Received: 12 November 2021 | Revised: 21 February 2022 | Accepted: 25 February 2022

DOI: 10.1002/alz.12663

Alzheimer's & Dementia®
THE JOURNAL OF THE ALZHEIMER'S ASSOCIATION

FEATURED ARTICLE

Dementia risk predictions from German claims data using methods of machine learning

Constantin Reinke[1] | Gabriele Doblhammer[1,2] | Matthias Schmid[2,3] | Thomas Welchowski[3]

[1]Institute for Sociology and Demography, University of Rostock, Rostock, Germany

[2]German Center for Neurodegenerative Diseases, Bonn, Germany

[3]Institute of Medical Biometry, Informatics and Epidemiology (IMBIE), Medical Faculty, University of Bonn, Bonn, Germany

Correspondence
Constantin Reinke, Institute for Sociology and Demography, University of Rostock, Ulmenstr. 69, 18057 Rostock, Germany.
Email: constantin.reinke@uni-rostock.de

Abstract

Introduction: We examined whether German claims data are suitable for dementia risk prediction, how machine learning (ML) compares to classical regression, and what the important predictors for dementia risk are.

Methods: We analyzed data from the largest German health insurance company, including 117,895 dementia-free people age 65+. Follow-up was 10 years. Predictors were: 23 age-related diseases, 212 medical prescriptions, 87 surgery codes, as well as age and sex. Statistical methods included logistic regression (LR), gradient boosting (GBM), and random forests (RFs).

Results: Discriminatory power was moderate for LR (C-statistic = 0.714; 95% confidence interval [CI] = 0.708–0.720) and GBM (C-statistic = 0.707; 95% CI = 0.700–0.713) and lower for RF (C-statistic = 0.636; 95% CI = 0.628–0.643). GBM had the best model calibration. We identified antipsychotic medications and cerebrovascular disease but also a less-established specific antibacterial medical prescription as important predictors.

Discussion: Our models from German claims data have acceptable accuracy and may provide cost-effective decision support for early dementia screening.

KEYWORDS
calibration, dementia, discrimination, Germany, health claims data, machine learning, risk factors

1 | BACKGROUND

There are currently ≈47 million people worldwide living with dementia. This number is expected to increase to 78 million by 2030 and to 132 million by 2050.[1] In Germany, 1.6 million people are presently living with dementia, with an expected increase to 2.7 million by 2050.[2] Dementia creates high costs for society and the health care system, which increase significantly as the disease progresses.[3]

Because dementia is still incurable, prevention is the best strategy to delay its onset and to slow progression, with the goal of reducing the burden of dementia on those affected and on the health care system. For effective prevention, it is crucial to identify modifiable risk factors and to detect cognitive decline at an early stage prior to manifestation, even better, before the onset.

To achieve this, numerous dementia risk-prediction models have been developed with different target populations and outcomes. Most models predict the risk of late-life dementia for non-demented people, but there are also midlife risk models and models for the conversion from mild cognitive impairment (MCI) to dementia. Outcomes include Alzheimer's disease (AD), other dementia subtypes, combinations of subtypes, and all-cause dementia. However, reviews stress the need for further models in different populations[4,5] and point out that more recent and larger data sets are needed to overcome the lack of diversity in previous studies.[6] To address this limitation, routinely collected health care data has come more into focus in recent years.[7–11] These data are typically cost-effective because they do not need to be collected separately. Typical routinely collected data are electronic health records (EHRs) as well as administrative health claims data. Although the introduction of EHR in Germany has already been legally decreed, the implementation, collection, and availability of the data have not yet been established,[12] making them hardly accessible for research due to data protection regulations. In the German health care system, one of the largest health care systems worldwide, it is mandatory to participate in a health insurance fund, and nearly 90% of the population is covered by public health insurance. Although administrative health claims data are used primarily for the purpose of billing for health care services, these claims contain a large amount of clinically relevant information. They also include sociodemographic data, as well as all diagnoses made by physicians, operations performed, and medications (prescriptions filled). Therefore, claims data are used increasingly in public health research.

An illustrative example of the use of routinely collected data for disease prediction is the QRISK, a cardiovascular disease prediction risk score that complements the prevailing Framingham Risk Score.[13]

Early detection of dementia would require a life-course approach, with major vascular risk factors already becoming prevalent during midlife[14] and the diagnosis relying strongly on the medical history of the patient.[15] Health claims data have the potential to provide such long-term information.

Previous studies have used a number of statistical methods to predict dementia risk. In recent years the most common approaches have been logistic and Cox regression as well as an increasing number of machine learning (ML) techniques.[6] ML algorithms are well suited for the analysis of data sets with a large amount of information, as they usually contain automatic variable selection mechanisms and can include non-linear associations as well as complex interactions between variables.[16] Several studies have used ML algorithms for risk prediction with administrative claims data. For example, a recent study compared different ML methods and traditional models to predict heart failure outcomes, achieving the best performance with gradient-boosting models and logistic regression.[17] Another study found an improved accuracy of cardiovascular risk prediction using ML and electronic medical records.[18] Nori and colleagues identified incident dementia by applying ML algorithms to an administrative claims data set of privately insured individuals in the United States.[9] Moreover, logistic regression has been used to predict dementia diagnosis from administrative claims.[10] These studies identified neurological and psychological disorders and psychoactive medications as predictors with the greatest impact on dementia risk. Discriminatory power, as measured by the area under the curve and the concordance index, ranged between 0.63 and 0.76.

In this context we examined the following research questions: First, are German claims data a suitable data basis for individual dementia-risk prediction? Second, how do ML methods compare to classical regression methods in terms of predicting dementia risk? Third, which features are important predictors of dementia risk, and can new features be identified in addition to established risk factors?

