RESEARCH ARTICLE OPEN ACCESS

Improvement of an External Predictive Model Based on New Information Using a Synthetic Data Approach

Application to CADASIL

Henri Chhoa, MEng, Hugues Chabriat, MD, PhD, Adelina Joanita Anato, BS, Mamadou Bamba, BS, Florent Zittoun, BS, Sylvie Chevret, MD, PhD, and Lucie Biard, MD, PhD

Correspondence
Dr. Biard
lucie.biard@u-paris.fr

Neurol Genet 2023;9:e200091. doi:10.1212/NXG.0000000000200091

Abstract

Background and Objectives

Cerebral autosomal dominant arteriopathy with subcortical infarcts and leukoencephalopathy (CADASIL) is the most frequent hereditary cerebral small vessel disease. It is caused by mutations of the *NOTCH3* gene. The disease evolves progressively over decades leading to stroke, disability, cognitive decline, and functional dependency. The course and clinical severity of CADASIL seem heterogeneous. Predictive models are thus needed to improve prognostic evaluation and inform future clinical trials. A predictive model of the 3-year variation in the Mattis Dementia Rating Scale (MDRS), which reflects the global cognitive performance of patients with CADASIL, was previously proposed. This model made predictions based on demographic, clinical, and MRI data. We aimed to improve this existing predictive model by integrating a new potential factor, the location of the genetic mutation in the different epidermal growth factor (EGFr) domains of the *NOTCH3* gene, dichotomized into EGFr domains 1 to 6 or 7 to 34.

Methods

We used a new synthetic data approach to improve the initial predictive model by incorporating additional genetic information. This method combined the predicted outcomes from the previous model and 5 "synthetic" data sets with the observed outcome in a new data set. We then applied a multiple imputation method for missing data on the mutation location.

Results

The new data set included 367 patients who were followed up for 30 to 42 months. In the multivariable model with synthetic data, patients with *NOTCH3* mutations in EGFr domains 7 to 34 had an additional average decrease of −1.4 points (standard error 0.67, $p = 0.035$) in their MDRS score variation over 3 years compared with patients with mutations located in EGFr domains 1 to 6. Cross-validation results highlighted the improved predictive performance of the enhanced model. Moreover, the model estimation was found to be more robust than fitting a model without synthetic data.

Discussion

The use of synthetic data improved the predictive model of MDRS change over 3 years in CADASIL. The predictive performance and estimation robustness of the predictive model were enhanced using this approach, whether genetic information was used. A statistically significant association between the location of the mutation in the *NOTCH3* gene and the 3-year MDRS score variation was detected.

From the ECSTRRA Team (H. Chhoa, S.C., L.B.), Université Paris-Cité, UMR1153, INSERM; Translational Neurovascular Centre (H. Chabriat), GH Saint-Louis-Lariboisière, Assistance Publique des Hôpitaux de Paris APHP, Université Paris-Cité and DHU NeuroVasc Sorbonne Paris-Cité; UMR 1161 (H. Chabriat), INSERM; and ENSAI (A.J.A., M.B., F.Z.), Ecole d'ingénieur statistique, data science et big data, Bruz, France.

Go to Neurology.org/NG for full disclosures. Funding information is provided at the end of the article.

The Article Processing Charge was funded by INSERM.

Glossary

BPF = brain parenchymal fraction; **CADASIL** = cerebral autosomal dominant arteriopathy with subcortical infarcts and leukoencephalopathy; **EGFr** = epidermal growth factor repeat; **MAPE** = mean absolute prediction error; **MDRS** = Mattis Dementia Rating Scale; **mRS** = modified Rankin Scale; **MSPE** = mean squared prediction error; **RMSPE** = root mean squared prediction error.

Introduction

In clinical modeling studies, predicting disease progression from patient characteristics remains the primary goal. This is also true for cerebral autosomal dominant arteriopathy with subcortical infarcts and leukoencephalopathy (CADASIL), the most frequent hereditary cerebral small vessel disease. CADASIL is caused by stereotyped mutations of the *NOTCH3* gene, which encodes a transmembrane receptor of vascular smooth muscle cells and pericytes. These mutations lead to an odd number of cysteine residues within the epidermal growth factor repeat (EGFr) domains of the *NOTCH3* receptor. They result in a progressive accumulation of *NOTCH3* extracellular domains aggregating with multiple matrix proteins in the wall of cerebral arterioles and capillaries.[1]

However, both clinical manifestations and disabilities vary largely among patients with CADASIL. Thus, it is crucial to continuously refine the existing prediction models to improve future prognostic and therapeutic evaluation. Some studies have suggested that sex and cardiovascular risk factors, such as smoking or hypertension, might influence the clinical expression of the disease.[2] In 2016, Chabriat et al. proposed a multivariable model to predict how the Mattis Dementia Rating Scale (MDRS) score, a global measure of cognitive performance frequently used in patients with CADASIL, evolves over a 3-year period.[3] The model was built on data obtained from a prospective cohort of 290 patients recruited between September 2003 and April 2011 from 2 major referral centers for the disease (Lariboisière Hospital, Paris, France, and Ludwig Maximilians Universität, Munich, Germany). To predict the variation in the MDRS score, demographic (sex and age), clinical (modified Rankin score of 3 or above, presence of balance problems, and gait disturbances), and imaging parameters (number of lacunes,[4,5] microbleeds,[6] and brain parenchymal fraction[7,8]) were used.

More recently, an unexpectedly large number of mutations in the *NOTCH3* gene outside the EGFr 1–6 hotspot was identified in the general population.[9] This raised the hypothesis that the exact position of the mutation along the gene might also influence the clinical expression of the disease. Notably, a later stroke onset was reported in patients with *NOTCH3* mutations located in EGFr domains 7–34 compared with patients with a mutation inside EGFr domains 1–6.[10] Very recently, the mutation location was also shown to be strongly associated with the clinical severity of the disease, in addition to the effects of age, sex, hypertension, and hypercholesterolemia.[11] However, this association with the disease phenotype and the potential prognostic impact of the mutation location in predicting the clinical course of CADASIL remain undetermined.

