Original research

DIGITAL HEALTH

DIGITAL HEALTH
Volume 9: 1–23

DOI: 10.1177/20552076231178577
journals.sagepub.com/home/dhj

Sage

Benzodiazepine-related dementia risks and protopathic biases revealed by multiple-kernel learning with electronic medical records

Takashi Hayakawa[1,2], Takuya Nagashima[1,2], Hayato Akimoto[1,2], Kimino Minagawa[2], Yasuo Takahashi[2] and Satoshi Asai[1,2]

Abstract

Objectives: To simultaneously estimate how the risk of incident dementia nonlinearly varies with the administration period and cumulative dose of benzodiazepines, the duration of disorders with an indication for benzodiazepines, and other potential confounders, with the goal of settling the controversy over the role of benzodiazepines in the development of dementia.

Methods: The classical hazard model was extended using the techniques of multiple-kernel learning. Regularised maximum-likelihood estimation, including determination of hyperparameter values with 10-fold cross-validation, bootstrap goodness-of-fit test, and bootstrap estimation of confidence intervals, was applied to cohorts retrospectively extracted from electronic medical records of our university hospitals between 1 November 2004 and 31 July 2020. The analysis was mainly focused on 8160 patients aged 40 or older with new onset of insomnia, affective disorders, or anxiety disorders, who were followed up for 4.10 ± 3.47 years.

Results: Besides previously reported risk associations, we detected significant nonlinear risk variations over 2–4 years attributable to the duration of insomnia and anxiety disorders, and to the administration period of short-acting benzodiazepines. After nonlinear adjustment for potential confounders, we observed no significant risk associations with long-term use of benzodiazepines.

Conclusions: The pattern of the detected nonlinear risk variations suggested reverse causation and confounding. Their putative bias effects over 2–4 years suggested similar biases in previously reported results. These results, together with the lack of significant risk associations with long-term use of benzodiazepines, suggested the need to reconsider previous results and methods for future analysis.

Keywords

Dementia, benzodiazepines, electronic medical records, machine learning, kernel method

Submission date: 21 September 2022; Acceptance date: 6 May 2023

Introduction

Dementia is a currently incurable disorder that leads to a large socioeconomic burden, and thus, the importance of primary prevention of dementia by reducing its risk has been recognised. Among risk factors for dementia, long-term use of benzodiazepines (BZDs) has drawn attention as a modifiable factor in recent years. Numerous studies

[1]Division of Pharmacology, Department of Biomedical Sciences, Nihon University School of Medicine, Tokyo, Japan
[2]Division of Genomic Epidemiology and Clinical Trials, Clinical Trials Research Center, Nihon University School of Medicine, Tokyo, Japan

Corresponding author:
Takashi Hayakawa, Oyaguchi-Kamicho 30-1 Itabashi, Tokyo, 1738610, Japan.
hayakawa.takashi@nihon-u.ac.jp

have assessed the risk of dementia onset associated with the BZD use with increasingly elaborated study designs. In addition to the enrollment of larger numbers of patients to study cohorts, variables for finer risk assessment, such as cumulative doses of short-acting and long-acting BZDs, have been included in the analyses. Nevertheless, there is still inconsistency among the results of recent studies, and the causal role of BZDs in the development of dementia has remained controversial. While more than a half of previous observational studies up to 2020[1–14] from different countries and two meta-analyses[15,16] indicated a positive association between long-term BZD use and dementia risk, a recent meta-analysis and three recent large studies reported lack of statistical significance of the association.[13,14,17,18] Furthermore, the interpretation of previous results is particularly difficult because of potential bias due to reverse causation (protopathic bias) between dementia onset and prodromal symptoms with an indication for BZD use, such as insomnia. One of the recent large studies examined this bias by stratifying the studied population, and reported no significant risk.[18]

A technical limitation of these studies is their reliance on linear statistical analysis. In a linear statistical analysis, typically the logarithm of the hazard rate of dementia onset or odds ratio between treatment groups is assumed to depend on binary options such as current use of BZDs, or to grow linearly in proportion to the values of continuous variables such as cumulative dose of BZDs. Drug effects are, however, usually nonlinear, as represented by a typical drug's dose–response curve which saturates in the high-dose range. Bias effects that should be adjusted are also nonlinear, as represented by the increased risk of disease onset shortly, rather than long after the onset of its prodromal symptoms. Recent studies[11,19] partially resolved this issue by introducing a spline function that described nonlinear dependence of the hazard rate on the cumulative dose of BZDs, namely, a *'nonlinear risk function'*, and showing a positive risk associated with only low cumulative doses, which they discussed as protopathic bias. Although this approach potentially allows us to probe the nature of risks and biases associated with BZDs with greatly improved precision, it still suffers limitation. In the analysis of nonlinear effects of BZDs, one naturally asks fundamental questions, such as, whether dementia onset is best explained by a nonlinear function of the cumulative dose of BZDs, or is better explained by a nonlinear function of a confounding factor or a source of protopathic bias, such as the duration of the disorders for which BZDs may be prescribed. Or one may ask how the risk of dementia onset is nonlinearly associated with BZD use, if nonlinear protopathic biases are adjusted. To answer such a question, one needs to use multiple nonlinear risk functions in the statistical analysis. The use of multiple nonlinear risk functions, however, makes the problem harder, because it incurs larger statistical errors and raises another issue concerning the choice of the variables for which nonlinear dependency is considered. Previous approaches were severely limited in this regard.

In the present study, we addressed the above issue, using a framework developed for machine learning. In machine learning, the use of multiple nonlinear functions in statistical analysis and selection of a small number of relevant functions among many candidates have been studied intensively. *'Multiple kernel learning' (MKL)*[20] is an established framework for this problem that gives us a theoretically guaranteed bound on the statistical error incurred by the estimated functions. In this framework, we can automatically select a small number of relevant nonlinear functions from many candidates based on a relatively small dataset, where the statistical error grows only logarithmically with respect to the number of candidates.[21–24] Despite the availability of this established framework, thus far, only several pioneering studies have applied multiple kernel learning to clinical data,[25–29] and those studies did not perform a systematic analysis whose results can be directly compared with previous epidemiological results. In the present study, developing fast program codes that perform standard epidemiological procedures with models constructed with multiple kernels, and applying these to retrospective cohorts extracted from a large volume of anonymised electronic medical records, we evaluated the risk of dementia onset associated with BZDs in detail.

As we used only nonlinear risk functions for cumulative doses of BZDs and related drugs, we obtained a result similar to the previous result.[11] Further analysis with more candidate nonlinear functions, however, revealed more prominent, temporally varying risk over a rather long time associated with the administration period of short-acting BZDs and the durations of insomnia and anxiety disorders, while we detected no significant risk associated with long-term use of BZDs in the same analysis. These results suggested that analyses of dementia risk associated with BZD use performed in previous studies were likely to have been affected by confounding and reverse causation originating from the use of short-acting BZDs and prodromal symptoms, and that the causal role of long-term use of BZDs in the development of dementia is questionable.

Methods

Study design, data source and enrollment of patients

The present study was a population-based retrospective cohort study performed at the Department of Pharmacology and the Clinical Trial Research Centre of the School of Medicine, Nihon University between 1 February 2022 and 31 August 2022. Retrospective cohorts were constructed from Nihon University School of Medicine's Clinical Data Warehouse (NUSM's CDW), a centralised, anonymised

database that records demographic information, diagnoses, prescribed medications and laboratory data of patients who attended three hospitals affiliated with Nihon University between 1 November 2004 and 31 July 2020. Patients of our university hospitals were typically, but not always, local patients who were referred by a local clinician. Unlike patients of large hospitals in many other countries, a substantial proportion of these patients continued to be regularly and holistically cared for over a long period of time. We have observed previously reported clinical effects of drugs as well as unattended effects in over 10 previous analyses with this database (for reference, see the most recent ones[30–32]), and we therefore expected generalisable knowledge to be drawn from it.

As shown in Figure 1A, we extracted a study cohort from the electronic medical records of patients aged 40 years or older who had been regularly cared for over more than 2 years in our university hospitals, with intervals between prescriptions of less than 120 days. For each patient, the index date was defined as the later of the 180th day after the first visit and the 40th birthday. Patients who already had a record of diagnosed dementia (ICD10 code F00-03, G30, G31.0 and G31.8), a record of prescription of anti-dementia drugs, or a record of disorders that severely affect cognitive function (see Appendix 1), before the index date, were excluded. Then, the remaining patients were included in the study at the index date, and followed up until the patient stopped attending our hospitals regularly or was diagnosed with dementia. The date of diagnosis of dementia was determined from the record of diagnosis of dementia, with or without recorded prescription of anti-dementia drugs. If the dates of the diagnosis and prescription were different, the earlier of them was used.

For this cohort of patients, however, precise information about their exposure to BZDs was unavailable, because BZDs might have already been prescribed before the first visit. Thus, as we illustrated in Figure 1A, we further focused on a subcohort of patients who were newly diagnosed with a disorder for which BZDs are often prescribed, namely, insomnia (ICD10 code G47), affective disorders (ICD10 code F30–39) or anxiety disorders (ICD10 code F40–49). For this subcohort, we redefined the index date as the earlier of the date of the first diagnosis of these disorders and the date of the first chronic prescription of BZDs, Z-drugs or other hypnotics for over 14 days. Here, we included affective disorders, even though affective disorders are not a direct indication for BZD use, because these disorders are known to be a complication of dementia, and BZDs are often prescribed for associated anxiety symptoms. Since BZDs are unlikely to be used chronically for other disorders, we reasonably assumed that the patients in the subcohort had never used BZDs before the first visit, and hence, that precise information about their exposure to BZDs was available. It should be noted that we excluded patients with epilepsy, who might have used BZDs long term, at the time of the index date.

For both the cohort and subcohort defined above, the risk of dementia of Alzheimer type was assessed separately from that of dementia of any type. In the former case, the diagnosis of dementia of Alzheimer type was identified from the record of diagnosis with ICD10 code F00 or G30, and from the absence of diagnosis of dementia of any other type and Parkinson disease in any part of the patient record. In the analysis of dementia of Alzheimer type, the onset of Parkinson disease or dementia that did not satisfy the above diagnostic criteria was treated as censoring.

Design of statistical models for analysis

For the assessment of dementia risk, we used four statistical models. While these models determined the risk of dementia onset based on different sets of explanatory variables, they shared the same basic design. As illustrated in Figure 1B, explanatory variables for each model represented, for instance, doses and durations of prescribed medications, current and past results of laboratory tests, and presence and durations of disorders, and their values were obtained from the database for each time point of the observation period. Then, the model determined the logarithm of the hazard rate, namely, the instantaneous occurrence rate of dementia onset, for each patient and time point, as the sum of linear and nonlinear functions of one or two of these explanatory variables termed *'risk functions'*, whose functional forms were to be statistically estimated. Each of these estimated risk functions was associated with a function called *'kernel'* that acted as building blocks for the estimated risk function. In the present study, the kernels were chosen to be either a linear function or a Gaussian function. Although we refer readers interested in the technical details of the kernel method to the literature,[33] we note that a linear kernel leads to only a linear risk function, and that a Gaussian kernel leads to a nonlinear risk function of arbitrary functional form with a certain degree of smoothness. Thus, analysis based on our multiple-kernel model with only linear kernels reduced to a conventional linear analysis, while incorporating Gaussian kernels into this linear setting naturally extended the analysis to a nonlinear one. A precise, mathematical description of the above model settings is provided in Appendix 3.

