Statistical and machine learning techniques:

Which help in patient care and medical research?

Prof. Dr. Jörg Rahnenführer, Department of Statistics

technische universität
dortmund

41st Annual Conference of the
International Society of Clinical Biostatistics,
Kraków, August 25, 2020

Project of the STRATOS initiative

- Statistical and machine learning techniques: Which help in patient care and medical research?

  Medical goals: Diagnosis, prognosis and treatment selection for human diseases

In patient care, patient satisfaction is central.

Physicians should involve patients in their own care and explain their decision processes.

In medical research, clinical credibility of a model and evidence of its accuracy, generality and clinical effectiveness are central.

Physicians and model developers should work together.

Wyatt and Altman (1995): Prognostic models: clinically useful or quickly forgotten?

Project of the STRATOS initiative

- Statistical and machine learning techniques: Which help in patient care and medical research?

- Special focus on predictions (important in all 3 areas)
  - Diagnosis: associated with uncertainty that is relevant also for consequential treatment decisions
  - Prognosis: uncertainty part of prediction because development of disease depends also on unknown developments in the future
  - Treatment selection: aims to help a patient receive the treatment, which most likely leads to a positive outcome

- Acknowledgements:
  - Joint work with Matthias Schmid and Willi Sauerbrei
  - Comments from Federico Ambrogi, Riccardo de Bin, Anne-Laure Boulesteix, Ben van Calster, Mitch Gail, Frank Harrell, Marianne Huebner

Two cultures

- Data modelling vs. algorithmic modelling culture
  - Breiman, L. (2001) Statistical modeling: The two cultures, Statistical Science 16(3), 199-215
  - Raper, S. (2020): Leo Breiman's “Two Cultures”, Significance 17(1), 34-37

- Modelling the relationship between the inputs to a process or mechanism and the outputs of that process.
  - *Data modelling culture*: understanding the relationship means hypothesizing a mathematical model that explains the process
  - *Algorithmic modelling culture*: predicting correctly the output data given the input data with no constraints on how this is done
- Conflict between accuracy and interpretation?

Two cultures

- Breiman‘s view:
  - “The data and the problem must come first.”
  - “The great adventure of statistics is in gathering and using data to solve interesting and important real world problems.”
  - “Being a scientist is to be open to using a wide variety of tools.”
- Cox:
  - “Professor Breiman takes data as his starting point. I would prefer to start with an issue, a question or a scientific hypothesis.”
- Raper:
  - “We smile now at Breiman’s 2001 estimate of the size of the algorithmic modelling community: he puts it at just 2 per cent of all statisticians.”

Usefulness of predictions

- Focus of this talk
  - Not distinction between statistical modelling and machine learning (see invited session with Trevor Hastie and Frank Harrell on Wednesday)
  - Rather usefulness of predictions, applies to both fields
- 8 points to consider

RESEARCH

1. Clarity of goal
2. Suitability of data
3. Suitability of analysis methods and algorithms
4. Suitability of evaluation measures
5. Interpretability

PRACTICAL

6. Availability of the explanatory variables
7. Transportability across patient cohorts
8. Practical usefulness

Usefulness of predictions

1. Clarity of the goal

- Be clear about the goal of the analysis: What should be predicted and how should the predictions be used?
- Are predictions on a probability scale or dichotomous?
  - Make full use of available data, do not dichotomize if the probability can be used! Also not required by regression models for time to event.
  - Many ML algorithms try to discriminate between two or more groups, but this is not the same as predicting risk over time.
- What is the intention of the analysis, what is the outcome supposed to be used for, and does it deliver what was intended?

Usefulness of predictions

2. Suitability of the data

- Are the origin and the characteristics of the data appropriate to reach the goal?
  - Is sample size sufficient?
    - sample size calculation in planning phase
    - variance on estimated predictions
    - size of the validation data
  - Missing data
    - EHR (electronic health records) data from hospitals typically have a high rate of missing values, both for predictors and targets
  - Measurement error
  - Origin of the data, for what reason/goal were the data collected?

Usefulness of predictions

3. Suitability of the analysis methods and algorithms

- Are the analysis techniques appropriate to reach the goal?
  - Does the method fit to the data?
  - Properties of the method (strengths and limitations)
  - Robustness to violations of assumptions
  - Curse of dimensionality
  - Overfitting: Does overfitting occur? Are the used methods prone to overfitting?
  - Generalizability

Usefulness of predictions

4. Suitability of the evaluation measures

- Do the evaluation measures reflect the medical goals?
  - Are the measures relevant in practice?
    - AUC
    - C-Index
    - Calibration curve
  - When risk or life expectancy estimation are the real goals:
    - Proportion classified correctly, sensitivity, specificity, precision, and recall: all improper accuracy scoring rules

Usefulness of predictions

5. Interpretability

- Is the model interpretable w.r.t. the medical goal?
  - Explanation versus prediction
    - Interpretability of single variables or black-box prediction?
  - Isolation: What is the effect of one variable after accounting for the effects of other variables
  - Meaningfulness of the learned patterns: Are the learned patterns specifically disease-related?