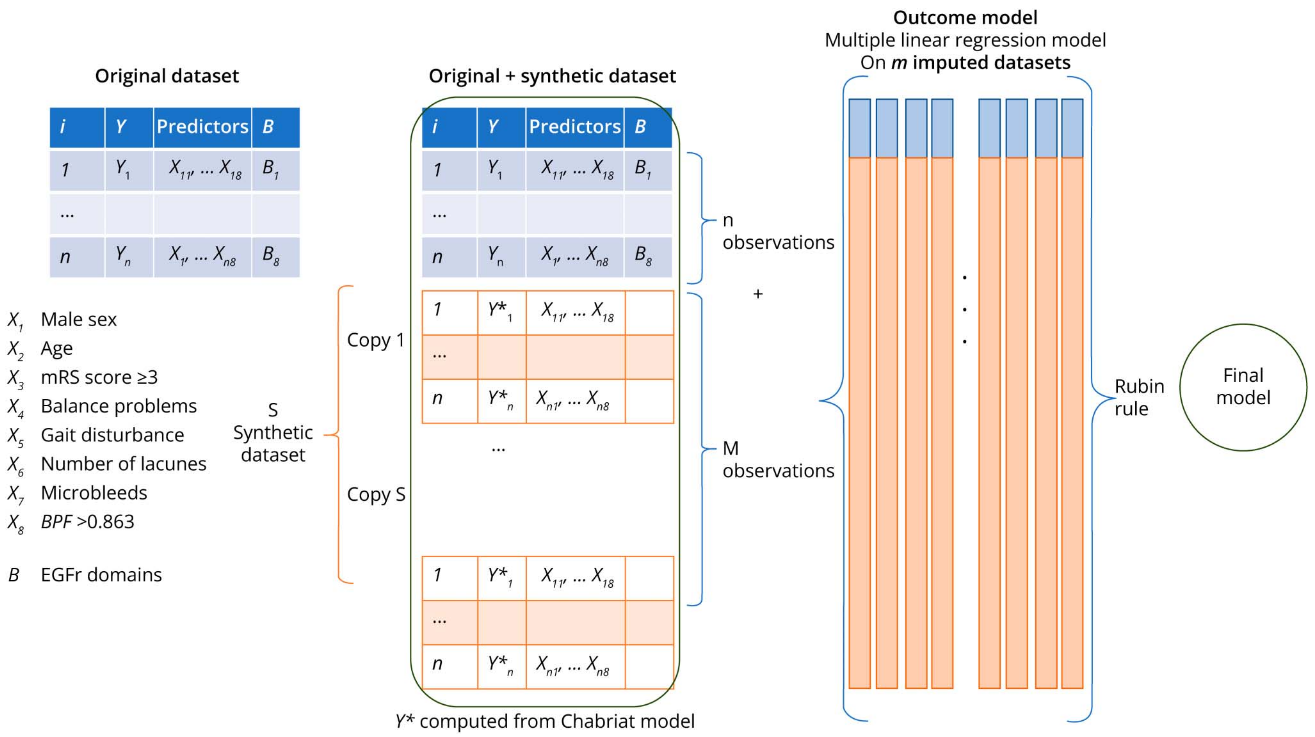
In that respect, we aimed to improve the prediction of clinical score changes in patients with CADASIL using both summary information from the prognostic model previously reported by Chabriat et al.[3] and newly available individual-level data. Such an approach is in line with the growing use of external data for treatment evaluation, notably, when sample sizes are low[12] or when making early stopping decisions.[13] In this study, the previously reported predictive model for the 3-year variation in the MDRS score[3] in CADASIL was used as external information. New information was obtained by creating a new data set including information related to the *NOTCH3* gene mutation location that could potentially improve the prediction accuracy. In settings, where an existing model that already includes several risk factors for predicting an outcome is available and a new study or data source that provides information on a new marker can be used, several methods based on Bayesian inference have been previously proposed.[14-17] These approaches are, however, based on a binary outcome measure, where the Bayes rule applies for updating the previous odds to the posterior odds through the likelihood ratio. We thus decided to use a new partially synthetic data approach[18] that consists of creating additional synthetic data observations from a previously reported model and then analyzing the combined data set to estimate the effect of the genetic information (here, the location of *NOTCH3* mutations in the different EGFr domains) on the 3-year variation in the MDRS score in patients with CADASIL.

Methods

Patients

A total of 482 patients with CADASIL aged older than 18 years were prospectively enrolled between June 03, 2003, and December 29, 2020, from the French National Referral Center for rare cerebrovascular diseases in France (cervco.fr). The diagnosis was confirmed by genetic testing showing a typical cysteine mutation in the *NOTCH3* gene. A follow-up interval of 30 to 42 months was chosen to obtain a final follow-up visit, approximately 3 years from enrollment. This time frame was considered to allow the detection of significant changes in clinical scores that are usually only observed after 2 years of follow-up.[19]

