Revista de psiquiatría y salud mental (Barcelona) 15 (2022) 167–175

Revista de Psiquiatría y Salud Mental

www.elsevier.es/saludmental

ORIGINAL ARTICLE

Estimation of the epidemiology of dementia and associated neuropsychiatric symptoms by applying machine learning to real-world data

Javier Mar[a,b,c,d,*], **Ania Gorostiza**[a,b], **Arantzazu Arrospide**[a,b,c,d], **Igor Larrañaga**[a,b], **Ane Alberdi**[e], **Carlos Cernuda**[e], **Álvaro Iruin**[f,c], **Mikel Tainta**[g,h], **Lorea Mar-Barrutia**[i], **Oliver Ibarrondo**[a,c,j]

[a] *Basque Health Service (Osakidetza), Debagoiena Integrated Healthcare Organisation, Research Unit, Arrasate-Mondragón, Gipuzkoa, Spain*
[b] *Kronikgune Institute for Health Service Research, Barakaldo, Bizkaia, Spain*
[c] *Biodonostia Health Research Institute, Donostia-San Sebastián, Gipuzkoa, Spain*
[d] *Health Services Research on Chronic Patients Network (REDISSEC), Bilbao, Bizkaia, Spain*
[e] *Mondragon Unibertsitatea, Faculty of Engineering, Electronics and Computing Department, Arrasate-Mondragón, Gipuzkoa, Spain*
[f] *Basque Health Service (Osakidetza), Gipuzkoa Mental Health Network, Donostia-San Sebastián, Gipuzkoa, Spain*
[g] *Basque Health Service (Osakidetza), Goierri-Urola Garaia Integrated Healthcare Organisation, Department of Neurology, Zumarraga, Gipuzkoa, Spain*
[h] *Fundación CITA-Alzheimer Fundazioa, Donostia-San Sebastián, Gipuzkoa, Spain*
[i] *Psiquiatry Service, Hospital Bellvitge, Hospitalet de Llobregat, Barcelona, Spain*
[j] *RS-Statistics, Arrasate-Mondragón, Gipuzkoa, Spain*

Received 15 February 2021; accepted 14 March 2021

KEYWORDS
Dementia;
Neuropsychiatric symptoms;
Machine learning;
Incidence;
Prevalence

Abstract
Introduction: Incidence rates of dementia-related neuropsychiatric symptoms (NPS) are not known and this hampers the assessment of their population burden. The objective of this study was to obtain an approximate estimate of the population incidence and prevalence of both dementia and NPS.
Methods: Given the dynamic nature of the population with dementia, a retrospective study was conducted within the database of the Basque Health Service (real-world data) at the beginning and end of 2019. Validated random forest models were used to identify separately depressive and psychotic clusters according to their presence in the electronic health records of all patients diagnosed with dementia.

* Corresponding author.
E-mail address: javier.marmedina@osakidetza.eus (J. Mar).

Results: Among the 631,949 individuals over 60 years registered, 28,563 were diagnosed with dementia, of whom 15,828 (55.4%) showed psychotic symptoms and 19,461 (68.1%) depressive symptoms. The incidence of dementia in 2019 was 6.8/1000 person-years. Most incident cases of depressive (72.3%) and psychotic (51.9%) NPS occurred in cases of incident dementia. The risk of depressive-type NPS grows with years since dementia diagnosis, living in a nursing home, and female sex, but falls with older age. In the psychotic cluster model, the effects of male sex, and older age are inverted, both increasing the probability of this type of symptoms.
Conclusions: The stigmatization factor conditions the social and attitudinal environment, delaying the diagnosis of dementia, preventing patients from receiving adequate care and exacerbating families' suffering. This study evidences the synergy between big data and real-world data for psychiatric epidemiological research.

PALABRAS CLAVE
Demencia;
Síntomas neuropsiquiátricos;
Aprendizaje automático;
Incidencia;
Prevalencia

Estimación de la epidemiología de la demencia y los síntomas neuropsiquiátricos asociados, mediante la aplicación del aprendizaje automático a los datos del mundo real

Resumen
Introducción: Se desconocen las tasas de incidencia de los síntomas neuropsiquiátricos (SN) asociados a la demencia, lo cual dificulta la evaluación de su carga para la población. El objetivo de este estudio fue obtener una estimación aproximada de la incidencia y prevalencia en la población tanto de la demencia como de los SN.
Métodos: Dada la naturaleza dinámica de la población con demencia, se realizó un estudio dentro de la base de datos del Servicio Vasco de Salud (datos del mundo real) a comienzos y finales de 2019. Se utilizaron modelos de bosques aleatorios validados para identificar por separado los clústeres depresivos y psicóticos, con arreglo a su presencia en los registros sanitarios electrónicos de todos los pacientes con diagnóstico de demencia.
Resultados: Entre los 631.949 individuos mayores de 60 años registrados, 28.563 fueron diagnosticados de demencia, de los cuales 15.828 (55,4%) mostraron síntomas psicóticos y 19.461 (68,1%) síntomas depresivos. La incidencia de la demencia en 2019 fue de 6,8/1.000 personas-años. Muchos de los casos incidentes de SN depresivos (72,3%) y psicóticos (51,9%) se produjeron en casos de demencia incidente. El riesgo de SN de tipo depresivo se incrementa con factores tales como los años transcurridos desde que se diagnostica la demencia, la residencia en un sanatorio, y el sexo femenino, pero desciende con la edad avanzada. En el modelo de clúster psicótico, los efectos del sexo masculino y la edad avanzada se invierten, incrementando ambos la probabilidad de este tipo de síntomas.
Conclusiones: El factor de estigmatización condiciona el entorno social y actitudinal, demorando el diagnóstico de la demencia, impidiendo que los pacientes reciban los cuidados adecuados, y exacerbando el sufrimiento de las familias. Este estudio evidencia la sinergia entre los grandes datos y los datos del mundo real para la investigación epidemiológica psiquiátrica.

