npj | Digital Medicine

www.nature.com/npjdigitalmed

COMMENT OPEN

Check for updates

Presenting machine learning model information to clinical end users with model facts labels

Mark P. Sendak[1]✉, Michael Gao[1], Nathan Brajer[1,2] and Suresh Balu[1,2]

There is tremendous enthusiasm surrounding the potential for machine learning to improve medical prognosis and diagnosis. However, there are risks to translating a machine learning model into clinical care and clinical end users are often unaware of the potential harm to patients. This perspective presents the "Model Facts" label, a systematic effort to ensure that front-line clinicians actually know how, when, how not, and when not to incorporate model output into clinical decisions. The "Model Facts" label was designed for clinicians who make decisions supported by a machine learning model and its purpose is to collate relevant, actionable information in 1-page. Practitioners and regulators must work together to standardize presentation of machine learning model information to clinical end users in order to prevent harm to patients. Efforts to integrate a model into clinical practice should be accompanied by an effort to clearly communicate information about a machine learning model with a "Model Facts" label.

npj Digital Medicine (2020)3:41 ; https://doi.org/10.1038/s41746-020-0253-3

INTRODUCTION

Recent advances in machine learning and artificial intelligence promise major improvements in medical diagnosis and prognosis[1]. Risk can now be estimated from a combination of pipelines of information from health records, patient reports and other sources, coupled with machine learning algorithms that produce probabilistic predictions. In the life of consumers, such algorithms underpin applications that enable the selection of routes of travel, restaurants and movies. In healthcare, however, the immediate stakes are higher, and algorithms can produce benefits and risks. Striking the right balance depends on how the algorithms are constructed and how they are used.

An interdisciplinary team including engineers, clinicians and quantitative scientists developed and validated a machine learning model to predict the risk of inpatient mortality at the time of hospital admission. The model was trained to predict the risk of death at any time during the inpatient stay. The model performed well on retrospective data, data from external hospitals, and prospectively after being integrated into the electronic health record. The team discussed workflows and agreed on the intended use of the model: to improve early alignment of goals of care, intensity of care and early engagement of palliative care for patients at high risk of inpatient mortality. During a workflow discussion, a seemingly benign question surfaced: can the model also be used to triage patients for the intensive care unit?

The potential harm to patients when using the model for a use case other than the one it was trained for was not immediately clear. Upon reflection, the 2015 experience of a team at Microsoft Research seemed pertinent. The team famously described a model developed to predict death amongst patients with pneumonia presenting to the hospital[2]. The goal was to identify which patients with pneumonia needed inpatient admission and which patients could be managed in the outpatient setting. The model found that patients with asthma were at lower risk of death, due to the fact that patients with asthma were admitted to the intensive care unit and received appropriately escalated care. If that model were integrated into clinical workflows without a clear indication for use, it's easy to imagine patients with pneumonia complicated by asthma inappropriately treated less intensively.

The clinical utility of models is widely questioned and the need to communicate the limitations of machine learning systems has been highlighted[3,4]. However, there has not been a systematic effort to ensure that front-line clinicians actually know how, when, how not, and when not to incorporate model output into clinical decisions. Nor is there an expectation that those who develop and promote models are responsible for providing instruction of model use and for the consequences of inappropriate use.

STANDARD REPORTING OF MACHINE LEARNING MODELS

In 2015, the Transparent Reporting of Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) statement was released to improve the reporting of prediction models in published literature[5]. A new initiative was recently announced to adapt the guidelines for machine learning models as well as to update clinical trial reporting for machine learning trials[6,7]. Unfortunately, models are often used without reference to the primary literature. If machine learning models are to be widely used in clinical practice, standard reporting of important model information should be coupled with use of the model, not publication of the model.

Even in published literature, model evaluations are often poorly conducted[8,9]. Model performance is often assessed using data that is easily available, rather than data that reflects the target population of actual model use. Model performance using statistical measures is often conflated with demonstrating clinical impact and utility in care delivery. Finally, clinical end users are often left ill-prepared to assess whether or not a model is generalizable to any particular setting[10,11]. Novel communication tools are needed to inform clinicians of the appropriate context and use of validated machine learning models.