Based on the above basic design, we defined four statistical models for analysis with increasing numbers of explanatory variables and risk functions. The smallest model, which we called *'linear model'*, was designed to be the counterpart for those employed in classical epidemiological studies. This model included 37 explanatory variables, which were either variables for demographic information, manually selected variables indicating the presence or absence of previously known risk factors, a variable indicating the serum level of low-density lipoprotein (LDL) cholesterol or variables indicating current

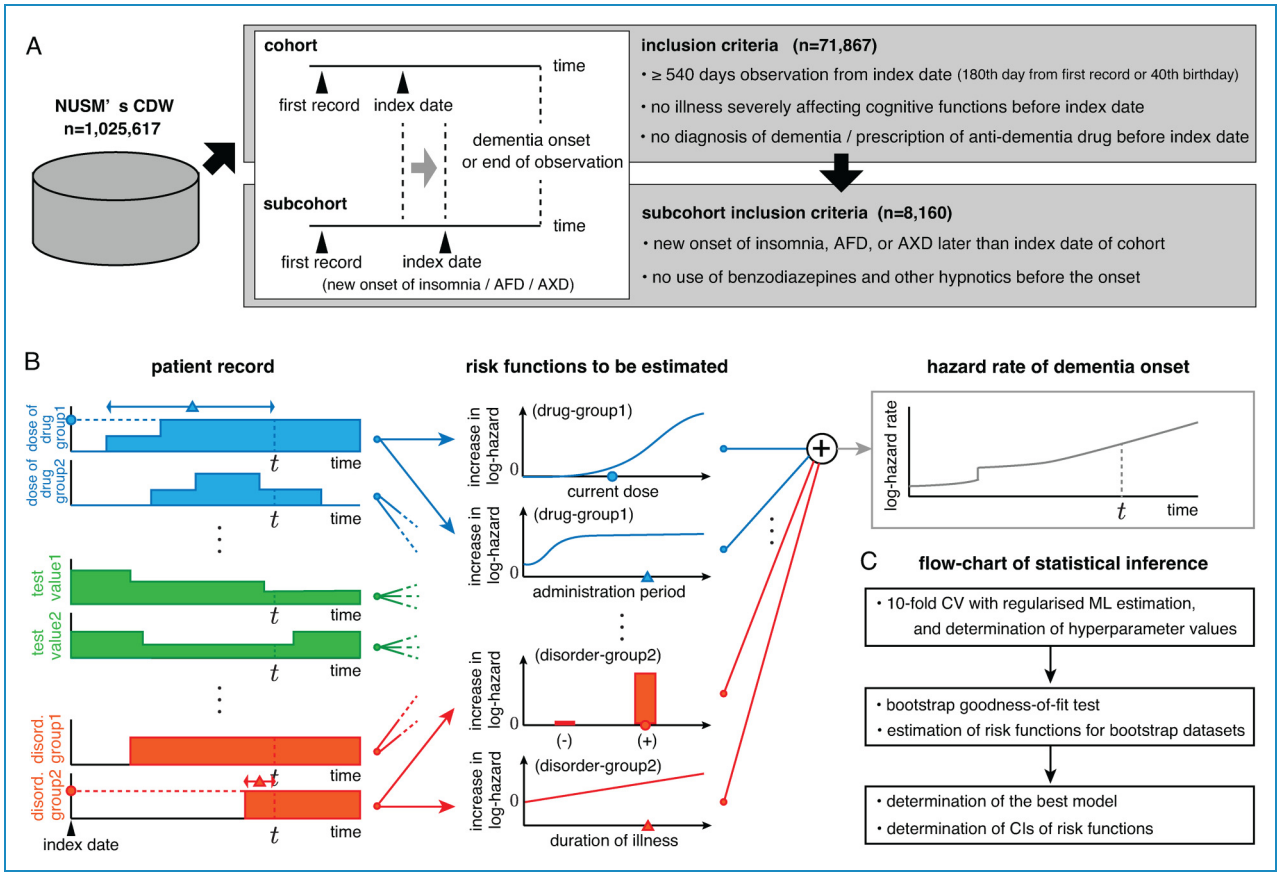

Figure 1. Study cohorts, design of statistical model for analysis and flowchart of inference. (A) Flowchart of the extraction of study cohorts from NUSM's CDW, a centralised database for anonymised electronic medical records of hospitals affiliated with Nihon University. In addition to basic demographic information of patients, information about drug prescriptions, laboratory test results and diagnoses of disorders were stored in the database. (B) Design of statistical model used for analysis. As illustrated in the panel, the logarithm of the hazard rate of dementia onset was modelled as the sum of linear and nonlinear risk functions of explanatory variables that represented, for instance, doses and administration periods of groups of drugs, values of laboratory test results and presence and durations of groups of disorders. The forms of the risk functions were statistically estimated, using the framework of multiple-kernel learning. (C) Flowchart of concrete procedures for statistical inference. AFD: affective disorder; AXD: anxiety disorder; CV: cross-validation; ML: maximum-likelihood; CI: confidence interval; NUSM's CDW: Nihon University School of Medicine's Clinical Data Warehouse.

doses or ever (or non-) use of BZDs, Z-drugs and other hypnotics. The demographic information included age, age at entry, sex, date and body-mass index. The previously known risk factors included past history of cerebrovascular disease, ischaemic heart disease, Parkinson disease, hypertension, dyslipidaemia, epilepsy, affective disorders, anxiety disorders, diabetes mellitus and insomnia, and current use of muscarinic cholinergic antagonists with two different burden levels, anti-hypertensive agents, statins, anti-diabetic drugs, anti-platelet drugs, anti-coagulant drugs, anti-depressants and anti-Parkinson-disease drugs. As we defined variables for BZDs, we grouped BZDs into four categories whose member drugs had more or less distinct biological half-lives from those of the other categories [Table 1]. Hereafter, we called these BZDs, very-short-acting, short-acting, intermediate-acting and long-acting BZDs, respectively. Although we used more categories than the three categories used in previous studies, we did not expect this finer categorisation to seriously affect the detection power of the statistical analysis. Because most patients had used short-acting or long-acting BZDs [Table 1], the detection of significant risks associated with these two categories of drugs was a statistical problem of comparable scale to the one dealt with in previous studies, while we expected risks associated with the other two categories to be harder to detect. From this point of view, since the number of patients who had used hypnotics other than BZDs and Z-drugs was also smaller, we expected less significant results for this category as well.

As well as the explanatory variables and risk functions included in the linear model, the *'MKL1' model* included nonlinear risk functions of cumulative doses of BZDs, Z-drugs and other hypnotics, being expected to serve as the counterpart for the model in the previous study.[11] To

Table 1. Defined daily doses (DDs) and biological half-lives of active metabolites of BZDs and other hypnotics authorised by the Pharmaceutical and Medical Devices Agency of Japan (summary).

Drug	Category ($T_{1/2}$ [hour])	DD [mg]	DD (ATC) [mg]	Cohort	Subcohort
Triazolam	Very short (2.9)	0.375	0.25	3478	583
Flutazolam	Very short (3.5)	12	–	14	1
Clotiazepam	Short (6.3)	22.5	–	1584	295
Etizolam	Short (6.3)	3	–	8589	1748
Brotizolam	Short (7)	0.25	0.25	8156	1531
Lormetazepam	Intermediate (10)	1.5	1	174	18
Rilmazafone hydrochloride hydrate	Intermediate (11)	1.5	–	1147	180
Lorazepam	Intermediate (12)	2	2.5	1550	198
Alprazolam	Intermediate (14)	1.2	1	2024	302
Bromazepam	Long (20)	10.5	10	1042	102
Fludiazepam	Long (23)	0.75	0.75	0	0
Estazolam	Long (24)	2.5	3	2472	284
Flunitrazepam	Long (24)	1.25	1	0	0
Flurazepam hydrochloride	Long (24)	20	30	10	0
Nitrazepam	Long (25–27)	7.5	5	5487	702
Clonazepam	Long (27)	4	8	1391	298
Clobazam	Long (30)	20	20	31	7
Quazepam	Long (37)	20	15	573	58
Haloxazolam	Long (-)	7.5	–	29	0
Chlordiazepoxide	long (-)	40	30	80	12
Oxazolam	Long (-)	45	–	47	1
Mexazolam	Long (-)	2.25	–	26	0
Clorazepate dipotassium	Long (-)	19.5	20	0	0
Diazepam	Long (-)	10.5	10	2404	425
Cloxazolam	Long (-)	7.5	–	255	19
Medazepam	Long (-)	20	20	76	13
Ethyl loflazepate	Long (122)	2	2	2164	366

(continued)

Table 1. Continued.

Drug	Category ($T_{1/2}$ [hour])	DD [mg]	DD (ATC) [mg]	Cohort	Subcohort
Flutoprazepam	Long (190)	3	–	72	5
Zolpidem tartrate	Z-drug (2.1)	7.5	10	11,765	2513
Zopiclone	Z-drug (3.7)	8.75	7.5	4058	545
Eszopiclone	Z-drug (5.1)	2	2	3475	684
Ramelteon	Other (2.46)	1	–	1537	320
Suvorexant	Other (10)	20	–	634	122
Lemborexant	Other (31–56)	5	–	0	0

For comparison with other studies, DDs authorised by the anatomical therapeutic chemical (ATC) classification of drugs[36] are also shown, if available. In the two right columns, the numbers of patients in the cohort and subcohort who used the drugs are summarised.

investigate the possibility that other variables related to BZD use better account for dementia onset, we included in the *'MKL2' model*, nonlinear functions of administration periods and current doses of BZDs, Z-drugs and other hypnotics, nonlinear functions of the durations of insomnia, affective disorders and anxiety disorders, and a nonlinear function of the administration period of anti-depressant. The largest model, *'MKL3' model*, included as many variables and nonlinear risk functions available from the database as possible.

For all of the four models, each explanatory variable was clipped to the narrowest range of values in which values of the variable except for the largest and smallest 1% potential outliers fell. Then, we standardised these clipped explanatory variables to have mean of zero and unit variance, and used them in further analysis. Further details of the explanatory variables and risk functions (kernels) included in each model are provided in Appendices 2, 4 and 5 and Table 4.

Statistical analysis

To estimate the risk functions of the statistical models described above, we used the regularised maximum-likelihood approach. In this approach, we computed risk functions that maximised the sum of the log-likelihood of dementia onset in the dataset and a regulariser, the latter of which controlled the complexity of the risk functions. MKL was originally formulated as a model with constraints equivalent to a 1-norm regulariser[34] that allowed only a small number of candidate risk functions to take non-zero values, while forcing others to be zero. This sparsifying regulariser was appropriate for our purpose, because our purpose was to discern a small number of the most relevant explanatory variables that nonlinearly account for dementia onset. Previous statistical theories[21–24] guaranteed that MKL with a 1-norm regulariser is likely to recover such a set of explanatory variables with a dataset of sufficient size, which is known as support consistency. Although estimation consistency, namely, the convergence of risk functions to the optimal ones in the infinite-sample-size limit, was shown for both 1-norm and 2-norm regularisers, estimation with a 1-norm regulariser was shown to control statistical errors more effectively than estimation with a 2-norm regulariser in a problem with a small sample size and a large number of explanatory variables and kernels.[24] In the present study, a few tens to a few hundreds of kernels were used for analysis of a dataset containing only 184 dementia patients and 85 Alzheimer-type dementia patients, and thus, estimation with a 1-norm regulariser was expected to be more suitable. However, because the number of kernels above which the 1-norm regulariser outperformed the 2-norm regulariser was not known in advance, we adopted both the 1-norm and 2-norm regularisers and compared their results. Although previous theories showed more favourable statistical properties for mixtures of 1-norm and 2-norm regularisers and adaptive 1-norm or non-convex regularisers than for simple 1-norm and 2-norm regularisers in certain settings, the use of these regularisers markedly increases the computational cost for estimation and tuning of hyperparameters. Thus, we restricted our analysis to only cases with a simple 1-norm or 2-norm regulariser.

Although the maximum-likelihood approach has been recognised as a straightforward approach in medical statistics,[35] most previous epidemiological studies used classical or generalised Cox models that estimate risks by maximising a partial likelihood. The advantage of employing a Cox model lies in its properties that an arbitrary nonlinear function of time can be implicitly used as a baseline hazard, and that focusing on relative risks of patients only at the times of event occurrence reduces computational complexity. The present study, however, contrasted with previous epidemiological studies in that precise values of explanatory

variables for each time point throughout the observation period could be found in our database. This precise temporal information is not utilised by Cox models. Furthermore, as MKL allowed us to estimate multiple nonlinear functions, the implicit use of a single nonlinear baseline hazard in the Cox models did not merit the loss of temporal information. For these reasons, we employed the more straightforward maximum-likelihood approach.

The concrete procedures in the regularised maximum-likelihood approach are summarised in the flowchart in Figure 1C and described in detail in the following sections. We repeatedly followed these procedures for each of the combinations of the statistical model, the type of target dementia and the choice of cohort.

Determination of values of hyperparameters based on 10-fold cross-validation. Our models had three hyperparameters whose values needed to be determined empirically. Two of these hyperparameters represented the strengths of the regularisers for risk functions constructed with linear and Gaussian kernels, respectively. The other hyperparameter represented the bandwidth of the Gaussian kernels for standardised explanatory variables, which controlled smoothness of the constructed risk functions. For determination of the values of the hyperparameters and further evaluation, we followed the standard 10-fold cross-validation (CV) procedure. Patients were randomly partitioned into fold 0–9. For each $i \in \{0 \dots 9\}$, fold i was used for testing, fold $i + 1$ (mod 10) was used for validation, and the remaining folds were used for estimation (training). For each of the different sets of values for the three hyperparameters in a grid search, the sum of the log-likelihoods of the validation data across the 10 folds were computed as CV score, using the estimated risk functions, and the values of the hyperparameters giving the best score were selected for use in further analysis. With the selected values of the hyperparameters, the sum of the log-likelihoods of the test data across the 10 folds was computed as a test score.

Bootstrap test for test scores of two models and determination of best model. We determined the best model by systematically comparing the test scores described above. For comparison of the test scores of two models, we performed a bootstrap goodness-of-fit test. In this test, we repeatedly calculated the test scores of a tested model (larger model with a higher test score) and a null model (smaller model with a lower test score) for bootstrap datasets generated from the hazard rate estimated with the null model. Concretely, we probabilistically generated fictitious dementia onsets for all patients based on the hazard rate estimated with the null model, and prepared 10 bootstrap datasets. With each of these bootstrap datasets, we repeatedly performed 10-fold CV, using the same assignment of patients to the 10 folds as that used for the tuning of the hyperparameters. Then, comparing the distribution of the difference of the bootstrap test scores of the two models, we calculated p-value of the difference of the test scores for the original dataset. Since repeated CV is computationally demanding, we calculated the p-value, approximating the distribution of the differences of the bootstrap test scores with a Gaussian distribution. This Gaussian approximation was justified by a Shapiro-Wilk test.

