Title A causal framework for assessing the transportability of clinical prediction models

Jana Fehr*[1,2], Marco Piccininni[3,4], Tobias Kurth[3], Stefan Konigorski*[1,2,5]

1 Digital Health and Machine Learning, Hasso-Plattner-Institute, Potsdam, Germany
2 Digital Engineering Faculty, University of Potsdam, Germany
3 Institute of Public Health, Charité – Universitätsmedizin Berlin, Berlin, Germany
4 Center for Stroke Research Berlin, Charité – Universitätsmedizin Berlin, Berlin, Germany
5 Hasso Plattner Institute for Digital Health at Mount Sinai, Icahn School of Medicine at Mount Sinai, New York, NY, USA.

* Corresponding authors: Jana Fehr & Stefan Konigorski
Hasso-Plattner-Institut für Digital Engineering gGmbH
Prof.-Dr.-Helmert-Str. 2-3
14482 Potsdam, Germany
e-Mail: jana.fehr@hpi.de; stefan.konigorski@hpi.de

†Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in the analysis or writing of this report. A complete listing of ADNI investigators can be found at: https://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

Abstract

BACKGROUND: Machine learning promises to support the diagnosis of dementia and Alzheimer's Disease, but may not perform well in new settings. We present a framework to assess the transportability of models predicting cognitive impairment in external settings with different demographics.

METHODS: We mapped and quantified relationships between variables associated with cognitive impairment using causal graphs, structural equation models, and data from the ADNI study. These estimates generated datasets for training and validating prediction models. We measured transportability to external settings with interventions on age, APOE ε4, and sex, using calibration metric differences.

RESULTS: Models predicting with causes of the outcome were 1.3-12.8 times more transportable than those predicting with consequences. Logistic and lasso models had better calibration in internal validation settings than random forest and boosted models.

DISCUSSION: Applying a framework considering causal relationships is crucial to assess transportability. Future research could investigate more interventions and methods to quantify causal relationships in risk prediction.

Keywords: Alzheimer's Disease, clinical risk prediction, DAG, causality, transportability

Research in context

1. Systematic Review: Machine learning models supporting the diagnosis of cognitive impairment may not perform well in external validation settings. Theoretical research established that models can be more transportable to external settings when predictors are causes of the outcome. Causal frameworks and practical examples to assess transportability are needed.
2. Interpretation: We developed and applied a causal framework to assess the transportability of models predicting cognitive impairment to settings with different demographics using a causal graph and interventions on semi-synthetic data. Our results add a practical example showing that models are more transportable when predicting with causes of the outcome rather than with its consequences. This supports using causal frameworks in prediction models to improve transportability.
3. Future directions: Our framework can be extended to include more complex semi-synthetic data generation methods to quantify causal relationships. Further applications to risk prediction models could assess transportability under different interventions that simulate complex differences between populations.

Introduction

Alzheimer's disease (AD) and other forms of dementia are the second leading cause of death globally [1] and more than 55 million people currently suffer from dementia. Detecting dementia at an early stage of cognitive impairment is important to give affected individuals adequate care and eventually administer disease-modifying treatments [2]. In recent years,

several machine learning (ML) models have been proposed to support clinical decision making by predicting the diagnosis of AD and cognitive impairment [3–8]. One obstacle for deploying such prediction models in clinical practice is that they are often developed in a particular setting (e.g., one hospital or region), but might not generalize well when being transported (i.e., being applied) to other settings (e.g., in another hospital or regions with different patient demographics). One reason for reduced transportability may be that ML models learn non-causal associations between input and output variables, which might be different in external settings [9,10]. This scenario can occur especially when models predict a diagnosis based on clinical consequences of the disease (e.g., when prediction is in the anti-causal direction) [11–13].

Two approaches have the potential to improve transportability for prediction models. First, causal relationships can be incorporated in prediction models *a priori* for learning relationships that are more stable across settings and can therefore avoid systematic failures in external settings [10,14–17]. To this aim, directed acyclic graphs (DAGs) are a useful tool to map assumed causal relationships between variables, represent differences and commonalities between settings [18,19], and select variables for transportable health prediction tasks [20–22]. Second, the causal validity of learned relationships can be assessed through guided interventions on data distributions, to simulate differences between internal and external validation settings [17,23,24].

Few previous studies have employed causal thinking and DAGs to develop transportable clinical prediction models [25–29]. Piccininni et al. described the use of DAGs for selecting predictors in a hypothetical clinical risk prediction model for AD [25]. They discussed that prediction models for AD are more likely to transport well to different settings when the selected predictor variables are causes of AD and not consequences. However, their study included only three variables and a theoretical simulation.

While theories on transportability exist, statistical frameworks are needed to assess the transportability of trained prediction models in new settings in practice. In this work, we apply

a causal framework to assess the transportability of clinical prediction models for cognitive impairment in external validation settings with different demographics.

Methods

Overview

As first step in our framework, we mapped assumptions about causes and consequence of cognitive impairment in a DAG (Figure 1) and estimated the associations using a structural equation model (SEM) applied to data from the Alzheimer's Disease Neuroimaging Initiative (ADNI). With the estimates, we generated datasets to train and validate four ML algorithms (logistic regression, lasso regression, random forest and generalised boosted regression). We assessed the transportability of ML models between internal and external validation datasets, using the difference in calibration. The external settings contained interventions on the distributions of age, APOE ε4 prevalence and sex. All prediction algorithms and data-simulations were implemented using R version 4.0.3.

Data source and data preprocessing

We used the TADPOLE grand challenge dataset (https://tadpole.grand-challenge.org/Data/), which was derived from ADNI [30,31]. The ADNI study acquired multiple, longitudinal, measurements from elderly subjects across more than 50 clinics in USA and Canada. We selected individuals who had measurements at baseline (n=1,737) and selected baseline variables that have a reported association with AD and had less than 30% of missing entries. A complete list of the variables is provided in Supplementary Table S1. Detailed information on data preprocessing is provided in Supplementary Text 1. Missing data was imputed using the R package 'mice' with default settings, and one imputed dataset was generated. All numeric variables were normalized by z-transformation.

DAG creation

In DAGs, nodes represent variables and directed edges represent causal relationships pointing from the cause to the effect [18,32,33]. We reviewed scientific literature to identify causal relationships between variables in our dataset that are involved in cognitive impairment and AD processes (Supplementary Table S2) and mapped them in a first DAG (Supplementary Figure 1). Then, we tested if the generated DAG was a good fit to the ADNI dataset, using conditional independence testing with the R package 'dagitty' [34]. We reviewed test results with low p-values and large point estimates, which indicated causal relationships that violated conditional independence. We added 13 causal connections (Supplementary Text 2) according to the test results and our domain expertise to create the final DAG (Figure 2).