Figure 1 Proposed Synthetic Data Set Approach


Original dataset
i | Y | Predictors | B
1 | Y_1 | $X_{11}, \dots X_{18}$ | B_1
...
n | Y_n | $X_1, \dots X_{n8}$ | B_8
X_1 Male sex
X_2 Age
X_3 mRS score ≥3
X_4 Balance problems
X_5 Gait disturbance
X_6 Number of lacunes
X_7 Microbleeds
X_8 BPF >0.863
B EGFr domains
S Synthetic dataset
Copy 1
Copy S
Original + synthetic dataset
i | Y | Predictors | B
1 | Y_1 | $X_{11}, \dots X_{18}$ | B_1
...
n | Y_n | $X_1, \dots X_{n8}$ | B_8
1 | $Y*_1$ | $X_{11}, \dots X_{18}$
...
n | $Y*_n$ | $X_{n1}, \dots X_{n8}$
...
1 | $Y*_1$ | $X_{11}, \dots X_{18}$
...
n | $Y*_n$ | $X_{n1}, \dots X_{n8}$
Y* computed from Chabriat model
n observations
+
M observations
Outcome model
Multiple linear regression model
On m imputed datasets
Rubin rule
Final model

Notably, of these 482 patients, 178 (37%) individuals enrolled before April 2011 were included in the cohort from which the first prediction model was derived.[3] However, these 178 patients were previously analyzed jointly with an additional 112 German patients from Munich. The present analysis was only based on the French cohort, which has grown since 2016, including 304 new patients.

Standard Protocol Approvals, Registrations, and Patient Consents

This study was approved by an independent ethics committee (updated agreement CEEI-IRB-17/388) and conducted per the Declaration of Helsinki and guidelines for Good Clinical Practice and General Data Protection Regulation in Europe. Informed consent was obtained from all participants included in the study.

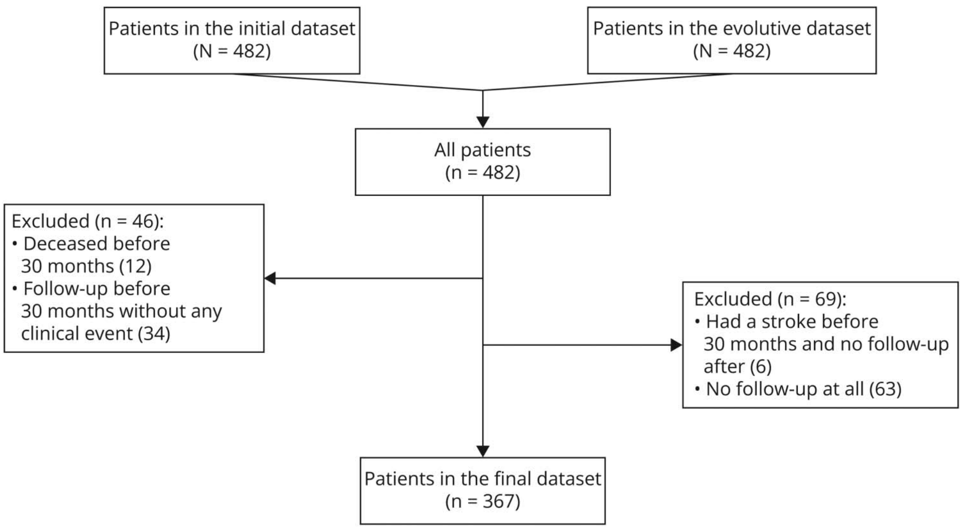
Measurements

Clinical data were collected prospectively by board-certified neurologists during individual consultations using a standardized questionnaire and a detailed neurologic assessment. Several clinical scores were systematically recorded for each individual at cohort entry to evaluate the following: (1) global cognitive performances using the Mini-Mental-State Examination score[20] ranging from 0 (worst score in patients with severe dementia) to 30 (best score) and the MDRS ranging from 0 (worst performance in patients with severe dementia) to 144 (best performance); (2) disability with the modified Rankin Scale (mRS)[21] ranging from 0 (no disability) to 6 (death), with 5 indicating severe disability and bedridden status; and (3) functional dependency using the Barthel index[22] ranging from 0 (most dependence of the patient) to 100 (total independence). Finally, the patients were assessed for occurrences of stroke (ischemic and hemorrhagic) and the presence of dementia (according to DSM IV criteria). In addition, we also recorded variables previously considered in the first multivariable model[3]: sex (male or female), age, presence of balance problems (defined on the basis of patient complaints and/or neurologic examination), gait disturbances (defined as the presence of any difficulty during walking presumably related to the disease confirmed by neurologic examination), and 3 imaging parameters obtained from brain MRI, namely, the number of lacunes, defined as small cavities of diameter less than 15 mm secondary to small deep ischemic lesions,[4,5] the presence of microbleeds, defined as rounded hypointensities of diameter less than 10 mm on susceptibility-weighted images,[6] and the brain parenchymal fraction (BPF), where the brain volume was calculated based on 3D-T1-weighted images using SIENAX methods and was divided by the intracranial volume.[7,8,23,24] For comparison purposes, the BPF variable for multivariable analyses was defined following the initial model and dichotomized using the baseline median value as <0.863 or ≥0.863. Finally, we considered the exposure of interest, that is, the mutation location in the *NOTCH3* EGFr domains (from domains 1 to 6, or 7 to 34).

Model

We considered the reported prediction model[3] as drawn from an external population without any individual-level data. The aim of this external model was to predict the 3-year mean

Figure 2 Flowchart of Patient Enrollment in the Study


Patients in the initial dataset (N = 482)
Patients in the evolutive dataset (N = 482)
All patients (n = 482)
Excluded (n = 46):
• Deceased before 30 months (12)
• Follow-up before 30 months without any clinical event (34)
Excluded (n = 69):
• Had a stroke before 30 months and no follow-up after (6)
• No follow-up at all (63)
Patients in the final dataset (n = 367)

absolute variation in the MDRS score from the baseline value using a multivariable linear regression model. Thus, the predicted outcome was the weighted sum of the 8 selected risk factors, namely, sex, age, mRS of 3 or above, balance problems, gait disturbance, number of lacunes, and MRI findings (microbleeds and BPF). Each variable was weighted by the parameters derived from the linear model. In contrast, our new cohort was assumed to have been drawn from a new population, with individual-level data available on the same 8 risk factors and the genetic mutation location considered a new marker.

