Trends in the development and validation of patient-level prediction models using electronic health record data: a systematic review

Cynthia Yang, Jan A. Kors, Solomon Ioannou, Luis H. John, Aniek F. Markus, Alexandros Rekkas, Maria de Ridder, Tom Seinen, Ross D. Williams, Peter R. Rijnbeek

Background

The wide implementation of the electronic health record (EHR) in recent decades drastically increased the availability of data for patient-level prediction modelling. This has resulted in the development of many patient-level prediction models using EHR data. Before recommending a prediction model for clinical practice, it is important to assess whether and for which patients the model works well by externally validating it across various centers and settings (1-3). External validation can also be done by other investigators in a separate study. However, to allow other investigators to interpret results and externally validate a model, model development and validation need to be reported in full detail.

Previous systematic reviews of clinical prediction studies covered different individual years or periods prior to 2015. They found room for improvement in the conduct and reporting of model development and validation: most studies made no explicit mention of how missing data were handled, model calibration was rarely assessed, and external validation was uncommon (1, 4, 5). To encourage improvement in the conduct and reporting of model development and validation, the Transparent Reporting of a multivariable clinical prediction model for Individual Prognosis or Diagnosis (TRIPOD) Statement was simultaneously published in 11 leading journals in January 2015 (6). The aim of this systematic review is to provide further insights in the development of the field over time, with a focus on the transparent reporting of model development and validation using EHR data to enable external validation by other investigators and whether transparent reporting has improved since the TRIPOD Statement was published.

Methods

To identify relevant papers, we searched Embase, Medline, Web-of-Science, Cochrane Library and Google Scholar. The search was limited to papers written in English and published between January 1, 2009, and November 15, 2019. Studies that were not original research (e.g., comments, letters, editorials) or had no abstract were excluded. We included all papers that described the development of one or more multivariable prediction models using EHR data to estimate the probability of a particular clinical outcome occurring within a certain period in the future (i.e., prognostic prediction). One reviewer (C.Y.) screened all titles and abstracts to identify potentially eligible studies. The same reviewer assessed eligibility of all potentially eligible studies based on the full text. Data extraction was completed by multiple reviewers (J.A.K., S.I., L.H.J., A.F.M., A.R., M.R., T.S., R.D.W.) and verified by a second reviewer (C.Y.). Data extraction was based on the Checklist for critical Appraisal and data extraction for systematic Reviews of prediction Modelling Studies (CHARMS) Checklist (7) and the TRIPOD Statement (6). To investigate trends, we assessed differences in the extracted items between the periods 2009-2014 and 2015-2019.

Results

Our literature search resulted in a total of 9,942 papers. After duplicates were removed, 6,235 titles and abstracts were screened. From this, 1,075 potentially eligible papers were identified. Upon full text inspection, 422 papers were eventually included for synthesis. In total, we extracted items for 579 models from 422 studies (with 1 to 6 models per study). We observed an increase from 135 models in 101 studies in the period 2009-2014 to 444 models in 321 studies in the period 2015-2019.

First, we assessed whether code lists were provided to define the target population, outcome, and candidate predictors. For all three prediction components, the percentage of models for which code lists were provided was very low and remained below 20% over time (Figure 1). In both periods, the prediction horizon was reported for 84% of all models. The percentage of models for which the time window for candidate predictor measurement was reported increased from 46% to 50%, while the percentage of models for which the final model was completely presented decreased from 49% to 39%.

