictions. For example, it will also require doctors to expend additional effort to validate more input data and then interpret its pervasive output in order to discern potentially biased results, spurious correlations and confounders (43).

Is AI worth it?

As hinted above, pragmatic and ecological validity relate to the replicability of good statistical and relational results. The need for replication, as a core principle of science experimentation, also explains why US drug regulators require at least two adequate and well-controlled trials to support drug effectiveness (44). However, Coiera *et al.* (45) recently noted that studies in the health informatics fields are seldom replicated, and when they are, the results are often varied. The lack of relevant external replication is not specific to computational science and has already been described in many medical disciplines. In the authors' view, different outcomes between similar studies can obviously relate to different contextual conditions but also to methodological flaws that fail to take context (including necessary implementation changes, existing workload and users' attitudes towards digital innovation) into account. In either case, relying on the results of a single-facility study would be insufficient for assessing the clinical validation of an AI solution.

However, pragmatic and ecological validity go beyond mere replicability. Laboratory (46) or Hawthorne effects can be common in different experimental settings, while pragmatic and ecological validity require testing and validation in real-world conditions and hospital routines. Thus, evaluating pragmatic and ecological validity also requires addressing the final, and most important, question posed by Haynes (35): is AI worth the efforts to obtain, use and maintain it? Obtaining proofs to address this question is obviously the most difficult task of validation. For instance, with regard to effectiveness (pragmatic validity), Moja *et al.* (47) showed that there is little evidence that decision support systems (mostly rule-based ones), when integrated with EMR data, can improve morbidity outcomes and other surrogate endpoints, and there is no evidence that they can affect mortality (or survival). To our knowledge, no study has been published on a specific (ML-based) AI to determine whether it achieves general (i.e., cross-sectional) validity in different contexts (i.e., its local validity is replicated in different settings). This has been done, for example, in the case of computer-aided mammography (48), for which no significant statistical difference was found in comparison to unaided mammography. Importantly, the authors pointed out that, although computer-aided detection (CAD) was not found to be beneficial for mammographical interpretation (cf. statistical and pragmatic validity), it might "offer advantages beyond interpretation, such as improved workflow or reduced search time for faint calcifications" (48), hinting at the concepts of relational and ecological validity. More recently, an ML-based AI system (developed by Google researchers and certified in Europe under the Verily name), which proved to be statistically valid in detecting diabetic retinopathy from retinal fundus photographs (49), has been used at the Aravind Eye Hospital in India (50). This system was trained on clear, unobstructed images of the retina, and the researchers are now struggling to make it valid with lower-quality images (i.e., pragmatic valid) and to integrate AI-based detection into routine care in India (ecological validity). A relatively simple way to assess ecological validity would be to compare the performance-, outcome- and practice-oriented measures (e.g., satisfaction) exhibited by medical teams that adopt the AI support compared to those exhibited by unaided teams (or that use traditional technology) in the same clinical setting (ceteris paribus). This type of comparison is often advocated (43) but seldom performed (51).

A further crucial difference between statistical/relational validity and pragmatic/ecological validity is based not only on replicability but also on sustainability. The former could be certified once and for all through well-designed and well-conducted user studies (possibly of a prospective nature, as advocated by multiple authors) (2,7). The latter, instead, calls for continuous monitoring over time to ensure that the initially valid system continues to deliver net benefits and requires periodic impact assessments and continual monitoring of ethical issues in the settings where the algorithms are used to support medical practice.