We wished to provide a predictive model, referred to as the "internal" or "enhanced" model,

$$E(Y|X,B) = \tilde{\gamma}_1 X_1 + \tilde{\gamma}_2 X_2 + \cdots + \tilde{\gamma}_8 X_8 + \gamma_B B. \qquad (1)$$

where X is the vector containing the 8 risk factors for the first reported model, $\tilde{\gamma}_{j\ (j=1,\ldots,8)}$ are the estimated parameters from the published model, B represents which EGFr domains are affected (distinguishing domains 1–6 and 7–34), and Y is the 3-year variation in MDRS score.

To estimate the model defined by Equation (1), the synthetic data approach proposed by Gu in 2019[18] was applied, under the assumption that the initial and proposed models were identical in the external and new populations. Briefly, this approach consisted of creating a large number of new observations samples to create additional records (called "synthetic data"), generating pseudodata for the outcomes of these new records based on the existing prediction model. Then, to estimate the model parameters from the combined data set, missing values of the EGFr domain were handled through multiple imputations (Figure 1). All these processes are detailed in eAppendix 1 (links.lww.com/NXG/A622).

Statistical Analysis

Summary statistics, the mean (SD) for quantitative variables, and percentages for binary variables were reported unless otherwise specified. To assess whether selection biases were introduced, comparisons between the enrolled and excluded patients from the whole population were performed using the Wilcoxon nonparametric test for quantitative variables and the exact Fisher test for binary variables.

The variations in the MDRS score were computed by subtracting the baseline value from the score at the 3-year follow-up. For patients who did not have a 3-year visit (i.e., between 30 and 42 months) and whose MDRS score after 42 months was at least 140, the 3-year MDRS score was set at that score. For all other patients, MDRS scores were considered missing.

To evaluate the predictive performance of the model based on the synthetic data approach and compare the resulting model to the initial model, we performed 10-fold cross-validation.[25] Hence, each fold was used for modeling and testing to reduce the variability introduced when using only a simple train/test split. We used cross-validation to tune the number of times the original data set was replicated. Several metrics were used to compare the predictive performance of the different models; the mean squared prediction error (MSPE), root mean squared prediction error (RMSPE), and mean absolute prediction error (MAPE). We evaluated prediction errors using parameter mean estimates and parameter values sampled from a normal distribution centered on their mean estimates with SD equal to their standard errors to consider the variability of the estimation.

Finally, we checked the assumption that the external model did not differ in the previous and new populations by fitting a multivariable linear regression with multiple imputations but ignoring synthetic data.

Table 1 Comparison of Baseline Characteristics of the Study Cohort

Characteristics	Study cohort, n = 367	External cohort, from the report of Chabriat et al.
Age (y), median (IQR)	53.1 (16.4)	50.6 (11.4)
Men, n (%)	167 (45.5)	130 (44.8)
Moderate or severe disability[a], n (%)	48 (13.1)	51 (18.0)
Balance problems, n (%)	101 (27.5)	86 (29.7)
Gait disturbance, n (%)	94 (25.6)	87 (30.0)
Number of lacunes, median (IQR)	5 (11)	4.9 (6.1)
Microbleeds, n (%)	123 (33.5)	
BPF %, median (IQR)	81.4 (5.1)	85.3 (0.6)
MDRS score at baseline, median (IQR)	141.0 (10.0)	
EGFr domains 7–34, n (%)	122 (33.2)	—

Abbreviations: BPF = brain parenchymal fraction; EGFr = epidermal growth factor receptor; IQR = interquartile range; MDRS = Mattis Dementia Rating Scale.
[a] Modified Rankin Scale score ≥3.

All statistical analyses were performed in R 4.1.1 (R-project.org/). To implement multiple imputation, we use the R package MICE.[26] Two-sided *p* values of 0.05 or less denoted statistical significance.

Data Availability

The data that support the findings of this study are available on request from the corresponding author, LB. The data are not publicly available, under the French regulation for data protection policy, because of their containing information that could compromise the privacy of research participants.

Results

Characteristics of the Study Sample

Of the 482 patients included in the cohort study, 115 individuals did not have a follow-up of at least 3 years. Thus, the study sample consisted of the remaining 367 (76%) patients (Figure 2). There was no marked evidence of selection bias. The enrolled and excluded populations were very similar in age, sex, balance problems, gait disturbances, number of lacunes, presence of microbleeds, and mutation location (eTable 1, links.lww.com/NXG/A623). Only differences in the brain parenchymal fraction (with median values of 81% in the included patients vs 79.5% in the excluded patients) and modified Rankin score (with an increased proportion of moderate and severe disability among the excluded patients) were observed between the included and excluded patients.

Table 2 Multivariable Linear Regression of 3-Y MDRS Score Variation With and Without the EGFr Domain Variable in 367 Patients With CADASIL Based on the Internal Model From Synthetic and Observed Data

	Without EGFr domain		**With EGFr domain**	
Parameter	**Estimate β (SE)**	***p* Value**	**Estimate β (SE)**	***p* Value**
Sex (male)	0.90 (0.36)	0.012	0.77 (0.39)	0.049
Age	−0.14 (0.01)	<0.0001	−0.13 (0.01)	<0.0001
mRS score ≥3	−5.11 (0.48)	<0.0001	−5.18 (0.50)	<0.0001
Balance problems	−3.27 (0.47)	<0.0001	−3.42 (0.47)	<0.0001
Gait disturbance	3.31 (0.57)	<0.0001	3.41 (0.59)	<0.0001
Number of lacunes	−0.40 (0.02)	<0.0001	−0.40 (0.02)	<0.0001
Microbleeds	−0.58 (0.40)	0.143	−0.41 (0.42)	0.332
BPF <0.863	1.08 (0.54)	0.046	1.25 (0.56)	0.027
EGFr domain 1–6			1.44 (0.67)	0.035

Abbreviations: BPF = brain parenchymal fraction; EGFr = epidermal growth factor receptor; mRS = modified Rankin Scale.

