External validation of existing dementia prediction models on observational health data

Luis H. John (l.john@erasmusmc.nl)
Erasmus University Medical Center
Jan A. Kors
Erasmus University Medical Center
Egill A. Fridgeirsson
Erasmus University Medical Center
Jenna M. Reps
Janssen Research and Development
Peter R. Rijnbeek
Erasmus University Medical Center

Research Article

Keywords: patient-level prediction, prognostic model, external validation, transportability, dementia, alzheimer

Posted Date: July 15th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1742342/v1

Abstract

Background

Many dementia prediction models have been developed, but only few have been externally validated, which hinders clinical uptake and may pose a risk if models are applied to actual patients regardless. Replicating and externally validating a prediction model is a difficult task, where we mostly rely on the completeness of model reporting in a published article.

In this study, we aim to externally validate existing dementia prediction models. To that end, we define replicability criteria, review published models, and externally validate three selected models using routinely collected health data from administrative claims and electronic health records.

Methods

We identified dementia prediction models that were developed between 2011 – 2020 and assessed if they could be replicated given a set of external validation criteria. In addition, we replicated three of these models (Walters' Dementia Risk Score, Mehta's RxDx-Dementia Risk Index, and Nori's ADRD dementia prediction model) and externally validated them on a network of six observational health databases from the United States, United Kingdom, Germany and the Netherlands, including the original development databases of the models.

Results

We reviewed 59 dementia prediction models. All models reported the prediction method, development database, and target and outcome definitions. Less frequently reported by these 59 prediction models were predictor definitions (46 models) including the time window in which a predictor is assessed (21 models), predictor coefficients (19 models), and the time-at-risk (39 models). The replicated model by Walters (development c-statistic: 0.84) showed moderate transportability (0.67 – 0.76 c-statistic). The Mehta model (development c-statistic: 0.81) transported well to some of the external databases (0.69 – 0.79 c-statistic). The Nori model (development AUROC: 0.69) transported well (0.62 – 0.68 AUROC), but performed modestly overall. Recalibration showed improvements for the Walters and Nori models, while recalibration could not be assessed for the Mehta model due to unreported baseline hazard.

Conclusion

We observed that reporting is mostly insufficient to fully replicate and externally validate published dementia prediction models, and therefore, it is uncertain how well these models would work in other clinical settings. We emphasise the importance of following established guidelines for reporting clinical prediction model. We recommend that reporting should be more explicit and have external validation in mind if the model is meant to be applied in different settings.

Key Points

- Many dementia prediction models have been developed, but only few have been externally validated, which may limit clinical uptake.
- We assess whether reported information on existing dementia prediction models allows for external validation.
- We replicate three of the dementia prediction models and externally validate them across a network of observational databases, including the original development databases.

1 Background

Dementia is an umbrella term to describe various illnesses that affect cognition and may lead to mental degradation. Early diagnosis of individuals at high risk of dementia allows for improved care and risk-factor targeted intervention (1). Many patient-level prediction models for identifying individuals who are at risk of dementia have been developed (2–4).

Earlier models were mostly developed on data from cohort studies where data were recorded during health checkups using a variety of questionnaires and cognition tests (5–10). In recent years models have increasingly been developed on observational health data (11–16). These routinely collected data from administrative claims and electronic health records are considered to enhance a model's applicability at the point of care (15). Although observational health data generally do not include known predictive variables such as education level, cognitive test results and genetic information (15), various studies have shown good internal validation performance when developing models on this kind of data. Notable examples are Walters et al. who developed dementia prediction models using electronic health record data from the THIN database, and Albrecht et al. who developed predictive models for Alzheimer's disease and related dementias (ADRD) using administrative claims data from the OptumLabs Data Warehouse (14, 15).

However, the systematic reviews of Hou et al. and Goerdten et al. conclude that although many dementia risk prediction models have been developed, only a handful of them have been externally validated (3, 4). External validation assesses a model's reliability for clinical use in external data sources that have not been used for model development. A lack of external validation can lead to a plethora of proposed models with little evidence about which are reliable and under what circumstances (17).

External validation can be a cumbersome process due to the difficulty of replicating a prediction model, e.g., replicating a cohort and predictor definitions from a manuscript. We hypothesize that succesful model replication largely depends on completeness of model reporting. Insufficient reporting may prevent efficient and large-scale external validation, potentially resulting in small clinical uptake of published models (18).

In this study, we aim to externally validate existing dementia prediction models. To that end, we define replicability criteria, review published models, and externally validate three selected models using routinely collected health data from administrative claims and electronic health records.

2 Methods

2.1 Article selection

Our literature search for existing dementia prediction models was based on the search query presented in a systematic review on dementia risk prediction modelling by Tang et al. from 2015 (2). The search interval was extended from 1 January 2011 to 31 December 2020.

Articles were included if they met the following criteria: (1) the sample was population-based; (2) the risk model predicts the risk of dementia in non-demented individuals; (3) measurements of discrimination are provided, e.g. the area under the receiver operating characteristic curve (AUROC) or c-statistic.

2.2 Replicability criteria

This study does not develop or propose a prediction model, but merely applies existing models. If the prediction models to be applied are not presented in the form of a calculator, e.g., as a nomogram or chart score, it is necessary to replicate them from the accompanying documentation, such as the research paper and supplemental material. The replicability criteria that a study must report are presented in Table 1 and can be directly inferred from our prediction approach (Fig. 1). Among a population at risk, we predict which patients at a defined moment in time (the index) will experience some outcome during a time-at-risk. Prediction is done using only information about the patients in an observation window prior to the index.

Replicability criteria can be broadly categorized into population settings and statistical analysis settings (Table 1) (19).

Table 1
Model replicability criteria that prediction studies should report to enable external validation.