Introduction

Dementia is a progressive disease whose expression in cognitive, behavioural, and functional dimensions changes over time.[1] Among them, the behavioural component measured in terms of care required for neuropsychiatric symptoms (NPS) and disruptive behaviours places the greatest burden on caregivers.[2] The dynamic nature of dementia means that the prevalence of NPS varies enormously depending on the origin of the sample (community or institution), subtype (Alzheimer's disease or vascular dementia), and the time since onset of the condition.[3] Furthermore, most studies have analyzed small samples that are not representative of populations.[4] Nonetheless, the existing evidence indicates that the prevalence increases as dementia progresses and that almost all patients experience some form of NPS at some point in the course of the disease.[5]

While the literature on the prevalence of NPS is abundant,[2–6] there is a paucity of data on incidence. As for dementia, however, estimating the incidence of NPS requires efforts to diagnose them over time through serial cross-sectional studies.[7–10] While cohort studies require contact with the sample each time data are gathered, population registries based on real-world data (RWD) allow

us to efficiently obtain automated and repeated information whenever needed for psychiatric epidemiological research.[11,12] The lack of recording of specific NPS codes in electronic medical records (EHRs) is a barrier to understanding their epidemiology but this can be overcome by using machine learning tools to predict them.[13,14] Although predicting them cannot be deemed the same as the symptom assessment achieved using scales such as the Neuropsychiatric Inventory (NPI),[15] it provides an approach that can be used to monitor their incidence and prevalence in entire populations over time.[14] Though NPS are varied, authors have attempted to cluster them into distinct groups that predictive models should take into account.[2,5] In general, four groups are distinguished: psychotic symptoms, depressive state, apathy, and agitation[5]; and these can be reduced to two clusters by summing psychotic symptoms and agitation, on the one hand, and depressive symptoms and apathy, on the other.[12,14]

The objective of this study was to obtain an approximate estimate of the incidence and prevalence of dementia and associated NPS in a population database covering the whole population of the Basque Country (2.2 million).

Methods

Design

Given the dynamic nature of populations with specific health problems, to measure the incidence and prevalence of dementia and dementia-related NPS, we conducted a retrospective study with two cross-sectional datasets (31 December 2018 and 2019). By ''dynamic population'', we mean one in which the members vary over time with individuals entering or leaving as a function of the occurrence of events.[16]

Data were obtained from the organization-wide database of the Basque Health Service, called Oracle Business Intelligence (OBI), which stores anonymized administrative and clinical records from all primary care, hospitalization, emergency care, and outpatient consultations.[8,12] The study protocol was approved by the Ethics Committee for Clinical Research of the Basque Country (with registration number PI2018143, EPA-OD).

Patients

Cases of dementia were identified by searching data in OBI on 31 December 2018 and 31 December 2019 for codes for this condition from the International Classification of Diseases, 9th and 10th Editions; this procedure has been shown to be adequate, achieving a positive predictive value of 95.1%, negative predictive value of 99.4%, sensitivity of 80.2%, and specificity of 99.9%.[8] Within the scope of NPS, we differentiated between a depressive cluster (depression, anxiety, and apathy) and a psychotic cluster (aggressiveness, irritability, restlessness, screaming, delusions and hallucinations).[14] As a way of approximately identifying NPS, we used two predictive algorithms based on random forest machine learning which indicates whether the clinical notes from EHRs of people with dementia contain mention of psychotic or depressive symptoms. The final predicted outcome is the result of combining all the variables included. An advantage of using a random forest algorithm is that it provides the weight of each variable in the classification of individuals as having NPS symptoms or not. The most important variables in the model of psychotic symptoms were the use of risperidone, level of sedation, use of quetiapine or haloperidol, and number of antipsychotics prescribed. In the depressive symptoms model, the most important variables were the number of antidepressants prescribed, use of escitalopram, level of sedation, and age. The validation of this procedure showed areas under the receiver operating characteristic curve of 0.80.[14]

Variables

The following variables were considered: age, sex, institutionalization status (whether the patient was living in a nursing home or similar), time since dementia diagnosis, concomitant diagnoses (diabetes mellitus, hypertension, dyslipidaemia, thyroid disease, Parkinson's disease, cerebrovascular accident, cardiovascular disease, head injury, depressive disorder, and psychotic disorder), and pharmacological treatment. Since the NPS identification algorithms use medications among other components, the data collected from the pharmacological registry included the following subgroups of the Anatomical Therapeutic Chemical Classification System: N06D (donepezil, rivastigmine, galantamine, and memantine), N06A (antidepressants), N05A, and NO6C (antipsychotics). Not only all prescriptions but also changes in prescriptions were recorded. The level of sedation produced by each drug was classified (0: none; 1: minimal; 2: mild; 3: moderate; or 4: deep) according to Table SM1 in the supplementary material.[14]

To be characterized as an incident case, in the case of both dementia and NPS, the rule applied was that a patient changed from not being diagnosed at the cut-off of 31 December 2018 to meeting the diagnostic criteria for the condition in 2019.

