Estimating disease prevalence from drug utilization data using the Random Forest algorithm

Slobbe, Laurentius C J; Füssenich, Koen; Wong, Albert; Boshuizen, Hendriek C; Nielen, Markus M J; Polder, Johan J; Feenstra, Talitha L; Van Oers, Hans A M

Published in:
European Journal of Public Health

DOI:
10.1093/eurpub/cky270

Publication date:
2019

Document Version
Publisher's PDF, also known as Version of record

16 Jayaraman J, Joseph K. Determinants of place of death: a population-based retrospective cohort study. *BMC Palliat Care* 2013;12:19.

17 Öhlén J, Cohen J, Håkanson C. Determinants in the place of death for people with different cancer types: a national population-based study. *Acta Oncol* 2017;56:455–61.

18 Costantini M, Balzi D, Garronec E, et al. Geographical variations of place of death among Italian communities suggest an inappropriate hospital use in the terminal phase of cancer disease. *Public Health* 2000;114:15–20.

19 Håkanson C, Öhlén J, Morin L, Cohen J. A population-level study of place of death and associated factors in Sweden. *Scand J Public Health* 2015;43:744–51.

20 Sleeman KE, Ho YK, Verne J, et al. Place of death, and its relation with underlying cause of death, in Parkinson's disease, motor neurone disease, and multiple sclerosis: a population-based study. *Palliat Med* 2013;27:840–6.

21 Ruiz-Ramos M, Garcia-Leon FJ, Mendez-Martinez C. Place of death in Andalusia: influence of age, gender and cause of death. *Rev Clin Esp* 2011;211:127–32.

22 Dominguez-Berjon MF, Esteban-Vasallo MD, Zoni AC, et al. Place of death and associated factors among patients with Amyotrophic Lateral Sclerosis in Madrid (Spain). *Amyotroph Lateral Scler Frontotemporal Degener* 2015;17:62–8.

23 Reyniers T, Deliens L, Pasman HR, et al. International variation in place of death of older people who died from Dementia in 14 European and non-European countries. *J Am Med Dir Assoc* 2015;16:165–71.

24 Organización Panamericana de la Salud, editor. *Clasificación Estadística Internacional de Enfermedades y Problemas Relacionados con la Salud*. Décima Revisión. Vol. 1, 1–1067. Pan American Health Org., 1995.

25 Murtagh FE, Bausewein C, Verne J, et al. How many people need palliative care? A study developing and comparing methods for population-based estimates. *Palliat Med* 2014;28:49–58.

26 Nelson W. Statistical methods for the ratio of two multinomial proportions. *Am Stat* 1972;26:22–7.

27 Sarmento VP, Higginson IJ, Ferreira PL, Gomes B. Past trends and projections of hospital deaths to inform the integration of palliative care in one of the most ageing countries in the world. *Palliat Med* 2016;30:363–73.

28 Kane PM, Daveson BA, Ryan K, et al. The need for palliative care in Ireland: a population-based estimate of palliative care using routine mortality data, inclusive of nonmalignant conditions. *J Pain Symptom Manage* 2015;49:726–33.

29 Woitha K, Garralda E, Martin-Moreno JM, et al. Ranking of palliative care development in the countries of the European Union. *J Pain Symptom Manage* 2016;52:370–7.

30 Escobar-Pinzón LC, Weber M, Claus M, et al. Factors influencing place of death in Germany. *J Pain Symptom Manage* 2011;41:893–903.

31 Houttekier D, Cohen J, Pepersack T, Deliens L. Dying in hospital: a study of incidence and factors related to hospital death using death certificate data. *Eur J Public Health* 2014;24:751–6.