RESEARCH IN CONTEXT

1. **Systematic review:** An increasing number of dementia risk- prediction models have been developed and the most common methodologies are machine learning (ML) and traditional regression methods. Despite the increasing availability of routinely collected health data in Germany, dementia risk predictions using these data are still rare.
2. **Interpretation:** German claims data are suitable for dementia risk prediction. We found moderate prediction accuracy for logistic regression and gradient boosting. In addition to some well-known dementia-related features, we identified the pharmacological subgroup of macrolides, lincosamides, and streptogramins (ATC-code: J01F) as an important predictor for dementia.
3. **Future directions:** Dementia risk-prediction models from German claims data may be useful in implementing cost-effective decision-support tools for early dementia screening. Data-driven approaches with claims data have the potential to identify new features or pathways affecting the risk of dementia.

HIGHLIGHTS

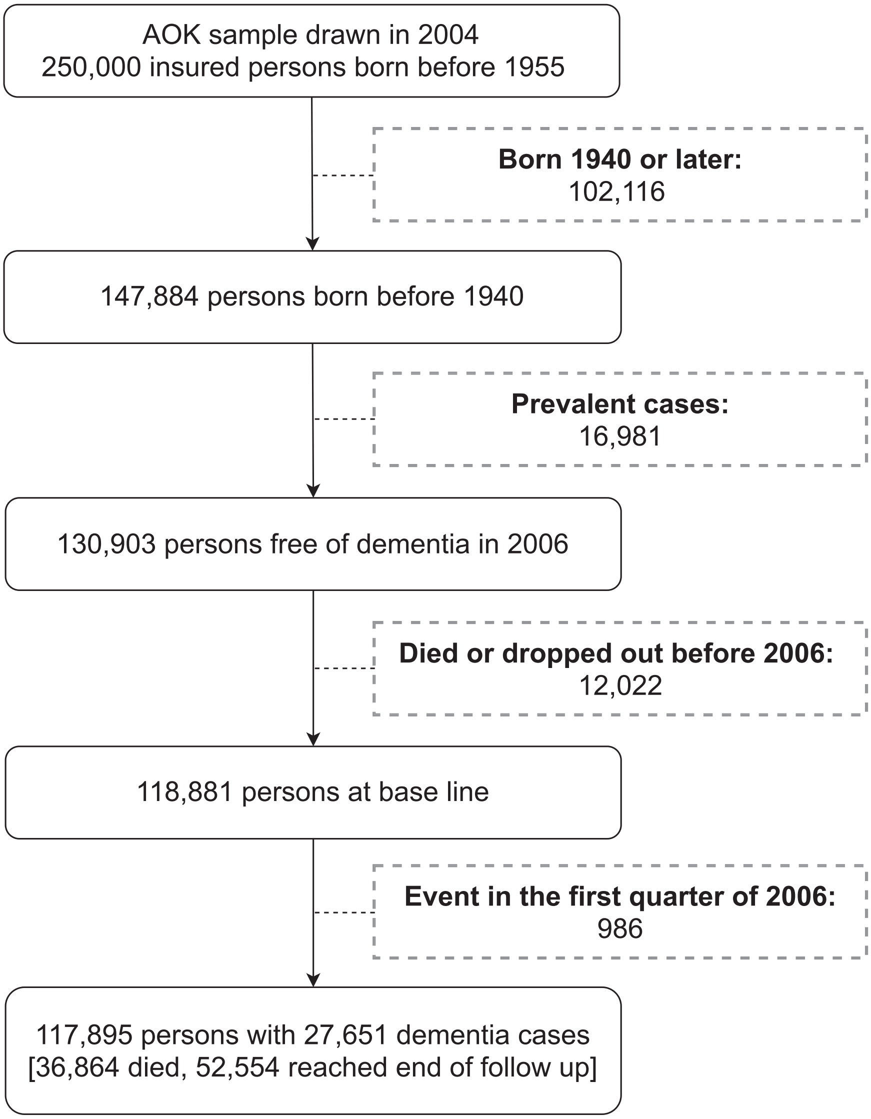
- Gradient boosting machine (GBM) showed the best model calibration.
- The prediction accuracy of GBM was comparable to that of classical logistic regression.
- Model accuracies were comparable and partly better than other studies using claims.
- Selected antibacterial medical prescriptions were an important predictor for dementia.

2 | MATERIAL AND METHODS

2.1 | Data

We used an age-stratified random sample of 250,000 persons insured in the largest German health insurance company "Allgemeine Ortskrankenkasse" (AOK). The sample was drawn in 2004 and included people born before 1955 with a follow-up to 2015. The AOK covers almost 30% of the German population and is representative of the German population aged 65+ in terms of mortality (Figure S1). The data contained the following information from the inpatient and outpatient sector: Diagnoses based on the International Classification of Diseases, Tenth Revision (ICD-10), all medical prescriptions according to the Anatomical Therapeutic Chemical (ATC) Classification

FIGURE 1 Selection of study cohort


AOK sample drawn in 2004
250,000 insured persons born before 1955
Born 1940 or later:
102,116
147,884 persons born before 1940
Prevalent cases:
16,981
130,903 persons free of dementia in 2006
Died or dropped out before 2006:
12,022
118,881 persons at base line
Event in the first quarter of 2006:
986
117,895 persons with 27,651 dementia cases
[36,864 died, 52,554 reached end of follow up]

System, surgeries based on the German procedure classification (OPS, www.dimdi.de/dynamic/en/classifications/ops/, accessed: November 1, 2021) an adaptation of the International Classification of Procedures in Medicine (ICPM), as well as sex, age, and time of death. All medical information was recorded on a quarterly basis. The data were anonymized claims data and did not require ethical review or patient consent.

2.2 | Study design

To predict incident dementia, we selected people born before 1940 because dementia before age 65 is extremely rare. Because we were interested in newly diagnosed dementia cases only, we excluded all people with prevalent dementia in 2004 or 2005. In the quarter of a dementia diagnosis, we assigned all predictors the values of the previous quarter because the exact timing of diagnoses, prescriptions, and surgeries or the chronological order of occurrence was unknown. Hence, we excluded all persons with an incident dementia diagnosis in the first quarter of 2006 (Figure 1), and the study period began with the first quarter of 2006 and ended with the last quarter of 2015. We assembled a training population by drawing a 60% random sample from our study cohort, stratified by dementia status (dementia diagnosis vs no dementia diagnosis). The remaining individuals were split into validation and test groups of equal size.