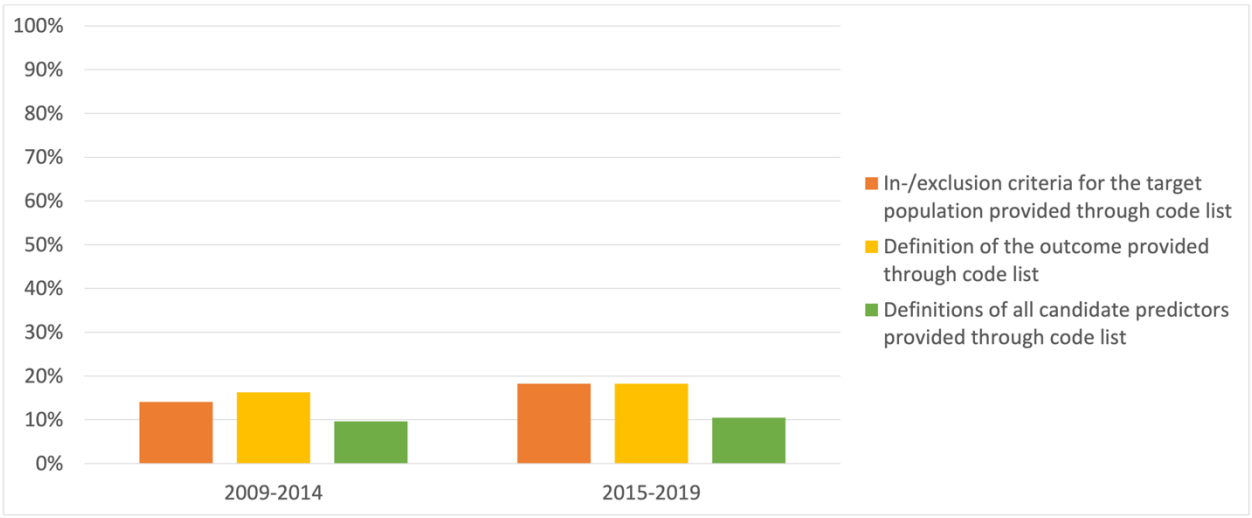

Figure 1. Reporting of code lists for defining the prediction problem.

Next, we grouped each model into one of the following three categories: 1) externally validated, when performance was assessed on data from a database other than the development set, 2) not externally but internally validated only, when performance was assessed on the development set by split-sample, cross-validation, temporal validation, or bootstrapping, and 3) not validated, when performance was not assessed or only assessed on the same data that were used to train the model. We found that external validation increased from 10% to 12%, internal validation only increased from 76% to 81%, and no validation decreased from 13% to 7% (Figure 2). The percentage of externally validated models that were validated using data from a different country increased from 7% to 9%.

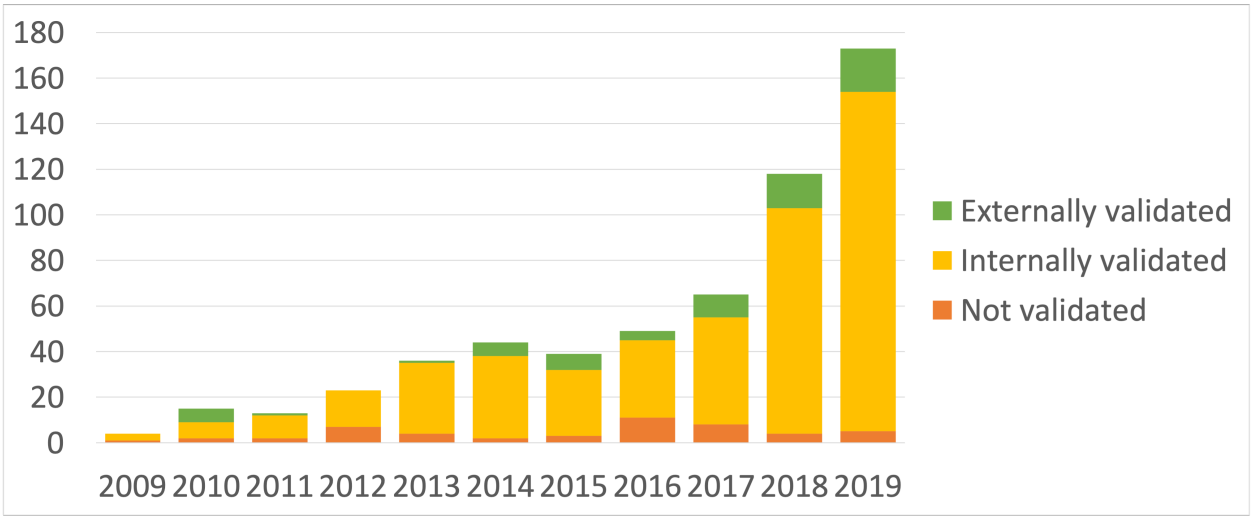

Figure 2. Reporting of validation results.

Finally, we investigated to what extent model discrimination and calibration were assessed internally (Figure 3). Internal validation results were reported for 525 models in 382 studies. This includes 64 models in 43 studies for which both internal and external validation results were reported. The area under the receiver operating characteristic curve (AUROC) was reported for more than 90% of all internal validations. The percentage of internal validations for which the ROC curve was presented increased from 28% to 47%. The percentage of internal validations for which the calibration plot was presented was slightly less than 30% in both periods. For about a quarter of all internal validations, other calibration measures such as the Hosmer-Lemeshow test or the calibration slope were reported.