The above goodness-of-fit tests have not been performed conventionally in the application of machine learning to medical data. However, such an evaluation of statistical errors in test scores is needed, in order to discuss the results of machine learning in comparison with those of conventional epidemiological studies. In machine learning studies, statistical tests are often performed to compare two sets of 10 likelihood scores for 10 test folds between the two compared models. This approach is sometimes convenient, but is not theoretically grounded. It was expected to have low power of detection of significance for our dataset and hence to be unsuitable, because the sizes of the partitioned test sets were too small as there were only 184 dementia patients in the analysed dataset, and the variance of the observation period and the number of dementia onsets among these test sets was relatively large. In conventional epidemiological studies, goodness-of-fit of two models is compared, using the likelihood ratio test or its bootstrap version. Our procedure described above is a straightforward extension of the bootstrap likelihood ratio test to regularised maximum-likelihood estimation with 10-fold CV, and is therefore suitable for our purpose.

Determination of confidence intervals of risk functions with bootstrap datasets. To evaluate the statistical error in the estimation of risk functions, we repeatedly performed the estimation using the values of the hyperparameters determined in the 10-fold CV, and with 500 bootstrap datasets generated by randomly resampling the same number of patients as in the original dataset. From the 500 estimates for each risk function, we calculated its 95% confidence intervals at different values of the argument variable.

Data and code availability

Most of the results obtained in the present study were summarised in Supplemental Figures 1 to 8. We have also attached the program codes used for the present study and a document describing how to use them. These will also be registered to the online code repository, Github, upon acceptance by the journal.

The original data used in the present study are not publicly available due to privacy and security concerns, especially because the detailed history of prescribed medication, laboratory tests and diagnoses in the input dataset for our program codes for analysis potentially allows identification of patients, even though patient identifiers were anonymised. Researchers who wish to perform

further analysis on our dataset should contact the corresponding author. If the requested investigation abides by pertinent guidelines in terms of privacy, security and scientific merit, and is approved by the ethics committee of our university, access to the dataset will be provided.

Results

Characterisation of the cohorts

The characteristics of the extracted cohorts are summarised in Table 2. The size of the entire cohort was moderately large and around one third of the sizes of the previous largest study populations,[14,18] and the size of the subcohort that focused on putative new users of BZDs was roughly the same as that of the previous study that performed nonlinear analysis of dementia risk of BZD use,[11] although the previous study did not focus on new users. The proportions of patients with comorbidities with known risks were comparable to those in previous studies, regardless of the difference that our study population was taken from medical records of university hospitals. These previously reported risk factors were found more frequently among patients who developed dementia during the observation period. The length of observation in our study was moderately long compared to previous studies, and, as a notable feature of our dataset, detailed history of prescriptions, diagnoses, and laboratory tests throughout the observation periods of the patients was available. We therefore expected more temporally precise information about dementia risk to be drawn from the data, within the limits determined by the length of observation.

Goodness-of-fit of models

The goodness-of-fit of the models is shown in Figure 2. From the results for the entire cohort [Figure 2 A and C], we confirmed that the linear model fitted the data significantly better than did a simple *'base model'* that predicted dementia onset based only on the date, age, age at entry, body-mass index and sex. The predictive performance of the linear model under two different types of regularisers showed almost no difference. From the results for the subcohort [Figure 2 B and D], we found that, regardless of the type of regulariser and target dementia, MKL2 significantly outperformed the linear and MKL1 models, while the performance of MKL1 was consistently poorer than that of the linear model. Concerning MKL3, we found that its predictive performance under 1-norm regularisation was much better than those of the other models in the risk estimation for any dementia onset, while it was poorer for the onset of dementia of Alzheimer type. Comparing the two types of regularisation, we found that 2-norm regularisation generally brought out better performance from the linear, MKL1 and MKL2 models than did 1-norm regularisation in the risk estimation for any dementia onset, while performances with the two regularisers showed the opposite for the onset of dementia of Alzheimer type. As expected from statistical theories,[21–24] we observed that the performance of the MKL3 model with a large number of risk functions was much poorer under 2-norm regularisation than under 1-norm regularisation.

Hazard ratios of use of BZDs, Z-drugs and other hypnotics estimated with linear model

For comparison with previous studies, we examined the hazard ratios of the current doses and ever use of BZDs with different half-lives, Z-drugs, and other hypnotics, and the hazard ratios of other previously known risk factors estimated with the linear model. From the estimated hazard ratios and their confidence intervals, we found that most of the previously known risk factors other than BZDs were associated with significant positive risks [Supplemental Figure 2]. In contrast, the observed patterns of risks associated with hypnotics shown in Figure 3 were neither consistent with previous results nor immediately interpretable (see Supplemental Figure 1 for results with 1-norm regularisation). In this result, we did not observe statistically significant, positive risks associated with use of BZDs, except for a positive risk of onset of any dementia associated with ever use of short-acting BZDs in the analysis of the subcohort [Figure 3B]. While the confidence intervals of the hazard ratios estimated with the entire cohort tended to be narrower than those estimated with the subcohort, we found that negative risk associated with Z-drugs and positive risk associated with short-acting BZDs were significant only in the analysis of the subcohort. Since the entire cohort was almost ten times larger than the subcohort, and since statistically significant risks are hence generally harder to detect for the subcohort, these results were extraordinary. The risks associated with these hypnotics were therefore suggested to be amplified by the design of the subcohort containing many new users of hypnotics.

Nonlinear risks associated with cumulative doses and administration periods of BZDs and durations of baseline disorders

Next, for comparison with the previously reported nonlinear risks of BZDs, we examined the nonlinear risk functions for the cumulative doses of BZDs estimated with the MKL1 model. Although the estimated risk functions shown in Figure 4A were only for short-acting BZDs and were drawn over a wider range of cumulative doses, 0–1000 defined daily dose (DD), than for the previous results which were for all BZDs and Z-drugs over 0–400 DD, the two showed good agreement over the shared range, both taking positive values for 0–180 DD and negative values for larger doses. We note that the negative peak of

Table 2. Characteristics of the populations in the cohort and subcohort are summarised in the table.

Characteristics	Dem(C)	DAlz(C)	All(C)	Dem(SC)	DAlz(SC)	All(SC)
	1955	885	71,867	184	85	8160
Female	1040 (52.1)	478 (54.0)	32,797 (45.6)	110 (59.8)	55 (64.7)	4380 (53.7)
Age at ID [year]	73.6±7.8	74.6±7.03	64.3±11.4	76.3±7.7	77.3±6.6	67.4±11.2
BMI	21.9±3.8	22.2±3.69	22.9±3.9	21.3±3.9	22.0±4.1	22.7±3.8
Observation [year]	7.69±3.80	4.95±3.40	6.86±4.03	5.81±2.33	3.31±2.56	4.10±3.47
Comorbidity						
Insomnia [ID]	190 (9.7)	99 (11.1)	5396 (7.5)	85 (46.2)	38 (44.7)	4607 (56.5)
Insomnia [EO]	702 (35.9)	309 (34.9)	17584 (24.5)	116 (63.0)	51 (60)	5436 (66.6)
Affective disorder [ID]	80 (4.1)	31 (3.5)	1738 (2.4)	36 (19.6)	20 (23.5)	1041 (12.8)
Affective disorders [EO]	349 (17.9)	144 (16.3)	5714 (8.0)	53 (28.8)	27 (31.8)	1385 (17.0)
Anxiety disorders [ID]	119 (6.1)	52 (5.9)	3276 (4.6)	70 (38.0)	33 (38.8)	2984 (36.6)
Anxiety disorders [EO]	538 (27.5)	230 (26.0)	11083 (15.4)	89 (48.4)	45 (52.9)	3531 (43.3)
Depression [ID]	77 (3.9)	31 (3.5)	1651 (2.3)	36 (19.6)	20 (23.5)	982 (12.0)
Depression [EO]	329 (16.8)	139 (15.7)	5375 (7.5)	51 (27.7)	26 (30.6)	1309 (16.0)
Treated DM [ID]	232 (11.9)	100 (11.3)	7714 (10.7)	27 (14.7)	13 (15.3)	1226 (15.0)
Treated DM [EO]	355 (18.2)	156 (17.6)	12311 (17.1)	33 (17.9)	15 (17.6)	1489 (18.2)
Treated DL [ID]	387 (19.8)	174 (19.7)	10690 (14.9)	61 (33.2)	27 (31.8)	2113 (25.9)
Treated DL [EO]	632 (32.3)	280 (31.6)	19911 (27.7)	78 (42.4)	36 (42.4)	2793 (34.2)
Treated HTN [ID]	807 (41.3)	387 (43.7)	23723 (33.0)	120 (65.2)	58 (68.2)	4348 (53.3)
Treated HTN [EO]	1223 (62.6)	549 (64.2)	36057 (50.2)	141 (76.6)	67 (78.8)	5224 (64.0)
CVD [ID]	196 (10.0)	103 (12.0)	3677 (5.1)	46 (25.0)	25 (29.4)	1163 (14.3)
CVD [EO]	869 (44.5)	417 (48.8)	12655 (17.6)	93 (50.5)	45 (52.9)	1872 (22.9)
IHD [ID]	189 (9.7)	87 (10.2)	5867 (8.2)	45 (24.5)	24 (28.2)	1429 (17.5)
IHD [EO]	577 (29.5)	253 (29.6)	16162 (22.5)	61 (33.2)	34 (40.0)	2056 (25.2)
PD [ID]	104 (5.3)	0 (0.0)	582 (0.7)	20 (10.9)	0 (0.0)	139 (1.5)
PD [EO]	225 (11.5)	0 (0.0)	1501 (1.8)	29 (15.8)	0 (0.0)	226 (2.5)
Epilepsy [ID]	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)

(continued)

Table 2. Continued.

Characteristics	Dem(C)	DAlz(C)	All(C)	Dem(SC)	DAlz(SC)	All(SC)
Epilepsy [EO]	163 (8.3)	59 (6.7)	2561 (3.6)	12 (6.5)	6 (7.1)	184 (2.3)
Prescription profile						
BZD user [ID]	537 (27.5)	232 (27.1)	11395 (15.9)	87 (47.3)	34 (40.0)	3388 (41.5)
BZD user [EO]	951 (48.6)	480 (54.2)	22348 (31.1)	110 (59.8)	45 (52.9)	4313 (52.9)
Per. of BZD use [year]	2.75±3.01	2.70±2.93	2.75±3.34	1.52±1.71	2.00±2.40	1.54±2.01
TDD of BZDs	1397±4887	1222±4897	1093±5242	331±642	431±716	346±1413
Z-drug user [ID]	191 (9.6)	81 (9.2)	6069 (8.4)	39 (21.2)	21 (24.7)	2356 (28.9)
Z-drug user [EO]	580 (29.7)	232 (26.2)	16,264 (22.6)	68 (37.0)	32 (37.6)	3259
Per. of Z-drug use [year]	1.60±2.18	1.65±2.05	1.70±2.22	0.92±0.89	1.06±1.00	1.29 ±1.62
TDD of Z-drugs	365±798	366±738	290±742	121±284	130±358	193±473
OH users [ID]	2 (0.1)	2 (0.2)	247 (0.3)	12 (6.5)	1 (1.2)	168
OH users [EO]	110 (5.5)	34 (3.8)	1991 (2.8)	12 (6.5)	3 (3.5)	399
Per. of OH use [year]	0.63±0.68	0.93±0.89	1.01±1.17	0.49±0.47	0.77±0.41	0.80±1.09
TDD of OHs	140±213	200±272	210±484	83±142	248±202	171±399

Number of patients (and its percentage) or mean value ± standard deviation for the characteristics in the left column are shown. The characteristics of the patients who developed any dementia and of those who developed dementia of Alzheimer type are shown in separate columns. Dem: (Any) Dementia; DAlz: Dementia of Alzheimer type; C: cohort; SC: subcohort; ID: at index date; EO: at the end of observation period; BMI: body-mass index; DM: diabetes mellitus; DL: dyslipidaemia; HTN: hypertension; CVD: cerebrovascular diseases; IHD: ischaemic heart diseases; PD: Parkinson disease; TDD: total defined daily dose; per.: period; OH: other hypnotics.

the risk function for the onset of any dementia in Figure 4A was statistically significant, while that of the previous result was not. Apart from the above similarity, the estimated risk functions for BZDs with other ranges of half-lives did not show significant risk-dose relationships [Supplemental Figure 4]. These results suggested that the previously reported risk nonlinearly correlated with the cumulative dose of all BZDs and Z-drugs was mainly due to shorter-acting hypnotics.