Statistical analysis

All statistical and random forest processing and analysis were performed using R version 3.6.0. First, cases with depressive and psychotic NPS were identified. It was also analyzed whether the cases from 2018 were still alive in 2019 and whether the cases from 2019 were present in 2018. In this way, the cases of dementia and NPS were classified into four groups: alive in 2019 and diagnosis already present in 2018; alive in 2019 and diagnosis absent in 2018; dead in 2019 and diagnosis already made in 2018; and dead in 2019 and diagnosis absent in 2018.

The dementia cases identified as of 31 December 2019 that were not present in the 2018 dataset were used as the numerator of the incidence of dementia. To calculate the rate per 1000 person-years, the denominator was estimated with the population in 2018 from which we subtracted the years of observation lost due to deaths and the cases of dementia already identified in 2018. It was assumed that both events occurred uniformly throughout the year, and hence, they were assigned a duration of 0.5 years.

To estimate the prevalence and incidence of dementia-related NPS, the two random forest algorithms were applied to each dataset to identify cases on 31 December 2018 and 2019. Their prevalence disaggregated by age on 31 December 2019 was calculated by dividing the number of cases of dementia-related NPS by the total number of cases of dementia. To measure the incidence rate per 1000 person-years, the exposed population and observation time determined the denominator. This included the population with dementia in 2018, having subtracted the cases with NPS diagnosed before 31 December 2018, and the deaths during 2019 set to have been observed at 0.5 years. In addition, the observation time associated with the dementia cases incident during that year (2019), and not, therefore, present in the 2018 dataset, was added to the denominator. As these incident cases occur throughout the year, the diagnosis was set to have been observed at an average of 0.5 years.[16] The cases identified as of 31 December 2019 that were not present in the 2018 dataset were used as the numerator in the NPS incidence calculation. The calculation of the incidence and prevalence of dementia and dementia-related NPS is fully described in the supplementary material.

Logistic regression models were used to investigate the relationship between the diagnosis of dementia and the probability of NPS, taking presence of NPS as the dependent variable and time since the diagnosis of dementia as the independent variable. Adjustment covariates included age, sex, and living in a nursing home.

Results

From the 631,949 individuals over 60 years registered in the Basque Health Service database and alive on 31 December 2019, 28,563 had a diagnosis of dementia. Their characteristics are summarized in Table 1. They had a median age of 85 years and had been diagnosed a median of 4 years earlier, while 70% were women, 24% lived in nursing homes, and a third had been diagnosed by neurologists. The psychotic cluster was observed in 55.4% of cases and the depressive cluster in 68.1%.

The same table (Table 1) shows the dynamic components of the populations diagnosed with dementia and with dementia and psychotic or depressive NPS. The incidence of NPS has been disaggregated by whether dementia was diagnosed that same year or in previous years. It is striking that 72.3% and 51.9% of incident cases of depressive NPS and psychotic NPS respectively occur in individuals diagnosed with dementia in the same year. Nonetheless, most of the cases of dementia (85.9%) and dementia with psychotic NPS (80.0%) or depressive NPS (82.8%) on 31 December 2019 were from the prevalent cohort composed of those already diagnosed with dementia before 31 December 2018.

The prevalence and incidence of dementia by age group appear in Table 2, Figs. 1 and 2, and Tables SM2 and SM3 with confidence intervals. Both increase exponentially with age until the oldest group (over 90 years of age), in which the incidence was 25 cases per 1000 person-years and prevalence 21.3%. Figs. 1 and 2 show comparisons with the results in the literature on the prevalence and incidence of dementia by age group.

The prevalence of NPS remains stable with age up to 85 years, with rates of over 70% in the depressive cluster and over 50% in the psychotic one (Table 2, Fig. 3, and Tables SM4 and SM5 with confidence intervals). From that age, the prevalence of the psychotic type increases and that of depressive type decreases. The incidence of psychotic-type NPS also increases steadily with age, exceeding 60 cases per 100 person-years over 90 years. In contrast, the distribution of the incidence of the depressive cluster shows a peak between 75 and 80 years of age, decreasing in the oldest groups (Table 2, Fig. 3, and Table SM4 with confidence intervals). While the percentage of patients with dementia using the most common antidepressant drug (trazodone) increases with age, the use of escitalopram decreases in over-85-year-olds (Table SM6).

The risk of depressive-type NPS grows with years since dementia diagnosis, living in a nursing home, and female sex, but falls with older age (Table 3). In the psychotic cluster model, the effects of male sex and older age are inverted, both increasing the probability of this type of symptoms (Table 3).