Semi-synthetic data generation using structural equation models

We fitted a structural equation model (SEM) using the ADNI dataset to quantify the causal relationships specified in our DAG. The SEM was implemented using the 'sem' function in the R package 'lavaan' with default parameters [35]. For numeric endogenous (dependent) variables, the function computes weighted least squares estimates. For categorical endogenous variables, the function automatically uses a diagonally weighted least squares estimator and assumes that a conditionally normally distributed latent variable underlies the categorical variable (and estimates the thresholds).

We then used the SEM parameter estimates to generate five semi-synthetic datasets with each 10,000 individuals: One for training, one for internal validation and three for external validation of ML models (Figure 1). We bootstrapped exogenous (independent) variables (age, APOE ε4 and sex) together 10,000 times without replacement from the original data and used those to generate the endogenous variables for training and internal validation sets, using the linear equations from the SEM. We then generated three external validation sets implemented by interventions on the exogenous variables to reflect three different populations with 1) a

younger mean age, compared to the original data (73 years ⇒ 35 years) 2) lower prevalence of the APOE ε4 gene compared to the original data (46.9% ⇒ 10.0%), and 3) increased percentage of females compared to the original data (52.6% ⇒ 90.0%). For the external age-intervention data, we sampled the age variable from a normal distribution with a mean age of 35 and standard deviation of 10 and bootstrapped together APOE ε4 and sex. For the APOE ε4 intervention, we sampled from a Bernoulli distribution with 10% probability. For the sex intervention, we sampled from a Bernoulli distribution with 90% probability.

The generated exogenous variables age, APOE ε4 and sex were then used as input to generate endogenous variables, using the SEM estimates.

Prediction algorithms

We applied logistic regression, lasso regression, random forest and generalized boosted regression (GBM) to predict the cognitive state of an individual as either cognitive normal or cognitive impairment. Logistic regression was performed using the glm function in the 'stats' R package. Lasso regression was implemented using the 'glmnet' R package [36]. The lasso model was initialized with an optimized penalization hyperparameter obtained from a grid-search with 10-fold cross validation that selected the minimum value of lambda for minimum deviance. The random forest is an ensemble of regression trees which aims at improving the generalizability compared to a single regression tree [37]. Previous works demonstrated the strengths of random forests for diagnostic prediction modelling of AD [3–5]. The random forest algorithm was applied from the 'randomForest' R package, using 500 trees (as per default). GBM implements boosting by adding regression trees sequentially with respect to the error of the current tree ensemble. This boosting approach increases robustness and generalizability compared to a single regression tree [38–40]. The GBM algorithm was applied using the 'gbm' R package with 100 trees (as per default).

Based on the causal assumptions in our DAG, we defined three predictor sets that included either all variables or only those which are direct causes of the outcome (parent nodes), or

only direct consequences of the outcome (children nodes) (Table 1). Each ML model was trained and validated with each predictor set. We performed 30,000 repetitions, in which the five datasets (one for training, one for internal validation and three for external validation) were generated and used for training and validating all prediction models.

Calibration metric

We assessed the transportability of all prediction models using calibration metrics. We measured the calibration of all trained prediction models in the internal validation setting and in each external validation setting. Calibration was measured using the Integrated Calibration Index (ICI) [41] and the calibration component of a three-way decomposed Brier score. Low ICI and Brier scores indicate better calibration. The Brier calibration component was obtained from the bias-corrected 'BrierDecomp' function of the 'SpecsVerification' R package using quantile bins of predicted probabilities in 10% steps. ICI and Brier scores are given with 95% confidence intervals (95%CI).

We calculated the calibration difference (ICI or Brier score) between the internal setting and each external validation setting to assess transportability. Differences of zero indicate equal calibration in both internal validation and external settings and therefore good transportability. Negative values indicate decreased calibration from internal validation to the intervention setting and therefore decreased transportability. We calculated the median and 95%CI for calibration metrics across all 30,000 repetitions.

Results

Description of the participants' characteristics

The ADNI study, represented in TADPOLE, recorded a total of 1737 participants at baseline together with their diagnosis, demographic information (age, sex, and education), behavioural information (smoking and alcohol abuse history), clinical measurements (BMI, FDG-PET, brain volumetric measurements with MR imaging, Aβ and tau concentrations in cerebrospinal

fluid (CSF), Minimental State Cognitive Exam (MMSE) and medical history (history of hypertension and cardiovascular events) (Table 2). Among all participants, 1214 (69.9%) had cognitive impairment.

Semi-synthetic data generation

The SEM was able to estimate all parameters quantifying the causal relationships in our DAG. We reviewed the estimated parameters and found that many were in agreement with existing neurology domain knowledge. For example, age had a positive coefficient and therefore increased CSF-tau (0.37), the likelihood of hypertension (0.11) and the likelihood of cardiovascular events history (0.13) (Supplementary Table S1). Some estimated relationships however, were controversial to domain knowledge. For example, increasing age decreased CSF-Aβ (-0.14) and the likelihood of cognitive impairment (-0.13). The SEM additionally indicated a small correlation between sex and age (Pearson correlation = 0.06) and between age and APOE ε4 (Pearson correlation = -0.05).

We compared endogenous variable distributions between the original ADNI data and generated validation datasets and found that the percentage of cognitive impairment was underestimated in the internal validation set (40.7%, 95%CI [39.8, 41.7]) in comparison to the original ADNI data (69.9%) (Supplementary Table S3). We further compared endogenous variable distributions between internal and external datasets. Lowering the mean age from the internal validation setting to the external setting under age intervention, decreased the prevalence of cognitive impairment from 40.7% to 26.0%, increased the smoking prevalence from 23.8% to 34.7% and alcohol abuse history from 0.6% to 35.0%, decreased the prevalence of hypertension from 23.0% to 7.3% and previous cardiovascular events from 60.5% to 37.3%. Intervening on age increased the mean of Aβ from 1658.1 to 2092.9 pg/ml, shrank the mean of tau from 265.4 to 20.1 pg/ml and increased the MMSE from 28.0 to 30.0, in comparison to the internal validation data.

In the APOE ε4 intervention, lowering the prevalence of the APOE ε4 gene from the internal setting to the external setting decreased the prevalence of cognitive impairment from 40.7% to 33.8%, increased the mean of CSF-Aβ from 1658.1 to 1843.1 pg/ml and decreased the mean CSF-tau from 265.4 to 237.1 pg/ml.

In the sex intervention, increasing the percentage of females from the internal setting to the external setting decreased the frequency of cognitive impairment from 40.7% to 32.4%, while other variable distributions were similar to the internal validation setting.

Transportability

Transportability of logistic regression, lasso regression, random forest and gradient boosting was measured by ICI and Brier calibration differences (Supplementary Table S4) between internal validation and intervention settings.