Here, it should be mentioned that we occasionally found a discrepancy between the results of 1-norm and 2-norm regularisations, apart from the clear results described above. In such cases, we typically observed small, but significant, risk variations under 2-norm regularisation, and zero risk variations with a high frequency under 1-norm regularisation. Although there is no way to judge with certainty which observation is more reliable, the following difference in the properties of 1-norm and 2-norm regularisations should be noted: if a confounder with no influence on the outcome is strongly correlated with a risk factor, and if there is no hidden risk factor correlated with both, use of a 1-norm regulariser is likely to result in estimation of a zero risk for the confounder, while use of a 2-norm regulariser possibly results in estimation of a significant risk due to the colinear effect. Based on this consideration, we conservatively dismissed small significance under 2-norm regularisation that had not been confirmed under 1-norm regularisation.

With confirmation of the previous result, we moved on to examination of the risk functions estimated with the MKL2 model, in order to answer the question we raised in the introduction of this article, identifying which of the current and cumulative doses and administration periods of BZDs and the durations of baseline disorders is most relevant to the onset of dementia. The goodness-of-fit of the linear, MKL1 and MKL2 models under 1-norm and 2-norm regularisations in Figure 2 B and D indicated that the risk functions included only in MKL2 contributed more to the prediction of dementia onset than those for cumulative doses of BZDs. Among those risk functions, the risk functions for the administration period of short-acting BZDs and the duration of insomnia and anxiety disorders [Figure 4B] were estimated to be associated with

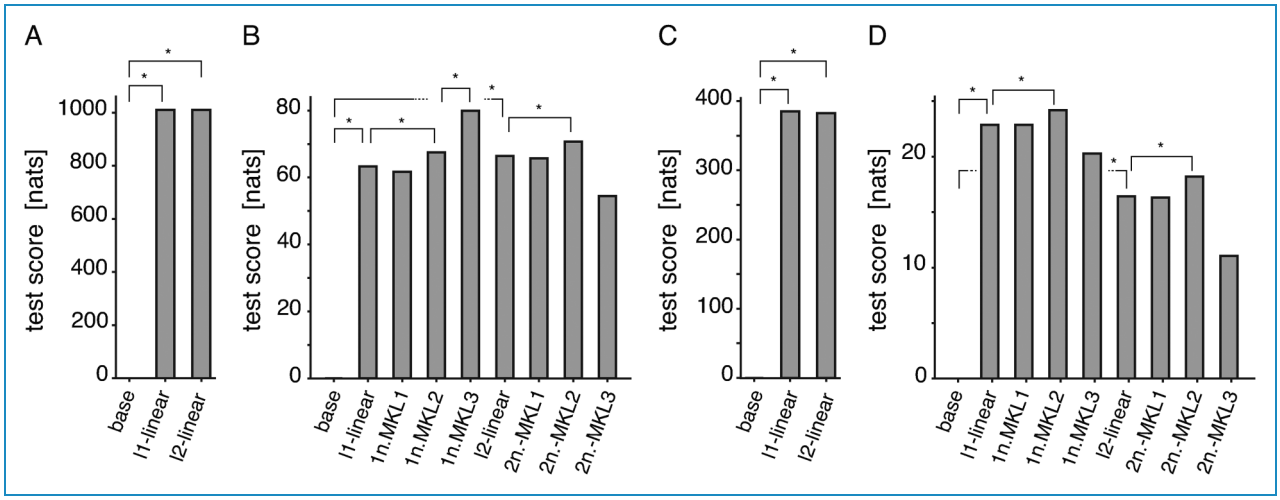

Figure 2. Test scores of the linear and MKL1–3 models under 1-norm and 2-norm regularisations. The sum of the log-likelihoods of test sets across 10 folds as a measure of goodness-of-fit was calculated for both the entire cohort (panels A and C) and subcohort (panels B and D), and for both the onset of any dementia (panels A and B) and the onset of dementia of Alzheimer type (panels C and D). $\ell 1/2$: regularisation by $\ell 1/2$ norm, $1/2n.$: regularisation with $1/2$ norm. Evaluating the significance of the differences of the test scores for pairs of models by performing bootstrap tests, we obtained the following p-values: (panel A) base versus $\ell 1$-linear: $<1.0 \times 10^{-16}$, base versus $\ell 2$-linear: $<1.0 \times 10^{-16}$, (panel B) base versus $\ell 1$-linear: $<1.0 \times 10^{-16}$, $\ell 1$-linear versus 1n.MKL2: 5.7×10^{-14}, 1n.MKL2 versus 1n.MKL3: $<1.0 \times 10^{-16}$, base versus $\ell 2$-linear: $<1.0 \times 10^{-16}$, $\ell 2$-linear versus 2n.MKL2: 6.8×10^{-5}, (panel C) base versus $\ell 1$-linear: $<1.0 \times 10^{-16}$, base versus $\ell 2$-linear: $<1.0 \times 10^{-16}$, (panel D) base versus $\ell 1$-linear: $<1.0 \times 10^{-16}$, $\ell 1$-linear versus 1n.MKL2: 4.6×10^{-6}, base versus $\ell 2$-linear: $<1.0 \times 10^{-16}$, $\ell 2$-linear versus 2n.MKL2: 9.8×10^{-4}. The Gaussian approximations used in the bootstrap goodness-of-fit tests were justified by the following p-values of Shapiro-Wilk tests: (panel A) base versus $\ell 1$-linear: 0.986, base versus $\ell 2$-linear: 0.847, (panel B) base versus $\ell 1$-linear: 0.428, $\ell 1$-linear versus 1n.MKL2: 0.148, 1n.MKL2 versus 1n.MKL3: 0.666, base versus $\ell 2$-linear: 0.457, $\ell 2$-linear versus 2n.MKL2: 0.198, (panel C) base versus $\ell 1$-linear: 0.177, base versus $\ell 2$-linear: 0.587, (panel D) base versus $\ell 1$-linear: 0.177, $\ell 1$-linear versus 1n.MKL2: 0.138, base versus $\ell 2$-linear: 0.955, $\ell 2$-linear versus 2n.MKL2: 0.519.

significant temporal variations in the risk of onset of any dementia. Particularly, the estimated risk functions for the administration period of short-acting BZDs and the duration of anxiety disorders showed prominent, diphasic temporal variations with positive and negative peaks. Less prominently, the estimated risk function for the duration of insomnia showed a significant negative peak around the 40th month. Notably, the estimated risk functions for the cumulative dose of short-acting BZDs showed profiles similar to those obtained with the MKL1 model, but with reduced significance indicated by wider confidence intervals estimated under 2-norm regularisation and, more remarkably, by a lower percentage of estimating non-zero risk functions under 1-norm regularisation (from 99% to 28%). These results suggested the possibility that the previously reported risk associated with the cumulative dose of BZDs was partly due to the effects caused by confounding between the cumulative dose and the administration period. Also, notably, we did not observe a significant positive risk linearly associated with the current dose of short-acting BZDs in the results with the MKL2 model, as we observed with the linear model [Supplemental Figure 3]. This suggested that some part of the positive risk observed with the linear model is explained by the time-varying risks we described above (as shown by comparison of Figure 3B and Supplemental Figure 1B with Supplemental Figure 3A and C). On the other hand, the positive risk associated with affective disorders and negative risks associated with the current doses of very-short-acting BZDs and Z-drugs persisted in the results with the MKL2 model [Supplemental Figures 3 and 7], the former of which did not appear to be largely adjusted by the simultaneous estimation of the non-significant risk variation associated with the administration period of anti-depressants [Supplemental Figure 8]. In contrast to all these results, we detected no significant increase in risk associated with long-term BZD use from the risk functions of administration periods and cumulative doses of BZDs [Supplemental Figures 5 and 6].

The results of analysis for the onset of dementia of Alzheimer type showed qualitatively similar results to those for the onset of any dementia, but with reduced statistical significance. This was partly expected from the small number of patients with both BZD use and disease onset. Nevertheless, the risk functions estimated with the MKL2 model showed diphasic temporal profiles similar to those for the onset of any dementia, as seen in the risk functions for short-acting BZDs [Figure 4C], while the diphasic risk variation seemed overlayed with another temporal risk variation that increases over time. Concerning this overlayed risk variation, it might be worth mentioning that, although

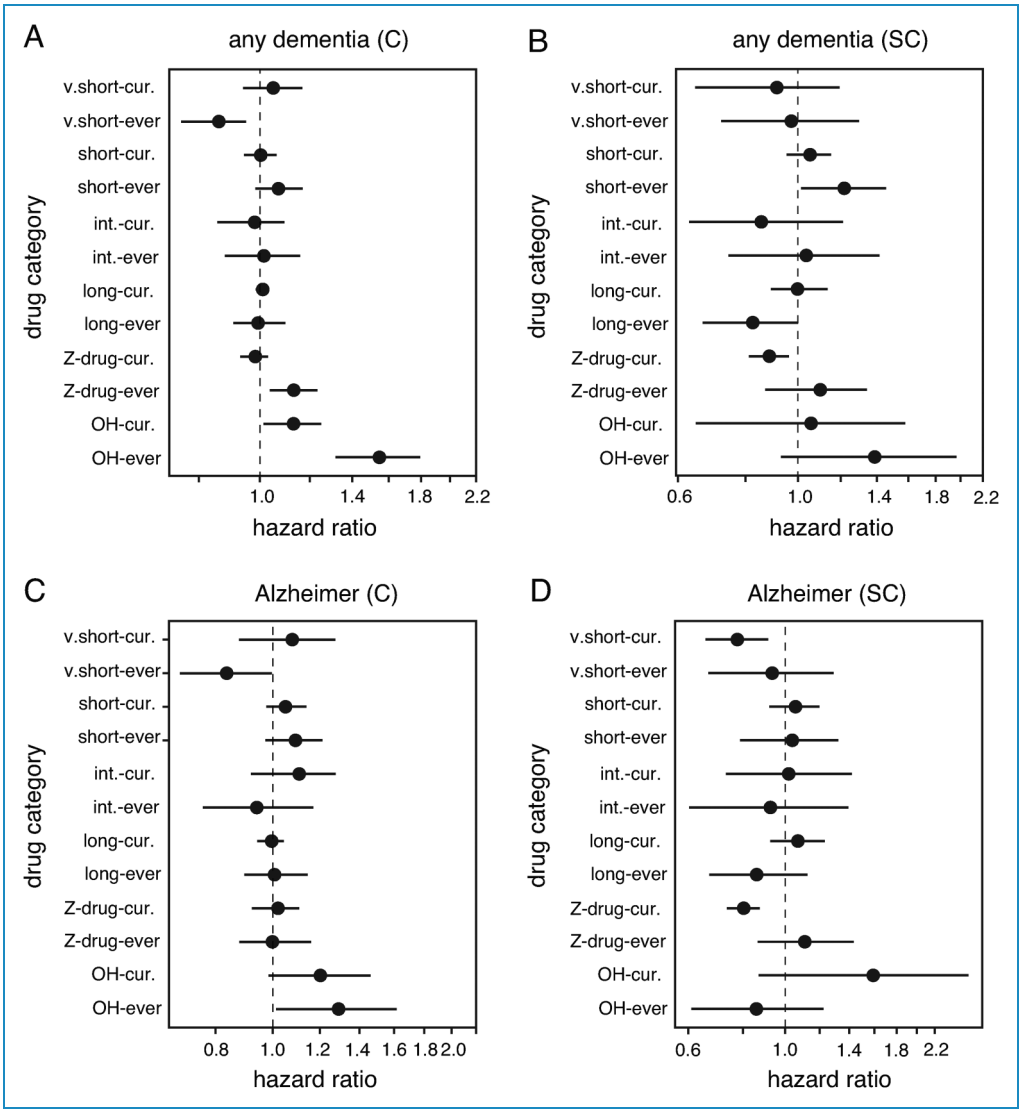

Figure 3. Hazard ratios and 95% confidence intervals for current doses and ever use of hypnotics estimated with linear model. The hazard ratios for the onset of any dementia (panels A and B) and for the onset of dementia of Alzheimer type (panels C and D), estimated with the linear model under $\ell 2$-norm regularisation, and with the entire cohort (panels A and C) or the subcohort (panels B and D) are shown. For continuous explanatory variables, hazard ratios for a unit difference in standardised variables with zero mean and unit variance are shown. C: the entire cohort; SC: the subcohort; v.short: very-short-acting BZDs; short: short-acting BZDs; int.: intermediate-acting BZDs; long: long-acting BZDs; OH: other hypnotics; -cur.: current dose; -ever: ever use.

we again did not observe a significant increase in risk associated with long-term use of BZDs, the risk functions for the administration periods of short-acting and long-acting BZDs, together with the risks of these drugs linearly estimated in the same analysis, indicated a slight, but nonsignificant increase in risk over time [Figure 4C, Supplemental Figures 3B and D, 5 and 6].