In other words, clinical validity, that is, the overall (i.e., statistical + relational) validity of a system that is proven in different clinical settings and is a necessary precondition for its adoption in others, cannot be decoupled from the periodic assessment and continuous monitoring of its appropriateness in clinical practice, as this can change and evolve over time, and of its capability to keep delivering a positive balance between the clinical and other (even

Table 1 Summary of the concepts discussed

	Statistical validity	Relational validity	Pragmatic validity	Ecological validity
Paradigm	Objectivity (the system)	Inter-subjectivity	Consequentialist (with respect to data)	Consequentialist (with respect to work)
Focus	Efficacy	Usability	Effectiveness	Cost-effectiveness (unintended consequences)
Main requirements	Replicability/optimal accuracy	Optimal performance	Better outcomes/noninferiority	Net benefits resilience
Scope	Internal (local lab)	Internal (local lab or real-world setting)	External (mainly cross-sectional)	External (mainly longitudinal)
Standards available	Yes (e.g., ISO 5725)	Yes (e.g., ISO 9241)	Yes (e.g., ISO 14155)	Not yet
Question	Can it work?	Can it work? Does it work?	Does it work?	Is it worth it?

intangible) benefits and the potential risks and costs. Thus, AI is similar to any other medical technology, such as medical devices, diagnostic tests and drugs (52), which are all "expected to do better than harm for a patient with a given indication or set of indications" (53) and for which cost-effectiveness analyses are often proposed and undertaken. In the case of AI, this kind of analysis entails considering alternatives, such as whether or not to invest in AI equipment, choosing from among different ways to integrate AI into existing hospital workflows and considering both tangible and intangible costs (e.g., opportunity costs and those related to the erosion of human responsibility, control and self-determination) (54). To date, the medical community has developed and tested many sound methodologies and techniques for assessing the benefits of healthcare interventions. The transferability of these techniques to AI-driven interventions is a challenge that should be given high priority. This includes the difficulty in isolating the relative advantage of a single technology from its socio-technical environment, the difficulty in isolating the opportunity costs that are related to the underuse of AI below its full potential due to fear, ignorance, misplaced concerns or excessive reactions, which Floridi *et al.* (54) call the "wrong reasons", and, as simple as it may sound, the difficulty in pinpointing the very concept of technology success, which is more a social and situated achievement than a technical, objective and transferrable property (34,55).

In praise of AI technovigilance

We strongly encourage regulators and all relevant stakeholders to keep in mind the different kinds of validation that oracular AI products need to demonstrate before allowing their widespread adoption. To contribute to this effort, we distinguished between different forms of validity (*Table 1*): the statistical validity of AI systems, which is an abstract and context-independent measure of its performance; relational validity, which regards the extent to which the system provides a sample of physicians with meaningful point-of care advice and is usable; pragmatic validity, which regards the capability of the system to perform well with real-world data and in real-world conditions; and ecological validity, which regards the ecological fit of the system within a network of interactions between humans (including the patients), which are seen as joint cognitive systems (56), and therefore the extent to which the system contributes positively to the specific socio-technical agency in which it is embedded and exerts its positive effects continuously.

While statistical validation, which has been the predominant concern so far in the medical AI community, refers to the system isolated from human interaction, the other kinds of validations are all grounded in the interaction of practitioners with AI and must therefore be assessed with the involvement of physicians. Therefore, since these forms of validity regard practices, we advocate that validation be performed in a consequentialist manner. This approach should not be confused with a fatalist one. Outcomes can be predicted; their nature, be it positive or negative, can be assessed, and the outcome likelihood can be either increased or decreased with specific interventions if the net benefits are found to be higher or lower than the costs, respectively. Pragmatic validity should be assessed using established standards that include meaningful endpoints of clinical benefit and appropriate benchmarks (7), while

ecological validations should result in an increasing number of guidelines and best practices being shared across multiple research and hospital settings. This will push the medical community to change its attitude towards validation.

To date, however, none of the studies reviewed in (10) have demonstrated that their methods were indeed ready for clinical use or adopted design features that the authors recommend for robust validation of the real-world clinical performance of AI algorithms: diagnostic cohort design, the inclusion of multiple institutions, verification that the data reflect relevant variations in patient demographics and diseases and prospective data collection for external and independent validation.

In summary, in proposing four types of validity corresponding to different perspectives to evaluate true clinical validity, we do not mean to make a short story long. On the contrary, we are aware that effectiveness is hard to prove in medicine, where only one treatment and intervention out of ten is clearly proven to be beneficial (57). We make the case that other dimensions beyond effectiveness must be considered, and methods besides trials, more pragmatic and grounded in continuous monitoring, must be adopted to guarantee validity over time.