Usefulness of predictions

5. Interpretability

- Problem of complex methods (e.g., deep learning)
  - Usage of too many (unclear) features
  - Exploitation of information beyond specific disease-related findings (e.g. on imaging data from x-rays)
  - Example (UCSF)
    - Data set with chemical features for molecules (dipole moments, NMR shifts, calculated electrostatic)
    - Feature values replaced with random Gaussian numbers
    - Machine learning algorithm applied to noise
    - Resulting model whose predictions are nearly as good as the original.

Usefulness of predictions

6. Availability of the explanatory variables in practice

- Are the variables used for the predictions available for future patients?
  - Costs of collecting variables
  - Privacy
  - Non-ethical variables (e.g., in feature selection on EHR data, e.g. a proxy for private insurance coverage)
  - Timing
    - EHR data are typically produced at multiple time points during encounters with patients, and variables can have a time stamp
    - Can be problematic to uniquely match a variable with a time stamp
  - Is the algorithm even a black box (e.g., for deep learning)?
    - Then it is often not provided and not usable by others

Usefulness of predictions

7. Transportability across patient cohorts

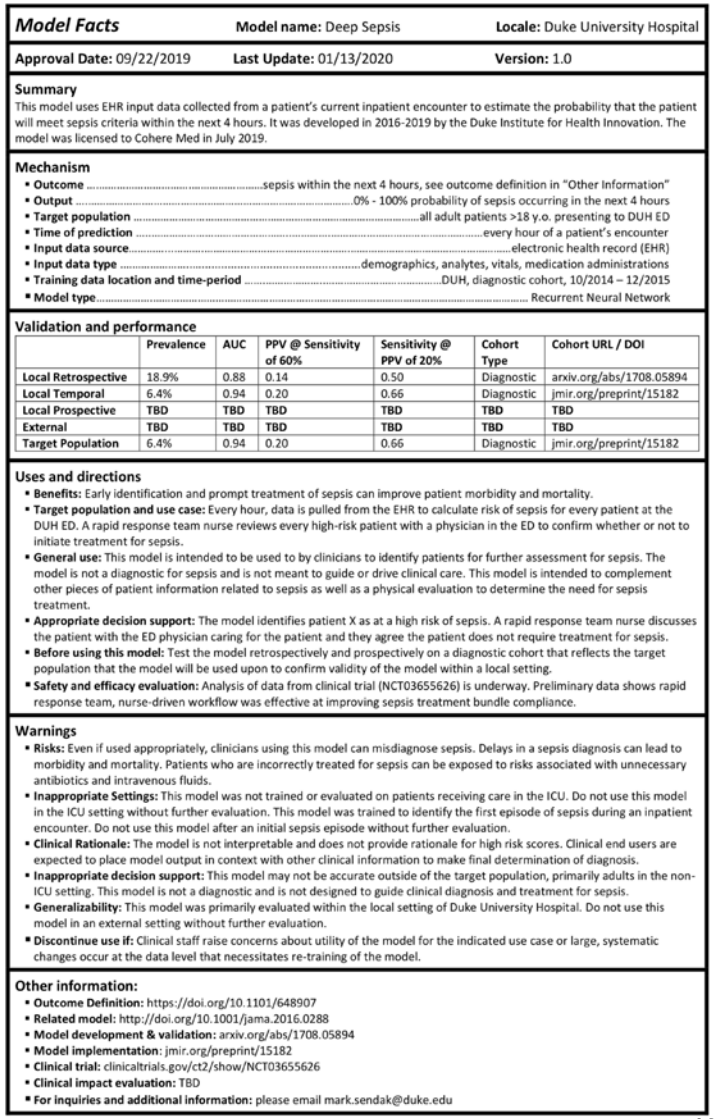
- How easily can the predictions be calculated also by other researchers?
  - Transportability of the prediction rules
  - Problem of sustainability
  - Software often rapidly becomes obsolete

8. Practical usefulness

- Are the predictions meaningful for therapy decisions
  - Gain in prediction accuracy (sensitivity, specificity,..) significant but not medically relevant
  - Relevant effect size, or only statistical significance?

Translation of ML model into clinical care

- „Model Facts“ labels
  - Translation of machine learning models into clinical care
  - “Clinical end users are often unaware of the potential harm to patients.”
  - “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.”
  - Sendak, M.P et al. (2020): Presenting machine learning model information to clinical end users with model facts labels. npj Digit. Med. 3, 41.
  - Note: This was developed for ML models, but important for all models



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

„Model Facts“ labels

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

“Model Facts“ labels

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.