2.3 | Outcome

Our outcome was a binary variable that indicated a validated incident dementia diagnosis. Dementia was defined by ICD-10 codes (Table S1). To address the problem of false-positive diagnosis, we applied a two-stage validation strategy[19] (Description S1).

2.4 | Predictors

We included 23 major age-related diseases[20] and risk factors for dementia according to the factors reported by the Lancet Commission[21] (see Table S1 for the respective ICD codes). In addition, we included all medical prescriptions coded by the German version of the ATC classification system on the 3rd level (pharmacological subgroup), and all surgeries based on three-digit classes from Chapter 5 of the OPS. Further predictors were sex and age at baseline.

All predictors except age and sex were included as time-dependent binary "ever" variables, with the value 1 from the first occurrence of a particular code onward and zero otherwise.

2.5 | Statistical analysis

The data were structured quarterly, with one observation per person and quarter. Therefore, we defined a discrete time process starting in the first quarter of 2006, with time intervals given by quarters, resulting in a data structure with binary outcome.[22] Accordingly, time measurements ranged between 1 and 40, referring to the number of quarters observed for each individual. To predict dementia risk, we built prediction models using logistic regression, gradient boosting machines (GBMs), and random forests (RFs). The input data for these models consisted of one row per quarter, thereby allowing for time-dependent predictors and accounting for right-censoring ("discrete hazard models," cf.[22]) We excluded predictors with near-zero-variance using the nearZeroVar function from,[23] and predictors with fewer than five observations per cell in a cross-tabulation with dementia. Surgery codes 502 and 583 were excluded due to collinearity. In total, we included 324 features: 212 ATC codes, 87 OPS codes, 23 diagnoses, and age and sex.

As a benchmark model, we considered a logistic regression model (including all predictors) that was fitted to the combined training and validation data, and the process time was included as a categorical predictor.

We used the R package xgboost version 1.1.1.1[24] to train a GBM with a learning rate of 0.01 and a maximum of 10,000 iterations on the training data. The algorithm was stopped if there was no improvement in log-loss in the last 100 iterations evaluated on the validation data.[24] We used a grid search on the validation data to find optimal values for the parameters *max depth* and *min child weight*. Finally, we fitted a GBM with the optimal parameter values to the combined training and validation data.

We used the R package ranger version 0.12.1[25] to train RF with ntree = 1000 and ntree = 10,000 trees on the training data. All other parameters were set to default. Because logistic regression and GBM clearly outperformed RF, we did not perform any further parameter tuning here.

Because a detailed description of the GBM and RF methodology is beyond the scope of this applied study, we included our source code in the supplements (Supplementary_source_code). This code provides detailed information about our analysis and preprocessing methods.

2.6 | Model evaluation

We evaluated the performance of our prediction models in terms of accuracy, discriminatory power, and calibration.

The overall accuracy was evaluated by the integrated prediction error, an adaptation of the Brier score, which is based on weighted quadratic differences between predicted and observed survival functions.[22] We calculated the integrated prediction error using the R package discSurv[26]; 95% confidence intervals (95% CIs) were calculated using 5000 bootstrap replications from the test data.

To evaluate the discriminatory power, we calculated a time-independent version of the concordance index (C statistic).[27] As with the prediction error, we used the discSurv package and calculated 95% CIs using bootstrapping.

Because a strong discriminatory power is not sufficient to assess a model for clinical usability,[28] we additionally examined calibration plots (to graphically assess model calibration) and calculated an intercept and slope to test whether the predicted risks were systematically overestimated or underestimated.[29]

To extract the explainable information from the GBM model, we identified the 20 most influential features for prediction, using a permutation approach to calculate a relative importance score for each predictor.[30] To present explainable information for logistic regression, we report the corresponding odds ratios, ranked by absolute z values. Because GBM clearly outperformed RF, we did not calculated variable importance for RF.

3 | RESULTS

3.1 | Study cohort

Our study cohort consisted of 117,895 individuals, and we observed 27,651 incident dementia cases. During the study period, 63,864 individuals died (details in Figure S2) and 52,554 reached the end of follow-up (Figure 1). The training data consisted of 70,737 people, and the validation data included 23,579 people, as did the test data. At baseline, the mean age was 74.8 years (SD = 6.6), and the mean age at dementia diagnosis was 82.3 years (SD = 6.3); 38% of the individuals were men (See Table 1). The mean follow-up time was 17.8 quarters (SD = 11.3).

3.2 | Model evaluation

Logistic regression indicated the strongest discriminatory power on the test data, with a C statistic of 0.714 (95% CI = 0.708–0.720), closely followed by GBM with a C statistic of 0.707 (95% CI = 0.700–0.713). Although the CIs overlapped (Summary in Table S2), pairwise differences in C values were different from zero (mean = 0.007, 95% CI = 0.005–0.009). The discriminatory power of RF (ntree = 1000) was considerably lower, with a C statistic of 0.636 (95% CI = 0.628–0.643). The same model-ranking sequence appeared for the integrated prediction error, where the lowest error was found for the logistic regression model (0.044, 95% CI = 0.044–0.045), followed by GBM (0.046, 95% CI = 0.046–0.047) and RF (0.105, 95% CI = 0.104–0.107). Looking at the calibration plots (Figure 2), the GBM appeared to be a well-calibrated model with an intercept near zero (0.097) and a slope close