Category	Replicability criteria	Description
Population settings	Target population definition	Definition or description of the population for which predictions are made.
	Index date	Date at which a patient qualifies for inclusion in the target population.
	Time-at-risk	Time window in which a model's predictions are valid relative to the index date.
	Outcome definition	Definition or description of the outcome to be predicted during the time-at-risk.
Statistical analysis settings	Prediction method	Prediction methods in this study are limited to logistic regression and Cox proportional hazard for predicting a binary outcome.
	Predictor definitions	Predictor descriptions or definitions in terms of data source codes.
	Predictor time window	Time window in which the predictor is assessed.
	Model specifications	The prediction model, e.g., parameters to construct the model given a prediction method. Alternatively, a risk calculator or nomogram could be reported. We also distinguish here between fully and partially specified models. For example, if no intercept is reported in the case of a logistic regression model, we are still able to construct a simple linear model using only coefficient values. However, this method does not consider the baseline risk of the original model and (re-)calibration will not be assessed.

2.3 Data sources

For external validation, we selected a diverse set of electronic health record (EHR) and claims observational databases from different countries (Table 2).

IBM MarketScan Medicare Supplemental Database (MDCR) includes data from the health services of retirees in the United States with Medicare supplemental coverage through employer-sponsored plans. Optum De-Identified Clinformatics Data Mart – Socio-Economic Status (OPSES) Database is derived from administrative health claims for members of large commercial and Medicare Advantage health plans in the United States. The Iqvia Disease Analyzer Germany (IQGER) database consists of mostly primary care physician data collected from German practices and medical centers for all ages. Optum de-identified Electronic Health Record Dataset (OPEHR) represents longitudinal EHR data derived from dozens of healthcare provider organizations in the United States. The Clinical Practice Research Datalink

(CPRD) is a governmental, not-for-profit research service consisting of data collected from UK primary care for all ages. The Integrated Primary Care Information (IPCI) database is a Dutch database containing the complete medical record of patients provided by around 350 general practitioners (GPs) geographically spread over the Netherlands (20). Iqvia Medical Research Database (IMRD), incorporating The Health Improvement Network (THIN), is a longitudinal patient database collected from primary care practices in the UK.

All data sources have been mapped to the Observational Medical Outcome Partnership (OMOP) Common Data Model (CDM) version 5, which provides a standardized data structure and vocabulary (21).

Table 2
Data sources selected for external validation of replicated dementia prediction models.

Database	Acronym	No. of patients (million)	Country	Data type
IBM MarketScan® Medicare Supplemental Database	MDCR	10	United States	Claims
Iqvia Germany DA	IQGER	30	Germany	GP, EHR
Optum SES	OPSES	85	United States	Claims
Optum EHR	OPEHR	94	United States	EHR
Clinical Practice Research Datalink	CPRD	13	United Kingdom	GP
Integrated Primary Care Information	IPCI	2.5	Netherlands	GP
Iqvia Medical Research Database (incorporating THIN)	IMRD	18	United Kingdom	GP

2.3.1 Model selection for external validation

From the reviewed studies, we select models that were developed on one of our included data sources (Table 2). We replicate these models and apply them to their original development database, which allows us to approximate quality of model reporting. In the optimal case a similar discrimination performance as in the research paper should be achieved. A significantly lower performance could indicate poor replicability. This performance of the replicated model on the original development database will be referred to as “round-trip” performance. In addition, the selected models were externally validated on the remaining data sources.

If predictor coefficients are reported, we can often construct a risk stratification model that identifies high/low risk patients but does not assign an absolute risk estimate (22). This is achieved by scaling the

maximum achievable score of $\theta^T X$, θ being the coefficient vector, variable importance vector, or point vector, and X being the feature vector, with values between 0 and 1. Because there is no parameter that indicates baseline risk of the development population, we cannot assign a risk estimate and will not assess calibration for this kind of model.

For external validation we use the standardized patient-level prediction framework (PLP), which was developed by the Observational Health Data Science and Informatics (OHDSI) network (19). This framework enables the development of analysis packages in R that can be shared across sites that have access to data sources OMOP CDM. Our validation packages are populated on-site through computer-executable cohort and predictor definitions using SQL queries. The patient-level-prediction framework is based on best practices proposed by the Prognosis Research Strategy (PROGRESS) and follows the recommendations of the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) statement (23, 24).

2.4 Performance evaluation

External validation of a prediction model typically involves quantifying a model's discrimination and calibration performance.

A model's predicted risks must discriminate well between those participants who will and will not have the outcome of interest. Discrimination is generally reported by the area under the receiver operating characteristic curve (AUROC) or when censoring is considered by the concordance statistic (c-statistic), which in practice will take the same value for a binary prediction problem (25). We will use the AUROC to measure discrimination performance, which is computed from the area under the receiver operating curve, the plot of sensitivity vs. 1-specificity as the value of the cut-off point moves from 0 to 1.

Calibration examines the agreement between predicted and observed risks. Calibration can be visualized graphically in various ways, for example by plotting observed versus predicted risks across deciles of predicted risk or age groups. For this study, we decided on the latter method using age groups.

Transported models may also benefit from re-calibration so that predicted risk better matches the proportion of subjects that actually have the dementia outcome in the external data source. We will use slope and intercept re-calibration. To assess the relative improvement of a model through recalibration, we will assess E_{avg}, a single value metric, which is the average absolute difference between observed and predicted probabilities (26).

3 Results

3.1 Model reporting

The inclusion criteria of our literature search were met by 35 studies, which described a total of 59 prediction models (5, 7–16, 27–50). Table 3 summarizes the reporting of our replicability criteria in the

included articles.

Table 3
Reported model criteria for replicability by included studies.

Category	Validation criteria	No. of models (%)
Population settings	Target population definition	59 (100)
	Index date	23 (39)
	Time-at-risk	39 (66)
	Outcome definition	59 (100)
Statistical analysis settings	Prediction method	59 (100)
	Predictor definitions	46 (78)
	Predictor time window	21 (36)
	Model specifications: Full model	8 (14)
	Model specifications: Partial model	19 (32)

Criteria that were reported by all included studies were the target population definition, the outcome definition, and the prediction method. Various prediction methods were used, including Cox proportional hazard, single tests, logistic regression, linear discrimination analysis, competing risk regression, disease state index, random forest, and support vector machine. Most frequently reported prediction methods were Cox proportional hazard (13 studies, 21 models) and logistic regression (8 studies, 14 models).

Frequently reported criteria were the time-at-risk (66% of the models) and the predictor definitions (78% of the models). Most reported time-at-risk was between three and five years. Studies that did not explicitly state the time-at-risk or predicted over the full follow-up time of a patient were considered to not report this criterion.