MKL with hundreds of explanatory variables

As we explained above, the results of the analysis with the MKL2 model revealed a few factors that accounted for onset of dementia better than the cumulative doses of BZDs. Then, we naturally considered whether we could find an even more informative set of explanatory variables from the hundreds of candidates available in our database. We investigated this point by performing analysis with the MKL3 model. As shown in Figure 2, we found that the MKL3 model accounted for the onset of any dementia much better than did the MKL2 model, while this was not the case for the onset of dementia of Alzheimer type. We, however, found that the estimated risk function for the administration period of short-acting BZDs and the duration of anxiety disorders in Figure 4D showed the same, almost

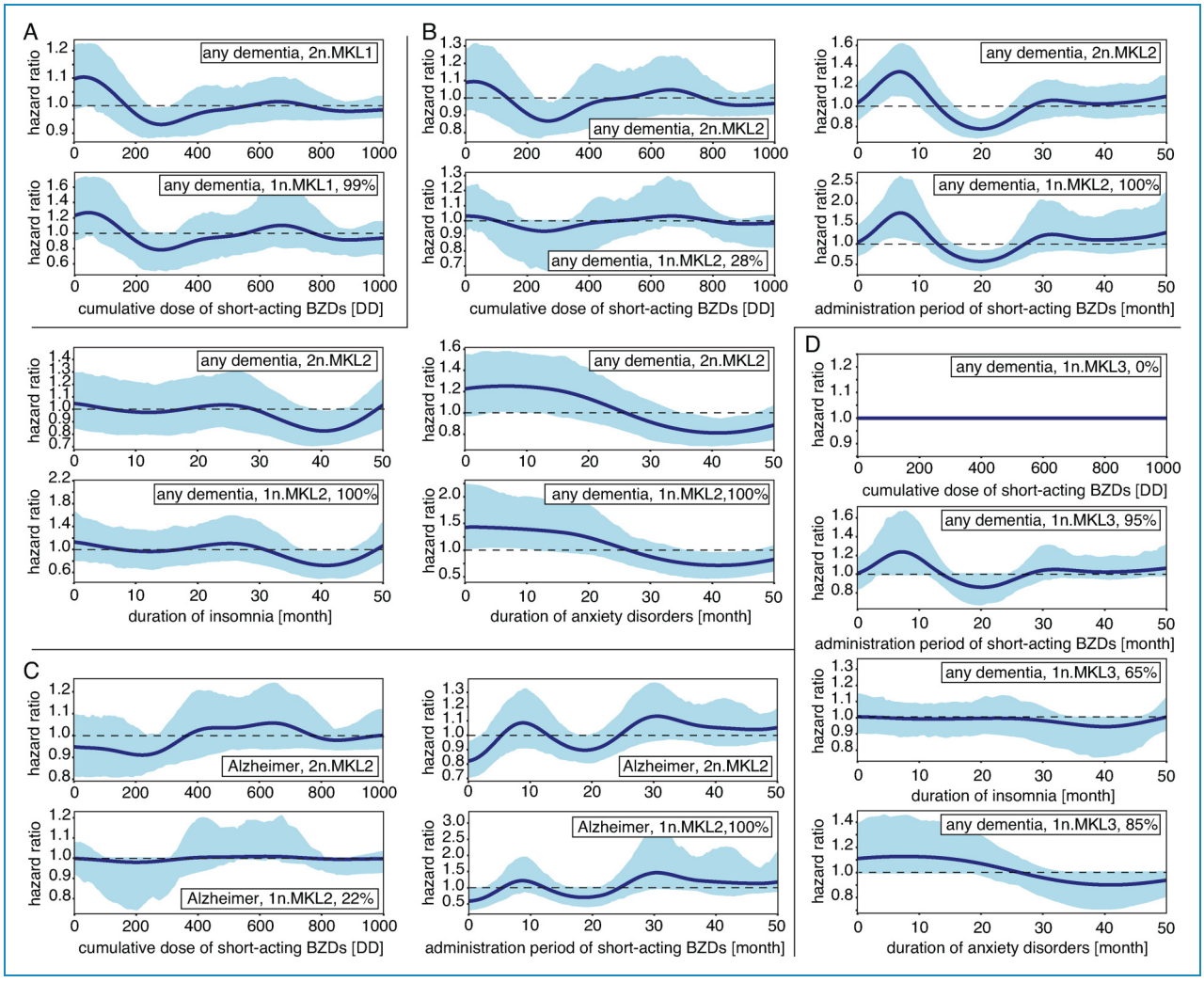

Figure 4. Nonlinear risk functions estimated with MKL1–MKL3 models. Insets in panels indicate the type of target dementia (any dementia or dementia of Alzheimer type) and the type of regularisation used in the analysis. For the results with 1-norm regularisation, the percentage of estimating non-zero risk functions in the bootstrap procedure is also shown with its subdecimal value rounded. The labels of the horizontal axes indicate the explanatory variables for which the risk functions are shown in the panels. (A) Risk functions for onset of any dementia estimated with MKL1 model. (B) Risk functions for onset of any dementia estimated with MKL2 model. (C) Risk functions for onset of dementia of Alzheimer type estimated with MKL2 model. (D) Risk functions for onset of any dementia estimated with MKL3 model. 1/2n.: 1/2-norm regularisation; DD: defined daily dose.

significant, diphasic risk variations as those estimated with the MKL2 model, confirming the robustness of the result. Most of the other nonzero risks detected with the MKL3 model under 1-norm regularisation were either variables related to risk factors investigated with the MKL2 model, to less frequent causes of cognitive impairment that were not investigated with the MKL2 model, or to systemic conditions of patients, as summarised in Table 3. The lack of improved prediction by the MKL3 model for onset of dementia of Alzheimer type was understandable from these results, since patients with secondary cognitive impairment related to many of the detected factors were presumably not diagnosed with dementia of Alzheimer type. Concerning the negative risks associated with current doses of very-short-acting BZDs and Z-drugs estimated with the linear and MKL2 models, we could not speculate on the sources of these risks from the results with the MKL3 model, because the optimal regularisation hyperparameter value for linear kernels was such a large value that most linear risks were estimated to be zero. This suggested that 1-norm regularisation sacrificed the estimation of less contributing, but significant risk factors in predicting dementia onset, as we had expected as its behaviour with data of relatively small size.

Discussion

In the present study, we estimated the risks of dementia onset in retrospective cohorts extracted from the electronic

Table 3. Nonzero risk functions detected with MKL3 model.

Explanatory variable	Kernel type	Percentage
Cerebrovascular diseases (I60–69, G45–46)	D1	100
Antipsychotics (N05A)	P3	99.2
Short-acting BZDs	P1	94.8
HDL-cholesterol	T2	90.2
Body-mass index	-	90.2
Choline esterase	T3	89.4
Head injury	D1	89.0
Potassium	T3	88.8
Antipsychotics (N05A)	P1	88.0
LDL-cholesterol	T3	87.8
Antiinflammatory and antirheumatic products, non-steroids (M01A)	P1	86.0
Anxiety disorders	D1	85.8
Calcium	T3	84.0
Free T4	T3	82.6
Vitamin B12	T3	82.0
Thyroid stimulating hormone	T3	80.2
Total cholesterol	T3	79.6
Ischaemic heart diseases	D1	76.8
LDL-cholesterol/HDL-cholesterol ratio	T2	76.2
Diabetes mellitus (E10–14, G59.0, G63.2, H28.0, H36.0, M14.2, N08.3)	D1	75.2
Anti-cholinergic drugs	P1	74.8
Hypertension	D1	74.2
Mean corpuscular volume of red blood cells	T3	72.4
Anti-depressant	P1	72.0
Age	-	71.8
Serum uric acid	T3	69.4
Total bilirubin	T3	69.4
Dyslipidaemia (E78)	D1	68.4
Blood haemoglobin	T2	67.4
Vitamin B1, plain and in combination with B6 and B12 (A11D)	P1	66.8
Anti-thrombotic agents (B01A)	P1	66.8
Anti-hypertensive agents	P4	66.2
Insomnia (G47)	D1	65.4
Free T3	T3	65.0
Lactate dehydrogenase	T3	64.2
Total protein	T3	63.6
Cardiac stimulants excluding cardiac glycosides	P1	63.4
Anti-Parkinson drugs (N04B)	P1	63.2

Risk functions estimated to be nonzero with higher frequencies in the repeated estimation with bootstrap datasets and the MKL3 model under 1-norm regularisation are listed in descending order of the percentage of nonzero estimation. Linear and Gaussian kernels were not distinguished in these statistics.

medical records of our university hospitals, using a hazard model extended using the techniques of multiple kernel learning. Our statistical model coincided with the conventionally used linear hazard model, as we used only linear kernels, while it described hazard models with multiple nonlinear risk functions varying with different explanatory variables, as we used Gaussian kernels. Among 71,680 regularly cared for patients in the cohort, we focused detailed analyses on a subcohort of 8160 patients who had newly developed insomnia, affective disorders or anxiety disorders during the observation period.

Interpretation of results

Our results clearly showed that the administration period of short-acting-BZDs and the duration of anxiety disorders, among other variables related to BZD use and potential sources of bias, diphasically affected the risk of dementia onset in the medium-term of 2–4 years. Less prominently, it was also shown that insomnia was associated with a

Table 4. Kernels used for the linear and MKL1–3 models.

Category of explanatory var.	Kernel type	Linear	MKL1	MKL2	MKL3
Age	L/G	+	+	+	+
Date, age at entry, BMI	L	+	+	+	+
Date, age at entry, BMI	G	–	–	–	+
Sex	L	+	+	+	+
X-acting BZDs, Z-drugs, OH	P3/5-L	+	+	+	+
X-acting BZDs, Z-drugs, OH	P2-G	–	+	+	+
X-acting BZDs, Z-drugs, OH	P1-G	–	–	+	+
X-acting BZDs, Z-drugs, OH	P1/2-L, P3-G	–	–	–	+
Anti-depressants	P3/P5-L	+	+	+	+
Anti-depressants	P1-G	–	–	+	+
Anti-depressants	P1/4-L, P4-G	–	–	–	+
Other known risk-drug groups	P3/5-L	+	+	+	+
Other known risk-drug groups	P1/4-L/G	–	–	–	+
All other drug groups	P1/4-L/G, P3/5-L	–	–	–	+
Disorder groups with BZDs indication	D3-L	+	+	+	+
Disorder groups with BZDs indication	D1-G	–	–	+	+
Disorder groups with BZDs indication	D1-L, D2-L/G	–	–	–	+
Other known risk-disorder groups	D3-L	+	+	+	+
Other known risk-disorder groups	D1/2-L/G	–	–	–	+
All other disorder groups	D3-L, D1/2-L/G	–	–	–	+
LDL-Chol.	T1-L	+	+	+	+
LDL-Chol.	T1/2/3-G, T4-L	–	–	–	+
All other laboratory tests	T1/4-L, T1/2/3-G	–	–	–	+

The symbols + and – indicate the use and non-use of the kernel indicated in the left column, respectively. The details of the listed categories of explanatory variables are given in Appendix 5. The kernel types are described using the P1–P5, T1–T4 and D1–D3 types defined in Appendix 4. X-acting BZD: X is replaced by either very short, short, intermediate, or long, OH: other hypnotics; LDL-Chol.: low-density-lipoprotein cholesterol; L: (modified) linear kernel; G: (modified) Gaussian kernel; BMI: body-mass index.

negative risk around the 40th month from its onset, and that affective disorders were associated with an increased risk regardless of their duration. These findings contrasted with the absence of any significant risk associated with long-term use of BZDs. As we considered the complicated overlap among users of different types of BZDs and patients suffering different baseline disorders, such a clear dissection of their effects was striking. The most valid interpretation of the diphasic risk variations would be that the short-acting BZDs acted as a loading test that detected subclinical

dementia patients, and that anxiety disorders of some patients appeared as a prodromal symptom of dementia. This interpretation accounted for the negative phase of the diphasic risk variation, indicating that patients who did not develop dementia during the first positive phase were more likely to be free from subclinical dementia. The risk associated with the duration of insomnia could not be interpreted straightforwardly. It is conceivable, however, that the increased risk in the early phase after the onset of insomnia was overridden by the earlier increase in risk due to use of short-acting BZDs. This hypothesis was supported by the periods of the risk variations associated with insomnia and short-acting BZDs, together with the fact that onset of insomnia and initiation of short-acting BZDs were often simultaneous. Since insomnia is known to complicate dementia in its early phase, we reasonably suspect the presence of diphasic risk variation similar to that associated with anxiety disorders. Whether affective disorders are prodromal symptoms or risk factors for dementia is controversial.[37] Although our results did not allow us to infer the mechanisms behind the risk associations with affective disorders, the lack of detection of temporal risk variations is consistent with both hypotheses, if the prodromal onset of an affective disorder occurs on a fairly long time scale before the onset of dementia.