[1]Duke Institute for Health Innovation, Durham, NC, USA. [2]Duke University School of Medicine, Durham, NC, USA. ✉email: mark.sendak@duke.edu

Measures of model performance must also be meaningful within the context of care delivery to clinical end users. For machine learning models that discriminate between normal and abnormal states, a commonly used metric is the area under the receiver operator characteristic curve, also known as AUC[12]. AUC is a single measure of discrimination that can be interpreted as the probability of correctly ranking a randomly selected patient with the outcome as higher risk than a randomly selected patient without the outcome. The metric does not take prevalence of the outcome into account, making it difficult to interpret for rare events, and does not provide any information about calibration. Accordingly, models with improvements in AUC may be inaccurate in populations with different underlying risks or may not be anchored to appropriate absolute risk predictions. For a clinical end user receiving an alert prompted by a machine learning model, AUC is a measure that provides no actionable guidance.

RELATED WORK

The "Model Facts" label is an example of risk communication, defined by the United States Food and Drug Administration (FDA) as "the term of art used for situations when people need good information to make sound choices"[13]. As machine learning innovations progress through different stages of diffusion, risk communication needs to be developed for different audiences and distributed via different channels[14]. Risk communication that is important during the decision stage to approve and adopt an innovation include FDA device approval summaries and medication guides as well as academic manuscripts. The "Model Facts" label specifically serves the audience of clinical end users at the implementation stage and is distributed via channels that are closely integrated with the clinical decision support.

Transparency in machine learning model reporting is not enough. As Onora O'Neill describes, "it is easy to place information in the public domain, but hard to ensure that it is in practice accessible to those for whom it might be valuable, intelligible to them if they find it, or assessable by them if they find and understand it"[15]. Ensuring that risk communication is accessible, intelligible, and assessable requires clear understanding of the objectives of the model, close collaboration with end users, and rigorous evaluation[16]. While the US FDA provides guidance on risk communication, it also acknowledges that there is no one-size-fits-all approach[13].

Two instructive examples of risk communication research within health care are shared decision making aids and "Drug Facts" boxes. International expert consensus groups have gathered to synthesize the research and propose best practices for designing decision aids for patients[16,17]. Notable examples include https://knowyourchances.cancer.gov in the United States and https://breast.predict.nhs.uk in the United Kingdom. "Drug Facts" boxes have been rigorously evaluated in multiple randomized controlled trials[18,19], culminating in recommendations from Congress for US FDA to consider implementing "Drug Facts" boxes[20]. Outside of health care, preliminary efforts have begun to standardize documentation to accompany a trained machine learning model[21]. There is an urgent need to design machine learning product labels that address the context-specific challenges of health care.

THE "MODEL FACTS" LABEL

Shortly after the experience described above, an interdisciplinary team including developers, clinicians, and regulatory experts designed the "Model Facts" label. The target audience is clinicians who make decisions supported by a machine learning model. The purpose is to collate relevant, actionable information in 1-page to ensure that front-line clinicians know how, when, how not, and when not to incorporate model output into clinical decisions. The "Model Facts" label is not meant to be comprehensive and individual sections may need to be populated over time as information about the model becomes available. For example, a model may be used in a local setting before it has been externally validated in a distinct geographical setting. There is also important information about the model, such as the demographic representation of training and evaluation data, that may need to be immediately available to an end user preceding full publication of a model.

Figure 1 illustrates an example "Model Facts" label designed for a sepsis prediction model. The major sections of the "Model Facts" label include the model name, locale, and version, summary of the model, mechanism of risk score calculation, validation and performance, uses and directions, warnings, and other information. The structure is meant to mirror product information for food, drugs and devices. Publication hyperlinks in the "Validation and performance" and "Other information" section point to additional details.