The European Journal of Public Health, Vol. 29, No. 4, 615–621

doi:10.1093/eurpub/cky270 Advance Access published on 3 January 2019

Estimating disease prevalence from drug utilization data using the Random Forest algorithm

Laurentius C.J. Slobbe[1,2,*,†], Koen Füssenich[1,3,*], Albert Wong[1], Hendriek C. Boshuizen[1,4], Markus M.J. Nielen[1,5], Johan J. Polder[1,2], Talitha L. Feenstra[1,3], Hans A.M. van Oers[1,2]

1 National Institute for Public Health and the Environment (RIVM), Bilthoven, The Netherlands
2 Tilburg University, Department Tranzo, Tilburg, The Netherlands
3 Groningen University, University Medical Center, Department of Epidemiology, Groningen, The Netherlands
4 Wageningen University and Research, Wageningen, The Netherlands
5 Netherlands Institute for Health Services Research (NIVEL), Utrecht, The Netherlands

*Both authors contributed equally to this work.

†Deceasd 27 February 2018.

Correspondence: Koen Füssenich, RIVM, Department of Quality of Health Care and Health Economics, PO Box 1, 3720 BA Bilthoven, The Netherlands, Tel: +31 (0) 30 2743696, Fax: +31 (0)30 274 29 71, e-mail: koen.fussenich@rivm.nl

Background: Aggregated claims data on medication are often used as a proxy for the prevalence of diseases, especially chronic diseases. However, linkage between medication and diagnosis tend to be theory based and not very precise. Modelling disease probability at an individual level using individual level data may yield more accurate results. **Methods:** Individual probabilities of having a certain chronic disease were estimated using the Random Forest (RF) algorithm. A training set was created from a general practitioners database of 276 723 cases that included diagnosis and claims data on medication. Model performance for 29 chronic diseases was evaluated using Receiver-Operator Curves, by measuring the Area Under the Curve (AUC). **Results:** The diseases for which model performance was best were Parkinson's disease (AUC = .89, 95% CI = .77–1.00), diabetes (AUC = .87, 95% CI = .85–.90), osteoporosis (AUC = .87, 95% CI = .81–.92) and heart failure (AUC = .81, 95% CI = .74–.88). Five other diseases had an AUC >.75: asthma, chronic enteritis, COPD, epilepsy and HIV/AIDS. For 16 of 17 diseases tested, the medication categories used in theory-based algorithms were also identified by our method, however the RF models included a broader range of medications as important predictors. **Conclusion:** Data on medication use can be a useful predictor when estimating the prevalence of several chronic diseases. To improve the estimates, for a broader range of chronic diseases, research should use better training data, include more details concerning dosages and duration of prescriptions, and add related predictors like hospitalizations.

Introduction

Information on disease prevalence is important for assessing the health needs of populations.[1] Several sources can deliver population disease prevalence estimates, such as surveys,[2–4] dedicated epidemiologic studies using diagnostics[5–7] or administrative data sources.[8–12] Drug use data, especially on prescription drugs, has also frequently been used to estimate disease prevalence.[13,14] In many countries insurers or providers maintain extensive prescription databases, allowing easy access to national drug use data.

Drug use has several advantages over other sources. Surveys are costly to execute on a large scale. Hospital discharge registers are large, but involve hospital-related events only. In addition, in the Netherlands, the GP-serves as a gatekeeper, implying that patients—except in emergencies—can only visit a medical specialist with a referral of the GP. This means that the GP sees both patients that see only the GP and those he refers to specialist care. Hospital data is therefore more likely that GP data to underestimates the prevalence.[15,16] GP-registers containing diagnosis codes are not readily available in all countries. Furthermore, GPs may have different coding habits, hindering comparisons between GPs.[17]

While drug use data is often recorded without a diagnosis, some studies base disease prevalence estimates on direct links of specific drug use to the presence of certain diseases.[9,13,14,18,19] The links are based on literature or medical guidelines. For two reasons, this procedure is problematic. First, many drugs are used for the treatment of multiple diseases; assuming that all patients who take a specific drug do have a specific disease will then lead to overestimation. Second, some patients with a disease are not prescribed the specific drug, and this will lead to underestimation of prevalence.