First, we compared the transportability of models based on parent variables with the transportability of models based on children variables. In all intervention settings, we found that models predicting with parent nodes were more transportable than those predicting with children nodes (ICI: Figure 3, Brier: Figure 4). Models predicting with parents had good transportability in intervention settings, indicated by a similar calibration between the internal validation and intervention setting. For example, the difference in ICI between the internal validation and age intervention for logistic regression was very small (median ICI -0.007, 95%CI [-0.038, 0.007]). Models predicting with children had low transportability in intervention settings, as indicated by negative calibration differences. For example, logistic regression predicting with children had a median ICI difference between the internal and age intervention setting of -0.094, 95%CI [-0.108, -0.078], which was 12.8-fold lower compared to predicting with parents. Only the GBM model showed better transportability in the age intervention setting when predicting with children (median ICI -0.004, 95%CI [-0.020, -0.010]) than with parents (median ICI -0.028, 95%CI [-0.050, -0.009]). We compared the transportability difference between logistic and lasso models predicting with parents and children across intervention

settings. While logistic and lasso models had very similar calibrations, we found that the median ICI difference between parents and children was largest in the age intervention (0.087) and sex intervention (0.052) and more subtle in the APOE ε4 setting (0.013).

Second, we compared the transportability between all predictors and parent predictors. Logistic regression and lasso regression models predicting with all and parent variables were similarly transportable in all intervention settings. For example, the logistic model had close to zero median ICI differences between internal and intervention settings (age intervention: all predictors -0.009, 95%CI [-0.045, 0.006], parent predictors: -0.007, 95%CI [-0.038, 0.007]; APOE ε4 intervention: all predictors 0.000, 95%CI [-0.008, 0.008], parent predictors 0.000, 95%CI [-0.009, 0.009]; sex intervention: all predictors 0.000, 95%CI [-0.009, 0.009], parent predictors 0.000, 95%CI [-0.010, 0.009]). The random forest model had similar transportability (ICI difference) in APOE ε4 intervention and sex intervention settings when predicting with all and with parent predictors (APOE ε4 intervention: all variables 0.001, 95%CI [-0.009, 0.011]; parent variables: -0.001 [-0.014, 0.011]). In the age intervention setting, the random forest model predicting with parents (median ICI difference: 0.009, 95%CI [-0.033, 0.031]) was similarly transportable than predicting with all variables (-0.029, 95%CI [-0.058, -0.008]), with different medians but overlapping confidence intervals. GBM models predicting with all and parent variables had similarly small median ICI differences in APOE ε4 (all: 0.001, 95%CI [-0.009, 0.011], parents: 0.001, 95%CI [-0.011, 0.012]) and sex intervention settings (all: -0.001, 95%CI [-0.012, 0.010], parents: -0.001, 95%CI; [-0.014, 0.011]). In the age-intervention setting, however, GBM models was similarly transportable (larger negative median ICI differences) when predicting with parents (-0.028, 95%CI [-0.050, -0.009]) compared to all (-0.019, 95%CI [-0.037, -0.002]) and children predictors (-0.004, 95%CI [-0.020, 0.010]). One explanation for these inconsistencies could be that in the internal validation setting, random forest and GBM models with parent predictors had three to four times poorer median ICI calibration than lasso and logistic regression with parents (logistic and lasso: 0.009, 95%CI

266 [0.004, 0.017]; random forest: 0.037, 95%CI [0.026, 0.048]; GBM: 0.028, 95%CI [0.018,
267 0.039], Supplementary Table S5).
268
269 We observed that transportability measured by Brier score differences (Figure 4, Table S4)
270 between internal validation and external settings supported the same trends as the ICI
271 differences. The scale of the Brier scores were exactly zero or closer to zero in the internal
272 validation setting, compared to the ICI scores (Supplementary Table S5).

273 ## Discussion

274 In this study, we have presented a general framework for assessing the transportability of ML
275 models and applied the framework to evaluate the transportability of prediction models for
276 cognitive impairment in external settings with different distributions of age, APOE ε4 allele
277 frequency and sex.
278
279 Our application shows that causal thinking is important when selecting predictors for clinical
280 prediction models. We demonstrated that, under a specific set of interventions, transportability
281 remained stable when ML models predicted with all variables or only causes (parent nodes),
282 but was reduced when predicting with consequences (children nodes) of the diagnostic
283 outcome 'cognitive impairment'. We demonstrated this in all prediction models (logistic
284 regression, lasso regression, random forest, and GBM) with one exception, when validating
285 GBM models in the age intervention setting. A closer investigation of the models showed
286 miscalibration of random forest and GBM models in the internal validation setting. Calibration
287 of GBM models predicting with children nodes was low in the internal validation setting and
288 remained low in all intervention settings, which serves as an explanation of the inconsistent
289 results of GBM models in the age intervention setting. This miscalibration might have
290 happened because we generated data assuming linear relationships in the SEM, whereas
291 random forest and GBM are designed to capture non-linear relationships [42].
292

Our findings can guide the process of selecting predictor variables. Predictors are often selected based on how much they increase model performance in the development setting. Previously developed prediction models for dementia and AD have used brain volumetric measures or cognitive assessment scores as predictors because they reduced prediction errors [5,43,44]. We assumed that cognitive status increases the likelihood of seeing MR images with volumetric changes in brain morphology and we mapped brain volumetric measurements and cognitive assessment scores as consequences. Another work similarly assumed that the brain morphological state (measured by regional brain volumes) influences the performance in cognitive tests [45]. We showed, in agreement with theoretical work [20,25], that the transportability of models predicting with information on the consequences of the outcome of interest (i.e., the anti-causal direction) is more likely to be reduced in settings with different underlying demographic or genetic distributions. Another work suggested that predictors derived from medical images may often predict in the anti-causal direction as they depict the consequences of a disease, which may raise a caveat towards transportability [15].

Limitations

Our framework and its application has limitations in each step. First, our application focused on the prediction of cognitive impairment within the AD continuum, while there are also other types of impairments, e.g. from brain injury, which were not considered. It cannot be empirically verified if DAGs map causal relationships correctly and if all relevant factors were included. We only included observed variables (other than latent variables for factors) and it is likely that there are unobserved variables within the causal processes of cognitive impairment. Strong domain expertise is crucial to increase the correctness of DAGs [32]. We included neurological and epidemiological domain expertise for creating our DAG, but our knowledge may not be complete from other perspectives. Conditional independence tests can test if there is evidence against a given DAG in a dataset [12]. We applied conditional independence tests to add directed connections between variables, but unexplainable violations were present. One study suggested that causal relationships should generally be

assumed to exist between any two variables and they should only be omitted when evidence is available [22]. We ensured that our assumptions in the DAG correctly represent the data by using semi-synthetic data so that any possible misspecification of the DAG did not affect the evaluation of the model transportability.