TABLE 1 Cohort characteristics at baseline

Training	Number	%	2.5% Quantile	97.5% Quantile
N	70,737			
Age at baseline, mean (SD)	72.8 (6.6)		64	89
Men	26,997	38.2		
Dementia cases (not at baseline)	16,522	23.4		
Antipsychotics (ATC: N05A)	3,528	5.0		
Cerebrovascular disease	14,528	20.5		
Anti-dementia drugs (N06D)	966	1.4		
Depression	13,547	19.2		
Parkinson disease	1,489	2.1		
Injuries to the head	3,734	5.3		
Antidepressants (N06A)	8,729	12.3		
Cardiomyopathy and heart failure	17,075	24.1		
Test				
N	23,579			
Age at baseline, mean (SD)	72.9 (6.6)		64	88
Men	9,063	38.4		
Dementia cases (not at baseline)	5,506	23.4		
Antipsychotics (N05A)	1,090	4.6		
Cerebrovascular disease	4,900	20.8		
Anti-dementia drugs (N06D)	330	1.4		
Depression	4,458	18.9		
Parkinson disease	479	2.0		
Injuries to the head	1,200	5.1		
Antidepressants (N06A)	2,848	12.1		
Cardiomyopathy and heart failure	5,794	24.6		
Total				
N	117,895			
Age at baseline, mean (SD)	72.8 (6.6)		64	89
Men	45,038	38.2		
Dementia cases (not at baseline)	27,651	23.5		
Antipsychotics (N05A)	5,827	4.9		
Cerebrovascular disease	24,242	20.6		
Anti-dementia drugs (N06D)	1,639	1.4		
Depression	22,548	19.1		
Parkinson disease	2,491	2.1		
Injuries to the head	6,176	5.2		
Antidepressants (N06A)	14,601	12.4		
Cardiomyopathy and heart failure	28,575	24.2		

Table 1 shows the baseline cohort characteristics for the training, test, and full data sets. In addition to age and sex, we report only the 10 most influential predictors in terms of variable importance (Figure 2). In total, we included 212 ATC codes, 87 OPS codes, 23 diagnoses, and age and sex. Source: AOK data 2004-2015, own calculations.

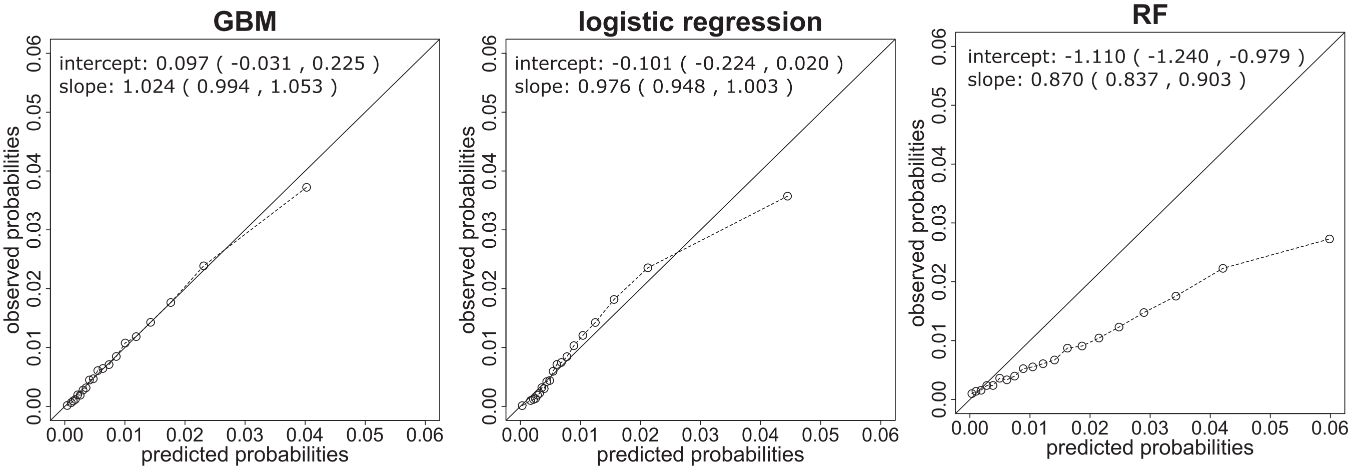


FIGURE 2 Model calibration on test data for GBM (gradient boosting machine), logistic regression, and RF (random forest). Intercepts and slopes were calculated by logistic calibration (95% confidence intervals in parentheses). The models showed weaker performance in the higher risk segments compared to the medium and lower segments

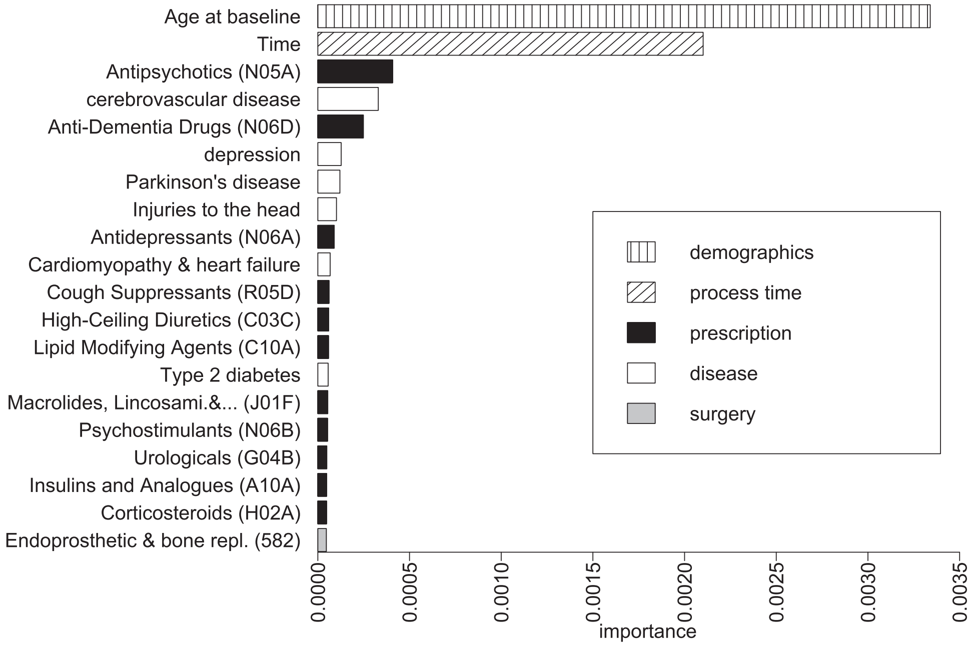


FIGURE 3 Variable importance of 20 of the most influential features from the GBM (gradient boosting machine) model. *Source*: AOK data 2004-2015, own calculations

to 1 (1.024). The logistic regression model was calibrated slightly worse (intercept: −0.101; slope: 0.976) and the RF showed substantial issues with calibration (intercept: −1.110; slope: 0.870). Predictive performance of the RF with ntree = 10,000 was very similar to the performance of RF with ntree = 1000 (Table S2). The performance of RF may be improved further by hyper-parameter tuning.