We are not the first researchers to notice important similarities between AI validation and drug validation and to shed light on the importance of considering a form of pharmacovigilance for the "software as a medical device" (SaMD), a concept for which a specific term has been proposed: technovigilance (52,58). We also warn against the simplistic view of taking for granted the effectiveness of surveillance infrastructures in the case of AI. In doing so, we agree with Parikh and colleagues (7), who point out that "unlike a drug or device, algorithms are not static products [as] their inputs […] can change with context". As most editorials and viewpoints end with the great potential of AI for outcome improvements for all, we do not need to reiterate that perspective. Indeed, such optimism is welcome for its role in attracting funding for AI development and creating a positive terrain for its widespread adoption. However, what we advocate now is a culture demanding the responsible assessment of the benefits and costs of AI and the realistic management of the inevitable risks, which should not offset the immense potential of AI or be overlooked—or, worse yet, removed from sight. Most pharmacovigilance advances have been made in reaction to drug accidents. We do not need to wait for serious problems to occur to create a reliable system of AI technovigilance.

Acknowledgements

None.

Footnote

Conflicts of Interest: Dr. Zeitoun is an advisor for several consulting firms in link with pharmaceutical industry (Cepton, Oliver Wyman, Roland Berger, McCann Healthcare, Omnicom, Grey Healthcare, TBWA, Havas). He also reports speaking fees from a manufacturers' professional association, consulting fees from Mayoly-Spindler, Ferring, Pierre Fabre, AbbVie, and Johnson & Johnson, Biogen, Astra-Zeneca, unpaid consultancy for EY. He is a personal investor in approximately 20 digital companies, medtech companies or biotech companies, and as a limited partner in an investment fund. He is also a shareholder and advisory board member in several medtech companies. He reports being cofounder and shareholder of Inato, a digital company involved in clinical research and whose customers are pharmaceutical companies. Prof. Cabitza reports being cofounder and shareholder of Visemio, a digital company involved in the visualization of medical data and whose customers are health care providers. He also reports speaking fees from Watson Towers Willis, MedTronic, Bayer and Sudler, and consulting fees from Deloitte.

References

1. Saria S, Butte A, Sheikh A. Better medicine through machine learning: What's real, and what's artificial? PLoS Med 2018;15:e1002721.
2. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019;25:44-56.
3. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017;542:115.
4. Hannun AY, Rajpurkar P, Haghpanahi M, et al. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. Nat Med 2019;25:65-9.
5. Abràmoff MD, Lavin PT, Birch M, et al. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. NPJ Digit Med 2018;1:39.