To overcome these two problems, it is better to estimate the probability of having a specific disease given all different medications a person uses. Avoiding any *a priori* assumption on the relationship between the drugs and diagnoses, machine learning algorithms can be used to estimate this relation from data. In this paper more specifically the Random Forest (RF) algorithm will be applied, as this method yielded the best results in comparison with others.[20,21]

This algorithm requires a test set with both diagnosis data and drug use. This diagnosis data could also be used directly to estimate disease prevalence. This is the case particularly when it is possible to assume that the set containing diagnosis data is representative for the population of interest. However, using the diagnosis data in combination with drug use as proposed alters the assumption. Rather than that the diagnosis data should be representative for the population of interest, the relationship between diagnosis and drug use should be similar as in the population of interest. This might be a more reasonable assumption in many cases, as medical professionals are influenced by standardized prescription guidelines. Countries which do have a prescription registration, but lack population surveys on disease prevalence, as is often the case, can use the relation derived in comparable countries to obtain prevalence estimates.

Existing applications of RF analysis to the problem of disease prevalence estimates have some limitations. Chaudhry[20] used RF to predict the population prevalence of diabetes and dementia from administrative data in GP and hospital records. However, his choice of predictors was informed by *a priori* knowledge. Khalilia et al.[21] predicted the presence of eight diseases with RF from hospital in-patient data, but did not make any population prevalence estimates.

In contrast, we apply the RF approach to a broad range of 29 diseases. The RF algorithm allows us to select important predictors from the full range of possible drug use predictors. Afterwards, we have a list of predictors for comparison with existing theory based lists of predictors, e.g. the Dutch Pharmacotherapeutic compass.[22]

The objective of this paper hence is to examine for which diseases the prevalence can be estimated using the RF algorithm, and if so, to see which drug groups should be used.

Methods

Random Forest

Estimating the probability that an individual has a certain disease could be considered a mathematical classification problem. RF is a non-parametric method to address classification problems.[23] For implementation the R-package 'Random Forest'[24] was used.

Data

Drug use data of the entire Dutch population is available from the National Health Care Institute (ZiN).[25] The ZiN claims database covers all outpatient prescriptions reimbursed under the Dutch mandatory Health Insurance scheme. Drugs were classified in 204 pharmaceutical groups according to the four position ATC-code. To these groups, age and gender were added as predictors. The dataset contained 47 million individual prescription records in 2010, covering a population of 16.7 million, of which 70% had at least one prescription.

A training set with disease information was obtained from the primary care database of the Netherlands Institute for Health Services Research (NIVEL).[26] As every citizen is required to have a GP—with the exception of those living institutionalized—this means the dataset is likely to cover the whole Dutch population, with the exception of the 80+ population of which in 2010 a significant part lived institutionalized.[27] All patient contacts were labelled with a diagnostic code, ICPC.[28] A person was defined to have a disease when he/she had at least one contact with a GP for this disease over a period of 3 years. All GP-patients with full data available over 2008–2010 were selected. This resulted in a training set of 276 723 individuals. The selection of 29 diseases was based on a list provided by O'Halloran et al.[29] See Supplementary file S1 for details. We combined the available data (drug utilization, age and gender, and ICPC codes) at Statistics Netherlands within the System of Social Statistical Datasets (SSD). The SSD allows data from different administrative registers to be combined using an anonymous patient identifier for research purposes.[30,31]

Implementation of RF

Usually, all observations in a training set and all predictors are combined in one RF-analysis. However, within the SSD system, computing power is limited, and analysis with our dataset (276 723 records with 206 variables) proved to be difficult. We therefore used a two-step approach. First, for each chronic disease, persons with the disease were randomly selected, up to a maximum of 5000 patients. To this set, an equal number of persons without the disease was randomly selected and added. For each of these smaller sets, the RF algorithm was applied. The variable importance measure,[24] defined as the average decrease in accuracy when a predictor is left out of the analysis, was evaluated. For each disease, the 10 drug groups with the highest variable importance were selected. By selecting 10 drug groups, the most important predictors were included for all diseases, while limiting the computing times. Second, a new dataset was created for each disease based on the full training set, but only age, gender and the drug groups selected in the first step were added as predictors (276 723 records with 12 variables for each disease), and we applied RF a second time. For each disease this second RF-model was then applied to obtain the probability of having this disease for each individual in the prescriptions database, hence for the 11.6 million Dutch inhabitants that were reimbursed a prescription drug in 2010. The model was also applied to the remaining 5. million Dutch individuals without any prescription. They received for each of the 29 diseases a probability equivalent to the age and gender specific probability in the training set for those diagnosed with the disease, but not receiving any prescription.