Subsequent analyses included only 367 patients enrolled with a 3-year follow-up. Their baseline characteristics are shown in Table 1. Their characteristics were close to those reported in the initial cohort. Patients with mutations in EGFr domains 1–6 represented 66.8% of the sample.

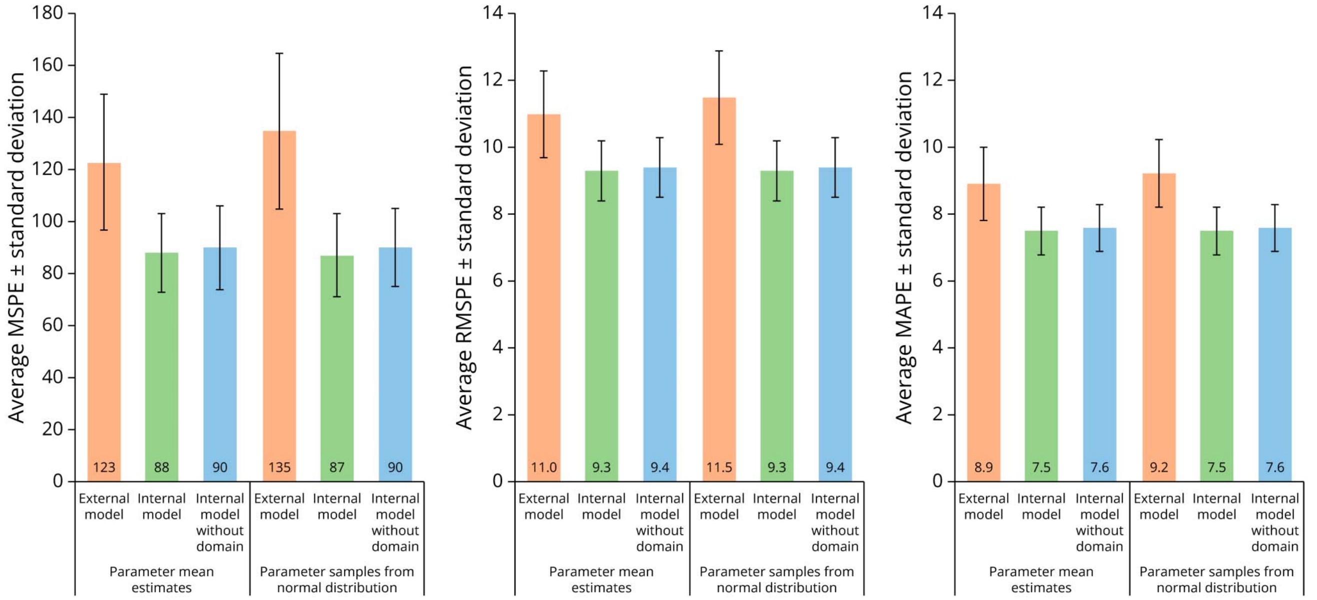
The median follow-up duration was 37 months (interquartile range, 20–75). At 3 years, the mean MDRS score was estimated at 135.9 (SD, 15.5), with a mean variation from baseline estimated at −1.7 (SD, 7.8).

Model Estimation

After generating 5 synthetic samples and deriving the predicted outcomes from the previous model, we used the whole data set to fit the multivariable linear regression model with all data available, synthetic and observed, with and without the EGFr domain variable (Table 2). The estimates of covariate effects were very close in the model not including the EGFr domain variable compared with those of the model with the new genetic marker, suggesting that prognostic information achieved from the EGFr domain is somewhat independent of that of the other predictors. All 9 predictors, except microbleeds, were associated with the 3-year MDRS variation. Notably, the genetic mutation in EGFr domains 7–34 was associated with a larger mean decrease of 1.4 in the MDRS score variation over 3 years compared with EGFr domains 1–6.

Moreover, using 10-fold cross-validation, the model incorporating the EGFr domain was selected as the best model, followed by that without the mutation domain. Both models

Figure 3 Tenfold Cross-Validation Average of MSPE, RMSPE, and MAPE Scores With SD



For each metric, the 2 categories correspond to the prediction with parameter mean estimates and parameter samples from a normal distribution. Red bars: model by Chabriat; Green bars: internal model; Blue bars: internal model without EGFr domain covariate. Error bars are 2 SD. MAPE = mean absolute prediction error; MSPE = mean squared prediction error; RMSPE = root mean squared prediction error.

derived from the synthetic data approach outperformed the external first model (Figure 3). Adding variability by drawing parameter values from normal distributions did not modify these findings, with a clear improvement in prediction errors for the internal model based on the synthetic data and additional genetic information.

Ignoring synthetic data, we estimated the same model on the new sample, without checking the underlying assumption that the external model applied in both populations. As tabulated in Table 3, there were some differences in the estimated effects, notably, regarding age, balance, and gait problems.

Discussion

CADASIL causes cognitive decline which is associated with a reduction in the MDRS score with disease progression. In this study, we built a new model based on a previously reported external model for predicting 3-year MDRS score variations. The predictive performance was improved when using this new model that included information on the *NOTCH3* gene mutation location.