Rarely reported criteria include the index date, the predictor time window, and the full model specifications. Non-demographic predictors were most commonly measured in a time window between one year and five years before index.

Of the included studies, three reported all nine replicability criteria for a total of seven proposed models (27). The median number of reported criteria across all included models was five.

3.2 Externally validated models

We selected one of the seven fully reported models, Walters' Dementia Risk Score for persons aged 60–79, for replication and dismissed the other six for various reasons: Walters et al. did not endorse their second model aimed at persons aged 80–95 for clinical use due to low discrimination performance (15);

four models which used detailed education variables (0 to 5 years of primary school, Vocational school certificate, French junior-school diploma, French high school diploma, Graduate studies) that were unavailable in the validation databases (29); one model was developed on data from a prospective cohort study, Adult Changes in Thought (ACT), which is currently not available in the OMOP CDM format (27). Of the partially reported models, there were two for which the development data and predictors were available in the OMOP CDM, and for which missing criteria, such as the baseline hazard and time-at-risk could be left out or approximated under reserve. Therefore, we selected the following three models for external validation: (1) Walters' Dementia Risk Score which predicts 5-year risk of first recorded dementia diagnosis among patients aged 60–79 using a Cox proportional hazard model and was developed on THIN/IMRD (15); (2) Mehta's RxDx-Dementia Risk Index predicts risk of incident dementia among patients diagnosed with type 2 diabetes mellitus and hypertension using a Cox proportional hazard model developed on CPRD (16); and (3) Nori's ADRD prediction model which predicts Alzheimer's disease and related dementias (ADRD) among patients aged 45 and older using a logistic regression model developed on OptumLabs (13).

Of the externally validated models, the Mehta model did not report the baseline hazard and the time-at-risk and the Nori model did not report the time-at-risk. Because missing information could not be provided by authors, we decided to use a 5-year time-at-risk as used by the Walters model and many of the other reviewed models.

3.2.1 Walters model

The Walters model was found to report all replicability criteria defined in Table 1. It was developed on the THIN database and had several notable modeling decisions. For example, during development data imputation has been used for various numeric variables. Of the six imputed variables (smoking, height, total cholesterol, HDL cholesterol, systolic blood pressure, and weight) only smoking remained in the final model. In the replicated model the smoking status is imputed by assuming that patients that have neither a code for "smoker" nor "ex-smoker" are considered "non-smokers". This demonstrates a general shortcoming of observational data, as the absence of a code does not guarantee the absence of a condition, drug, or in this case smoking, and the code may simply not have been recorded despite the patient being a smoker or ex-smoker.

The Walters models uses a variable called "social deprivation score", which ranges from 1 to 5 indicating social deprivation. The information in this variable has been established through a linkage of the UK postal (zip) code recorded in patient notes to UK Population Census data. However, this linkage is no longer available, unlikely to exist in other databases across the world, and establishing the linkage may not be possible or feasible.

The index date (and start of follow-up) of the Walters model is the latest of four entry events: (1) 1 January 2000, (2) when the individual turned 60, (3) one year following new registration with a THIN practice, and (4) one year after the practice met standard criteria for accurate recording of deaths,

consultation, health measurements and prescribing. Only the index date of a patient turning 60 could be fully replicated. The start of follow-up on 1 January 2000 was not applicable to any of the data sources as it lies too far in the past. The remaining two index events are THIN-specific and could not be replicated in other databases, including IMRD. Considering that the model is meant for patients aged 60–79, we extended the inclusion for the index event to patients aged 60–79 with the latest visit in their patient record as entry event. Visits are suitable for defining index dates, because they indicate interaction with a healthcare provider that may be qualified to apply a model and interpret its predictions.

The paper mentions that Read codes were used for development, which is a hierarchical coding system that maps onto ICD-10 codes. The authors provide literal names of predictors, for which the corresponding code could be determined by us.

3.2.2 Mehta model

The research paper does not report the full model, which is a Cox proportional hazard model. While the coefficients are reported, the baseline hazard and the time-at-risk are missing. We have contacted the authors of this study, but they were unable to provide us with this information. We are still able to replicate the model for an estimated time-at-risk of 5 years and by normalizing the values of $\theta^T X$ to a risk score between 0 and 1, where θ and X are the coefficient vector and feature vector, respectively. However, without the baseline hazard we are unable to assess calibration and will report discrimination performance only.

In addition, no data source codes or vocabularies are provided for the predictors, so that predictors needed to be replicated from the medical terms.

3.2.3 Nori model

The Nori model did not explicitly report the time-at-risk. As with the Mehta model, we are still able to replicate the model for an estimated time-at-risk of 5 years.

The paper provides ICD-9-CM codes for diagnoses and CPT-4 codes for procedures that are used as predictors in the final model. The OMOP-CDM uses CPT-4 as the standard vocabulary for procedures and SNOMED CT for diagnoses. However, a mapping table from ICD-9-CM to SNOMED CT is available. Therefore, we could replicate predictors using exact code definitions for the Nori model.

A characteristic of the Nori model was a complex target population definition with multiple entry events and various observation windows. Given a written definition and graphical representation (Fig. 1 in original paper) of the target population, replication was notably more difficult than for the other replicated models (13).

3.2.4 External validation performance

Table 4 provides the discrimination and Table 5 the recalibration performance of the replicated models. Calibration and re-calibration in terms of the E_{avg} was only assessed if the model's authors provided the

baseline risk, for example in the form of the intercept or baseline hazard. In Fig. 2 we present “round-trip” calibration as observed versus expected risks.

Walters’ Dementia risk score performed best during its development on THIN and worst after model replication and validation on CPRD, MDCR, and IMRD. Interestingly, IMRD, which incorporates THIN, presents the best approximation of the development data and still shows a significant performance deterioration for the round-trip. Figure 2a shows the Walters model round-trip calibration of the original Walters model on IMRD indicating moderate agreement between observed and predicted risk for the entire target population.

Mehta’s RxDx-Dementia Risk index performed best during development on CPRD and almost equally well in the three primary care databases CPRD, IPCI, and IMRD.