Second, we applied a SEM to the ADNI data to quantify the causal relationships in our DAG. While SEMs are widely applied for this purpose [16], their methodology comes with limitations, for example, when using categorical variables [46–48]. In our application, we had seven categorical variables and found a small correlation between sex and age and between age and APOE ε4. We assume their correlation might stem from biased selection in the ADNI study or an unknown common cause, which we did not consider in our DAG. Additionally, we found that some parameter estimates in the SEM were controversial to domain knowledge. For example, the relationship between age and cognitive impairment was estimated to be -0.13, whereas the prevalence of cognitive impairment increases with age. We hypothesize that these incorrect estimates explain why the SEM underestimated the percentage of cognitive impairment, and suggest that the model misspecified some relationships in the underlying ADNI data generation process.

Third, we simulated external validation data by intervening on one exogenous variable (age, sex, and APOE ε4) at a time. These interventions may simplify differences between populations in real-world applications. For example, it is possible that multiple variables including endogenous variables such as BMI vary jointly from one validation setting to another. Moreover, it is also possible that causal relationships themselves change from one setting to the other.

Lastly, we applied our framework to assess transportability using structured data. Applications using high-dimensional unstructured data (such as images) require novel methods due to the difficulty of mapping causal relationships for unstructured data.

Outlook

Our framework to assess the transportability of models predicting cognitive impairment can be extended to overcome the described limitations. Future work could adapt the framework to the work of Pölsterl et al. and include unobserved variables in the DAG [45]. To ensure that the estimated parameters of causal relationships follow biological laws, future refinements could include the SEM with prior distributions as implemented in the *blavaan* R package [49]. Future research could adapt our framework to assess the transportability of prediction models when intervening on endogenous variables and causal relationships, because real-world populations likely do not only differ by exogenous variables. Recent ML models predicted AD from medical images [7,50], which requires new approaches to identify the causal structure in complex data [51].

Ethics

The ADNI study was approved by the Institutional Review Boards of all of the participating institutions. Informed written consent was obtained from all participants at each site as described in [30,31].

Contributions

MP and SK conceptualized the study. JF implemented the methods in R, ran the analyses, and drafted a first version of the manuscript. JF, MP, SK analysed the results. SK supervised the study. TK consulted the creation of the DAG. All authors provided critical input to the manuscript and approved the final version.

Conflict of interest

TK reports outside the submitted work to have received research grants from the German Joint Committee and the German Ministry of Health. He further reports personal compensation from Eli Lilly and Company, Teva, TotalEnergies S.E. and the BMJ. MP reports having received partial funding for a self-initiated research project from Novartis Pharma. MP further reports being awarded a research grant from the Center for Stroke Research Berlin (private donations) for a self-initiated project. All other authors declare no conflicts of interest.

Acknowledgements

We thank Prof. Dr. Terrence Jorgensen (Assistant Professor, Methods and Statistics Research Institute for Child Development and Education, the University of Amsterdam) for providing support for the R lavaan package.

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer’s Association; Alzheimer’s Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company

Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

References

[1] GBD 2016 Neurology Collaborators VL, Nichols E, Alam T, Bannick MS, Beghi E, Blake N, et al. Global, regional, and national burden of neurological disorders, 1990-2016: a systematic analysis for the Global Burden of Disease Study 2016. *Lancet Neurol*. 2019;18(5):459-480. doi:10.1016/S1474-4422(18)30499-X

[2] Sabbagh MN, Boada M, Borson S, Doraiswamy PM, Dubois B, Ingram J, et al. Early Detection of Mild Cognitive Impairment (MCI) in an At-Home Setting. *J Prev Alzheimer's Dis*. 2020;7(3):171-178. doi:10.14283/jpad.2020.22

[3] Weiner MW, Veitch DP, Aisen PS, Beckett LA, Nigel J, Green RC, et al. Recent publications from the Alzheimer's Disease Neuroimaging Initiative: Reviewing progress toward improved AD clinical trials. *Alzheimers Dement*. 2017;13(4):1-85. doi:10.1016/j.jalz.2016.11.007

[4] Sarica A, Cerasa A, Quattrone A. Random forest algorithm for the classification of neuroimaging data in Alzheimer's disease: A systematic review. *Front Aging Neurosci*. 2017;9(OCT):1-12. doi:10.3389/fnagi.2017.00329

[5] Moore PJ, Lyons TJ, Gallacher J. Random forest prediction of Alzheimer's disease using pairwise selection from time series data. *PLoS One*. 2019;14(2):1-14. doi:10.1371/journal.pone.0211558

[6] Al-Amyn Valliani A, Ranti D, Oermann KE. Deep Learning and Neurology: A Systematic Review. *Neurol Ther*. 2019;8:351-365. doi:10.6084/m9.figshare.9272951

[7] Kang MJ, Kim SY, Na DL, Kim BC, Yang DW, Kim EJ, et al. Prediction of cognitive impairment via deep learning trained with multi-center neuropsychological test data. *BMC Med Inform Decis Mak*. 2019;19(1):1-9. doi:10.1186/s12911-019-0974-x

[8] Grueso S, Viejo-Sobera R. Machine learning methods for predicting progression from mild cognitive impairment to Alzheimer's disease dementia: a systematic review. *Alzheimers Res Ther*. 2021;13(1):1-29. doi:10.1186/s13195-021-00900-w

[9] Siontis GCM, Tzoulaki I, Castaldi PJ, Ioannidis JPA. External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination. *J Clin*

Epidemiol. 2015;68(1):25-34. doi:10.1016/j.jclinepi.2014.09.007

[10] Steyerberg EW. *Clinical Prediction Models: A Practical Approach to Development, Validation and Updating. Second Edition.* Second Edi. Springer Nature; 2019.

[11] Schölkopf B, Janzing D, Peters J, Sgouritsa E, Zhang K, Mooij JMOOIJ J. *On Causal and Anticausal Learning*.

[12] Peters J, Janzing D, Schölkopf B. *Elements of Causal Inference: Foundations and Learning Algorithms*. Vol 88.; 2018. doi:10.1080/00949655.2018.1505197

[13] Prosperi M, Guo Y, Sperrin M, Koopman JS, Min JS, He X, et al. Causal inference and counterfactual prediction in machine learning for actionable healthcare. *Nat Mach Intell*. 2020;2:369–375. doi:10.1038/s42256-020-0197-y

[14] Kilbertus N, Parascandolo G, Schölkopf B, De BM. Generalization in anti-causal learning. *arXiv*. Published online December 3, 2018. Accessed January 26, 2022. https://arxiv.org/abs/1812.00524v1

[15] Castro DC, Walker I, Glocker B. Causality matters in medical imaging. *Nat Commun*. 2020;11(1):1-10. doi:10.1038/s41467-020-17478-w

[16] Richens JG, Lee CM, Johri S. Improving the accuracy of medical diagnosis with causal machine learning. *Nat Commun*. 2020;11(1):3923. doi:10.1038/s41467-020-17419-7

[17] Schölkopf B, Locatello F, Bauer S, Ke NR, Kalchbrenner N, Goyal A, et al. Toward Causal Representation Learning. *Proc IEEE*. Published online 2021:1-24. doi:10.1109/JPROC.2021.3058954

[18] Pearl J. Causal diagrams for empirical research. *Biometrika*. 1995;82(4):669-688.