6. Lynn LA. Artificial intelligence systems for complex decision-making in acute care medicine: a review. Patient Saf Surg 2019;13:6.
7. Parikh RB, Obermeyer Z, Navathe AS. Regulation of predictive analytics in medicine. Science 2019;363:810-2.
8. Chen JH, Asch SM. Machine Learning and Prediction in Medicine—Beyond the Peak of Inflated Expectations. N Engl J Med 2017;376:2507-9.
9. Oakden-Rayner L, Palmer LJ. Artificial Intelligence in Medicine: Validation and Study Design. In: Ranschaert ER, Morozov S, Algra PR. editors. Artificial Intelligence in Medical Imaging. Cham: Springer International Publishing, 2019:83-104.
10. Kim DW, Jang HY, Kim KW, et al. Design Characteristics of Studies Reporting the Performance of Artificial Intelligence Algorithms for Diagnostic Analysis of Medical Images: Results from Recently Published Papers. Korean J Radiol 2019;20:405-10.
11. Hollnagel E, Woods DD. Joint cognitive systems: foundations of cognitive systems engineering. Boca Raton, FL CRC, 2005:223.
12. Ripley BD. Pattern recognition and neural networks. Cambridge: Cambridge University Press, 1996.
13. Rose S. Machine Learning for Prediction in Electronic Health Data. JAMA Netw Open 2018;1:e181404.
14. Lyell D, Magrabi F, Raban MZ, et al. Automation bias in electronic prescribing. BMC Med Inform Decis Mak 2017;17:28.
15. Komorowski M, Celi LA. Will Artificial Intelligence Contribute to Overuse in Healthcare? Crit Care Med 2017;45:912-3.
16. Corey KM, Kashyap S, Lorenzi E, et al. Development and validation of machine learning models to identify high-risk surgical patients using automatically curated electronic health record data (Pythia): A retrospective, single-site study. PLoS Med 2018;15:e1002701.
17. Miller T. Explanation in artificial intelligence: Insights from the social sciences. arXiv:1706.07269.
18. Mullainathan S, Obermeyer Z. Does Machine Learning Automate Moral Hazard and Error? Am Econ Rev 2017;107:476-80.
19. Cabitza F, Locoro A, Alderighi C, et al. The elephant in the record: On the multiplicity of data recording work. Health Informatics J 2019:1460458218824705.
20. Wachter S, Mittelstadt B, Russell C, et al. Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR. Harv J Law Technol 2018;841-87.
21. Holzinger A. Interactive machine learning for health informatics: when do we need the human-in-the-loop? Brain Inform 2016;3:119-31.
22. Verghese A, Shah NH, Harrington RA. What This Computer Needs Is a Physician: Humanism and Artificial Intelligence. JAMA 2018;319:19-20.
23. Doward LC, Gnanasakthy A, Baker MG. Patient reported outcomes: looking beyond the label claim. Health Qual Life Outcomes 2010;8:89.
24. Panagioti M, Geraghty K, Johnson J, et al. Association Between Physician Burnout and Patient Safety, Professionalism, and Patient Satisfaction: A Systematic Review and Meta-analysis. JAMA Intern Med 2018;178:1317.
25. Kane RL, Maciejewski M, Finch M. The relationship of patient satisfaction with care and clinical outcomes. Med Care 1997;35:714-30.
26. Chen Q, Beal EW, Okunrintemi V, et al. The Association Between Patient Satisfaction and Patient-Reported Health Outcomes. J Patient Exp 2018. doi:10.1177/2374373518795414.
27. Ancker JS, Edwards A, Nosal S, et al. Effects of workload, work complexity, and repeated alerts on alert fatigue in a clinical decision support system. BMC Med Inform Decis Mak 2017;17:36.
28. Ash JS, Sittig DF, Campbell EM, et al. Some unintended consequences of clinical decision support systems. AMIA Annu Symp Proc 2007;26-30.
29. Shanafelt TD, Dyrbye LN, Sinsky C, et al. Relationship Between Clerical Burden and Characteristics of the Electronic Environment with Physician Burnout and Professional Satisfaction. Mayo Clin Proc 2016;91:836-48.
30. Gregory ME, Russo E, Singh H. Electronic Health Record Alert-Related Workload as a Predictor of Burnout in Primary Care Providers. Appl Clin Inform 2017;8:686-97.
31. Kansoun Z, Boyer L, Hodgkinson M, et al. Burnout in French physicians: A systematic review and meta-analysis. J Affect Disord 2019;246:132-47.
32. Abran A, Khelifi A, Suryn W, et al. Usability Meanings and Interpretations in ISO Standards. Software Quality Journal 2003;11:325-38.
33. Dopp AR, Parisi KE, Munson SA, Lyon AR. A glossary of user-centered design strategies for implementation experts. Transl Behav Med 2018. [Epub ahead of print].
34. Ammenwerth E, Gräber S, Herrmann G, et al. Evaluation of health information systems-problems and challenges. Int J Med Inform 2003;71:125-35.