Discussion

Our study presents, for the first time, the incidence rates of dementia-related NPS through a comprehensive analysis of the whole population with dementia and considering the dynamic nature of their epidemiology in a region of 2.2 million inhabitants. Understanding the actual burden that NPS represent requires placing these symptoms in the context of all the individuals with dementia. That is, generating real world evidence from analysing RWD.[10]

Most NPS prevalent cases come from patients diagnosed with dementia in previous years, given the relationship of NPS with the progression of the disease.[3] Nonetheless, it is noteworthy that new diagnoses of dementia and NPS appear in the same year in more than 50% of cases. Our explanation for this ''anomaly'' is that NPS act as a trigger for contact with the health system in patients with dementia. While the progression is limited to the cognitive dimension, families tend to keep patients at home, managing their needs without medical support. Further, in societies with traditional family models, such as the Basque Country, there are barriers to early recognition of dementia due to its stigmatizing nature,[17] that is, being diagnosed with dementia is linked to a devalued social identity.[17] The stigmatization factor conditions the social and attitudinal environment, delaying the diagnosis, preventing patients from receiving adequate care and exacerbating families' suffering.[17,18] Nonetheless, when NPS appear, families tend to give up, due to the associated stress, and seek social and health care for the patient, this enabling the registration of the dementia code in the EHR. At the same time, NPS are often written in the EHR notes but without including their specific codes. This does not prevent them from being treated pharmacologically, however, and hence, it possible to use machine learning algorithms to identify them.[14]

The estimated incidence rate of dementia by age group in our study is almost the same as that in other studies also carried out using RWD, by Ponjoan et al. in the Catalan population, van Bussel et al. in the Dutch population,

Table 1 Characteristics of patients diagnosed with dementia and neuropsychiatric symptoms on 31 December 2019.

	Dementia *N* = 28,563	Psychotic cluster *N* = 15,828 (55.4%)	Depressive cluster *N* = 19,461 (68.1%)
Gender: female	20,110 (70.4%)	10,722 (67.7%)	14,367 (73.8%)
Age	85.0 [79.0;89.0]	86.0 [80.0;90.0]	84.0 [79.0;88.0]
Living in a nursing home	6,920 (24.2%)	5,458 (34.5%)	5,213 (26.8%)
Dementia diagnosis in neurology clinics	9,429 (33.0%)	5,501 (34.8%)	6,599 (33.9%)
Years since dementia diagnosis, median [IQR]	4.00 [1.00;7.00]	4.00 [2.00;7.00]	4.00 [2.00;7.00]
Arterial hypertension	17,043 (59.7%)	9,417 (59.5%)	11,521 (59.2%)
Diabetes	7,159 (25.1%)	3,974 (25.1%)	4,613 (23.7%)
Dyslipidaemia	14,526 (50.9%)	7,633 (48.2%)	10,388 (53.4%)
Thyroid disease	5,494 (19.2%)	2,868 (18.1%)	3,957 (20.3%)
Parkinson's disease	1,736 (6.08%)	1,369 (8.65%)	1,412 (7.26%)
Stroke	8,729 (30.6%)	4,751 (30.0%)	5,887 (30.3%)
Cardiovascular disease	5,678 (19.9%)	3,474 (21.9%)	3,783 (19.4%)
Traumatic brain injury	5,534 (19.4%)	3,524 (22.3%)	4,025 (20.7%)
Antipsychotic treatment	16,843 (59.0%)	14,132 (89.3%)	12,552 (64.5%)
N. antipsychotic treats	1.00 [1.00;2.00]	1.00 [1.00;2.00]	1.00 [1.00;2.00]
Changes from antipsychotic to antidepressant			
No changes	3,200 (11.2%)	2,311 (14.6%)	1,250 (6.42%)
Some change	13,643 (47.8%)	11,821 (74.7%)	11,302 (58.1%)
No medication	11,720 (41.0%)	1,696 (10.7%)	6,909 (35.5%)
N. changes from antipsychotic to antidepressant	2.00 [1.00;2.00]	2.00 [1.00;3.00]	2.00 [1.00;3.00]
Antidepressant treatment	21,313 (74.6%)	13,057 (82.5%)	18,167 (93.4%)
N. antidepressant treats.	1.00 [1.00;2.00]	2.00 [1.00;2.00]	2.00 [1.00;2.00]
Changes from antidepressant to antipsychotic			
No changes	9,413 (33.0%)	2,539 (16.0%)	7,747 (39.8%)
Some change	11,900 (41.7%)	10,518 (66.5%)	10,420 (53.5%)
No medication	7,250 (25.4%)	2,771 (17.5%)	1,294 (6.65%)
N. changes from antidepressant to antipsychotic	2.00 [1.00;3.00]	2.00 [1.00;3.00]	2.00 [1.00;3.00]
Sedation level			
None	6,699 (23.5%)	405 (2.56%)	3,053 (15.7%)
Minimal	1,822 (6.38%)	111 (0.70%)	786 (4.04%)
Mild	10,224 (35.8%)	5,656 (35.7%)	7,618 (39.1%)
Moderate	9,589 (33.6%)	9,430 (59.6%)	7,793 (40.0%)
Deep	229 (0.80%)	226 (1.43%)	211 (1.08%)

Dynamic features	Dementia	Psychotic cluster	Depressive cluster
Prevalence (31/12/2018)	28,474	15,460	18,687
Prevalent cohort (31/12/2019)	24,535	12,667	16,106
Mortality (2019)	3,939	2,793	2,581
Incidence (2019)	4,028	3,161	3,355
Dementia diagnosis pre-2019		1,520 (48.1%)	929 (27.7%)
Dementia diagnosis 2019		1,641 (51.9%)	2,426 (72.3%)

N.: number of; prevalent cohort: cases already diagnosed in 2018.

and Perera et al. in various European populations.[7,19,20] As would be expected, a collaborative study with population cohort design yielded higher figures.[9] The comparison of the dementia prevalence by age group found in our study with that in the literature shows heterogeneous results. Although our estimated figures exactly match values obtained in the Catalan population,[19] it is higher than that obtained in the group of European populations.[7] This less satisfactory comparison of the prevalence may be due to databases used in these other studies being from PC or hospital settings, while ours includes PC, hospitalization, and outpatient databases to identify cases of dementia. In contrast, in population-based studies, the prevalence of dementia reached significantly higher figures.[21,22] This underreporting of dementia in clinical databases, especially in age groups older than 85 years, is explained by the trivialization of symptoms of cognitive decline in advanced ages and their attribution to normal ageing. The consistency found when comparing our dementia epidemiology indicators with those of other RWD studies supports the validation of the incidence and prevalence figures for depressive or psychotic NPS derived from them. Since no such RWD studies estimating NPS incidence are available, it was not possible to make comparisons with the literature.