[19] Pearl J, Bareinboim E. Transportability of causal and statistical relations: A formal approach. In: *Proceedings of the Twenty- Fifth National Conference on Artificial Intelligence (AAAI 2011).* AAAI Press; 2011:247–54. doi:10.1109/ICDMW.2011.169

[20] Schölkopf B, Janzing D, Peters J, Sgouritsa E, Zhang K, Mooij J. On causal and anticausal learning. *Proc 29th Int Conf Mach Learn ICML 2012*. 2012;2:1255-1262.

[21] Pearl J, Bareinboim E. External validity: From do-calculus to transportability across populations. *Stat Sci*. 2014;29(4):579-595. doi:10.1214/14-STS486

[22] Tennant PW, Murray EJ, Arnold KF, Berrie L, Fox MP, Gadd SC, et al. Use of directed acyclic graphs (DAGs) to identify confounders in applied health research: review and recommendations. *Int J Epidemiol*. 2021;50(2):620-631. doi:10.1093/ije/dyaa213

[23] Moons KGM, Kengne AP, Woodward M, Royston P, Vergouwe Y, Altman DG, et al. Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio)marker. *Heart*. 2012;98(9):683-690. doi:10.1136/HEARTJNL-2011-301246

[24] Moons KGM, Kengne AP, Grobbee DE, Royston P, Vergouwe Y, Altman DG, et al. Risk prediction models: II. External validation, model updating, and impact assessment. *Heart*. 2012;98(9):691-698. doi:10.1136/HEARTJNL-2011-301247

[25] Piccininni M, Konigorski S, Rohmann JL, Kurth T. Directed acyclic graphs and causal thinking in clinical risk prediction modeling. *BMC Med Res Methodol*. 2020;20(1):179. doi:10.1186/s12874-020-01058-z

[26] Ganopoulou M, Kangelidis I, Sianos G, Angelis L. Prediction Model for the Result of Percutaneous Coronary Intervention in Coronary Chronic Total Occlusions. *Proc 3rd Int Conf Stat Theory Appl*. 2021;2018. doi:10.11159/icsta21.129

[27] Gebremedhin AT, Hogan AB, Blyth CC, Glass K, Moore HC. Developing a prediction model to estimate the true burden of respiratory syncytial virus (RSV) in hospitalised children in Western Australia. *Sci Rep*. 2022;12(332):1-12. doi:10.1038/s41598-021-04080-3

[28] Sperrin M, Martin GP, Pate A, Van Staa T, Peek N, Buchan I. Using marginal structural models to adjust for treatment drop-in when developing clinical prediction models. Published online 2018. doi:10.1002/sim.7913

[29] Dickerman BA, Dahabreh IJ, Cantos K V, Roger ·, Logan W, Lodi S, et al. Predicting counterfactual risks under hypothetical treatment strategies: an application to HIV. *Eur J Epidemiol*. 123AD;1:3. doi:10.1007/s10654-022-00855-8

[30] Mueller SG, Weiner MW, Thal LJ, Petersen RC, Jack C, Jagust W, et al. The Alzheimer's disease neuroimaging initiative. *Neuroimaging Clin N Am*. 2005;15(4):869-877. doi:doi: 10.1016/j.nic.2005.09.008.

[31] Petersen RC, Aisen PS, Beckett LA, Donohue MC, Gamst AC, Harvey DJ, et al. *Alzheimer's Disease Neuroimaging Initiative (ADNI) Clinical Characterization*.; 2010. www.neurology.org

[32] Pearl J. *Causality: Models, Reasoning and Inference*. Cambridge University Press; 2000.

[33] Hernán MA, Robins JM. Causal Inference. *Causal Inference What If*. Published online 2019:235-281. doi:10.1017/9781316831762.008

[34] Ankan A, Wortel IMN, Textor J. Testing Graphical Causal Models Using the R Package "dagitty." *Curr Protoc*. 2021;1(2):1-22. doi:10.1002/cpz1.45

[35] Rosseel Y. Lavaan: An R package for structural equation modeling. *J Stat Softw*. 2012;48(2). doi:10.18637/jss.v048.i02

[36] Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J Stat Softw*. 2010;33(1):1-22.

[37] Breiman L. Random forests. *Mach Learn*. 2001;45(1):5-32. doi:10.1023/A:1010933404324

[38] Friedman JH. Greedy function approximation: A gradient boosting machine. *Ann Stat*. 2001;29(5):1189-1232. doi:10.1214/aos/1013203451

[39] Friedman JH. Stochastic gradient boosting. *Comput Stat Data Anal*. 2002;38(4):367-378. doi:10.1016/S0167-9473(01)00065-2

[40] Hastie T, Tibshirani R, Friedman JH. 10. Boosting and Additive Trees. In: *The Elements of Statistical Learning*. 2nd ed. Springer; 2009:337-384.

[41] Austin PC, Steyerberg EW. The Integrated Calibration Index (ICI) and related metrics for quantifying the calibration of logistic regression models. *Stat Med*. 2019;38(21):4051-4065. doi:10.1002/sim.8281

[42] Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*. Vol 27. Second Edi. Springer Science + Business Media; 2009. doi:10.1007/b94608

[43] Lebedev A V., Westman E, Van Westen GJP, Kramberger MG, Lundervold A, Aarsland D, et al. Random Forest ensembles for detection and prediction of Alzheimer's disease with a good between-cohort robustness. *NeuroImage Clin*. 2014;6:115-125. doi:10.1016/j.nicl.2014.08.023

[44] Guest F, Kuzma E, Everson R, Llewellyn DJ, David Llewellyn CJ. Identifying key features for dementia diagnosis usingmachine learning. *Alzheimer's Dement*. 2020;16:e046092. doi:10.1002/alz.046092

[45] Pölsterl S, Wachinger C. Estimation of Causal Effects in the Presence of Unobserved Confounding in the Alzheimer's Continuum. In: *Information Processing in Medical Imaging*. Springer International Publishing; 2021:45-57. doi:10.1007/978-3-030-78191-0_4

[46] Sass DA, Schmitt TA, Marsh HW. Evaluating Model Fit With Ordered Categorical Data Within a Measurement Invariance Framework: A Comparison of Estimators. *Struct Equ Model*. 2014;21(2):167-180. doi:10.1080/10705511.2014.882658