Nori’s ADRD dementia prediction model performed best during development on OptumLabs and almost equally well in the remaining data sources. Interestingly, the round-trip performance on OPEHR was the worst. In Fig. 2c we learn that the model overpredicts the round-trip risk in the target population of CPRD.

Almost all models show improvements of the E_{avg} after recalibration (Table 5). Recalibration for the Mehta model was not assessed because no baseline hazard was provided.

Table 4. Internal and external discrimination performance in AUROC of externally validated models. The round-trip performances for each model are presented in the shaded cells.

Model	Development database	MDCR	IQGER	OPSES	OPEHR	CPRD	IPCI	IMRD
Walters	0.84 (THIN)	0.69 (0.69 – 0.69)*	0.75 (0.75 – 0.75)*	0.74 (0.74 – 0.74)*	0.73 (0.73 – 0.73)*	0.67 (0.66 – 0.67)*	0.76 (0.75 – 0.77)*	**0.68 (0.68 – 0.69)***
Mehta	0.81 (CPRD)	0.69 (0.69 – 0.70)	0.72 (0.71 – 0.72)	0.71 (0.70 – 0.71)	0.73 (0.73 – 0.73)	**0.79 (0.78 – 0.80)**	0.78 (0.76 – 0.80)	0.79 (0.78 – 0.80)
Nori	0.69 (Optum)	0.66 (0.66 – 0.67)	0.67 (0.66 – 0.68)	0.67 (0.66 – 0.68)	**0.62 (0.62 – 0.63)**	0.68 (0.67 – 0.69)	0.64 (0.62 – 0.67)	0.68 (0.68 – 0.69)

*Discrimination AUROC (95% confidence intervals)

Table 5
External calibration and recalibration performance in E_{avg} of externally validated models. Calibration of Mehta's RxDx-Dementia Risk Index was not assessed due to missing baseline hazard.

Model	MDCR	IQGER	OPSES	OPEHR	CPRD	IPCI	IMRD
Walters	0.060 (0.002)*	0.025 (0.011)	0.064 (0.011)	0.057 (0.032)	0.073 (0.015)	0.024 (0.011)	0.065 (0.001)
Mehta	-	-	-	-	-	-	-
Nori	0.164 (0.001)	0.142 (0.001)	0.258 (0.001)	0.170 (0.001)	0.198 (0.001)	0.790 (0.0002)	0.19 (0.001)

* Calibration E_{avg} (recalibrated E_{avg})

4 Discussion

We assessed reporting of published dementia prediction models and found shortcomings in reporting essential information that would allow for full model replication.

Our results showed that while reporting was complete for some criteria such as target and outcome definitions, reporting of statistical analysis criteria are mostly insufficient to fully replicate the dementia prediction models. Moreover, our external validation of three selected models showed that even if reporting was sufficient for replication, it does not guarantee that replication and external validation becomes non-trivial, because predictors had to be present, and inclusion and exclusion criteria of target and outcome had to be generalizable to other data sources. Performance across external data sources showed substantial differences in discrimination performance as compared to the reported development performance.

4.1 Model reporting

All studies reported the target population and the outcome. However, only 23 of 59 models reported the index date. The problem arises that although it is clear for which (sub) population a risk model is meant to be used, it often remains unclear at what point in time the model is to be applied. A better solution for choosing an overall index date is using a visit or a condition diagnosis, which are associated with an individual date per patient. Additionally, a visit or a diagnosis date most likely involves interaction with a healthcare provider who is qualified to apply a model and interpret its results.

The time-at-risk is anchored to the index date and determines during which time the predictions of a model are valid. There were 39 of 59 models that explicitly reported the time-at-risk, while for the remaining studies it was unclear. Some studies would use the full follow-up of each individual patient, however, it remains unclear what the valid time frame is following the index date, for which reason these models cannot be applied reliably.

A majority of the studies reported predictor definitions or at least names that can be interpreted to replicate a predictor. However, only 21 of 59 models provided the time window in which the predictor is measured. Predictor definitions or descriptions without time window are not useful for non-demographic predictors. It could make a significant difference whether a predictor was recorded recently or 20 years in the past. More importantly, the replicated model should match the original model settings, which cannot be achieved if predictor time windows are not reported. In addition, only 19 of 59 models provided a partial model, for example only coefficient values, and 8 of 59 models provided the full model. Therefore, our results suggest that while population settings are moderately well reported, there is lack of reporting statistical analysis settings, which in many cases make replication and external validation impossible.

Calibration is essential to assess if a model underestimates or overestimates outcome risk in an external population. Original calibration can be computed, if the intercept/baseline hazard or similar baseline risk parameters are reported. Only 10 of 59 models provided this information. Recalibration should generally be done, which yields best results if the original baseline risk is known.

4.1.1 Walters model

The performance of the Walters model in the external databases was good.

The “round-trip” performance after validating on IMRD is low compared to the development performance on THIN, yet agrees with the performance on CPRD, which is also a UK primary care database of similar structure. The reason for this performance deterioration is not immediately evident. Possible causes may be the entry events that could not be replicated, the visit entry event we defined ourselves, the missing social deprivation predictor, or inaccuracy when matching the literal predictor names to Read codes.

4.1.2 Mehta model

The Metha model saw a drop off in performance across all external data sources. The “round-trip” performance after validating on CPRD was 0.79 as compared to 0.81 during development. This model was explicitly reported, which made replication easier, for example predictors were provided in the form of code lists. We had to assume the missing time-at-risk to be 5 years, which is a value commonly used across the reviewed models. Due to the good performance on CPRD, we are confident that the model is mostly well replicated, despite not having the baseline hazard. However due to incomplete model specification, calibration is unknown.

Generally, a model that is not completely reported should not be considered for clinical use. In this case, the 5-year time-at-risk appears to work well since we based it on the design choice of other reviewed dementia models. However, instead of replicating an incomplete model we suggest to take explicit target cohort, outcome or predictor definitions as a starting point to build new models directly on the validation databases.

4.1.3 Nori model

The Nori model was the lowest performing model with a 0.69 AUC during development. This dropped to 0.62 on OPEHR, after the “round-trip” while maintaining good calibration (Fig. 2c). The target population definition appeared complex, with four different cohort entry events causing difficulties during replication. Replication would have benefitted from a more verbose and systematic presentation of such a complex target population.