Two sections of the "Model Facts" label that are rarely discussed in machine learning model publications are "Uses and directions" and "Warnings". Every machine learning model is trained for a specific task and the boundary lines around that task must be clearly communicated. In our example, warnings are provided to only use the model within settings in which the model was evaluated, to not use the model after a patient develops a first episode of sepsis, and to not use the model in an intensive care unit without further evaluation. There is also a warning against automated treatment assignment.

"Model Facts" labels need to be localized and need to be updated over time. Similar to how antimicrobial sensitivity data guide use of antibiotics within a local population, "Model Facts" labels include information about model performance within the local population. If a model is adopted in a new setting, a new "Model Facts" label needs to be generated and distributed to clinical end users. The target population of model use is also specified in both the "Uses and directions" and "Validation and performance" sections. The version of the "Model Facts" label is documented and version control with documentation of changes should be accessible to all end users[22]. Use of the model and the "Model Facts" label also needs to be approved by governance structures that function similarly to pharmacy and therapeutics committees that monitor use of medications and adverse outcomes.

The structure of our "Model Facts" label presented in Fig. 1 requires rigorous testing and evaluation. It is not meant to be immediately adopted, but to spark dialogue and to be iterated upon and critiqued by a broad group of stakeholders. Risk communication research advises against only using words in communication material[16] and we hope that other teams implementing machine learning tools create their own versions of "Model Facts" labels.

Many questions remain about the design of the "Model Facts" label and how to make this information accessible, intelligible, and assessable to clinicians. Should the information be accessible within the electronic health record, software applications, an online registry, or some combination? And how is information presented to an end user when it's not immediately clear that a model was involved, for example with a text notification? Despite unanswered questions, without bringing together practitioners and regulators to standardize presentation of machine learning model information to clinical end users, we risk significant harm to patients. Any effort to integrate a model into clinical practice should be accompanied by an effort to clearly communicate how, when, how not, and when not to incorporate model output into clinical decisions.

Model Facts

Model name: Deep Sepsis **Locale:** Duke University Hospital

Approval Date: 09/22/2019 **Last Update:** 01/13/2020 **Version:** 1.0

Summary

This model uses EHR input data collected from a patient's current inpatient encounter to estimate the probability that the patient will meet sepsis criteria within the next 4 hours. It was developed in 2016-2019 by the Duke Institute for Health Innovation. The model was licensed to Cohere Med in July 2019.

Mechanism

- **Outcome** .....sepsis within the next 4 hours, see outcome definition in "Other Information"
- **Output** .....0% - 100% probability of sepsis occurring in the next 4 hours
- **Target population** .....all adult patients >18 y.o. presenting to DUH ED
- **Time of prediction** .....every hour of a patient's encounter
- **Input data source**.....electronic health record (EHR)
- **Input data type** .....demographics, analytes, vitals, medication administrations
- **Training data location and time-period** .....DUH, diagnostic cohort, 10/2014 – 12/2015
- **Model type**..... Recurrent Neural Network

Validation and performance

	Prevalence	AUC	PPV @ Sensitivity of 60%	Sensitivity @ PPV of 20%	Cohort Type	Cohort URL / DOI
Local Retrospective	18.9%	0.88	0.14	0.50	Diagnostic	arxiv.org/abs/1708.05894
Local Temporal	6.4%	0.94	0.20	0.66	Diagnostic	jmir.org/preprint/15182
Local Prospective	**TBD**	**TBD**	**TBD**	**TBD**	**TBD**	**TBD**
External	**TBD**	**TBD**	**TBD**	**TBD**	**TBD**	**TBD**
Target Population	6.4%	0.94	0.20	0.66	Diagnostic	jmir.org/preprint/15182