[47] Bandalos DL. Relative Performance of Categorical Diagonally Weighted Least Squares and Robust Maximum Likelihood Estimation. *Struct Equ Model*. 2014;21(1):102-116. doi:10.1080/10705511.2014.859510

[48] DiStefano C, Morgan GB. A Comparison of Diagonal Weighted Least Squares Robust Estimation Techniques for Ordinal Data. *Struct Equ Model*. 2014;21(3):425-438. doi:10.1080/10705511.2014.915373

[49] Merkle EC, Rosseel Y. Blavaan: Bayesian Structural Equation Models Via Parameter Expansion. *J Stat Softw*. 2018;85(4). doi:10.18637/jss.v085.i04

[50] Nigri E, Ziviani N, Cappabianco F, Antunes A, Veloso A. Explainable Deep CNNs for MRI-Based Diagnosis of Alzheimer's Disease. *arXiv*. Published online 2020. http://arxiv.org/abs/2004.12204

[51] Pawlowski N, Castro DC, Glocker B. Deep structural causal models for tractable counterfactual inference. In: *34th Conference on Neural Information Processing Systems (NeurIPS 2020)*. ; 2020.

Tables

Table 1: Predictor sets with corresponding lists of variable names.

Predictor set	Variable names
All nodes	age, APOE ε4, sex, education, BMI*, history of hypertension, history of alcohol abuse, history of smoking behaviour, history of cardiovascular event, CSF†-tau, CSF†-Aβ‡, hippocampus, ventricles, intracranial volume, FDG-PET§, MMSE¶
Parent nodes	age, APOE ε4, sex, education, BMI*, history of cardiovascular event, CSF†-tau, CSF†-Aβ‡
Children nodes	hippocampus, ventricles, intracranial volume, FDG-PET§, MMSE¶

*BMI: Body Mass Index; †CSF: Cerebrospinal fluid; ‡Aβ: Amyloid β; §FDG-PET: fluorodeoxy-glucose-positron emission tomography; ¶MMSE: Mini Mental State Exam score

551 Table 2: Participant characteristics of ADNI dataset at baseline stratified by cognitive state.

Variables	Cognitive normal n=523	Cognitive impairment n=1214	Total n=1737
age#	73.7 (70.5, 78.0)	74.0 (68.3, 79.3)	73.9 (69.2, 78.9)
APOE ε4	149 (28.6%) Nmiss: 2	660 (54.8%) Nmiss: 10	809 (46.9%) Nmiss:12
sex (male)	253 (48.4%)	704 (58.0%)	957 (55.1%)
education# (years)	16 (14, 18)	16 (14, 18)	16 (14, 18)
CSF†-Aβ‡# (pg/ml)	1271 (820.8, 1734.0) Nmiss: 156	741.5 (559.1, 1130.3) Nmiss: 366	854.2 (596.2, 1395.5) Nmiss: 522
CSF†-tau# (pg/ml)	214.3 (175.2, 287.8) Nmiss: 156	281.6 (210.4, 379.2) Nmiss: 366	257.8 (193.4, 349.7) Nmiss: 522
history of alcohol abuse	16 (3.1%) Nmiss: 6	27 (2.3%) Nmiss: 27	43 (2.5%) Nmiss: 33
history of smoking	115 (26.8%) Nmiss: 94	223 (23.2%) Nmiss: 254	338 (24.3%) Nmiss: 348
BMI*#	28.6 (25.8, 32.5) Nmiss: 4	28.17 (25.5, 31.2) Nmiss: 5	28.31 (25.5, 31.6) Nmiss: 9
history of hypertension	130 (24.9%)	460 (38.0%) Nmiss: 5	590 (34.1%) Nmiss: 5
history of cardiovascular events	343 (65.6%)	835 (68.8%)	1178 (67.8%)
MMSE¶#	29 (29.0, 30.0)	27 (25.0, 29.0)	28 (25.0, 29.0)

552 *BMI: Body Mass Index; †CSF: Cerebrospinal fluid; ‡Aβ: Amyloid β; ¶MMSE: Mini Mental State
553 Exam score

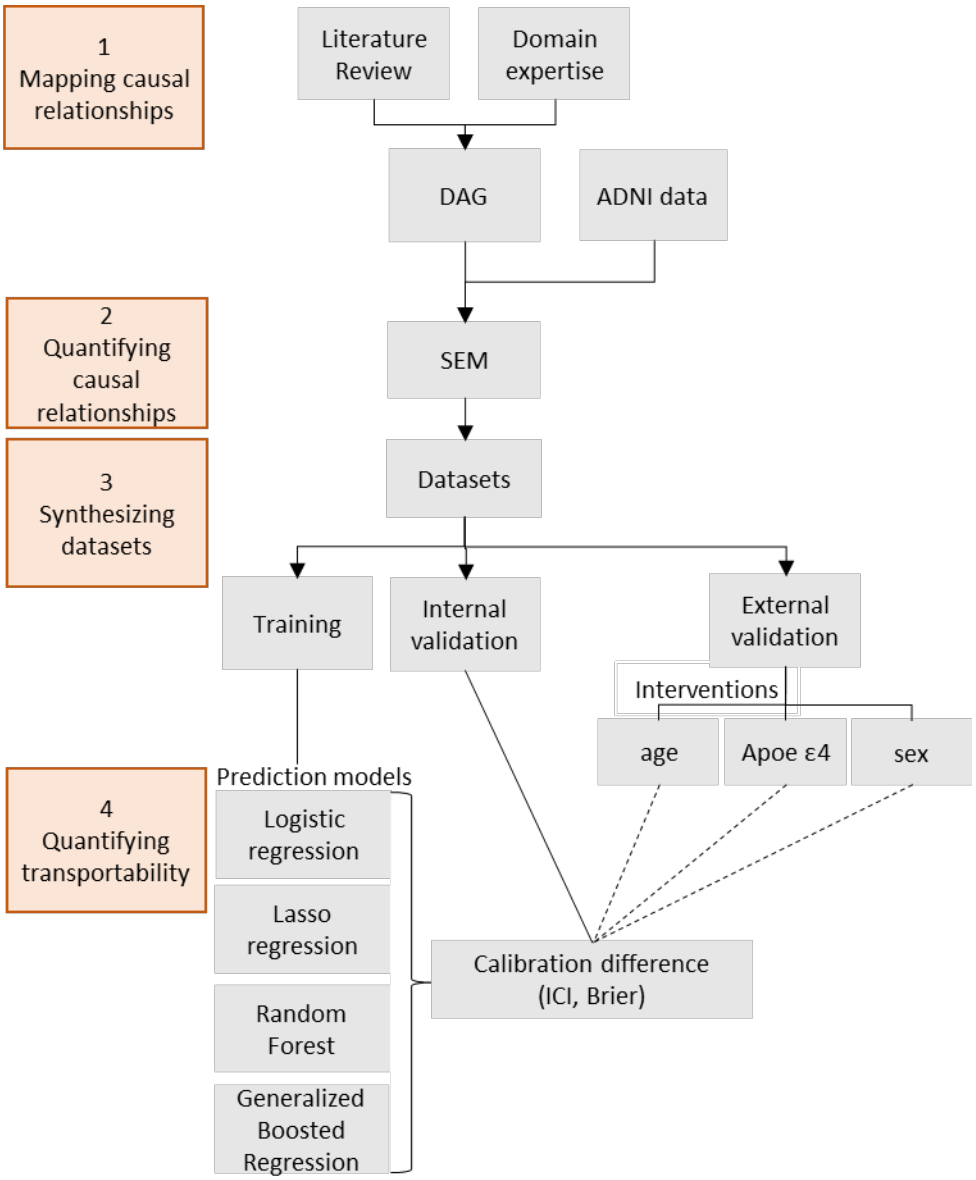
554 NOTE. Numeric variables are indicated with # and are given with median and 25-75%
555 interquartile range (IQR). All other variables are categorical variables with two categories and
556 the absolute number and the column wise percentage of the reference category is given.
557 Absolute numbers of missing values (Nmiss) are given.