3.3 | Most important predictors

By far the most important predictor in the GBM model was age at baseline (Figure 3). The most important medical prescriptions were antipsychotics (N05A), anti-dementia drugs (N06D), and antidepressants (N06A), all of which were among the top 10 features. Most interestingly, 11 of the top 20 features were medical prescriptions; among these some medications associated with diseases linked to a high risk of dementia (insulin for diabetes mellitus, diuretics for high blood pressure). However, we also found medication that has not been described in the context of dementia prediction, such as macrolides, lincosamides and streptogramins (J01F), and medication that has been described as being protective in previous studies, for example, corticosteroids (H02A).

Among the diseases, cerebrovascular diseases were most important, followed by depression, Parkinson disease, and injuries of the

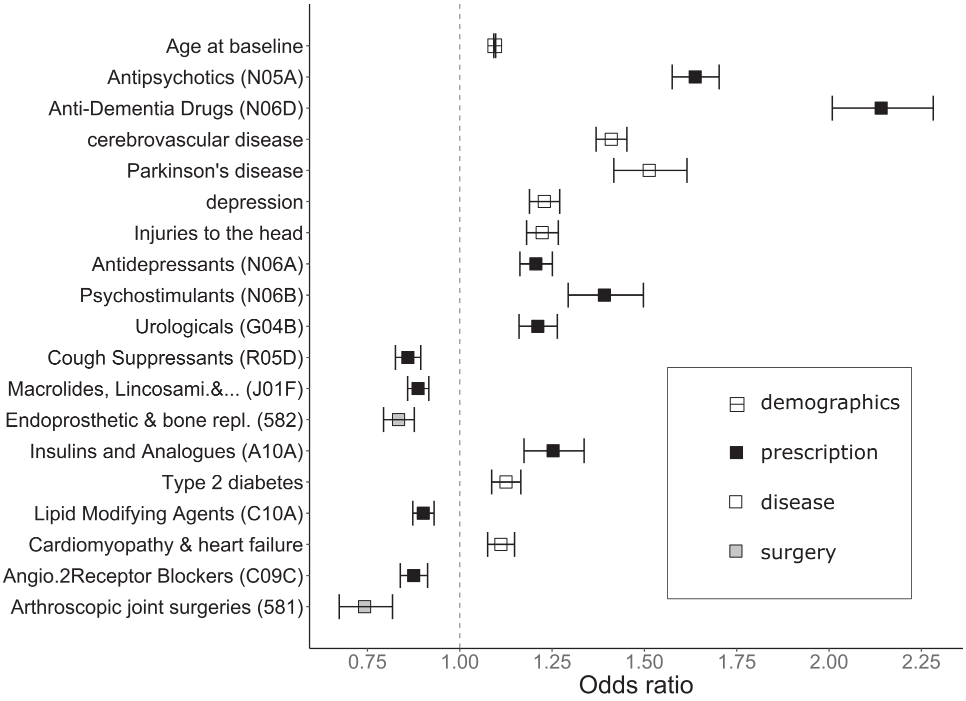


FIGURE 4 Odds ratios with 95% confidence intervals from logistic regression according to the highest absolute z values (except time). *Source*: AOK data 2004-2015, own calculations

head. In addition, cardiomyopathy and heart failure and diabetes mellitus were among the top 20. For surgeries, endoprosthetic and bone replacement was an important feature.

The odds ratios (Figure 4) show a slightly different ranking than the variable importance, however, most variables are identical. Angiotensin-2 receptor blockers and arthroscopic joint surgeries occurred here instead of high-ceiling diuretics and corticosteroids.

4 | DISCUSSION

We applied and compared classical and ML methods to develop dementia risk-prediction models using German health claims data. To our knowledge, this is the first study using German claims data for dementia-risk prediction.

With C statistics higher than 0.70, both logistic regression and GBM indicated acceptable discriminatory power. When compared to clinical studies, which used information on biomarkers, cognitive test scores, or laboratory results, this discriminatory power is somewhat lower.[4] This may be because much of this information is more closely related to diagnostic approaches for dementia than individual medical history. For example, cognitive tests are usually part of the diagnostic process. Nevertheless, our results are comparable to or partly better than the results of other studies, which used claims data only.[9,10,31]

The ML methods did not outperform classical logistic regression in terms of discriminatory power, as indicated by our test on pairwise differences. However, GBM showed a better calibration. Regarding interpretability, odds ratios obtained from logistic regression are usually easier to interpret than variable importance measures obtained from ML methods. Specifically, odds ratios come along with a sign (indicating risks vs protective effects) and are well suited for explanatory purposes (provided that the model has been specified correctly). On the other hand, variable importance measure provides more insight in the *predictive* ability of variables, incorporating effects due to non-linearity and variable selection.