Table 1 Pharmaceutical utilization in dataset

	Persons with at least one recorded episode for each chronic disease in 2008–2010	Total number of pharmaceutical groups utilized in 2010	Average number of pharmaceutical groups utilized per person
Persons without disease	184 826	328 385	*1.8*
Persons with 1 chronic disease	60 065	235 032	*3.9*
Persons with 2 chronic diseases	20 090	125 335	*6.2*
Persons with 3 chronic diseases	7609	63 064	*8.2*
Persons with 4 chronic diseases	2697	27 190	*9.9*
Persons with 5 or more chronic diseases	1436	17 219	*11.6*
Total	276 723	796 225	*2.9*
Percentage with at least one chronic disease:	33.2%		
Percentage study population with multiple diseases:	11.5%		

Legend: Training set population has been divided into six strata, based on the number of chronic diseases present. First column presents stratum. Second column gives population size. Third column gives total number of pharmaceutical groups utilized. Pharmaceutical groups have been defined in terms of an ATC 4 position code: A01A, A02A, etc. Last column gives average utilization in stratum.

Table 2 Model outcome AUC with confidence interval, ordered by mean AUC

Disease	AUC (95% conf. interval)	Prevalence in training set per 10 000 persons
Parkinson's disease	*.89 (.77–1.00)*	15
Diabetes mellitus	*.87 (.85–.90)*	421
Osteoporosis	*.87 (.81–.92)*	103
Heart failure	*.81 (.74–.88)*	82
Chronic obstructive pulmonary disease	*.79 (.75–.83)*	209
Chronic enteritis/colitis ulcerosa	*.79 (.68–.90)*	31
HIV/AIDS	*.78 (.39–1.00)*	4
Asthma	*.77 (.74–.80)*	424
Epilepsy	*.77 (.66–.87)*	41
Coronary heart disease	*.70 (.66–.74)*	255
Visual disorder	*.69 (.64–.73)*	191
Schizophrenia	*.69 (.48–.89)*	10
Rheumatoid arthritis	*.68 (.60–.76)*	66
Dementia	*.67 (.54–.80)*	28
Congenital neurological anomaly	*.67 (.01–1.00)*	3
Multiple sclerosis	*.66 (.42–.90)*	9
Cancer	*.60 (.56–.64)*	264
Chronic alcohol abuse	*.59 (.49–.69)*	45
Depressive disorder	*.58 (.54–.63)*	253
Stroke (including TIA)	*.57 (.52–.63)*	137
Congenital cardiovascular anomaly	*.57 (.37–.77)*	7
Chronic back or neck disorder	*.56 (.53–.60)*	432
Osteoarthritis	*.56 (.52–.60)*	282
Anxiety disorder, neurosis, PTSS	*.56 (.50–.61)*	154
Mental retardation	*.55 (.36–.74)*	13
Hearing disorder	*.52 (.44–.61)*	62
Anorexia	*.52 (.33–.71)*	8
Gastric or duodenal ulcer	*.50 (.39–.62)*	25
Tuberculosis	*.50 (.06–.94)*	2

Legend: First column gives name of chronic disease. Second column gives model outcome of RF-analysis as AUC with 95% confidence interval, in order of decreasing AUC. Third column states prevalence of chronic disease or condition in the training set. (n = 276 723).

Outcome measures

For each disease, the most important drugs according to the variable importance were compared with theoretical drug classifications included in relevant guidelines. For 13 of the 29 chronic diseases pharmaceutical groups used in the Dutch insurance system were available.[32] In addition, for four other diseases the drugs found were compared with Dutch treatment guidelines: tuberculosis,[33] MS,[34] chronic back or neck disorder[35,36] and gastric or duodenal ulcer.[37]

To measure the performance of the final RF-models, the area under the Receiver-Operator Curve (AUC) was measured for the training set for each disease separately. An AUC-value above .7 is generally considered useful.[38] To prevent overfitting, 10-fold cross validation was applied.