4.2 Implications

The lack of external validation in dementia prediction literature can to some extend be attributed to the insufficient reporting of models. Models should be developed with external validation in mind. This could for example mean to report all aspects of the model explicitly. Such transparency is best achieved programmatically through code lists and underlying logic rather than literal descriptions, for example by providing a full description of the model (development) in code, ideally against a common data model. This approach will likely eliminate ambiguity as a source of error. For example, Nori’s ADRD prediction model uses two variables named “Diabetes Mellitus”, which originate from ICD9CM codes 250.00 and 250.02. If these codes were not provided by the authors, it would not have been possible to verify hat the former codes specifies “not stated as uncontrolled” and the latter “uncontrolled”.

Development choices should not rely on properties unique to the development database, e.g., the Walters model contained criteria to define the target population and predictors that did not exist in the external data sources, for example the cohort entry event “one year following new registration with a THIN practice”.

In general, authors should avoid uncommon predictors during model development to guarantee replicability, if the model is meant to be applied in external healthcare settings. Instead of building a single model with multiple, complex cohort entry events, it can be beneficial to build a model for each entry event, which may be easier to interpret and replicate. The Nori model suffered from this problem as it had a complex target population definition with multiple entry events. Defining the time-at-risk window is crucial to indicate in which time window a model’s predictions are valid. Using the full follow-up of a population is not a valid approach, as follow-up can vary per person.

Recalibration showed improvements in the E_{avg} across most models and databases, however, this can only be observed when the intercept or baseline hazard is reported. To perform recalibration in an external setting, an annotated dataset is required. If such a dataset is available, the question arises, whether developing a new model altogether, potentially using definitions from existing models, may be an even better approach. Recalibration performed during our external validations shows improvements, but if no intercept/baseline hazard is reported possible recalibration improvements cannot be assessed.

Recalibration needs to be performed, if discrepancy between expected and observed risk is large, which can be assessed through visualizations such as Fig. 2.

4.3 Limitations

We reviewed studies for replicability criteria to the best of our ability, however, due to vague descriptions, that leave room for interpretation, our general approach was to consider criteria as not reported once uncertain. Moreover, this study is purely methodological and assesses the quality of model reporting in existing dementia prediction literature. We did? not assess the clinical usefulness of any of the replicated models, nor do we endorse any of the models for clinical use.

The “round-trip” is meant to approximate the performance of the replicated model on the original development data. However, over time the composition of people in the databases may have changed.

5 Conclusion

Many dementia risk prediction models have been developed, but only a handful have been externally validated (3, 4). We reviewed 35 studies that proposed a total of 59 dementia risk models. We observed that reporting is mostly insufficient to fully replicate and externally validate published dementia prediction models, and therefore, it is uncertain how well these models would work in other clinical settings. In addition, we replicated and externally validated three existing dementia prediction models and encountered difficulties beyond our replicability criteria, such as ambiguous cohort or predictor definitions. We emphasise the importance of following established guidelines for reporting clinical prediction model. We recommend that reporting should be more explicit and have external validation in mind if the model is meant to be applied in different settings.

List Of Abbreviations

ADRD	Alzheimer’s disease and related dementias
OHDSI	Observational Health Data Sciences and Informatics
OMOP	Observational Medical Outcomes Partnership
PROGRESS	Prognosis Research Strategy
TRIPOD	Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis
PLP	Patient-level prediction R-package framework
AUROC	Area under the Receiver Operating Characteristic curve

Declarations

Ethics approval and consent to participate

All patient data included in this study were deidentified.

The New England Institutional Review Board determined that studies conducted in Optum, IBM MDCR, Iqvia Germany are exempt from study-specific IRB review, as these studies do not qualify as human subjects research.

IMRD (study reference 21SRC025), CPRD (study reference 22_001724) and IPCI (study reference 2/2022) had institutional review board approval for the analysis, or used deidentified data, and thus the analysis was determined not to be human subjects research and informed consent was not deemed necessary at any site. IMRD incorporates data from THIN, A Cegedim Database. Reference made to THIN is intended to be descriptive of the data asset licensed by IQVIA.

Consent for publication

Not applicable

Availability of data and materials

The Optum and IBM MDCR data that support the findings of this study are available from IBM MarketScan Research Databases (contact at: http://www.ibm.com/us-en/marketplace/marketscan-research-databases) and Optum (contact at: http://www.optum.com/solutions/data-analytics/data/real-world-data-analytics-a-cpl/claims-data.html) but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available.

Due to ethical concerns, supporting data cannot be made openly available for the CPRD, IMRD, IPCI and Iqvia Germany datasets.

The replicated models that support the findings of this study are available in a study package on the mi-erasmusmc GitHub repository (https://github.com/mi-erasmusmc/EmcDementiaModelValidation). The study package require the OHDSI environment (https://github.com/ohdsi), notably the Patient-Level Prediction framework (https://github.com/OHDSI/PatientLevelPrediction). More information can be found on the network homepage (https://OHDSI.org).

Competing interests

Jenna M. Reps is an employee of Janssen Research & Development and shareholder of Johnson & Johnson. Peter R. Rijnbeek, Egill A. Fridgeirsson, Luis H. John, Jan A. Kors work for a research group who received unconditional research grants from Boehringer-Ingelheim, GSK, Janssen Research & Development, Novartis, Pfizer, Yamanouchi, Servier. None of these grants result in a conflict of interest to the content of this paper.

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.

Authors' contributions

L.H.J. lead and J.A.K., E.A.F., J.M.R. and P.R.R. contributed to the conception and design of the work. J.A.K. prepared the literature review, J.M.R assisted in preparing the technical analysis and L.H.J prepared the data and implemented and carried out the technical analysis. L.H.J., J.A.K., E.A.F., J.M.R, and P.R.R. contributed to the interpretation of the results. L.H.J. took the lead in writing the manuscript. All authors provided critical feedback and helped shape the research, analysis and manuscript.

Acknowledgements

The authors would like to thank Sarah Seager and Steven Salama from IQVIA for executing study packages on the IQVIA Medical Research Database (IMRD).