Table 2 Incidence in 2019 and prevalence (on 31 December 2019) of dementia and neuropsychiatric symptoms by age group.

Age group	Population	Dementia prevalence	(%)	Prevalence psychotic NPS	(%)	Prevalence depressive NPS	(%)
Total	631,949	28,563	4.5	15,828	55.4	19,461	68.1
[60;65)	145,211	383	0.3	204	53.3	302	78.9
[65;70)	127,039	776	0.6	403	51.9	584	75.3
[70;75)	117,218	1,938	1.7	981	50.6	1,428	73.7
[75;80)	84,954	4,047	4.8	2,019	49.9	3,068	75.8
[80;85)	72,987	6,485	8.9	3,301	50.9	4,860	74.9
[85;90)	54,553	8,536	15.6	4,776	56.0	5,662	66.3
[90;105)	29,987	6,398	21.3	4,144	62.7	3,557	58.1

Age group	Person-years	Dementia incidence	IR/1000 person-years	Incidence psychotic NPS	IR/100 person-year	Incidence Depressive NPS	IR/100 person-year
Total	593,486	4028	6.8	3,161	24.5	3,355	36.3
[60;65)	144,079	54	0.4	32	9.6	48	31.6
[65;70)	125,316	123	1.0	87	15.2	106	35.4
[70;75)	113,731	413	3.6	248	17.5	342	44.8
[75;80)	79,419	756	9.5	504	20.3	698	59.8
[80;85)	62,998	1028	16.3	697	17.3	820	35.3
[85;90)	44,710	1068	23.9	919	31.0	840	29.9
[90;105)	23,234	586	25.2	674	60.8	501	28.8

NPS: neuropsychiatric symptoms; IR: incidence rate.

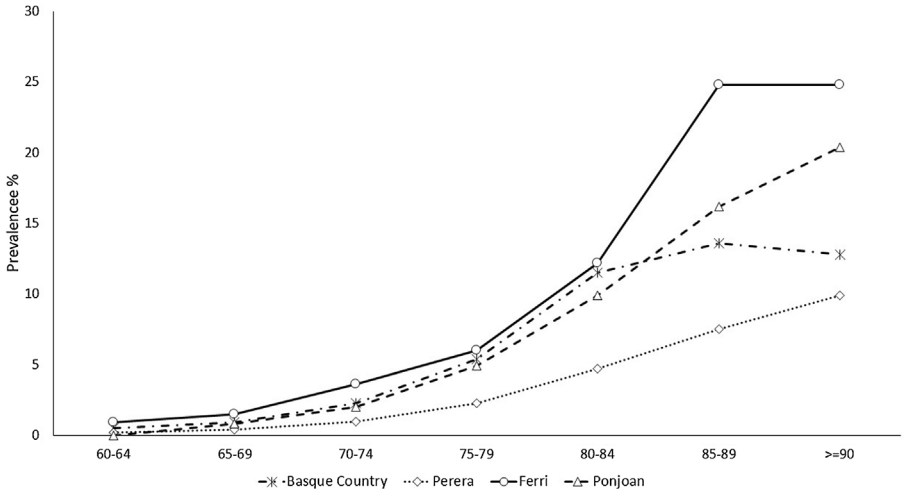

Figure 1 Prevalence of dementia by age group compared with the literature.

Table 3 Effect of time since dementia diagnosis on the probability of identification of neuropsychiatric symptoms.

	Psychotic cluster identification			Depressive cluster identification		
	OR	2.5% (OR)	97.5% (OR)	OR	2.5% (OR)	97.5% (OR)
Intercept	0.23	0.17	0.30	246.52	178.18	341.83
Years since dementia diagnosis	1.03	1.02	1.04	1.03	1.03	1.04
Age	1.02	1.02	1.02	0.94	0.93	0.94
Women	0.63	0.59	0.66	1.90	1.80	2.01
Living in a nursing home	3.87	3.62	4.13	1.74	1.63	1.86

OR: odds ratio.

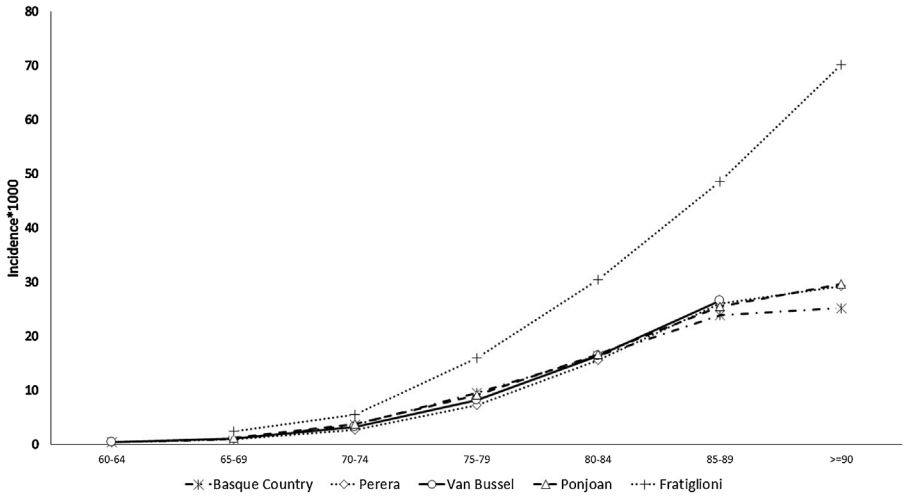

Figure 2 Incidence of dementia by age group compared with the literature.