Uses and directions

- **Benefits:** Early identification and prompt treatment of sepsis can improve patient morbidity and mortality.
- **Target population and use case:** Every hour, data is pulled from the EHR to calculate risk of sepsis for every patient at the DUH ED. A rapid response team nurse reviews every high-risk patient with a physician in the ED to confirm whether or not to initiate treatment for sepsis.
- **General use:** This model is intended to be used to by clinicians to identify patients for further assessment for sepsis. The model is not a diagnostic for sepsis and is not meant to guide or drive clinical care. This model is intended to complement other pieces of patient information related to sepsis as well as a physical evaluation to determine the need for sepsis treatment.
- **Appropriate decision support:** The model identifies patient X as at a high risk of sepsis. A rapid response team nurse discusses the patient with the ED physician caring for the patient and they agree the patient does not require treatment for sepsis.
- **Before using this model:** Test the model retrospectively and prospectively on a diagnostic cohort that reflects the target population that the model will be used upon to confirm validity of the model within a local setting.
- **Safety and efficacy evaluation:** Analysis of data from clinical trial (NCT03655626) is underway. Preliminary data shows rapid response team, nurse-driven workflow was effective at improving sepsis treatment bundle compliance.

Warnings

- **Risks:** Even if used appropriately, clinicians using this model can misdiagnose sepsis. Delays in a sepsis diagnosis can lead to morbidity and mortality. Patients who are incorrectly treated for sepsis can be exposed to risks associated with unnecessary antibiotics and intravenous fluids.
- **Inappropriate Settings:** This model was not trained or evaluated on patients receiving care in the ICU. Do not use this model in the ICU setting without further evaluation. This model was trained to identify the first episode of sepsis during an inpatient encounter. Do not use this model after an initial sepsis episode without further evaluation.
- **Clinical Rationale:** The model is not interpretable and does not provide rationale for high risk scores. Clinical end users are expected to place model output in context with other clinical information to make final determination of diagnosis.
- **Inappropriate decision support:** This model may not be accurate outside of the target population, primarily adults in the non-ICU setting. This model is not a diagnostic and is not designed to guide clinical diagnosis and treatment for sepsis.
- **Generalizability:** This model was primarily evaluated within the local setting of Duke University Hospital. Do not use this model in an external setting without further evaluation.
- **Discontinue use if:** Clinical staff raise concerns about utility of the model for the indicated use case or large, systematic changes occur at the data level that necessitates re-training of the model.

Other information:

- **Outcome Definition:** https://doi.org/10.1101/648907
- **Related model:** http://doi.org/10.1001/jama.2016.0288
- **Model development & validation:** arxiv.org/abs/1708.05894
- **Model implementation**: jmir.org/preprint/15182
- **Clinical trial:** clinicaltrials.gov/ct2/show/NCT03655626
- **Clinical impact evaluation:** TBD
- **For inquiries and additional information:** please email mark.sendak@duke.edu

Fig. 1 Example "Model Facts" label for a sepsis machine learning model. This "Model Facts" label provides relevant information about a sepsis prediction model to clinical end users who use the model to assist with clinical diagnosis of sepsis. AUC Area Under the ROC Curve, PPV Positive Predictive Value, DOI Digital Object Identifier, EHR Electronic Health Record, ED Emergency Department.

Received: 16 November 2019; Accepted: 28 February 2020;
Published online: 23 March 2020