References

1. Stephan BC, Kurth T, Matthews FE, Brayne C, Dufouil C. Dementia risk prediction in the population: are screening models accurate? Nat Rev Neurol. 2010;6(6):318–26.
2. Tang EY, Harrison SL, Errington L, Gordon MF, Visser PJ, Novak G, et al. Current developments in dementia risk prediction modelling: an updated systematic review. PloS one. 2015;10(9):e0136181.
3. Hou XH, Feng L, Zhang C, Cao XP, Tan L, Yu JT. Models for predicting risk of dementia: a systematic review. Journal of Neurology, Neurosurgery and Psychiatry. 2019;90(4):373–9.
4. Goerdten J, Čukić I, Danso SO, Carrière I, Muniz-Terrera G. Statistical methods for dementia risk prediction and recommendations for future work: A systematic review. Alzheimer's and Dementia: Translational Research and Clinical Interventions. 2019;5:563–9.
5. Jacqmin-Gadda H, Blanche P, Chary E, Loubère L, Amieva H, Dartigues J-F. Prognostic score for predicting risk of dementia over 10 years while accounting for competing risk of death. American journal of epidemiology. 2014;180(8):790–8.
6. Derby CA, Burns LC, Wang C, Katz MJ, Zimmerman ME, L'Italien G, et al. Screening for predementia AD: time-dependent operating characteristics of episodic memory tests. Neurology. 2013;80(14):1307–14.
7. Restaino M, Matthews FE, Minett T, Albanese E, Brayne C, Stephan BCM. Predicting risk of 2-year incident dementia using the CAMCOG total and subscale scores. Age and ageing. 2013;42(5):649–53.
8. Mossaheb N, Zehetmayer S, Jungwirth S, Weissgram S, Rainer M, Tragl K-H, et al. Are specific symptoms of depression predictive of Alzheimer's dementia? The Journal of clinical psychiatry. 2012;73(7):1009–15.
9. Song X, Mitnitski A, Rockwood K. Nontraditional risk factors combine to predict Alzheimer disease and dementia. Neurology. 2011;77(3):227–34.
10. Ehreke L, Luppa M, König H-H, Villringer A, Riedel-Heller SG. Does the clock drawing test predict dementia? Results of the Leipzig longitudinal study of the aged (LEILA 75+). Dementia and geriatric

cognitive disorders. 2011;31(2):89–97.

11. Park JH, Cho HE, Kim JH, Wall MM, Stern Y, Lim H, et al. Machine learning prediction of incidence of Alzheimer's disease using large-scale administrative health data. NPJ digital medicine. 2020;3(1):1–7.
12. Park KM, Sung JM, Kim WJ, An SK, Namkoong K, Lee E, et al. Population-based dementia prediction model using Korean public health examination data: a cohort study. PloS one. 2019;14(2):e0211957.
13. Nori VS, Hane CA, Martin DC, Kravetz AD, Sanghavi DM. Identifying incident dementia by applying machine learning to a very large administrative claims dataset. PLoS ONE. 2019;14(7).
14. Albrecht JS, Hanna M, Kim D, Perfetto EM. Predicting diagnosis of Alzheimer's Disease and related dementias using administrative claims. Journal of Managed Care and Specialty Pharmacy. 2018;24(11):1138–45.
15. Walters K, Hardoon S, Petersen I, Iliffe S, Omar RZ, Nazareth I, et al. Predicting dementia risk in primary care: development and validation of the Dementia Risk Score using routinely collected data. BMC Med. 2016;14:6.
16. Mehta HB, Mehta V, Tsai C-L, Chen H, Aparasu RR, Johnson ML. Development and validation of the RxDx-Dementia risk index to predict dementia in patients with type 2 diabetes and hypertension. Journal of Alzheimer's Disease. 2016;49(2):423–32.
17. Riley RD, Ensor J, Snell KI, Debray TP, Altman DG, Moons KG, et al. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. BMJ. 2016;353:i3140.
18. Reps JM, Williams RD, You SC, Falconer T, Minty E, Callahan A, et al. Feasibility and evaluation of a large-scale external validation approach for patient-level prediction in an international data network: validation of models predicting stroke in female patients newly diagnosed with atrial fibrillation. BMC Med Res Methodol. 2020;20(1):102.
19. Reps JM, Schuemie MJ, Suchard MA, Ryan PB, Rijnbeek PR. Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data. J Am Med Inform Assoc. 2018.
20. de Ridder MA, de Wilde M, de Ben C, Leyba AR, Mosseveld BM, Verhamme K, et al. Data resource profile: the Integrated Primary Care Information (IPCI) database, the Netherlands. International Journal of Epidemiology. 2022.
21. Voss EA, Makadia R, Matcho A, Ma Q, Knoll C, Schuemie M, et al. Feasibility and utility of applications of the common data model to multiple, disparate observational health databases. Journal of the American Medical Informatics Association. 2015;22(3):553–64.
22. Hendrich AL, Bender PS, Nyhuis A. Validation of the Hendrich II Fall Risk Model: a large concurrent case/control study of hospitalized patients. Applied Nursing Research. 2003;16(1):9–21.
23. Steyerberg EW, Moons KG, van der Windt DA, Hayden JA, Perel P, Schroter S, et al. Prognosis Research Strategy (PROGRESS) 3: prognostic model research. PLoS Med. 2013;10(2):e1001381.

24. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): the TRIPOD Statement. Br J Surg. 2015;102(3):148–58.
25. Collins GS, Ogundimu EO, Cook JA, Manach YL, Altman DG. Quantifying the impact of different approaches for handling continuous predictors on the performance of a prognostic model. Statistics in medicine. 2016;35(23):4124–35.
26. Austin PC, Steyerberg EW. The Integrated Calibration Index (ICI) and related metrics for quantifying the calibration of logistic regression models. Statistics in medicine. 2019;38(21):4051–65.
27. Barnes DE, Zhou J, Walker RL, Larson EB, Lee SJ, Boscardin WJ, et al. Development and validation of eRADAR: a tool using EHR data to detect unrecognized dementia. Journal of the American Geriatrics Society. 2020;68(1):103–11.
28. Li CI, Li TC, Liu CS, Liao LN, Lin WY, Lin CH, et al. Risk score prediction model for dementia in patients with type 2 diabetes. European journal of neurology. 2018;25(7):976–83.
29. Mura T, Baramova M, Gabelle A, Artero S, Dartigues J-F, Amieva H, et al. Predicting dementia using socio-demographic characteristics and the Free and Cued Selective Reminding Test in the general population. Alzheimer's research & therapy. 2017;9(1):1–11.
30. Chouraki V, Reitz C, Maury F, Bis JC, Bellenguez C, Yu L, et al. Evaluation of a genetic risk score to improve risk prediction for Alzheimer's disease. Journal of Alzheimer's Disease. 2016;53(3):921–32.
31. Vuoksimaa E, Rinne JO, Lindgren N, Heikkilä K, Koskenvuo M, Kaprio J. Middle age self-report risk score predicts cognitive functioning and dementia in 20–40 years. Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring. 2016;4:118 – 25.
32. Kochan NA, Bunce D, Pont S, Crawford JD, Brodaty H, Sachdev PS. Reaction time measures predict incident dementia in community-living older adults: The Sydney Memory and Ageing Study. The American Journal of Geriatric Psychiatry. 2016;24(3):221–31.
33. Stephan BC, Tzourio C, Auriacombe S, Amieva H, Dufouil C, Alpérovitch A, et al. Usefulness of data from magnetic resonance imaging to improve prediction of dementia: population based cohort study. bmj. 2015;350.
34. Barnes DE, Beiser AS, Lee A, Langa KM, Koyama A, Preis SR, et al. Development and validation of a brief dementia screening indicator for primary care. Alzheimer's & Dementia. 2014;10(6):656–65. e1.
35. Hessler J, Tucha O, Förstl H, Mösch E, Bickel H. Age-correction of test scores reduces the validity of mild cognitive impairment in predicting progression to dementia. PloS one. 2014;9(8):e106284.
36. Ebbert MT, Ridge PG, Wilson AR, Sharp AR, Bailey M, Norton MC, et al. Population-based analysis of Alzheimer's disease risk alleles implicates genetic interactions. Biological psychiatry. 2014;75(9):732–7.
37. Verhaaren BF, Vernooij MW, Koudstaal PJ, Uitterlinden AG, van Duijn CM, Hofman A, et al. Alzheimer's disease genes and cognition in the nondemented general population. Biological psychiatry. 2013;73(5):429–34.

38. Exalto LG, Biessels GJ, Karter AJ, Huang ES, Katon WJ, Minkoff JR, et al. Risk score for prediction of 10 year dementia risk in individuals with type 2 diabetes: a cohort study. The Lancet Diabetes & Endocrinology. 2013;1(3):183–90.
39. Chary E, Amieva H, Pérès K, Orgogozo J-M, Dartigues J-F, Jacqmin-Gadda H. Short-versus long-term prediction of dementia among subjects with low and high educational levels. Alzheimer's & Dementia. 2013;9(5):562–71.
40. Okereke OI, Pantoja-Galicia N, Copeland M, Hyman BT, Wanggaard T, Albert MS, et al. The SIST-M: predictive validity of a brief structured clinical dementia rating interview. Alzheimer disease and associated disorders. 2012;26(3):225.
41. Jessen F, Wiese B, Bickel H, Eiffländer-Gorfer S, Fuchs A, Kaduszkiewicz H, et al. Prediction of dementia in primary care patients. PloS one. 2011;6(2):e16852.
42. Ohara T, Ninomiya T, Kubo M, Hirakawa Y, Doi Y, Hata J, et al. Apolipoprotein genotype for prediction of Alzheimer's disease in older Japanese: the Hisayama Study. Journal of the American Geriatrics Society. 2011;59(6):1074–9.
43. Cremers LG, Huizinga W, Niessen WJ, Krestin GP, Poot DH, Ikram MA, et al. Predicting Global Cognitive Decline in the General Population Using the Disease State Index. Frontiers in aging neuroscience. 2020;11:379.
44. Licher S, Leening MJ, Yilmaz P, Wolters FJ, Heeringa J, Bindels PJ, et al. Development and validation of a dementia risk prediction model in the general population: an analysis of three longitudinal studies. American Journal of Psychiatry. 2019;176(7):543–51.
45. Hall A, Pekkala T, Polvikoski T, van Gils M, Kivipelto M, Lötjönen J, et al. Prediction models for dementia and neuropathology in the oldest old: the Vantaa 85 + cohort study. Alzheimer's research & therapy. 2019;11(1):1–12.
46. Pekkala T, Hall A, Lötjönen J, Mattila J, Soininen H, Ngandu T, et al. Development of a late-life dementia prediction index with supervised machine learning in the population-based CAIDE study. Journal of Alzheimer's Disease. 2017;55(3):1055–67.
47. Downer B, Kumar A, Veeranki SP, Mehta HB, Raji M, Markides KS. Mexican-American Dementia Nomogram: Development of a Dementia Risk Index for Mexican-American Older Adults. Journal of the American Geriatrics Society. 2016;64(12):e265-e9.
48. Coupé P, Fonov VS, Bernard C, Zandifar A, Eskildsen SF, Helmer C, et al. Detection of Alzheimer's disease signature in MR images seven years before conversion to dementia: Toward an early individual prognosis. Human brain mapping. 2015;36(12):4758–70.
49. Exalto LG, Quesenberry CP, Barnes D, Kivipelto M, Biessels GJ, Whitmer RA. Midlife risk score for the prediction of dementia four decades later. Alzheimer's & Dementia. 2014;10(5):562–70.
50. Ewers M, Brendel M, Rizk-Jackson A, Rominger A, Bartenstein P, Schuff N, et al. Reduced FDG-PET brain metabolism and executive function predict clinical progression in elderly healthy subjects. NeuroImage: Clinical. 2014;4:45–52.