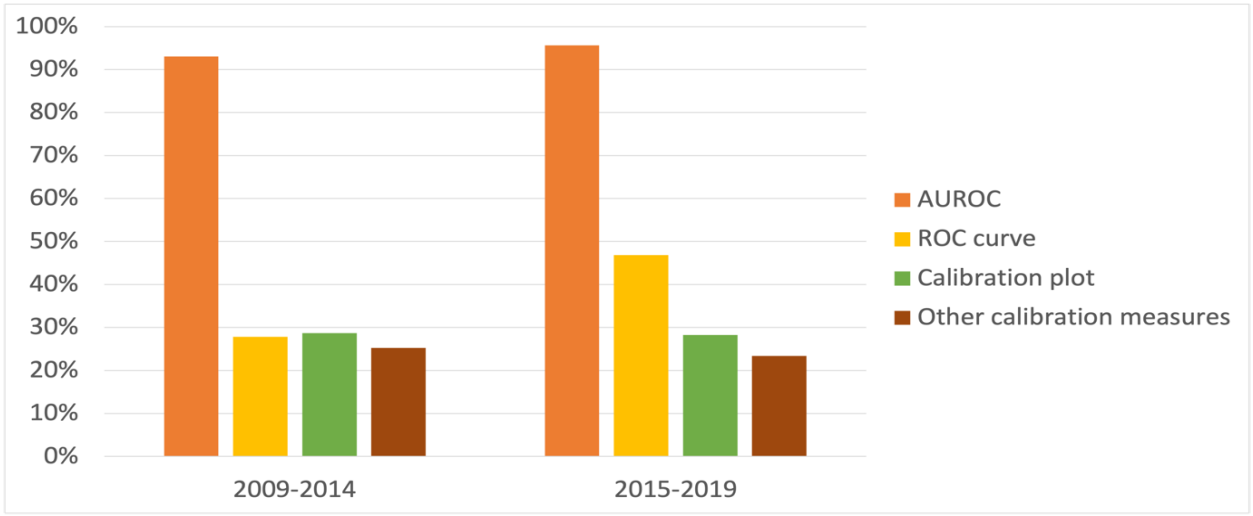

Figure 3. Reporting of discrimination and calibration on internal validation.

Conclusion

To allow other investigators to externally validate a patient-level prediction model, clearly reporting the prediction problem definition, as well as sharing the final model, is essential. In the same vein, models should be extensively validated, where not only discrimination but also calibration should be assessed. Overall, we found that despite existing reporting guidelines, there is still lack of awareness about the need for good methodological conduct and reporting of model development and validation. This review clearly shows the need for OHDSI's PatientLevelPrediction framework that enforces best practices. Improved conduct and reporting of model development and validation will enable more research on model transportability, towards wider proliferation and use in clinical practice of patient-level prediction models.

Funding

This project has received funding from the Innovative Medicines Initiative 2 Joint Undertaking (JU) under grant agreement No 806968. The JU receives support from the European Union's Horizon 2020 research and innovation programme and EFPIA.

References/Citations

1. Collins GS, de Groot JA, Dutton S, Omar O, Shanyinde M, Tajar A, et al. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. BMC Med Res Methodol. 2014;14:40.
2. Collins GS, Ogundimu EO, Altman DG. Sample size considerations for the external validation of a multivariable prognostic model: a resampling study. Stat Med. 2016;35(2):214-26.
3. Van Calster B, Wynants L, Timmerman D, Steyerberg EW, Collins GS. Predictive analytics in health care: how can we know it works? J Am Med Inform Assoc. 2019;26(12):1651-4.
4. Bouwmeester W, Zuithoff NP, Mallett S, Geerlings MI, Vergouwe Y, Steyerberg EW, et al. Reporting and methods in clinical prediction research: a systematic review. PLoS Med. 2012;9(5):1-12.
5. Goldstein BA, Navar AM, Pencina MJ, Ioannidis JP. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. J Am Med Inform Assoc. 2017;24(1):198-208.
6. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMJ. 2015;350:g7594.
7. Moons KG, de Groot JA, Bouwmeester W, Vergouwe Y, Mallett S, Altman DG, et al. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. PLoS Med. 2014;11(10):e1001744.