Mutation located in domain 7–34 was independently associated with a greater average decrease in MDRS over 3 years. The results differ from some previous studies with a delayed stroke onset recently reported in patients with *NOTCH3* mutations located in EGFr domains 7–34 compared with patients with mutations in EGFr domains 1–6.[10] Nevertheless, in a recent work, Hack et al. found that the genotype-phenotype correlation could be further delineated, rather than basically dichotomized, 1–6 vs 7–34.[27] Specifically, they classified domains 8, 11, and 26 as high risk and found them associated with greater disability, higher risk of stroke, and higher load of neuroimaging SVD markers, in a cohort of 434 patients with CADASIL.[27]

This prediction improvement was obtained by creating synthetic data. These data were generated by simulation, based on and mirroring properties of the original data set. The inclusion of synthetic data allows data utility to be optimized, which subsequently enhances the model prediction.[28] A key advantage of this method is that it naturally incorporates knowledge into the internal data by creating a large set of "synthetic" data compatible with the initial model. We found that the best predictive performance with 10-fold cross-validation was achieved using the approach combining data augmentation and genetic information leading to an improvement of 25%, 15%, and 16% in the RMSPE, MSPE, and MAPE scores, respectively. Of interest, although we found that mutations located in EGFr domains 7–34 were independently associated with poorer outcomes than those located in EGFr domains 1–6, repeating the process after excluding the genetic information still resulted in a similar improvement. This further supports the positive effect of using synthetic data. Notably, the synthetic data method also reduced the standard errors of regression coefficients compared with the direct regression (Table 3). In this study, we only used the aggregated data from the first model, which also illustrates how external information could be used together with new individual data when testing the

Table 3 Multivariable Linear Regression of 3-Y MDRS Score Variation Using Multiple Imputation Without the EGFR Domain: Original Published Model by Chabriat, Estimates on the New Cohort of 367 Patients With CADASIL With and Without Synthetic Data

	Original model by Chabriat	Present cohort including synthetic data	Present cohort excluding synthetic data
Parameter	**Estimate β (SE)**	**Estimate β (SE)**	**Estimate β (SE)**
Sex (male)	0.73 (1.6)	0.90 (0.36)	0.90 (1.66)
Age	−0.15 (0.08)	−0.14 (0.01)	0.01 (0.05)
mRS score ≥3	−12.6 (3.5)	−5.11 (0.48)	−2.81 (2.47)
Balance problems	−3.7 (2.2)	−3.27 (0.47)	−0.004 (2.22)
Gait disturbance	5.0 (2.7)	3.31 (0.57)	0.18 (2.83)
No. lacunes	−0.37 (0.17)	−0.40 (0.02)	−0.20 (0.12)
Microbleeds	−0.17 (0.07)	−0.58 (0.40)	−0.52 (1.89)
BPF <0.863	0.31 (0.16)	1.08 (0.54)	0.64 (2.40)

Abbreviations: BPF = brain parenchymal fraction; mRS = modified Rankin Scale.

additional value of a biomarker in predicting the same outcome.

This approach requires some assumptions, notably, that the external model can be applied to external and internal populations. One-third of the sample individuals participated in the previous cohort from which the initial model was derived.[3] Thus, this first assumption could be considered acceptable here. In this study, some differences were observed in the estimated effects of the initial model parameters. These differences are possibly related to differences in baseline patient features across the 2 cohorts. Patients from the latter cohort were older and had less severe disabilities. There were also some possible changes in the recruitment, diagnosis, and care over time.

One important limitation of our study is the relatively small size of the cohort, which may not represent the entire population of patients with CADASIL.[29,30] In addition, the mutation location was dichotomized into 2 groups, and the potential influence of more detailed genetic information could not be excluded.[27] We used synthetic data to incorporate new information into the established model. However, other approaches are possible, such as constrained maximum likelihood, partial regression, and Bayesian approaches.[31] Furthermore, the assumption that data were missing at random is also debatable. The imputation model imputed the missing variables using information from all available data, but other unknown features could not be excluded.[32,33] Other genetic data, demographic features such as the level of education, professional activity, daily activity, or even other cognitive scores such as those derived from the Brief Memory and Executive Test,[34] the Montreal Cognitive Assessment,[35] or the Trail Making Test version B[36] might be useful to consider for fulfilling the missing at random assumption.[37,38] Moreover, the purpose of the study was initially to evaluate the impact on model performance of adding genetic information to an existing predictive model in patients with CADASIL. Although the resulting model (Table 2) might be useful for clinical practice, formal development of a simplified prediction score as a clinical tool would require external validation on a different data set from another cohort. This was beyond the scope of the present work.

In summary, our approach, based on the creation of synthetic data, allowed us to evaluate the potential effect of additional genetic information related to the location of the *NOTCH3* gene mutation on the prediction of cognitive decline in CADASIL. Additional investigations are needed to further improve this type of model using additional covariates[39-42] and supplementary information from other cohorts.

An existing risk model predicting 3-year MDRS score variation was enhanced by synthetic data obtained from a new study cohort and using already established model coefficients along with multiple imputations by chained equations. Initially used for incorporating the genetic mutation location information into the analysis model, we observed that synthetic data creation could finally enhance the prediction model. The prediction performance and estimation robustness were improved regardless of whether genetic information was included.

Acknowledgment

The authors thank the cohort study participants for their time, assiduity, and consent to participate in this research work.

Study Funding

This work was funded by the Recherche Hospitalo-Universitaire "TRT-cSVD" project funded by the Agence Nationale de la Recherche (reference ANR-16-RHUS-004).

Disclosure

The authors report no relevant disclosures. Go to Neurology.org/NG for full disclosures.

Publication History

Received by *Neurology: Genetics* January 12, 2023. Accepted in final form June 7, 2023. Submitted and externally peer reviewed. The handling editor was Associate Editor Suman Jayadev, MD.