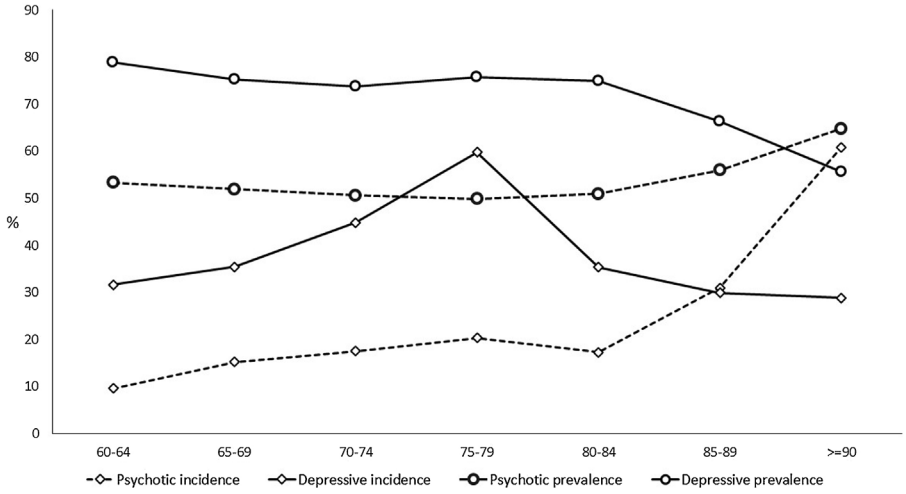

Figure 3 Incidence in 2019 and prevalence (on 31 December 2019) of neuropsychiatric symptoms by age group.

The NPS prevalence rates we have estimated are within the ranges observed in cohort studies.[2,6] On the other hand, the studies we found on the incidence of NPS did not allow comparison by age group.[23] Possibly this lack of incidence data is due to the small sample sizes in studies on NPS. In a systematic review, Borsje et al. analyzed the course of NPS, finding that the studies were based on a mean of 312 patients.[24] A strength of our work is that we have taken into account the dynamic nature of the database through a comprehensive epidemiological design in a population of 2.2 million, including 631,949 people over 60 years of age.[8] The population design includes both individuals living at home and those living in nursing homes, providing clinical information from all inhabitants. Notably, 20% of the population in our region receives care through the private health system because they have double insurance coverage (public and private). Nonetheless, all patients with dementia receive their medication from the public system because their private insurance does not cover pharmacy services.

The main weakness of our approach is that it does not allow analysis of the characteristics of symptoms other than distinguishing between the depressive and psychotic clusters.[14] For the sake of clarity, herein, we have used the same term (NPS) both for the outcome of the predictive algorithm (presence in EHR clinical notes) and that of validated scales (NPI). We acknowledge that the two outcomes are not the same, and that sensitivity and descriptive depth, though not specificity, are improved by using questionnaires such as the NPI[15] designed for clinical practice that proactively search for the presence of 12 symptoms.[14] Moreover, our algorithm searches for any previous prescriptions recorded in the EHR, not only for currently active prescriptions, as we wanted to identify patients that had experienced NPS at any point. The two approaches are not alternatives but rather complimentary, as they address different aims (clinical and epidemiological). Nonetheless, it is not feasible to administer instruments like the NPI to a population of more than 28,000 patients with dementia every year. Addressing NPS in terms of public health is only possible by applying big data tools to population databases such as those derived from EHRs.[13,25] But this approach requires measuring the error incurred, first, in the identification of dementia cases,[8] and,

second, when we applied the random forest algorithm to predict NPI.[14]

The profiles of the two clusters correspond to what could be expected according to the literature, namely, a higher prevalence of the depressive than the psychotic type.[6,12] The association of the psychotic cluster with dementia is relatively easy to interpret since both its incidence and its prevalence unequivocally increase with older age and male sex, indicating that psychotic symptoms are an expression of dementia. In contrast, the relationship of the depressive cluster with dementia and age follows a course that is complex to interpret, as already described in the literature.[26] Whether depression is a risk factor or a consequence of dementia is a matter of debate.[26,27] The design of our study allows us to indicate that both conditions are closely associated, but we are unable to draw conclusions about causality.[28,29] What we can ascertain is that they are registered in clinical databases at the same time in a very high percentage of cases. An excess mortality risk associated with depression and the lower incidence could explain the lower prevalence in dementia patients older than 85 years. On the other hand, observational studies have consistently pointed out that older patients are, in general, more likely to be prescribed with antidepressants.[30] The explanation may lie in the different meaning of being identified by the predictive model and the use of antidepressant treatment in general, drugs that cause drowsiness like trazodone and mirtazapine often being used as hypnotics rather than antidepressants. As our results also evidenced (Table SM6), these two drugs are currently the antidepressants most consumed by patients with dementia in our setting and their prescribing rates increase with age. On the contrary, the percentage of patients using escitalopram is lower in over-85-year-olds. Consistent with this, at that age, the incidence of depressive NPS is lower, its main drivers being use of escitalopram and multiple changes of antidepressant.[14]

In conclusion, in this study, we have found a notably high prevalence and incidence of NPS, with different age and sex patterns for the depressive and psychotic clusters, a correlation between incident cases of dementia and NPS, and an incidence of dementia measured in RWD consistent with that in the literature. Our results evidence the benefit of merging big data and RWD tools, to address the challenge of quantifying the behavioural dimension of dementia in epidemiological terms.