558

559

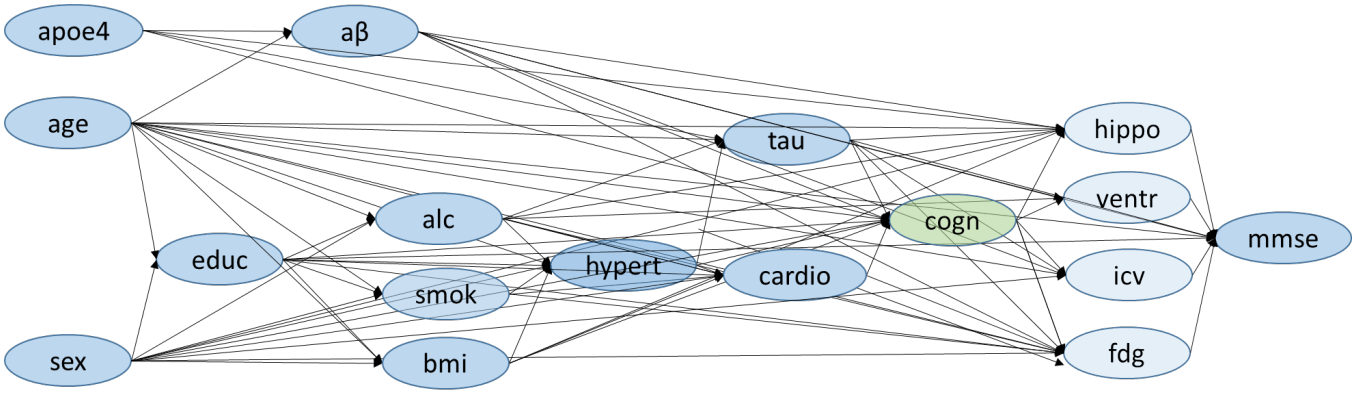
Figures

Figure 1: Framework to assess the transportability of machine learning models predicting cognitive impairment.



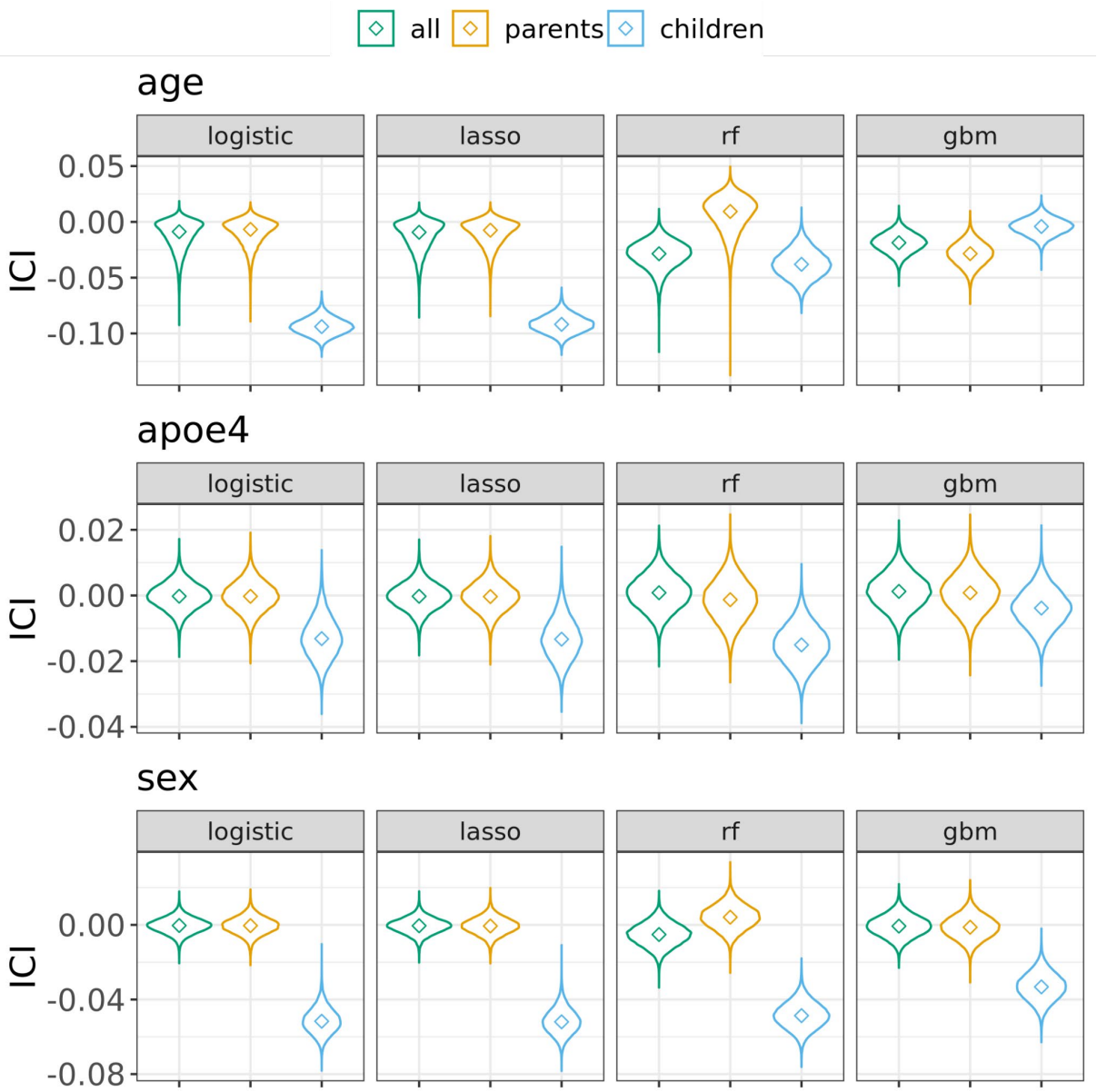
NOTE. Orange boxes mark the four general steps of this framework. We first mapped knowledge about cognitive impairment into a Directed Acyclic Graph (DAG) and quantified those using Structural equation modelling (SEM) and data from the Alzheimer's Disease Neuroimaging Initiative (ADNI). The estimates were used in linear equations to generate datasets for training, internal validation and three external validation datasets with interventions on age, APOE ε4 and sex. We trained four machine learning algorithms (logistic regression, lasso regression, random forest and generalized boosted regression) to predict

572 cognitive impairment. We measured transportability between internal and external settings
573 using calibration differences, measured by Integrated Calibration Index (ICI) and Brier score.
574 Steps 3 and 4 (data synthesis and model training and validation) were repeated 30,000 times.
575
576
577 Figure 2: Directed acyclic graph of variables influencing cognitive status.