35. Haynes B. Can it work? Does it work? Is it worth it? The testing of healthcareinterventions is evolving. BMJ 1999;319:652-3.
36. Obermeyer Z, Lee TH. Lost in Thought—The Limits of the Human Mind and the Future of Medicine. N Engl J Med 2017;377:1209-11.
37. Epner PL, Gans JE, Graber ML. When diagnostic testing leads to harm: a new outcomes-based approach for laboratory medicine. BMJ Qual Saf 2013;22:ii6-10.
38. Feddock CA. The Lost Art of Clinical Skills. Am J Med 2007;120:374-8.
39. Verghese A. Culture Shock—Patient as Icon, Icon as Patient. N Engl J Med 2008;359:2748-51.
40. Coiera E, Ash J, Berg M. The Unintended Consequences of Health Information Technology Revisited. Yearb Med Inform 2016;1:163-9.
41. West P, Giordano R, Van Kleek M, et al. The Quantified Patient in the Doctor's Office: Challenges & Opportunities. In: CHI Conference Committee. Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems-CHI '16. San Jose: ACM Press; 2016:3066-78.
42. Coiera E, Kocaballi B, Halamka J, et al. The digital scribe. NPJ Digit Med 2018;1:58.
43. Cabitza F, Rasoini R, Gensini GF. Unintended Consequences of Machine Learning in Medicine. JAMA 2017;318:517.
44. Gassman AL, Nguyen CP, Joffe HV. FDA Regulation of Prescription Drugs. N Engl J Med 2017;376:674-82.
45. Coiera E, Ammenwerth E, Georgiou A, et al. Does health informatics have a replication crisis? J Am Med Inform Assoc 2018;25:963-8.
46. Gur D, Bandos AI, Cohen CS, et al. The "Laboratory" Effect: Comparing Radiologists' Performance and Variability during Prospective Clinical and Laboratory Mammography Interpretations. Radiology 2008;249:47-53.
47. Moja L, Kwag KH, Lytras T, et al. Effectiveness of computerized decision support systems linked to electronic health records: a systematic review and meta-analysis. Am J Public Health 2014;104:e12-22.
48. Lehman CD, Wellman RD, Buist DSM, et al. Diagnostic Accuracy of Digital Screening Mammography With and Without Computer-Aided Detection. JAMA Intern Med 2015;175:1828.
49. Gulshan V, Peng L, Coram M, et al. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. JAMA 2016;316:2402.
50. Metz C. India Fights Diabetic Blindness With Help From A.I. The New York Times. March 10, 2019. Accessed on March 12, 2019. Available online: https://www.nytimes.com/2019/03/10/technology/artificial-intelligence-eye-hospital-india.html
51. Bien N, Rajpurkar P, Ball RL, et al. Deep-learning-assisted diagnosis for knee magnetic resonance imaging: Development and retrospective validation of MRNet. PLoS Med 2018;15:e1002699.
52. Harvey H, Cabitza F. Algorithms are the new drugs? Reflections for a culture of impact assessment and vigilance. MCCSIS 2018;281-5.
53. Lavis JN, Anderson GM. Appropriateness in health care delivery: definitions, measurement and policy implications. CMAJ 1996;154:321-8.
54. Floridi L, Cowls J, Beltrametti M, et al. AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. Minds Mach (Dordr) 2018;28:689-707.
55. Neff G. Why Big Data Won't Cure Us. Big Data 2013;1:117-23.
56. Nemeth C, Wears R, Woods D, et al. Minding the Gaps: Creating Resilience in Health Care. In: Henriksen K, Battles JB, Keyes MA, et al. editors. Advances in Patient Safety: New Directions and Alternative Approaches (Vol 3: Performance and Tools). Rockville: Agency for Healthcare Research and Quality, 2008.
57. Smith QW, Street RL, Volk RJ, et al. Differing levels of clinical evidence: exploring communication challenges in shared decision making. Introduction. Med Care Res Rev 2013;70:3S-13S.
58. Bates DW. Commentary: the role of "technovigilance" in improving care in hospitals. Milbank Q 2013;91:455-8.

Cite this article as: 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(8):161. doi: 10.21037/atm.2019.04.07