Ethics approval

The authors assert that all procedures contributing to this work comply with the ethical standards of the relevant national and institutional committees on human experimentation and with the Helsinki Declaration of 1975, as revised in 2008. All procedures involving human subjects/patients were approved by the Clinical Research Ethics Committee of the Basque Country (PI2018143).

Authors' contributions

JM conceived and designed the research. OI, AG and AA (Osakidetza) obtained the data and interpreted the data. AA (MU) and CC designed the machine learning methods, performed the analyses and interpreted the data. AI, LM-B and MT designed the research, interpreted the data, and critically revised the manuscript. OI, LM-B and JM drafted the manuscript and approved the final manuscript. AA (MU), CC, AA (Osakidetza), OI, II, AG, and CC revised the manuscript for important intellectual content and approved the final manuscript. All authors had full access to the full data in the study and accept responsibility to submit for publication.

Funding

The study was funded by two grants from the Basque Foundation for Health Innovation and Research (BIOEF) (grant number BIOD17/ND/015) and Gipuzkoa Regional Government (Adinberri program). The funding sources had no involvement in study design in the collection, analysis and interpretation of data, in the writing of the report; and in the decision to submit the article.

Conflicts of interest

The authors have no conflicts of interest to declare.

Acknowledgement

We would like to acknowledge the help of Ideas Need Communicating Language Services in improving the use of English in the manuscript.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.rpsm.2021.03.001.

References

1. Ballard C, Gauthier S, Corbett A, Brayne C, Aarsland D, Jones E. Alzheimer's disease. Lancet. 2011;377:1019–31, http://dx.doi.org/10.1016/S0140-6736(10)61349-9.
2. Rocca P, Leotta D, Liffredo C, Mingrone C, Sigaudo M, Capellero B, et al. Neuropsychiatric symptoms underlying caregiver stress and insight in Alzheimer's disease. Dement Geriatr Cogn Disord. 2010;30:57–63, http://dx.doi.org/10.1159/000315513.
3. Backhouse T, Camino J, Mioshi E. What do we know about behavioral crises in dementia? A systematic review. J Alzheimers Dis. 2018;62:99–113, http://dx.doi.org/10.3233/JAD-170679.
4. Lyketsos CG, Lopez O, Jones B, Fitzpatrick AL, Breitner J, DeKosky S. Prevalence of neuropsychiatric symptoms in dementia and mild cognitive impairment: results from the cardiovascular health study. JAMA. 2002;288:1475–83, http://dx.doi.org/10.1001/jama.288.12.1475.
5. Ford AH. Neuropsychiatric aspects of dementia. Maturitas. 2014;79:209–15, http://dx.doi.org/10.1016/j.maturitas.2014.04.005.
6. Zhao Q-F, Tan L, Wang H-F, Jiang T, Tan M-S, Tan L, et al. The prevalence of neuropsychiatric symptoms in Alzheimer's disease: systematic review and meta-analysis. J Affect Disord. 2016;190:264–71, http://dx.doi.org/10.1016/j.jad.2015.09.069.