Most of the features identified among the top 20 are consistent with previous results about important risk and predictive factors for dementia. By far the most important feature for predicting dementia in general and in our study is age. Also outstanding are antipsychotics (N05A), which are used frequently to treat a range of psychiatric symptoms[32] as well as for the pre-dementia stage of MCI.[33,34] Although the evidence for the benefit of antidementia drugs (N06D) in the pre-dementia stage is inconclusive,[35] they are widely used in clinical practice.[36] Note that the high importance of anti-dementia drugs may also indicate a high proportion of undiagnosed dementia cases, especially for individuals in the phase of conversion from MCI to dementia, which is a phase of high uncertainty.[37]

The high importance of cerebrovascular disease for incident dementia (feature ranks fourth after age and antipsychotics) may indicate a high proportion of vascular dementia. A further important cardiovascular-related feature is cardiomyopathy and heart failure, which underscores the importance of the association of cardiovascular diseases and dementia. Type 2 diabetes and insulin (A10A) can also be included in this category. Considering that drug utilization can be an indicator of diseases,[38] high-ceiling diuretics (C03C) may indicate hypertension. Neurodegenerative risk factors such as head injury, which is associated with an increased risk of dementia[39] and Parkinson disease,[40] are ranked seventh and eighth. Late-life depression is a known risk factor for dementia,[41] but may also be a prodromal symptom of dementia.[42] There is also evidence for an association between antidepressants and an increased risk of dementia,[43] but the prescriptions of antidepressants could also indicate undiagnosed depression. The only surgery code in the top 20 was endoprosthetic and bone replacement (OPS:582). There is evidence that although the risk of dementia was increased in the quarter of endoprosthetic and bone replacement surgery, it was lower in the postoperative period than for those without surgery.[44] Psychostimulants contain medication for the treatment of attention deficit/hyperactivity disorder (ADHD), which is associated with an increased risk of dementia.[45] Urologicals are partly associated with cognitive impairment[46] and are prescribed commonly for urinary incontinence, which has been identified as a predictor of dementia in previous studies.[9,10]

Features with evidence of a protective association with dementia risk were also included in the top 20 most important features, for example, cough suppressants,[47] lipid-modifying agents,[48] and corticosteroids.[49] However, we also found some less-established dementia-related features, such as medical prescriptions of the macrolides, lincosamides, and streptogramins (ATC:J01F). The association of inflammation and bacterial infections with cognitive decline and dementia are well established,[50] whereas the role of antibacterial medications and the risk of dementia is largely unclear.[51] Antibiotics, for one thing may damage the microbiome in the gut leading to a higher dementia risk.[52] Then again they may reduce inflammation leading to a lower risk.[53]

Although the discriminatory power of our approach suggests a rather moderate relevance for direct identification of dementia cases in clinical practice, it highlights the potential of health claims data as supporting tools for prediction of dementia risk. Dementia-risk predictions from claims data can be used as a cost-effective indication for the need for additional dementia screening. The combination of claims data with further information may improve the discriminatory power[17] and increase relevance in clinical practice.[54] One reason that ML did not perform better than logistic regression in our study could be the large number of 0-1 coded features (only age was continuous). By including additional information, such as laboratory values or test parameters, ML may perform better than logistic regression, as associations with dementia risk are not necessarily linear.

The use of data-driven ML methods in routinely collected data can be an important contribution to identifying or better understanding known pathways and new risk and preventive factors for dementia (such as the feature macrolides, lincosamides and streptogramins (J01F), which we found). The ability of ML methods to include complex interactions between several features may be key to the study of multimorbidity and effects of polypharmacy.[55] These strengths of ML methods are especially important against the background of increasing availability of high-dimensional data in health care.

4.1 | Strengths and limitations

The large longitudinal and population-based data containing information from the inpatient and outpatient sector as well as those living in nursing homes is representative of the older German population, which adds to the strength of the study conducted. Because the data are collected routinely in a standardized fashion, problems such as sample selection bias, attrition, and recall bias are less relevant than in other data sources. We applied an established strategy to validate dementia diagnosis. Three different statistical methods were used, which included more than 300 characteristics as time-dependent predictors. In addition to discriminatory power, we also explored the calibration of the models using a large internal test set.

We acknowledge some limitations in this study. Although claims data offer some advantages, it is important to note that this type of data is used primarily for billing purposes and the information available was generated for health care utilization only. For example, diagnoses can be identified only for people who went to a doctor, which limits the generalizability of our results. In addition, dementia may be underreported, especially in the early pre-clinical stages. Furthermore, a large proportion of dementia diagnoses are unspecific, which does not allow for an accurate distinction between dementia subtypes. The information about medical prescriptions is limited to the collection of the drug (redemption of the prescription), without any information about the actual intake. Moreover, the proportion of persons with low socioeconomic status is higher in the AOK than in other statutory health insurance companies and also in comparison with private health insurance companies.[56] Although these differences could influence both morbidity and the utilization of health care services, they can be explained partly by the different age structure of the AOK population, which is older than the German population. At the same age, the difference in the social structure of the AOK population is larger in younger age groups than in older.[56] Age-specific mortality[57] and age-specific prevalence and incidence of dementia in AOK data are similar to those shown for the total German population.[19] The diagnoses included here are limited to a manual selection of 23 age-related conditions, which seems counterintuitive in the context of a data-driven approach. However, consideration of all ICD-10 diagnosis codes at the disease group level (three digits) might not be accurate in medical terms, and inclusion of all diagnosis codes (more than 13,000) would exceed our computational resources and would lead to very sparse data. Because our models are limited to Germany and were validated with data from the same source, the generalizability of our results needs further investigation. Future studies should also investigate the effects of competing events (eg, death before dementia diagnosis).

5 | CONCLUSION

Our results from routinely collected claims data are not suitable for making diagnoses or replacing established tests in clinical practice, but they may be useful as an additional measure for risk detection. Specifically, they may be useful in implementing decision support for early dementia screening in a cost-effective manner if health care providers could continuously update physicians on current risk predictions. Ideally, the cornerstones of dementia diagnosis such as clinical assessment, laboratory testing, and imaging[15] should be combined into one data repository. Undoubtedly, more research on different types of health data is needed before any real benefit of such an approach at the public health level can be determined. Our results may also be relevant to dementia prevention research in order to identify new features or pathways that influence dementia risk. The combination of claims data with data-driven approaches may serve as a starting point for further research into the largely unknown association between dementia and characteristics such as certain types of antibacterial medical prescriptions identified in this study.