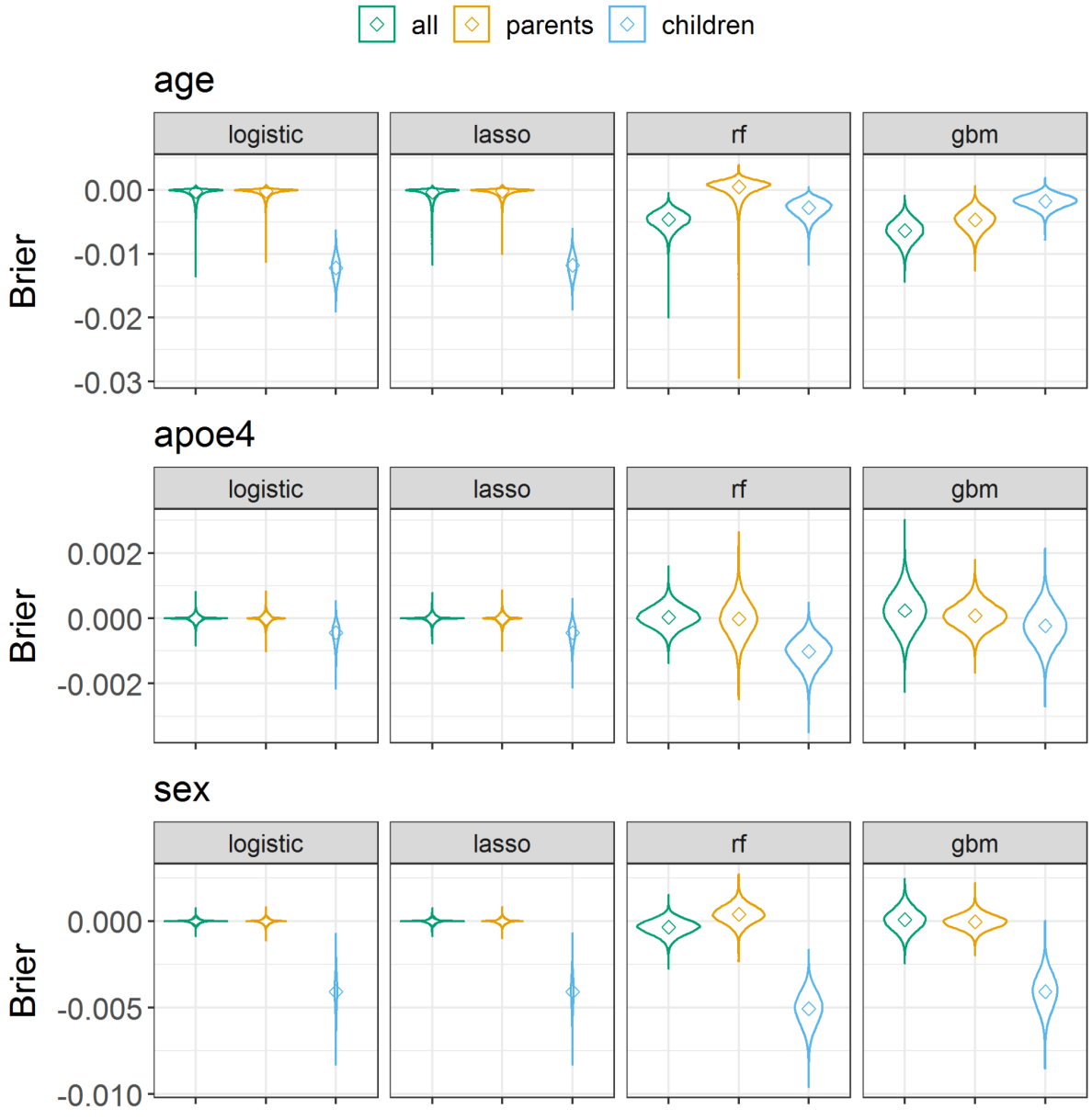
578
579 NOTE. Predictor variables are marked in blue and the outcome variable (cognitive status) in
580 green. Directed arrows indicate assumed causal relationships between variables. The
581 included variables are: APOE ε4 (apoe4), age, sex, education (educ), CSF-Aβ (aβ), history
582 of alcohol abuse (alc), history of smoking behaviour (smok), Body Mass Index (bmi), history
583 of hypertension (hypert), CSF-tau (tau), history of cardiovascular events (cardio), cognitive
584 status (cogn), hippocampus (hippo), ventricles (ventr), intracranial volume (icv), FDG-PET
585 (fdg), Mini-Mental State Exam score (mmse).

Figure 3: Transportability between internal validation and intervention test sets, measured by the difference of integrated calibration index (ICI).



NOTE. Three intervention test sets were created with 1) reducing the population mean age from 73 to 35 years, 2) reducing the APOE ε4 allele frequency from 46.6% to 10.0%, and 3) increasing the number of female individuals from 45.9% to 90.0%. Cognitive impairment was predicted using logistic regression, lasso regression random forest (rf) and generalized boosted regression (gbm) prediction models. Models were trained either with all predictor variables, only parent nodes (direct causes) of the diagnostic outcome, or only children nodes (consequences) of the diagnostic outcome.

596 Figure 4: Transportability between training and intervention test sets, measured by the
597 difference of the Brier calibration component.



598

599 NOTE. Three intervention test sets were created with 1) reducing the population mean age
600 from 73 to 35 years, 2) reducing the APOE ε4 allele frequency from 46.6% to 10.0%, and 3)
601 increasing the number of female individuals from 45.9% to 90.0%. Cognitive impairment was
602 predicted using logistic regression, lasso regression random forest (rf) and generalized
603 boosted regression (gbm) prediction models. Models were trained either with all predictor
604 variables, only parent nodes (direct causes) of the diagnostic outcome, or only children nodes
605 (consequences) of the diagnostic outcome.