7. Perera G, Pedersen L, Ansel D, Alexander M, Arrighi HM, Avillach P, et al. Dementia prevalence and incidence in a federation of European Electronic Health Record databases: The European Medical Informatics Framework resource. Alzheimers Dement. 2018;14:130–9, http://dx.doi.org/10.1016/j.jalz.2017.06.2270.
8. Perera G, Pedersen L, Ansel D, Alexander M, Arrighi HM, Avillach P, et al. Validity of a computerized population registry of dementia based on clinical databases. Neurologia. 2018, http://dx.doi.org/10.1016/j.nrl.2018.03.005.
9. Mar J, Arrospide A, Soto-Gordoa M, Machón M, Iruin Á, Martinez-Lage P, et al. Incidence of dementia and major subtypes in Europe: a collaborative study of population-based cohorts. Neurologic diseases in the elderly research group. Neurology. 2000;54:S10–5.
10. Berger ML, Sox H, Willke RJ, Brixner DL, Eichler H-G, Goettsch W, et al. Good practices for real-world data studies of treatment and/or comparative effectiveness: recommendations from the Joint ISPOR-ISPE Special Task Force on Real-World Evidence in Health Care Decision Making. Value Health. 2017;20:1003–8, http://dx.doi.org/10.1016/j.jval.2017.08.3019.
11. Steinhausen H-C, Jakobsen H. Incidence rates of treated mental disorders in childhood and adolescence in a complete nationwide birth cohort. J Clin Psychiatry. 2019;80, http://dx.doi.org/10.4088/JCP.17m12012.
12. Mar J, Arrospide A, Soto-Gordoa M, Iruin Á, Tainta M, Gabilondo A, et al. Dementia-related neuropsychiatric symptoms: inequalities in pharmacological treatment and institutionalization. Neuropsychiatr Dis Treat. 2019;15:2027–34, http://dx.doi.org/10.2147/NDT.S209008.
13. Obermeyer Z, Emanuel EJ. Predicting the future – big data, machine learning, and clinical medicine. N Engl J Med. 2016;375:1216–9, http://dx.doi.org/10.1056/NEJMp1606181.
14. Mar J, Gorostiza A, Ibarrondo O, Cernuda C, Arrospide A, Iruin Á, et al. Validation of random forest machine learning models to predict dementia-related neuropsychiatric symptoms in real-world data. J Alzheimers Dis. 2020, http://dx.doi.org/10.3233/JAD-200345.
15. Cummings JL, Mega M, Gray K, Rosenberg-Thompson S, Carusi DA, Gornbein J. The neuropsychiatric inventory: comprehensive assessment of psychopathology in dementia. Neurology. 1994;44:2308–14.
16. Vandenbroucke JP, Pearce N. Incidence rates in dynamic populations. Int J Epidemiol. 2012;41:1472–9, http://dx.doi.org/10.1093/ije/dys142.
17. Herrmann LK, Welter E, Leverenz J, Lerner AJ, Udelson N, Kanetsky C, et al. A systematic review of dementia-related stigma research: can we move the stigma dial? Am J Geriatr Psychiatry. 2018;26:316–31, http://dx.doi.org/10.1016/j.jagp.2017.09.006.
18. Gove D, Downs M, Vernooij-Dassen M, Small N. Stigma and GPs' perceptions of dementia. Aging Ment Health. 2016;20:391–400, http://dx.doi.org/10.1080/13607863.2015.1015962.
19. Ponjoan A, Garre-Olmo J, Blanch J, Fages E, Alves-Cabratosa L, Martí-Lluch R, et al. Epidemiology of dementia: prevalence and incidence estimates using validated electronic health records from primary care. Clin Epidemiol. 2019;11:217–28, http://dx.doi.org/10.2147/CLEP.S186590.
20. van Bussel EF, Richard E, Arts DL, Nooyens ACJ, Coloma PM, de Waal MWM, et al. Dementia incidence trend over 1992–2014 in the Netherlands: analysis of primary care data. PLoS Med. 2017;14:e1002235, http://dx.doi.org/10.1371/journal.pmed.1002235.
21. Ferri CP, Prince M, Brayne C, Brodaty H, Fratiglioni L, Ganguli M, et al. Global prevalence of dementia: a Delphi consensus study. Lancet. 2005;366:2112–7, http://dx.doi.org/10.1016/S0140-6736(05)67889-0.
22. Niu H, Álvarez-Álvarez I, Guillén-Grima F, Aguinaga-Ontoso I. Prevalence and incidence of Alzheimer's disease in Europe: a meta-analysis. Neurologia. 2017;32:523–32, http://dx.doi.org/10.1016/j.nrl.2016.02.016.
23. Steinberg M, Sheppard J-M, Tschanz JT, Norton MC, Steffens DC, Breitner JCS, et al. The incidence of mental and behavioral disturbances in dementia: the cache county study. J Neuropsychiatry Clin Neurosci. 2003;15:340–5, http://dx.doi.org/10.1176/jnp.15.3.340.
24. Borsje P, Wetzels RB, Lucassen PL, Pot AM, Koopmans RT. The course of neuropsychiatric symptoms in community-dwelling patients with dementia: a systematic review. Int Psychogeriatr. 2015;27:385–405, http://dx.doi.org/10.1017/S1041610214002282.
25. Gallacher J, de Reydet de Vulpillieres F, Amzal B, Angehrn Z, Bexelius C, Bintener C, et al. Challenges for optimizing real-world evidence in Alzheimer's disease: the ROADMAP Project. J Alzheimers Dis. 2019;67:495–501, http://dx.doi.org/10.3233/JAD-180370.
26. Gracia-García P, de-la-Cámara C, Santabárbara J, Lopez-Anton R, Quintanilla MA, Ventura T, et al. Depression and incident Alzheimer disease: the impact of disease severity. Am J Geriatr Psychiatry. 2015;23:119–29, http://dx.doi.org/10.1016/j.jagp.2013.02.011.
27. Meyers BS, Bruce ML. The depression–dementia conundrum: integrating clinical and epidemiological perspectives. Arch Gen Psychiatry. 1998;55:1082–3, http://dx.doi.org/10.1001/archpsyc.55.12.1082.
28. Brown EE, Rajji TK, Mulsant BH. Why do some older adults treated with antidepressants progress to dementia? J Clin Psychiatry. 2020;81, http://dx.doi.org/10.4088/JCP.20com13559.
29. Bartels C, Belz M, Vogelgsang J, Hessmann P, Bohlken J, Wiltfang J, et al. To be continued? Long-term treatment effects of antidepressant drug classes and individual antidepressants on the risk of developing dementia: a German case–control study. J Clin Psychiatry. 2020;81, http://dx.doi.org/10.4088/JCP.19m13205.
30. Loeb DF, Ghushchyan V, Huebschmann AG, Lobo IE, Bayliss EA. Association of treatment modality for depression and burden of comorbid chronic illness in a nationally representative sample in the United States. Gen Hosp Psychiatry. 2012;34:588–97, http://dx.doi.org/10.1016/j.genhosppsych.2012.07.004.