ACKNOWLEDGEMENT

The authors are grateful to the Scientific Research Institute of the Allgemeine Ortskrankenkassen (AOK), (German: Wissenschaftliches Institut der AOK, WIdO), for providing the data. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Open access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ORCID

Constantin Reinke https://orcid.org/0000-0003-3228-1794

REFERENCES

1. Prince MJ, Wimo A, Guerchet MM, Ali GC, Wu Y-T, Prina M. World Alzheimer Report 2015 - The Global Impact of Dementia: An analysis of prevalence, incidence, cost and trends. London: Alzheimer's Disease International, 2015.
2. Alzheimer Europe. Dementia in EuropeYearbook 2019: Estimating the Prevalence of Dementia in Europe. Luxembourg: Alzheimer Europe; 2019.
3. El-Hayek YH, Wiley RE, Khoury CP, et al. Tip of the Iceberg: assessing the Global Socioeconomic Costs of Alzheimer's Disease and related dementias and strategic implications for stakeholders. *J Alzheimers Dis.* 2019;70(2):323–341.
4. Hou X-H, Feng L, Zhang C, Cao X-P, Tan L, Yu J-T. Models for predicting risk of dementia: a systematic review. *J Neurol Neurosurg Psychiatry*. 2019;90(4):373–379.
5. Tang EYH, Harrison SL, Errington L, et al. Current developments in dementia risk prediction modelling: an updated systematic review. *PLoS One*. 2015;10(9):e0136181.
6. 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. *Alzheimers Dementia (New York, N Y).* 2019;5:563–569.
7. Ben Miled Z, Haas K, Black CM, et al. Predicting dementia with routine care EMR data. *Artif Intell Med*. 2020;102:101771.
8. Ford E, Greenslade N, Paudyal P, et al. Predicting dementia from primary care records: a systematic review and meta-analysis. *PLoS One.* 2018;13(3):e0194735.
9. 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):e0203246.
10. Albrecht JS, Hanna M, Kim D, Perfetto EM. Predicting diagnosis of Alzheimer's disease and related dementias using administrative claims. *J Manag Care Spec*. 2018;24(11):1138–1145.
11. Jain S, Rosenbaum PR, Reiter JG, et al. Using Medicare claims in identifying Alzheimer's disease and related dementias. *Alzheimers Dementia.* 2020;17(3):515–524.
12. Pohlmann S, Kunz A, Ose D, et al. Digitalizing health services by implementing a personal electronic health record in Germany: qualitative analysis of fundamental prerequisites from the perspective of selected experts. *J Med Internet Res*. 2020;22(1):e15102.
13. Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J, May M, Brindle P. Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: prospective open cohort study. *BMJ*. 2007;335(7611):136.
14. Reitz C, Brayne C, Mayeux R. Epidemiology of Alzheimer disease. *Nat rev Neurol*. 2011;7(3):137–152.
15. Gauthier S, Rosa-Neto P, Morais JA, Webster C. World Alzheimer Report 2021 - Journey Through the Diagnosis of Dementia. London: Alzheimer's Disease International, 2021.
16. Breiman L. Statistical modeling: the two cultures (with comments and a rejoinder by the author). *Statist Sci*. 2001;16(3):199–231.
17. Desai RJ, Wang SV, Vaduganathan M, Evers T, Schneeweiss S. Comparison of machine learning methods with traditional models for use of administrative claims with electronic medical records to predict heart failure outcomes. *JAMA network open*. 2020;3(1):e1918962.
18. Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PLoS One*. 2017;12(4):e0174944.
19. Doblhammer G, Fink A, Fritze T. Short-term trends in dementia prevalence in Germany between the years 2007 and 2009. *Alzheimers Dementia*. 2015;11(3):291–299.
20. Doblhammer G, Barth A. Prevalence of morbidity at extreme old age in germany: an observational study using health claims data. *J Am Geriatr Soc*. 2018;66(7):1262–1268.
21. Livingston G, Huntley J, Sommerlad A, et al. Dementia prevention, intervention, and care: 2020 report of the Lancet Commission. *Lancet North Am Ed*. 2020;396(10248):413–446.
22. Tutz G, Schmid M. *Modeling Discrete Time-to-Event Data*, Springer Ser. Statistics. 1, Cham: Springer International Publishing; 2016:1–247.
23. Kuhn M. Building Predictive Models in R Using the caret Package. *J Stat Soft*. 2008;28(5):1–26.
24. Zhang T, Yu B. Boosting with early stopping: convergence and consistency. *Ann Statist*. 2005;33(4):1538–1579.
25. Wright MN, Ziegler A. Ranger A fast implementation of random forests for high dimensional data in C++ and R. *J Stat Soft*. 2017;77(1): 1–17.
26. Welchowski T, Schmid M. discSurv: Discrete Time Survival Analysis. R package version 1.1.4; 2019. https://cran.r-project.org/web/packages/discSurv/discSurv.pdf
27. Schmid M, Tutz G, Welchowski T. Discrimination measures for discrete time-to-event predictions. *Econom Stat*. 2018;7:153–164.
28. van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. *BMC med*. 2019;17(1):230.
29. Miller ME, Langefeld CD, Tierney WM, Hui SL, McDonald CJ. Validation of probabilistic predictions. Medical decision making an international journal of the Society for. *Med Decis Making*. 1993;13(1):49–58.
30. Breiman L. Random Forests. *Mach Learn*. 2001;45(1):5–32.
31. Fukunishi H, Nishiyama M, Luo Y, Kubo M, Kobayashi Y. Alzheimer-type dementia prediction by sparse logistic regression using claim data. *Comput Methods Programs Biomed*. 2020;196:105582.
32. Gareri P, Segura-García C, Manfredi VGL, et al. Use of atypical antipsychotics in the elderly: a clinical review. *Clin Interv Aging*. 2014;9:1363–1373.
33. Söderlund ME, Guajardo ME, Cavagna M, et al. Prediction of antipsychotics use in patients with mild cognitive impairment and dementia. *Alzheimers Dementia.* 2020;16(S8). https://doi.org/10.1002/alz.046645
34. Dhikav V. Predictors of antipsychotic usage among patients with dementias and mild cognitive impairment. *Alzheimers Dementia*. 2020;16(S7). https://doi.org/10.1002/alz.043398