Figures

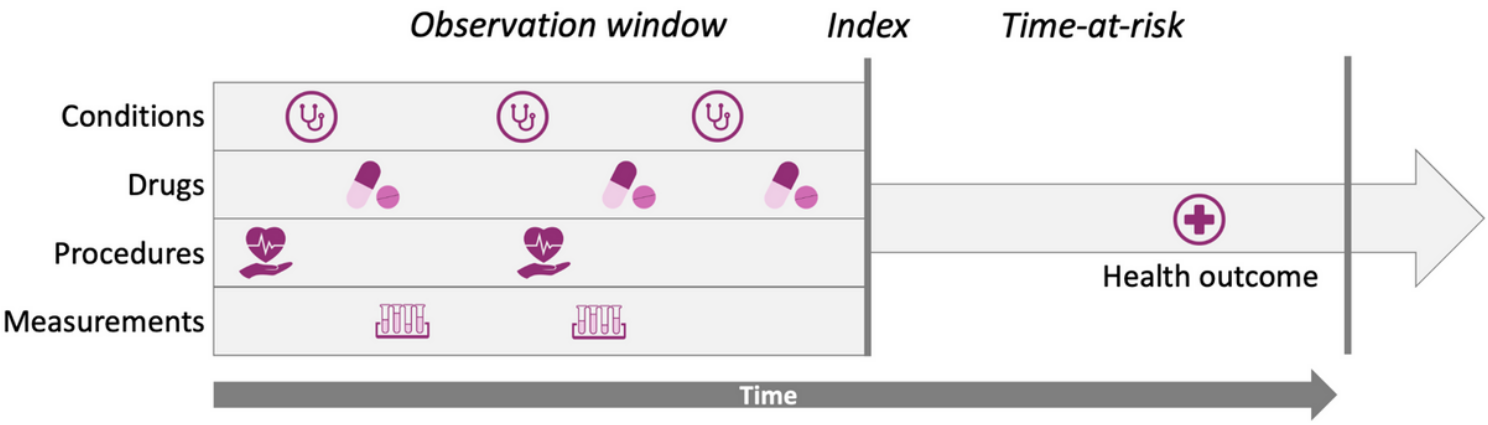
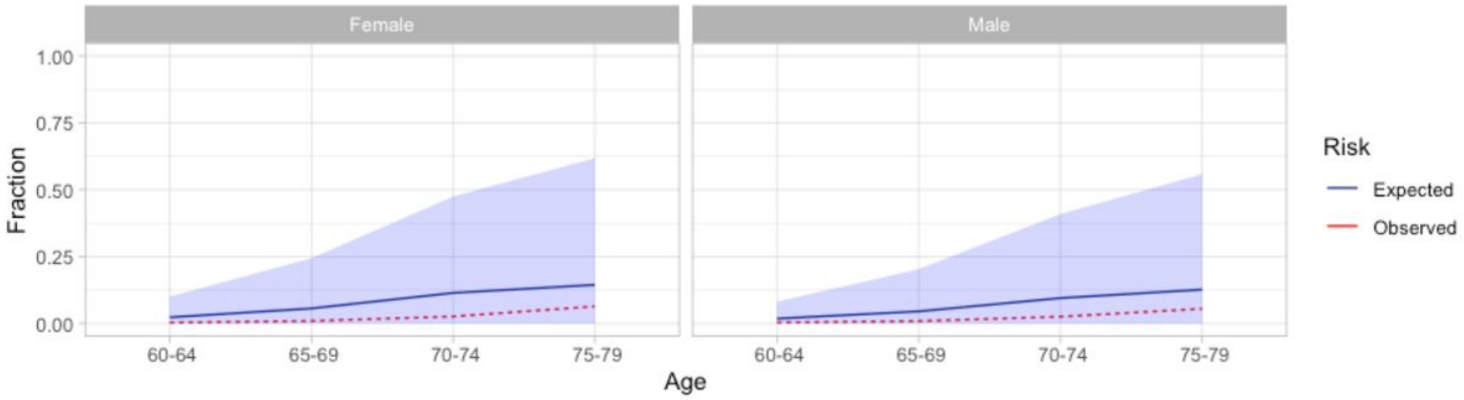
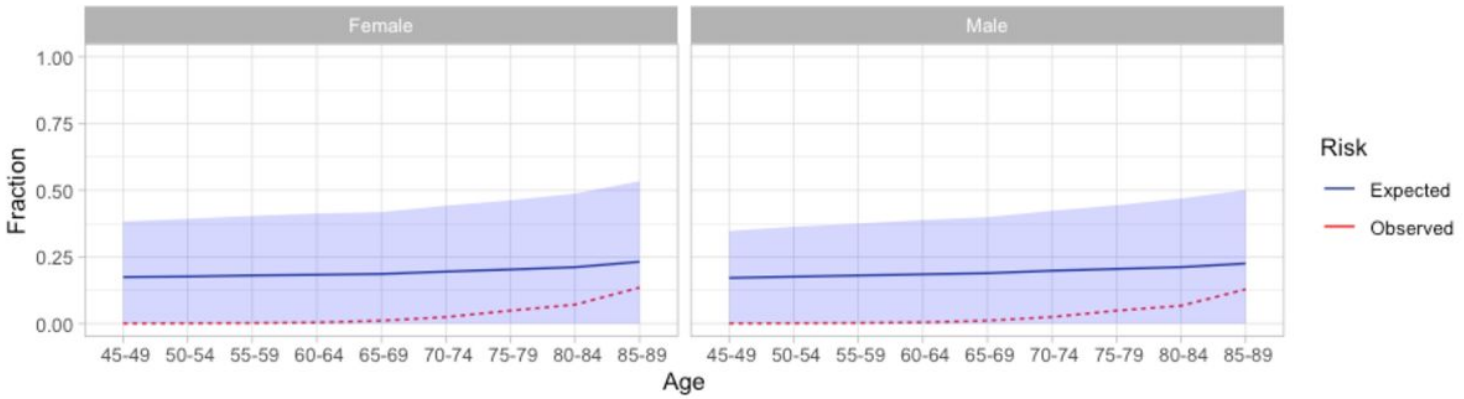


Figure 1

Patient-level prediction time windows and index date.



(a)



(b)

Figure 2

Round-trip calibration presented as observed versus expected risks across sex and age for non re-calibrated models: (a) Walters' Dementia Risk Score on IMRD; (b) Nori's ADRD prediction model on OPEHR. The shaded are presents the 95% confidence interval of the expected risk.