Appendix Authors

Name	Location	Contribution
Henri Chhoa, MEng	ECSTRRA Team, Université Paris-Cité, UMR1153, INSERM, France	Drafting/revision of the manuscript for content, including medical writing for content; analysis or interpretation of data
Hugues Chabriat, MD, PhD	Translational Neurovascular Centre, GH Saint-Louis-Lariboisière, Assistance Publique des Hôpitaux de Paris APHP, Université Paris-Cité and DHU NeuroVasc Sorbonne Paris-Cité; UMR 1161, INSERM, France	Drafting/revision of the manuscript for content, including medical writing for content; major role in the acquisition of data; study concept or design; analysis or interpretation of data
Adelina Joanita Anato, BS	ENSAI, Ecole d'ingénieur statistique, data science et big data, Bruz, France	Analysis or interpretation of data
Mamadou Bamba, BS	ENSAI, Ecole d'ingénieur statistique, data science et big data, Bruz, France	Analysis or interpretation of data
Florent Zittoun, BS	ENSAI, Ecole d'ingénieur statistique, data science et big data, Bruz, France	Analysis or interpretation of data
Sylvie Chevret, MD, PhD	ECSTRRA Team, Université Paris-Cité, UMR1153, INSERM, France	Drafting/revision of the manuscript for content, including medical writing for content; study concept or design; analysis or interpretation of data
Lucie Biard, MD, PhD	ECSTRRA Team, Université Paris-Cité, UMR1153, INSERM, France	Drafting/revision of the manuscript for content, including medical writing for content; analysis or interpretation of data

References

1. Wang MM. CADASIL. *Handb Clin Neurol.* 2018;148:733-743. doi:10.1016/B978-0-444-64076-5.00047-8
2. Adib-Samii P, Brice G, Martin RJ, Markus HS. Clinical spectrum of CADASIL and the effect of cardiovascular risk factors on phenotype. *Stroke.* 2010;41(4):630-634. doi:10.1161/STROKEAHA.109.568402
3. Chabriat H, Hervé D, Duering M, et al. Predictors of clinical worsening in cerebral autosomal dominant arteriopathy with subcortical infarcts and leukoencephalopathy: prospective cohort study. *Stroke.* 2016;47(1):4-11. doi:10.1161/STROKEAHA.115.010696
4. Benjamin P, Lawrence AJ, Lambert C, et al. Strategic lacunes and their relationship to cognitive impairment in cerebral small vessel disease. *Neuroimage Clin.* 2014;4:828-837. doi:10.1016/j.nicl.2014.05.009
5. Ling Y, Chabriat H. Incident cerebral lacunes: a review. *J Cereb Blood Flow Metab.* 2020;40(5):909-921. doi:10.1177/0271678X20908361
6. Yates PA, Villemagne VL, Ellis KA, Desmond PM, Masters CL, Rowe CC. Cerebral microbleeds: a review of clinical, genetic, and neuroimaging associations. *Front Neurol.* 2014;4:205. doi:10.3389/fneur.2013.00205
7. Rudick RA, Fisher E, Lee JC, Simon J, Jacobs L. Use of the brain parenchymal fraction to measure whole brain atrophy in relapsing-remitting MS. Multiple Sclerosis Collaborative Research Group. *Neurology.* 1999;53(8):1698-1704. doi:10.1212/wnl.53.8.1698
8. Vågberg M, Granåsen G, Svenningsson A. Brain parenchymal fraction in healthy adults—a systematic review of the literature. *PLoS One.* 2017;12(1):e0170018. doi:10.1371/journal.pone.0170018
9. Rutten JW, Dauwerse HG, Gravesteijn G, et al. Archetypal *NOTCH3* mutations frequent in public exome: implications for CADASIL. *Ann Clin Transl Neurol.* 2016;3(11):844-853. doi:10.1002/acn3.344
10. Rutten JW, Van Eijsden BJ, Duering M, et al. The effect of *NOTCH3* pathogenic variant position on CADASIL disease severity: *NOTCH3* EGFr 1-6 pathogenic variant are associated with a more severe phenotype and lower survival compared with EGFr 7-34 pathogenic variant. *Genet Med.* 2019;21(3):676-682. doi:10.1038/s41436-018-0088-3
11. Dupé C, Guey S, Biard L, et al. Phenotypic variability in 446 CADASIL patients: impact of *NOTCH3* gene mutation location in addition to the effects of age, sex and vascular risk factors. *J Cereb Blood Flow Metab.* 2023;43(1):153-166. doi:10.1177/0271678X221126280
12. Jiao F, Chen YF, Min M, Jimenez S. Challenges and potential strategies utilizing external data for efficacy evaluation in small-sized clinical trials. *J Biopharm Stat.* 2022;32(1):21-33. doi:10.1080/10543406.2021.2011906
13. Ventz S, Comment L, Louv B, et al. The use of external control data for predictions and futility interim analyses in clinical trials. *Neuro Oncol.* 2022;24(2):247-256. doi:10.1093/neuonc/noab141
14. Grill S, Fallah M, Leach RJ, Thompson IM, Hemminki K, Ankerst DP. A simple-to-use method incorporating genomic markers into prostate cancer risk prediction tools facilitated future validation. *J Clin Epidemiol.* 2015;68(5):563-573. doi:10.1016/j.jclinepi.2015.01.006
15. Grill S, Ankerst DP, Gail MH, Chatterjee N, Pfeiffer RM. Comparison of approaches for incorporating new information into existing risk prediction models. *Stat Med.* 2017;36(7):1134-1156. doi:10.1002/sim.7190
16. Ankerst DP, Koniarski T, Liang Y, et al. Updating risk prediction tools: a case study in prostate cancer. *Biom J.* 2012;54(1):127-142. doi:10.1002/bimj.201100062
17. Cheng W, Taylor JMG, Gu T, Tomlins SA, Mukherjee B. Informing a risk prediction model for binary outcomes with external coefficient information. *J R Stat Soc Ser C Appl Stat.* 2019;68(1):121-139. doi:10.1111/rssc.12306
18. Gu T, Taylor JMG, Cheng W, Mukherjee B. Synthetic data method to incorporate external information into a current study. *Can J Stat.* 2019;47(4):580-603. doi:10.1002/cjs.11513
19. Peters N, Herzog J, Opherk C, Dichgans M. A two-year clinical follow-up study in 80 CADASIL subjects: progression patterns and implications for clinical trials. *Stroke.* 2004;35(7):1603-1608. doi:10.1161/01.STR.0000131546.71733.f1
20. Cockrell JR, Folstein MF. Mini-mental state examination (MMSE). *Psychopharmacol Bull.* 1988;24(4):689-692.
21. Quinn TJ, Dawson J, Walters MR, Lees KR. Reliability of the modified Rankin Scale: a systematic review. *Stroke.* 2009;40(10):3393-3395. doi:10.1161/STROKEAHA.109.557256
22. Collin C, Wade DT, Davies S, Horne V. The Barthel ADL Index: a reliability study. *Int Disabil Stud.* 1988;10(2):61-63. doi:10.3109/09638288809164103
23. Viswanathan A, Guichard JP, Gschwendtner A, et al. Blood pressure and haemoglobin A1c are associated with microhaemorrhage in CADASIL: a two-centre cohort study. *Brain.* 2006;129(Pt 9):2375-2383. doi:10.1093/brain/awl177
24. Wardlaw JM, Smith EE, Biessels GJ, et al. Neuroimaging standards for research into small vessel disease and its contribution to ageing and neurodegeneration. *Lancet Neurol.* 2013;12(8):822-838. doi:10.1016/S1474-4422(13)70124-8
25. Browne MW. Cross-validation methods. *J Math Psychol.* 2000;44(1):108-132. doi:10.1006/jmps.1999.1279
26. van Buuren S, Groothuis-Oudshoorn K. Mice: multivariate imputation by chained equations in R. *J Stat Softw.* 2011;45:1-67. doi:10.18637/jss.v045.i03
27. Hack RJ, Gravesteijn G, Cerfontaine MN, et al. Three-tiered EGFr domain risk stratification for individualized *NOTCH3*-small vessel disease prediction. *Brain.* 2023;146(7):2913-2927. doi:10.1093/brain/awac486
28. James S, Harbron C, Branson J, Sundler M. Synthetic data use: exploring use cases to optimise data utility. *Discov Artif Intell.* 2021;1(1):15. doi:10.1007/s44163-021-00016-y
29. Kukull WA, Ganguli M. Generalizability. *Neurology.* 2012;78(23):1886-1891. doi:10.1212/WNL.0b013e318258f812
30. Murad MH, Katabi A, Benkhadra R, Montori VM. External validity, generalisability, applicability and directness: a brief primer. *BMJ Evid Based Med.* 2018;23(1):17-19. doi:10.1136/ebmed-2017-110800
31. Cheng W, Taylor JMG, Vokonas PS, Park SK, Mukherjee B. Improving estimation and prediction in linear regression incorporating external information from an established reduced model. *Stat Med.* 2018;37(9):1515-1530. doi:10.1002/sim.7600
32. Potthoff RF, Tudor GE, Pieper KS, Hasselblad V. Can one assess whether missing data are missing at random in medical studies? *Stat Methods Med Res.* 2006;15(3):213-234. doi:10.1191/0962280206sm448oa
33. Leacy FP, Floyd S, Yates TA, White IR. Analyses of sensitivity to the missing-at-random assumption using multiple imputation with delta adjustment: application to a tuberculosis/HIV prevalence survey with incomplete HIV-status data. *Am J Epidemiol.* 2017;185(4):304-315. doi:10.1093/aje/kww107
34. Brookes RL, Hollocks MJ, Khan U, Morris RG, Markus HS. The Brief Memory and Executive Test (BMET) for detecting vascular cognitive impairment in small vessel disease: a validation study. *BMC Med.* 2015;13(1):51. doi:10.1186/s12916-015-0290-y