Protopathic bias in previous studies expected from our results

A clear message from our results is that a large, positive or negative, protopathic bias can be introduced by variables related to short-acting BZDs, insomnia or anxiety disorders, if only linear effects of these variables are considered, as most previous studies did. This could be seen within our results that the current dose of short-acting BZDs was linearly associated with a significant positive risk in the analysis of the subcohort with the linear model, but not in the analysis of the entire cohort, and that this positive risk disappeared after nonlinear adjustment with the MKL2 model. Previous studies with many new initiators of BZDs would have also suffered this bias, but other studies would not have been free from this bias either. A retrospective study from Taiwan[6] investigated new onset of dementia within three years from the onset of insomnia and prescription of mainly short-acting BZDs. A bias is expected from the temporally varying risks associated with short-acting BZDs and insomnia that we observed. Of note, this bias can be positive, because the cumulative effects over the 3 years, but not the risk at the 36th month, should have been reflected in the previous result. The PAQUID study[7] was likely to be affected by the effects of both insomnia and anxiety disorders, because of its new-initiator design and lack of records of anxiety disorders. Since a recent Danish study[14] enrolled patients with newly developed affective disorders, their cohort was expected to contain many new BZD initiators. For such a study population, our results indicated that their design of a cohort study and nested case-control study with a two-year latency period was susceptible to the protopathic biases we observed. The potential protopathic bias was not pointed out for all of the previous studies. A case-control study based on the RAMQ database[8] reported a positive risk with a study design that would have reduced the bias effect by having an interval of 5 years between exposure assessment and dementia onset. The Caerphilly prospective study,[5] which compared the effects of exposure within 12 years of the outcome with exposure between 13 and 22 years prior to the outcome, would not be susceptible to the bias we observed. For these studies, we just note that we cannot exclude the possibility that the duration of affective disorders is associated with diphasic risk variation on a longer time scale, which could not be assessed in the present study because of our short observation period, and bias due to this temporal risk variation was introduced into their results.

The issue of protopathic bias was extensively investigated in a previous study[18] with the classical stratification approach. The results of this study and the current study cannot be directly compared, because the former analysed the association of BZD use more than 5 years and up to 20 years prior to dementia onset. From a technical point of view, however, our analysis contrasted with the repeated analysis with different stratifications and covariate measurements in the previous study, which made interpretation of the results complicated. In the previous study, simultaneous stratification with respect to all potential sources of bias was impossible, and thus, precise evaluation of their effects was not possible. In contrast, the approach we took here allowed us to directly estimate the effect of each source of bias.

Strengths and limitations

The strength of our study is the availability of temporally precise information about the administration of BZDs and other drugs, baseline disorders and laboratory tests throughout a moderately long observation period, accompanied by computationally efficient analytical tools based on multiple kernel learning and a regularised maximum-likelihood approach that fully takes advantage of this temporal information. As we compare our results with the previous result[11] obtained from a cohort of similar size, the strength of our study can be seen in the estimated, multiple nonlinear risk functions with narrower confidence intervals. The strength of our dataset and methodology, however, does not lie only in quantitative improvement. In fact, the diphasic protopathic biases and their likely influence on the interpretation of the previous results we discussed above demonstrate the need for such a method to disentangle multiple nonlinear risks. The approach to stratify the cohort

employed in the previous study[18] was not as effective as ours, because of the combinatorial issue with several confounders and sources of protopathic bias.

As an epidemiological study, however, our study suffered several limitations in addition to the retrospective study design. First, there was no guarantee that the complete clinical information of the patients was stored in the database. Although many patients are holistically and regularly cared for in our university hospitals, some patients in our cohort would have been cared for by other medical institutions as well, and the medical records of such institutions were not accessible in the present study. In addition to this, our analysis inevitably suffered incompleteness of laboratory test results, because laboratory tests are performed only for patients who need them. Second, dementia recorded in our system was not necessarily diagnosed by a specialist in neurology or psychiatry. Scores of a cognitive scale were not available, unlike some of the previous studies. Third, the present dataset allowed us to take only a 180-day run-in period, which means that bias due to medical history before the first visits might be larger in comparison with previous studies with longer run-in periods. Finally, it is fair to mention that the positive risks associated with hypnotics other than BZDs and Z-drugs, and the negative risks associated with very-short-acting BZDs and Z-drugs persisted even after the introduction of nonlinear risk functions for use of hypnotics and baseline disorders in the MKL2 model, and were hard to interpret. These previously unreported risks might suggest bias inherent in our study cohorts. However, we also point out the possibility that regularised maximum-likelihood estimation detected such a risk and bias more sensitively than did unregularised estimation in previous studies. We reasonably expect modern statistical techniques with an improved signal-to-noise ratio to inevitably detect more risks that cannot be immediately interpreted. To improve this point, we need to analyse different study populations with the same method, distinguishing universal effects from population-specific effects. With all these limitations in mind, we believe that the implication of the clear results obtained with our dataset is still worth consideration in designing future studies and making clinical decisions based on currently available literature.

From the point of view of machine learning, one may consider the application of other popular machine learning algorithms, such as algorithms based on neural networks or decision trees, that potentially predict the onset of events with higher precision. This line of study should be encouraged, but will still face technical and conceptual challenges. The first challenge is to resolve the issue that the results of estimation with ensembles of decision trees and neural networks are still hard to interpret. Although there have been attempts to develop a method with interpretable results,[38] further technical development is warranted before standard epidemiological practice is established. The second challenge is to overcome the computational intractability of these methods. Although a single estimation with these methods can be performed within a reasonable computation time, performing hundreds to thousands of repeated estimations with bootstrap datasets is still prohibitive, especially for a large amount of data, such as the 431,698 person-month data for our subcohort. We believe that such a bootstrap-based evaluation of statistical errors in goodness-of-fit scores and individual risk functions is needed, if we respect the standard of rigour maintained in conventional epidemiology. In this regard, noting that recent applications of neural networks and decision trees to medical data inevitably avoided such intensive statistical evaluation,[39] we point out the gap between the practice in recent machine-learning-based studies and conventional epidemiological practice. The third challenge is to acquire generalisable knowledge from a dataset with a small effective sample size. Only 184 patients in our subcohort developed dementia. Neural networks and decision tree ensembles are generally applied to datasets with a much larger effective sample size, and their predictive performance is believed to be poor if the sample size is small.[40] For these reasons, we did not attempt to perform analysis with popular algorithms based on neural networks and decision trees in the present study, restricting our development to the fast, large-scale, yet rigorous, statistical evaluation provided by the framework of multiple kernel learning. Previous epidemiological analyses based on multiple kernel learning[25–29] were focused on prediction of outcomes, and did not achieve this development.

Conclusions

In our analysis, the risk of dementia onset showed strong nonlinear associations with the durations of insomnia and anxiety disorders and the administration period of short-acting BZDs. Based on the pattern of risk variations, these associations were suggested to be due to protopathic biases and confounding. These putative confounding effects were observed over 2–4 years after the initiation of the drugs and the onset of the disorders, which suggested biases in many previously reported results. Under adjustment for these confounding factors, no significant risk association with long-term use of benzodiazepines was observed. Based on these results, we concluded that the role of BZDs in the development of dementia is questionable. The retrospective and observational nature of the present study, together with the strong limitations discussed above, prevent us reaching a firm conclusion on any clinical implications from our results. Thus, we propose reconsideration of previous results and a similar analysis of datasets in future and past studies. Even though observational studies cannot ultimately conclude a causal role of drugs, adjustment of the nonlinear bias effects we observed might resolve the inconsistency among the results of previous

studies. In this future perspective, Bayesian MKL based on Gaussian processes[41] might be even more suitable. Although there are a few technical challenges in carrying out estimation with Bayesian MKL, prodromal symptoms and subclinical dementia can be explicitly modelled by using latent variables in this Bayesian framework.

Acknowledgements: We would like to thank 4DIN Ltd. (Tokyo, Japan) and Phenogen Medical Corporation (Tokyo, Japan) for financial support. Both companies had no conflicts of interest and played no part in research design, analysis, data collection, interpretation of data, or review of the manuscript, and no honoraria or payments were made for authorship.

Contributorship: TH conceptualised the study, determined the methods to use, and wrote program codes that implement them. Then, TH and TN extracted data for analysis from the database, and all authors performed data analysis. All authors were involved in the discussion and interpretation of the results. TH wrote the first draught of the manuscript. All authors reviewed and edited the manuscript and approved the final version of the manuscript.

Declaration of conflicting interests: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Ethical approval: The experimental protocol was approved by the Ethical Committee of Nihon University School of Medicine (approval number: P30-2-3), and the study was conducted in compliance with the Ethical Guidelines for Medical and Health Research Involving Human Subjects of the Ministry of Education, Culture, Sports, Science, and Technology and the Ministry of Health, Labour, and Welfare of Japan. Written informed consent for participation was not required for this study in accordance with the national legislation. According to the guidelines, a retrospective observational study using only preexisting, anonymised information, not biological samples, cannot and does not need to obtain informed consent.

Funding: The author(s) received no financial support for the research, authorship and/or publication of this article: This research was supported by the Ministry of Education, Culture, Sports, Sciences and Technology (MEXT) of Japanese government and the Japan Agency for Medical Research and Development (AMED) under grant numbers JP18km0605001 and JP223fa627011.

Guarantor: TH

ORCID iD: Takashi Hayakawa https://orcid.org/0000-0002-8565-3486

Supplemental material: Supplemental material for this article is available online.

References

1. Lagnaoui R, Bégaud B, Moore N et al. Benzodiazepine use and risk of dementia: a nested case–control study. *J Clin Epidemiol* 2002; 55: 314–318.
2. Lagnaoui R, Tournier M, Moride Y et al. The risk of cognitive impairment in older community-dwelling women after benzodiazepine use. *Age Ageing* 2009; 38: 226–228.
3. Wu CS, Wang SC, Chang IS et al. The association between dementia and long-term use of benzodiazepine in the elderly: nested case–control study using claims data. *Am J Geriatr Psychiatry* 2009; 17: 614–620.
4. Wu CS, Ting TT, Wang SC et al. Effect of benzodiazepine discontinuation on dementia risk. *Am J Geriatr Psychiatry* 2011; 19: 151–159.
5. Gallacher J, Elwood P, Pickering J et al. Benzodiazepine use and risk of dementia: evidence from the caerphilly prospective study (caps). *J Epidemiol Community Health* 2012; 66: 869–873.
6. Chen PL, Lee WJ, Sun WZ et al. Risk of dementia in patients with insomnia and long-term use of hypnotics: a population-based retrospective cohort study. *PLoS ONE* 2012; 7: e49113.
7. de Gage SB, Bégaud B, Bazin F et al. Benzodiazepine use and risk of dementia: prospective population based study. *BMJ* 2012; 345: e6231.
8. de Gage SB, Moride Y, Ducruet T et al. Benzodiazepine use and risk of Alzheimer's disease: case-control study. *BMJ* 2014; 349: g5205.
9. Imfeld P, Bodmer M, Jick SS et al. Benzodiazepine use and risk of developing Alzheimer's disease or vascular dementia: a case-control analysis. *Drug Saf* 2015; 38: 909–919.
10. Shash D, Kurth T, Bertrand M et al. Benzodiazepine, psychotropic medication, and dementia: A population-based cohort study. *Alzheimers Dement* 2016; 12: 604–613.
11. Gray SL, Dublin S, Yu O et al. Benzodiazepine use and risk of incident dementia or cognitive decline: prospective population based study. *BMJ* 2016; 352: i90.
12. Biétry FA, Pfeil AM, Reich O et al. Benzodiazepine use and risk of developing Alzheimer's disease: a case-control study based on swiss claims data. *CNS Drugs* 2017; 31: 245–251.
13. Nafti M, Sirois C, Kröger E et al. Is benzodiazepine use associated with the risk of dementia and cognitive impairment—not dementia in older persons? The Canadian study of health and aging. *Ann Pharmacother* 2020; 54: 219–225.
14. Osler M and Jørgensen MB. Associations of benzodiazepines, z-drugs, and other anxiolytics with subsequent dementia in patients with affective disorders: a nationwide cohort and nested case-control study. *Am J Psychiat* 2020; 177: 497–505.
15. Zhong G, Wang Y, Zhang Y et al. Association between benzodiazepine use and dementia: a meta-analysis. *PLoS ONE* 2015; 10: 1–16.
16. Islam MM, Iqbal U, Walther B et al. Benzodiazepine use and risk of dementia in the elderly population: a systematic review and meta-analysis. *Neuroepidemiology* 2016; 47: 181–191.
17. Lucchetta RC, da Mata BPM and Mastroianni PdC. Association between development of dementia and use of benzodiazepines: a systematic review and meta-analysis. *Pharmacothera: J Human Pharmacol Drug Ther* 2018; 38: 1010–1020.
18. Richardson K, Mattishent K, Loke YK et al. History of benzodiazepine prescriptions and risk of dementia: possible bias due