35. Fink HA, Jutkowitz E, McCarten JR, et al. Pharmacologic interventions to prevent cognitive decline, mild cognitive impairment, and clinical Alzheimer-type dementia: a systematic review. *Ann Intern Med*. 2018;168(1):39–51.
36. Bertens D, Vos S, Kehoe P, Wolf H, et al. Use of mild cognitive impairment and prodromal AD/MCI due to AD in clinical care: a European survey. *Alzheimer's Res Ther*. 2019;11(1):74.
37. Bruscoli M, Lovestone S. Is MCI really just early dementia? A systematic review of conversion studies. *Int Psychogeriatr*. 2004;16(2):129–140.
38. Chini F, Pezzotti P, Orzella L, Borgia P, Guasticchi G. Can we use the pharmacy data to estimate the prevalence of chronic conditions? A comparison of multiple data sources. *BMC public health*. 2011;11:688.
39. Li Y, Li Y, Li X, Zhang S, et al. Head injury as a risk factor for dementia and Alzheimer's disease: a systematic review and meta-analysis of 32 observational studies. *PLoS One*. 2017;12(1):e0169650.
40. Aarsland D, Kurz MW. The epidemiology of dementia associated with Parkinson's disease. *Brain Pathol*. 2010;20(3):633–639.
41. Diniz BS, Butters MA, Albert SM, Dew MA, Reynolds CF. Late-life depression and risk of vascular dementia and Alzheimer's disease: systematic review and meta-analysis of community-based cohort studies. *Br J Psychiatry*. 2013;202(5):329–335.
42. Heser K, Fink A, Reinke C, Wagner M, Doblhammer G. The temporal association between incident late-life depression and incident dementia. *Acta Psychiatr Scand*. 2020;142(5):402–412.
43. Moraros J, Nwankwo C, Patten SB, Mousseau DD. The association of antidepressant drug usage with cognitive impairment or dementia, including Alzheimer disease: a systematic review and meta-analysis. *Depress Anxiety*. 2017;34(3):217–226.
44. Teipel SJ, Fritze T, Ellenrieder M, Haenisch B, Mittelmeier W, Doblhammer G. Association of joint replacement surgery with incident dementia diagnosis in German claims data. *Int Psychogeriatr*. 2018;30(9):1375–1383.
45. Tzeng N-S, Chung C-H, Lin F-H, et al. Risk of dementia in adults with ADHD: a nationwide, population-based cohort study in Taiwan. *J Atten Disord*. 2019;23(9):995–1006.
46. Kim YJ, Tae BS, Bae JH. Cognitive function and urologic medications for lower urinary tract symptoms. *Int Neurourol J*. 2020;24(3):231–240.
47. Hwang T-J, Chen J-J, Lin Y-T, Chan H-Y. Dextromethorphan for the treatment of agitation in dementia: a pilot study. *Alzheimers Dementia*. 2020;16(S9). https://doi.org/10.1002/alz.045791
48. Ancelin M-L, Carrière I, Barberger-Gateau P, et al. Lipid lowering agents, cognitive decline, and dementia: the three-city study. *J Alzheimers Dis*. 2012;30(3):629–637.
49. Nerius M, Haenisch B, Gomm W, Doblhammer G, Schneider A. Glucocorticoid therapy is associated with a lower risk of dementia. *J Alzheimers Dis*. 2020;73(1):175–183.
50. Muzambi R, Bhaskaran K, Brayne C, Davidson JA, Smeeth L, Warren-Gash C. Common bacterial infections and risk of dementia or cognitive decline: a systematic review. *J Alzheimers Dis*. 2020;76(4):1609–1626.
51. Kern DM, Cepeda MS, Lovestone S, Seabrook GR. Aiding the discovery of new treatments for dementia by uncovering unknown benefits of existing medications. *Alzheimers Dementia (New York, N Y)*. 2019;5:862–870.
52. Angelucci F, Cechova K, Amlerova J, Hort J. Antibiotics, gut microbiota, and Alzheimer's disease. *J Neuroinflammation*. 2019;16(1):108.
53. Balducci C, Forloni G. Doxycycline for Alzheimer's disease: fighting β-amyloid oligomers and neuroinflammation. *Front pharmacol*. 2019;10:738.
54. Rao A, Miller B, Kulman T, Pinho P, Aggarwal NT. Novel application of digital dementia phenotyping and risk classification for insurance and longevity risk modeling. *Alzheimers Dementia*. 2020;16(S10). https://doi.org/10.1002/alz.044372
55. Hassaine A, Canoy D, Solares JRA, et al. Learning multimorbidity patterns from electronic health records using non-negative matrix factorisation. *J Biomed Inform*. 2020;112:103606.
56. Epping J, Geyer S, Eberhard S, Tetzlaff J. Völlig unterschiedlich oder doch recht ähnlich? Die soziodemografische Struktur der AOK Niedersachsen im Vergleich zur niedersächsischen und bundesweiten Allgemein- und Erwerbsbevölkerung. *Gesundheitswesen*. 2021;83(S 02):77–86.
57. Nerius M, Fink A, Doblhammer G. Parkinson's disease in Germany: prevalence and incidence based on health claims data. *Acta Neurol Scand*. 2017;136(5):386–392.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Reinke C, Doblhammer G, Schmid M, Welchowski T. Dementia risk predictions from German claims data using methods of machine learning. *Alzheimer's Dement*. 2022;1–10. https://doi.org/10.1002/alz.12663