REFERENCES

1. Topol, E. J. High-performance medicine: the convergence of human and artificial intelligence. *Nat. Med.* **25**, 44–56 (2019).
2. Caruana, R. et al. *Intelligible Models for Healthare: Predicting Pneumonia Risk and Hospital 30-day Readmission*, 1721–1730 (ACM Press, New York, NY, 2015).
3. He, J. et al. The practical implementation of artificial intelligence technologies in medicine. *Nat. Med.* https://doi.org/10.1038/s41591-018-0307-0 (2019).
4. Wiens, J. et al. Do no harm: a roadmap for responsible machine learning for health care. *Nat. Med.* **343**, 1203–1204 (2019).
5. Collins, G. S., Reitsma, J. B., Altman, D. G. & Moons, K. G. M. Transparent reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): the TRIPOD statement. *Ann. Intern. Med.* **162**, 55–11 (2015).
6. Collins, G. S. & Moons, K. G. M. Reporting of artificial intelligence prediction models. *Lancet* **393**, 1577–1579 (2019).
7. Liu, X. et al. Reporting guidelines for clinical trials evaluating artificial intelligence interventions are needed. *Nat. Med.* https://doi.org/10.1038/s41591-019-0603-3 (2019).
8. Park, S. H. & Han, K. Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction. *Radiology* **286**, 800–809 (2018).
9. Park, S. H., Kim, Y.-H., Lee, J. Y., Yoo, S. & Kim, C. J. Ethical challenges regarding artificial intelligence in medicine from the perspective of scientific editing and peer review. *Sci. Editing* **6**, 91–98 (2019).
10. Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G. & King, D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med.* **17**, 44–49 (2019).
11. Van Calster, B., Wynants, L., Timmerman, D., Steyerberg, E. W. & Collins, G. S. Predictive analytics in health care: how can we know it works? *J. Am. Med. Inform. Assoc.* **320**, 27 (2019).
12. Shillan, D., Sterne, J. A. C., Champneys, A. & Gibbison, G. Use of machine learning to analyse routinely collected intensive care unit data: a systematic review. *Crit. Care Med*. https://doi.org/10.1186/s13054-019-2564-9 (2019).
13. Fischhoff, B., Brewer, N. T., & Downs, J. S. (2011). *Communicating Risks and Benefits: an Evidence-based User's Guide*. U.S. (Food Drug Administration, 2011).
14. Rogers, E. M. *Diffusion of Innovations*. 4 edn. (The Free Press, New York, NY, 1995).
15. O'Neill, O. Linking trust to trustworthiness. *Int. J. Philos. Studies*. https://doi.org/10.1080/09672559.2018.1454637 (2018).
16. Spiegelhalter, D. Risk and uncertainty communication. *Annu. Rev. Stat. Appl.* **4**, 31–60 (2017).
17. Trevena, L. J. et al. Presenting quantitative information about decision outcomes: a risk communication primer for patient decision aid developers. *BMC Med. Inform. Decis. Mak.* **13**, S7 (2013).
18. Schwartz, L. M., Woloshin, S. & Welch, H. G. Using a drug facts box to communicate drug benefits and harms. *Ann. Intern. Med.* **150**, 516–527 (2009).
19. Woloshin, S. & Schwartz, L. M. Communicating data about the benefits and harms of treatment. *Ann. Intern. Med.* **155**, 87–96 (2011).
20. Schwartz, L. M. & Woloshin, S. The drug facts box: improving the communication of prescription drug information. *Proc. Natl Acad. Sci. USA* **110**(Suppl 3), 14069–14074 (2013).
21. Mitchell, M. et al. Model cards for model reporting. In *Proc. ACM Conference on Fairness, Accountability, and Transparency in Machine Learning 2019*, 220–229 (ACM, New York, 2019).
22. Hwang, T. J., Kesselheim, A. S., & Vokinger, K. N. Lifecycle regulation of artificial intelligence- and machine learning-based software devices in medicine. *J. Am. Med. Assoc.* https://doi.org/10.1001/jama.2019.16842 (2019).

ACKNOWLEDGEMENTS

The authors thank Robert Califf, MD for his thoughtful review of this article and his many helpful comments. This effort was funded by the Duke Institute for Health Innovation. No external funding was supported this work.

AUTHOR CONTRIBUTIONS

M.P.S. wrote the first draft. All authors contributed to both the subsequent drafting and critical revision of the manuscript. N.B. designed the first iteration of the "Model Facts" label. All authors contributed to revisions of the "Model Facts" label.

COMPETING INTERESTS

M.P.S., M.G., N.B., and S.B. are named inventors of the Sepsis Watch deep-learning model, which was licensed from Duke University by Cohere Med, Inc. M.P.S., M.G., and S.B. do not hold any equity in Cohere Med, Inc.

ADDITIONAL INFORMATION

Correspondence and requests for materials should be addressed to M.P.S.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.