to prevalent users and covariate measurement timing in a nested case-control study. *Am J Epidemiol* 2019; 188: 1228–1236.
19. Goerdten J, Carrière I and Muniz-Terrera G. Comparison of Cox proportional hazards regression and generalized cox regression models applied in dementia risk prediction. *Alzheimer's Dement: Transl Res Clin Intervent* 2020; 6: e12041.
20. Lanckriet GR, Cristianini N, Bartlett P et al. Learning the kernel matrix with semidefinite programming. *J Mach Learn Res* 2004; 5: 27–72.
21. Bach FR. Consistency of the group Lasso and multiple kernel learning. *J Mach Learn Res* 2008; 9: 1179–1225.
22. Meier L, Van de Geer S and Bühlmann P. High-dimensional additive modeling. *Ann Stat* 2009; 37: 3779–3821.
23. Koltchinskii V and Yuan M. Sparsity in multiple kernel learning. *Ann Stat* 2010; 38: 3660–3695.
24. Suzuki T and Sugiyama M. Fast learning rate of multiple kernel learning: trade-off between sparsity and smoothness. *Ann Stat* 2013; 41: 1381–1405.
25. Daemen A, Valentin L, Fruscio R et al. Improving the preoperative classification of adnexal masses as benign or malignant by second-stage tests. *Ultrasound Obstet Gynecol* 2011; 37: 100–106.
26. Daemen A, Timmerman D, Van den Bosch T et al. Improved modeling of clinical data with kernel methods. *Artif Intell Med* 2012; 54: 103–114.
27. Soguero-Ruiz C, Hindberg K, Mora-Jiménez I et al. Predicting colorectal surgical complications using heterogeneous clinical data and kernel methods. *J Biomed Inform* 2016; 61: 87–96.
28. Fernández-Sánchez J, Soguero-Ruiz C, de Miguel-Bohoyo P et al. Clinical risk groups analysis for chronic hypertensive patients in terms of icd9-cm diagnosis codes. In *PhyCS*. pp.13–22.
29. Chakraborty P and Farooq F. A robust framework for accelerated outcome-driven risk factor identification from ehr. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. pp.1800–1808.
30. Takahashi Y, Yamazaki K, Kamatani Y et al. A genome-wide association study identifies a novel candidate locus at the dlgap1 gene with susceptibility to resistant hypertension in the Japanese population. *Sci Rep* 2021; 11: 1–11.
31. Akimoto H, Nagashima T, Minagawa K et al. Signal detection of potential hepatotoxic drugs: case-control study using both a spontaneous reporting system and electronic medical records. *Biol Pharm Bull* 2021; 44: 1514–1523.
32. Nagashima T, Hayakawa T, Akimoto H et al. Identifying antidepressants less likely to cause hyponatremia: triangulation of retrospective cohort, disproportionality, and pharmacodynamic studies. *Clin Pharmacol Ther* 2022; 111: 1258–1267.
33. Berlinet A and Thomas-Agnan C. *Reproducing kernel Hilbert spaces in probability and statistics*. Norwell, MA: Kluwer Academic Publishers, 2004.
34. Bach FR, Lanckriet GR and Jordan MI. Multiple kernel learning, conic duality, and the smo algorithm. In *Proceedings of the twenty-first international conference on machine learning*. p. 6.
35. Kearns B, Stevenson MD, Triantafyllopoulos K et al. Generalized linear models for flexible parametric modeling of the hazard function. *Med Decis Making* 2019; 39: 867–878.
36. For Drug Statistics Methodology WCC. Anatomical therapeutic chemical classification system. www.whocc.no.
37. Barnes DE, Yaffe K, Byers AL et al. Midlife vs late-life depressive symptoms and risk of dementia: differential effects for Alzheimer disease and vascular dementia. *Arch Gen Psychiatry* 2012; 69: 493–498.
38. Yang G, Ye Q and Xia J. Unbox the black-box for the medical explainable ai via multi-modal and multi-centre data fusion: a mini-review, two showcases and beyond. *Inform Fusion* 2022; 77: 29–52.
39. Wiemken TL and Kelley RR. Machine learning in epidemiology and health outcomes research. *Annu Rev Public Health* 2019; 41: 21–36.
40. van der Ploeg T, Austin PC and Steyerberg EW. Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. *BMC Med Res Methodol* 2014; 14: 1–13.
41. Suzuki T. Pac-Bayesian bound for Gaussian process regression and multiple kernel additive model. In *Conference on Learning Theory*. JMLR Workshop and Conference Proceedings, pp.8–1.
42. Grossi CM, Richardson K, Fox C et al. Anticholinergic and benzodiazepine medication use and risk of incident dementia: a UK cohort study. *BMC Geriatr* 2019; 19: 1–10.
43. Campbell N, Maidment I, Fox C et al. The 2012 update to the anticholinergic cognitive burden scale. *J Am Geriatr Soc* 2013; 61: S142–S143.
44. World Health Organization. *International Classification of Diseases 10th Revision*. https://icd.who.int/browse10/2019/en.
45. Lu C, Goeman J and Putter H. Maximum likelihood estimation in the additive hazards model. *arXiv preprint arXiv:200406156* 2020.
46. Martinussen T and Scheike TH. *Dynamic regression models for survival data*. New York, NY: Springer Science Business Media, 2006.
47. Suzuki T and Tomioka R. Spicymkl: a fast algorithm for multiple kernel learning with thousands of kernels. *Mach Learn* 2011; 85: 77–108.
48. Liu DC and Nocedal J. On the limited memory bfgs method for large scale optimization. *Math Program* 1989; 45: 503–528.
49. Bach F and Jordan M. Kernel independent component analysis. *J Mach Learn Res* 2002; 3: 1–48.

Appendix 1: Disorders for exclusion in enrollment

Patients who had records of the following disorders before the index date were excluded from the cohort: insomnia not under interest (G47.1, G47.3, G47.4 and G47.8), central tuberculosis (A17), congenital syphilis (A50), central viral diseases (A81–89, B00.4, B01.0, B01.1, B02.0, B02.1, B05.0 and B05.1), hepatic coma (B15.0, B16.0, B16.2 and B19.0), HIV infection (B20–24), central tumour (C69–72, C79-80, D33 and D43), addictive drugs (F11–18), schizoid disorders (F20–29), brain damage (F06 and F07), mental retardation (F70–89), central nervous system

inflammation (G00–09), central nervous system degeneration (G10–14), Parkinsonism after encephalitis (G21.3), central demyelinating diseases (G35–37), epilepsy (G40-41), cerebral palsy (G80), central disorders (G91–99), subarachnoid haemorrhage (I60 and I69.0), congenital diseases (Q00–99) and dementia (F00–03 and G30–32).

Appendix 2: Design of explanatory variables

In the present study, the logarithm of the hazard rate of dementia onset was modelled to be the sum of linear and nonlinear functions of explanatory variables that summarise the history of diagnosis, prescription and laboratory-test values up to the time point under consideration. In the following subsections, we describe the overall design of these explanatory variables, in detail. The design of kernel functions of these variables that construct risk functions is given in Appendix 4. The concrete set of the explanatory variables and kernels included in each model is given in Table 4.

Variables for prescribed medication

First, we created overlapped groups of related drugs for which explanatory variables were defined. In addition to automatically generated drug groups that correspond to the anatomical therapeutic chemical classification's (ATC's) categories of level 3[36], we manually created groups of benzodiazepines with very short, short, intermediate and long half-lives in blood [Table 1], Z-drugs, other hypnotics, anti-cholinergic drugs with two different burden levels (ACB12 and ACB3 defined in a previous study[42] according to the scale determined by another previous study[43]), antihypertensive drugs, lipid-modifying drugs, anti-diabetic drugs, anti-platelet drugs, anticoagulant drugs, anti-depressants and anti-Parkinson-disease drugs (see Appendix 5 for the list of the drug groups used to define explanatory variables). For drug group g, time t and patient p, we prepared four continuous variables, $u_{1,gpt}$–$u_{4,gpt}$ and one 0-1 binary variable, $u_{5,gpt}$. Three of the continuous variables, $u_{1,gpt}$–$u_{3,gpt}$, represented the administration period, the cumulative dose and the current dose of the drugs in group g, respectively. The binary variable, $u_{5,gpt}$, took value one if the patient had ever used a drug in group g at time t. The amounts of different drugs were compared by using the defined daily dose (DD) of the drugs authorised by the Pharmaceuticals and Medical Devices Agency of Japan as a unit. These DDs were defined as the mean of the largest and smallest recommended maintenance doses. For larger drug groups, such as those corresponding to ATC drug categories of level 3, doses of different member drugs could not be compared because their actions and indications were different. Thus, for such a group g, we did not use variables, $u_{2,gpt}$, and we redefined $u_{3,gpt}$ as a 0-1 binary variable that represented the use or non-use of the member drugs at time t. To reduce the bias due to the lack of information about the period of prescription before the first record, we introduced a continuous variable, $u_{4,gpt}$, that took the value of age at entry if a member drug of g was prescribed during the first 4 months of the record.

Variables for laboratory tests

For laboratory test g, patient p and time t, we defined three continuous variables, $v_{1,gpt}$–$v_{3,gpt}$, and one 0-1 binary variable, $v_{4,gpt}$. The continuous variables, $v_{1,gpt}$–$v_{3,gpt}$, represented the current value, the average of the past values and the value in the first examination of test g. For a time point at which the test item was not examined, the test result was assumed to take the average value of the previous and next examinations. For the periods before the first and after the last, the values of the first and last tests were used, respectively. The binary variable, $v_{4,gpt}$, took value one if the test was performed at least once. This variable was introduced to reduce bias due to the absence of examination records. See Appendix 5 for the list of the laboratory test items used to define explanatory variables.

Variables for diagnosis history

In a similar manner to the grouping of prescribed drugs, we created overlapped diagnosis groups of ICD10 codes. These groups varied in size, ranging from a single ICD10 code, to a group of nearly one hundred ICD10 codes (e.g. C00–97 for malignancy). We constructed these groups by either manually creating groups of related disorders, or automatically generating groups that correspond to categories and subcategories of the ICD10 coding scheme[44]. Since the number of thus constructed groups was beyond the limit determined by the available computational resources, we used only groups of disorders in which more than ten percent of the cohort suffered from a disorder, except for manually selected groups of disorders that were expected to influence the development of dementia. See Appendix 5 for the list of the disorder groups used to define explanatory variables. For diagnosis group g, time t and patient p, we defined two continuous variable and one 0-1 binary variables, $w_{1,gpt}$–$w_{3,gpt}$. The continuous variable, $w_{1,gpt}$, represented the duration of illness, namely, the length of the period between the date of the first record of disorders in g and time t. The binary variable, $w_{3,gpt}$, took value one if the patient p had already been diagnosed with a disorder in g at time t. To reduce bias due to the lack of information about the presence or absence of disorders before the first record, we introduced the continous variable, $w_{2,gpt}$, that took the value of the age at entry if the patient had already been diagnosed within the four months after the first visit.

Variables for demographic information

In addition to the above variables, for each patient p and time point t, we defined continuous variables, $z_{1,pt}$–$z_{4,pt}$, that indicate the age at time t, the age at entry, the date at time t and body-mass index recorded at an unknown timing, respectively. We also defined a 0-1 binary variable, $z_{5,pt}$, that indicated the sex of the patient.

Appendix 3: Mathematical description of statistical models and regularised maximum-likelihood estimation

Hazard model with multiple kernels

We modelled the hazard rate for onset of dementia, using the framework of multiple kernel learning. In this model, the logarithm of the hazard rate $h_p(t)$ at time t for patient p was nonhomogeneous with respect to time and was modelled as a non-linear function of explanatory variables collectively denoted by $x_p(t)$. Concretely, using the variables defined in Appendix 2, the vector of explanatory variables for time t and patient p was denoted by $x_p(t) = \{\{u_{i,gpt}\}_{i,g}, \{v_{i',g'pt}\}_{i',g'}, \{w_{i'',g''pt}\}_{i'',g''}, \{z_{i''',pt}\}_{i'''}, t\}^T$, where indices such as i, g ran through all possible values so that all of the explanatory variables of the model were included in $x_p(t)$, and the superscript T denoted vector transposition. Hereafter, we omit the range of indices similarly for the sake of notational simplicity, when it can be easily seen from the context. Given that the patient p had not suffered dementia until time t, the hazard rate $h_p(t)$ determined the conditional probability of dementia onset in the infinitesimal time interval $[t, t + dt)$ as $h_p(t)dt$. In the present study, we used the following exponential hazard function with function f to be estimated:

$$h_p(t) = \exp\left(f(x_p(t))\right). \tag{1}$$

In most previous epidemiological studies, the log-hazard function was defined as a linear function of explanatory variables, as usually defined in the Cox model. With our notations, this linear model was written as

$$f(x_p(t)) = \alpha^T x_p(t) + b_0, \tag{2}$$

with a vector of linear coefficients, α, and a bias parameter, b_0.

We nonlinearly extended the above model, introducing the following *'multiple-kernel model'*:

$$f(x_p(t)) = \sum_{1\le j\le K} f_j(x_p(t)) + b_0, \quad f_j \in \mathcal{H}_{k_j}, \tag{3}$$

where f_j for each index j is a linear or nonlinear risk function belonging to the reproducing-kernel Hilbert space $\mathcal{H}_{k_j}$ associated with a kernel function $k_j(x, x')$. Although we avoid going into the detail of the function space, $\mathcal{H}_{k_j}$, referring readers to the literature[33], we just note that it contains all functions of x of the following form:

$$\sum_{1\le i\le N} \alpha_i k(x, x^{(i)}) \tag{4}$$

with arbitrary number N and arbitrary values for α_i and $x^{(i)}$. In fact, $\mathcal{H}_{k_j}$ is mathematically defined as the completion of the set of functions of the above form. In the present study, the number of used kernels K was a few to several hundreds, and hence, the hazard rate for dementia onset was described as the exponential of the sum of multiple component risk functions. Each of the component risk functions was a linear or nonlinear function of one or two explanatory variables. The details of the design of kernels used in our analysis are described in Appendix 4.