35. Kang JM, Cho YS, Park S, et al. Montreal cognitive assessment reflects cognitive reserve. *BMC Geriatr.* 2018;18:261. doi:10.1186/s12877-018-0951-8
36. Kortte KB, Horner MD, Windham WK. The trail making test, part B: cognitive flexibility or ability to maintain set? *Appl Neuropsychol.* 2002;9(2):106-109. doi:10.1207/S15324826AN0902_5
37. Jouvent E, Duchesnay E, Hadj-Selem F, et al. Prediction of 3-year clinical course in CADASIL. *Neurology*. 2016;87(17):1787-1795. doi:10.1212/WNL.0000000000003252
38. Jolly AA, Nannoni S, Edwards H, Morris RG, Markus HS. Prevalence and predictors of vascular cognitive impairment in patients with CADASIL. *Neurology*. 2022;99(5):e453-e461. doi:10.1212/WNL.0000000000200607
39. Svetnik V, Liaw A, Tong C, Culberson JC, Sheridan RP, Feuston BP. Random forest: a classification and regression tool for compound classification and QSAR modeling. *J Chem Inf Comput Sci.* 2003;43(6):1947-1958. doi:10.1021/ci034160g
40. Brereton RG, Lloyd GR. Support vector machines for classification and regression. *Analyst.* 2010;135(2):230-267. doi:10.1039/b918972f
41. Elith J, Leathwick JR, Hastie T. A working guide to boosted regression trees. *J Anim Ecol.* 2008;77(4):802-813. doi:10.1111/j.1365-2656.2008.01390.x
42. Dasgupta A, Sun YV, König IR, Bailey-Wilson JE, Malley JD. Brief review of regression-based and machine learning methods in genetic epidemiology: the genetic analysis workshop 17 experience. *Genet Epidemiol.* 2011;35(suppl 1):S5-S11. doi:10.1002/gepi.20642