ML and AI for patient benefit

Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness

Sebastian Vollmer,[1,2] Bilal A Mateen,[1,3,4] Gergo Bohner,[1,2] Franz J Király,[1,5] Rayid Ghani,[6] Pall Jonsson,[7] Sarah Cumbers,[8] Adrian Jonas,[9] Katherine S L McAllister,[9] Puja Myles,[10] David Grainger,[11] Mark Birse,[11] Richard Branson,[11] Karel G M Moons,[12] Gary S Collins,[13] John P A Ioannidis,[14] Chris Holmes,[1,15] Harry Hemingway[16,17,18]

- "Clinically relevant research using modern statistical methods (such as machine learning and artificial intelligence) is too often limited by one or more of TREE concerns (transparency, reproducibility, ethics, and effectiveness); addressing these concerns can facilitate appropriate translation from computer bench to patient benefit"
- "Here we propose 20 critical questions that offer a framework for users and generators of ML/AI research"

ML and AI for patient benefit

Box 1: Critical questions for health related technology involving machine learning and artificial intelligence

Inception

1. What is the health question relating to patient benefit?
2. What evidence is there that the development of the algorithm was informed by best practices in clinical research and epidemiological study design?

Study

1. When and how should patients be involved in data collection, analysis, deployment, and use?
2. Are the data suitable to answer the clinical question—that is, do they capture the relevant real world heterogeneity, and are they of sufficient detail and quality?
3. Does the validation methodology reflect the real world constraints and operational procedures associated with data collection and storage?
4. What computational and software resources are required for the task, and are the available resources sufficient to tackle this problem?

Statistical methods

1. Are the reported performance metrics relevant for the clinical context in which the model will be used?
2. Is the ML/AI algorithm compared to the current best technology, and against other appropriate baselines?
3. Is the reported gain in statistical performance with the ML/AI algorithm justified in the context of any trade-offs?

Reproducibility

1. On what basis are data accessible to other researchers?
2. Are the code, software, and all other relevant parts of the prediction modelling pipeline available to others to facilitate replicability?
3. Is there organisational transparency about the flow of data and results?

Impact evaluation

1. Are the results generalisable to settings beyond where the system was developed (that is, results reproducibility/external validity)?
2. Does the model create or exacerbate inequities in healthcare by age, sex, ethnicity, or other protected characteristics?
3. What evidence is there that clinicians and patients find the model and its output (reasonably) interpretable?
4. How will evidence of real world model effectiveness in the proposed clinical setting be generated, and how will unintended consequences be prevented?

Implementation

1. How is the model being regularly reassessed, and updated as data quality and clinical practice changes (that is, post-deployment monitoring)?
2. Is the ML/AI model cost effective to build, implement, and maintain?
3. How will the potential financial benefits be distributed if the ML/AI model is commercialised?
4. How have the regulatory requirements for accreditation/approval been addressed?

ML and AI for patient benefit

- Selected critical questions
  - Inception:
    1. What is the health question relating to patient benefit?
       → Points to consider 1: Clarity of the goal
  - Statistical methods
    1. Are the reported performance metrics relevant for the clinical context in which the model will be used?
       → Points to consider 4: Suitability of the evaluation measures
  - Reproducibility
    1. On what basis are data accessible to other researchers?
    2. Are … code, software, and … available to others to facilitate replicability?
  - Impact evaluation
    1. On what basis are data accessible to other researchers?
    2. What evidence is there that clinicians and patient find the model and its output (reasonably) interpretable?

References

- Boulesteix, A-L; Schmid, M (2014): Machine learning versus statistical modeling, Biometrical Journal 56(4) 588-593.
- Sendak MP, Gao M, Brajer N, Balu S (2020): Presenting machine learning model information to clinical end users with model facts labels. NPJ Digit. Med. 3, 41.
- Shah NH, Milstein A, Bagley SC (2019): Making Machine Learning Models Clinically Useful. JAMA, published online 2019 Aug 8. doi: 10.1001/jama.2019.10306.
- Shmueli G (2010): Statistical Science 25(3), 289-310.
- Vollmer S, ..., Collins GS, Ioannidis J, Holmes C, Hemingway H (2020): Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness, BMJ 2020; 368 BMJ 368:l6927
- Watson DS, Krutzinna J, Bruce IN, Griffiths CE, McInnes IB, Barnes MR, Floridi L (2019): Clinical applications of machine learning algorithms: beyond the black box. BMJ 364:l886.
- Wyatt JC, Altman DG (1995): Commentary: Prognostic models: clinically useful or quickly forgotten? BMJ, 311:1539.

- Harrell F.: Lots of texts and blogs and comments…
- JAMA Internal Medicine: Many interesting editorials…