Regularised maximum-likelihood estimation

As a functional of the set of the component risk functions, $\{f_j\}_{1\le j\le K}$, the log-likelihood, $L(\{f_j\}_j)$, of the dementia onsets of the patients was written as

$$L(\{f_j\}_j) = \sum_{p\in\mathcal{D}} f(x_p(t_{p,\mathrm{dem}})) - \sum_{p\in\mathcal{A}} \int_{t_{p,\mathrm{idx}}}^{t_{p,\mathrm{end}}} \exp f(x_p(t))dt, \tag{5}$$

where $\mathcal{A}$ and $\mathcal{D}$ denoted the set of all patients and the set of patients who had developed dementia by the end of the observation, respectively, and $t_{p,\mathrm{dem}}$, $t_{p,\mathrm{idx}}$ and $t_{p,\mathrm{end}}$ denoted the date of diagnosis of dementia, index date and date of the end of the observation, respectively (the above formula can be found in the literature[45,46] in different ways of presentation). We numerically found a solution $\{f_j^*\}_j$ that maximises a regularised version of the above likelihood:

$$\{f_j^*\}_{1\le j\le K} = \arg\max_{\{f_j\}_j} \{L(\{f_j\}_j) - \Omega(\{f_j\}_j)\}, \tag{6}$$

where $\Omega(\{f_j\}_j)$ is a regularisation functional.

Historically, multiple kernel learning proposed by Lanckriet et al.[20] was formulated in a different form but later shown to be equivalent to the above regularised problem with the following 1-norm regulariser[34]:

$$\Omega(\{f_j\}) = \sum_{1\le j\le K} \lambda_j \|f_j\|_{\mathcal{H}_{k_j}}, \tag{7}$$

where $\|\cdot\|_{\mathcal{H}_{k_j}}$ denotes the norm of the argument function defined by kernel k_j, and $\lambda_j \in \mathbf{R}$ is a hyperparameter. Here, the value of λ_j was determined, depending on whether kernel k_j was linear or Gaussian. Hence, we performed two independent grid searches to determine these

values. In addition to the above 1-norm regulariser, we also used the following 2-norm regulariser:

$$\Omega(f) = \sum_{1 \le j \le K} \lambda_j \|f_j\|^2_{\mathcal{H}_{k_j}}. \quad (8)$$

As a solution for the problem in equation (6), we implemented a state-of-the-art algorithm called *'dual augmented Lagrangian (DAL) algorithm'*[47], approximating the integral in equation (5) by a summation with discretised timesteps of 60 days width. The optimisation appearing in DAL was solved with the limited-memory BFGS algorithm[48]. The entire program code was developed in C++ and OpenACC languages and designed in such a manner that most of the computations are offloaded to graphic processing units (GPUs). We also accelerated the computation by approximating Gram matrices, G_{k_j}, with their incomplete-Cholesky decompositions[49], $G_{k_j} \approx L_{k_j} L_{k_j}^T$. In practice, 1-percent tolerance of error with respect to the trace norm resulted in approximation with decomposed matrices of a few tens rank. These accelerations allowed us to obtain a single solution for equation (6) with the subcohort data of 431,698 person-month size and with the 59 kernels in the MKL2 model in 78.3±14.4 s on a computer with four Intel Xeon E5-2680 V4 CPU cores and four NVIDIA Tesla P100 16GB GPU cards. The same computation performed only with four Intel Xeon E5-2680 V4 CPU cores took 2187.3±552.4 s. Single estimation with the same dataset and 810 kernels in the MKL3 model was performed in 823.4±175.2 s with twice the computational resources. The other computations were performed similarly.

Appendix 4: Design of kernels

For the multiple-kernel model of the hazard function, we used linear and Gaussian kernel functions. For d variables, $\mathbf{q} = \{q_i\}_{1 \le i \le d}$, chosen from the set of explanatory variables, linear and Gaussian kernels were defined as $k_{\text{linear}}(\mathbf{q}_1, \mathbf{q}_2) = \mathbf{q}_1^T \mathbf{q}_2$ and $k_{\text{Gauss}}(\mathbf{q}_1, \mathbf{q}_2) = \exp(-\|\mathbf{q}_1 - \mathbf{q}_2\|^2 / 2d\sigma^2)$, respectively. Here, we used a bandwidth parameter σ for the Gaussian kernel, whose value was determined by cross-validation. In the above equation, the symbol $\|\cdot\|$ denoted the Euclidean norm.

In practice, we used a slightly extended version of the above kernels, because some patients had no history of prescribed medication, test results, or diagnosis represented by the variables in a kernel. To address this issue, we simply defined $k(\mathbf{q}_1, \mathbf{q}_2) = 0$, if no relevant record existed for either $\mathbf{q}_1$ or $\mathbf{q}_2$. This modification did not affect the positive-semidefiniteness of the kernel function. Using these modified kernels is equivalent to estimating the corresponding risk function only for the patients with relevant records, while setting it to zero for the other patients. The bias introduced by this treatment was separately estimated by introducing a linear risk function of a binary variable that indicated the presence or absence of the relevant record.

In the following, we described how we concretely defined kernels on the explanatory variables defined in Appendix 2. Using these definitions, the list of the kernels used in each model is shown in Table 4.

Kernels on variables for prescribed medication

For each drug group g, we considered modified linear and Gaussian kernels of five different types P1–P5. As argument variables, $\mathbf{q}_1$, $\mathbf{q}_2$, type-P1–3 kernels took administration period $u_{1,gpt}$, cumulative dose, $u_{2,gpt}$, and current dose, $u_{3,gpt}$, respectively. Type-P4 kernel was introduced to adjust the bias due to the unobserved prescription before the first record in an age-dependent manner. This kernel took age at entry, $u_{4,gpt}$, as its argument variable, but took a zero value if either one of the patients for the argument variables had not been prescribed any drug in g during the first 4 months. Type-P5 kernel was a linear kernel that took the binary variable, $u_{5,gpt}$, as its argument variable. The risk function generated from this kernel described the risk of ever-use of the drugs in g.

Kernels on variables for laboratory test results

For each laboratory test item g, we considered modified linear and Gaussian kernels of four different types T1–T4. For each of the argument variables, $\mathbf{q}_1$, $\mathbf{q}_2$, type T1 kernel took the current test result, $v_{1,gpt}$. Type-T2 kernel took the variable for past average test values and the length of the period between the first visit and time t under consideration, namely, $(v_{2,gpt}, t - t_{p0})$, with t_{p0} denoting the time of the first record of patient p. To adjust the bias due to the unobserved test values before the first record, we introduced type-T3 kernel that took the combination of the first test value and age at the first record, namely, $(v_{3,gpt}, z_{1,pt_{p0}})$. To reduce the bias due to the lack of test records, we also introduced type-T4 linear kernel that took the binary variables indicating the presence or absence of test results, $v_{4,gpt}$, as argument variables.

Kernels on variables for diagnosis history

For each diagnosis group g, we considered modified linear and Gaussian kernels of three different types D1–D3, for which the argument variables, $\mathbf{q}_1$ and $\mathbf{q}_2$, were continuous variables for the duration of illness, $w_{1,gpt}$, the age at the first record of an already diagnosed patient, $w_{2,gpt}$, and a binary variable indicating the presence or absence of illness, $w_{3,gpt}$, respectively. Type-D2 kernel was introduced to adjust the underestimation of the duration of illness for patients who had already been diagnosed in the first record. This kernel took value zero if either one of the patients represented by the two argument variables of the kernel was not diagnosed in the first 4 months of the record.

Kernels on variables for demographic information

In addition to the above kernels, we used modified linear or Gaussian kernels for each of the variables for demographic information.

Appendix 5: Drug groups, disorder groups and laboratory test items represented by explanatory variables

Groups of drugs

Anti-depressants denoted drugs belonging to ATC category N06A. Other known risk-drug groups (and their ATC categories or subcategories) included anti-hypertensive drugs (C02A,C02C, C02DB, C02K, C02L, C03A, C03B, C03D, C04A, C07A, C07B, C07C, C07D, C07E, C07F, C08C, C08DB, C08G, C09A, C09B, C09C, C09D and C09X), statin (C10AA, C10BA and C10BX), anti-diabetic drugs (A10A, A10B and A10X), anti-platelet drugs (B01AC), anti-coagulant drugs (B01AA, B01AE, B01AF and B01AX), and anti-Parkinson drugs (N04A, N04B and N04C). All other drug groups included only in the MKL3 model were those corresponding to the following ATC categories: insulin and analogues (A10A), other blood glucose lowering drugs (A10B), other drugs for diabetes mellitus (A10X), vitamins B1, B6 and B12 (A11D), antithrombotic agents (B01A), vitamin K and other haemostatics (B02B), cardiac glycosides (C01A), other cardiac stimulants (C01C), vasodilators for cardiac diseases (C01D), centrally acting, antiadrenergic, antihypertensive agents (C02A), peripherally acting, antiadrenergic, antihypertensive agents (C02C), agents acting on arteriolar smooth muscle (C02D), other hypertensive agents (C02K), thiazides (C03A), other low-ceiling diuretics (C03B), peripheral vasodilators (C04A), beta blockers (C07A), calcium blockers with mainly vascular effects (C08C), calcium blockers with direct cardiac effects (C08D), angiotensin converting enzyme inhibitors (C09A), angiotensin II receptor blockers (C09C), other agents acting on the renin-angiotensin system (C09X), antiinflammatory and antirheumatic products, non-steroids (M01A), opioids (N02A), antiepileptics (N03A), anticholinergic, anti-Parkinson drugs (N04A), dopaminergic, anti-Parkinson drugs (N04B), antipsychotics (N05A), anxiolytics (N05B), adrenergics for systemic use (R03C) and antihistamines for systemic use (R06A).

Groups of disorders

Groups of disorders in which BZDs are indicated (and their ICD10 codes) included insomnia (G47), affective disorders (F30–39) and anxiety disorders (F40–49). Other known-risk disorder groups included Parkinson disease (G20), diabetes mellitus (E10–14, G59.0, G63.2, H28.0, H36.0, M14.2 and N08.3), dyslipidaemia (E78), hypertension (I10–15), ischaemic heart diseases (I20–25), cerebrovascular diseases (I60–69, G21.4, G45 and G46) and epilepsy (G40 and G41). All other disorder groups included only in the MKL3 model were syphilis (A51–53, A65 and K67.2), malignancy (C00–99, G63.1, G73.1 and G73.2), vitamin B12 deficiency (D51), folate deficiency (D52), iodine deficiency (E00–03), disorders of thyroid gland (E00–07), type-I diabetes mellitus (E10), type-II diabetes mellitus (E11), obesity (E65–68), disorder of purine and pyrimidine metabolism (E79), epilepsy (G40,G41), glaucoma (H40–42), atrial fibrillation and flutter (I48), subarachnoid haemorrhage (I60 and I69.0), intracerebral haemorrhage (I61 and I69.1), intracranial haemorrhage (I62 and I69.2), cerebral infarction (I63 and I69.3), other cerebral vascular disorders (I65–69, G21.4 and G45–46), atherosclerosis (I70), disorders of arteries, arterioles and capillaries (I70–79), influenza (J09–11), chronic rhinitis, nasopharyngitis and sinusitis (J31–34), chronic sinusitis (J32), chronic lower respiratory diseases (J40–47), diseases of liver (K70–77), pruritus (L29), systemic connective tissue disorders (M30–36), chronic kidney disease (N18), chronic kidney disease stages 3 and 4 (N18.3 and N18.4), chronic renal dialysis (N18.5, Z99.2 and T80.9) and head injury (S00–09).

Laboratory test items

For the linear and MKL1-2 models, only the LDL-cholesterol level was used as an explanatory variable among laboratory test values. Other laboratory test values adopted in the MKL3 model included serum level or plasma level of haemoglobin, total bilirubin, total cholesterol, MDA-LDL, HDL-cholesterol, triglyceride, total protein, albumin, urea nitrogen, creatinine, sodium ion, potassium ion, chloride ion, calcium ion, zinc ion, magnesium ion, inorganic phosphorus ion, iron, ammonium, ferritin, uric acid, protein S, protein C, amyloid alpha protein, arachidonic acid, eicosapentaenoic acid, docosahexaenoic acid, dihomogammalinolenic acid, folic acid, thyroid stimulating hormone, free T3, free T4, glycoalbumin, haemoglobin A1c (NGSP), C-reactive protein, immunoglobulin-A/E/G, aspartate aminotransferase, alanine aminotransferase, gamma-glutamyl transpeptidase, choline esterase, lactate, lactate dehydrogenase, alkaline phosphatase, leucine aminopeptidase and vitamin B12, white blood cell count, eosinophil count, basophil count and platelet count in blood, mean corpuscular volume and mean corpuscular haemoglobin and its concentration of red blood cells, plasma total and unsaturated iron binding capacity, urinary albumin level, urinary albumin index, urinary protein, adjusted creatinine clearance, prothrombin time (INR), activated partial thromboplastin clotting time, qualitative test of urinary haemoglobin, LDL-HDL ratio, serologic test and latex agglutination test for Treponema pallidum, antibody titre for thyroglobulin and 50% haemolytic complement activity.