The AUC and a 95% confidence interval around the AUC-value were obtained using the R-package 'cvAUC'.[39] If the lower boundary of this interval was above .5, we considered the model to perform better than a random prediction.

The predicted population prevalence by age and gender for the Netherlands was graphically compared with a prevalence estimate based on direct extrapolation of the training set prevalence. Correlations were computed as well for the six diseases with lower confidence bound (95%) of the AUC >.70. The age range considered was 30–80 years, since the prevalence below 30 is very low for most chronic diseases and the 80+ population was not well covered in our training set.

For a binary classification of each individual, a cut-off needs to be chosen. This was done by setting an age and disease-specific cut-off value. All persons with a probability higher than the cut-off were classified as 'ill'. The cut-off was chosen to minimize the deviation between the observed and the predicted prevalence in the training set for each age, gender and disorder.

Results

Table 1 gives descriptives for the training set. The average annual number of different pharmaceutical drugs taken by patients in the training set was 2.9, which is very comparable with the utilization in the total Dutch population in the same year (2.8). Table 1 also shows that the number of ATC groups utilized by an individual patient rises proportionally with the number of chronic diseases present.

Table 2 lists the AUC values produced by our analysis, sorted by average AUC. For 17 diseases the lower boundary of the 95% AUC confidence interval was >.5. For 10 diseases the average AUC was .7 or higher, but for only six the lower boundary of the AUC 95% confidence interval was >=.7: Parkinson's disease, diabetes mellitus, osteoporosis, heart failure, asthma and chronic obstructive pulmonary disease (COPD).

There is some association between the frequency of the disease and the prediction of the AUC. For almost all 12 diseases with a prevalence in the training set higher than 100 per 10 000 persons, the prediction is better than a random assignment. The only exception is anxiety disorder (154 cases per 10 000 persons), with a very poor performance and AUC of .56 (95% cf. = .50–.61). For 11 out of 17 diseases with a frequency below 100 per 10 000 persons, performance is poor, i.e. the lower boundary of the AUC 95% confidence interval was below .5. A notable exception was Parkinson's disease which

Table 3 Predictors of chronic diseases in Random Forest analysis

Disease	ATC4 groups with strongest relation with disease in RF-analysis									
	[1]= strongest relation, [10]= weakest relation									
	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]
Parkinson's disease[a]	N04B	N04A	Birthyear	L04A	A07E	C10A	C09B	N05A	C03C	N06D
Diabetes mellitus[a]	A10A	A10B	C10A	B01A	H04A	C09A	C09B	C08C	C03C	C07A
Osteoporosis	M05B	A12A	A11C	L04A	C10A	B01A	C09D	D01A	C09C	D06A
Heart failure	C03C	C03D	A12B	C01A	C08D	C08C	C01D	C03A	C10A	C07B
Chronic obstructive pulmonary disease[a]	R03A	R03B	R06A	A01A	R01A	N06B	J01F	R05D	A06A	J01C
Chronic enteritis/colitis ulcerosa	A07E	L04A	L01B	B03B	A06A	J01M	A11C	A07D	N02A	M01A
HIV/AIDS[a]	J05A	J01E	J04A	J01F	N01B	J07B	D06B	A02B	J01C	J01A
Asthma[a]	R03A	R03B	R03C	Birthyear	H02A	A07A	S01G	R01A	R06A	R05D
Epilepsy[a]	N03A	N05B	Birthyear	A03F	N05A	B01A	N06A	D04A	D11A	N05C
Coronary heart disease[a]	C01D	C08D	C03C	C01B	C03A	C09A	D06A	C01E	C09C	B03A
Visual disorder	S01E	S01B	Birthyear	S01C	A10B	S01F	D02A	S01A	A10A	S01X
Schizophrenia[a]	N05A	N05B	N05C	N04A	N06A	N06B	Birthyear	N03A	A06A	N07C
Rheumatoid arthritis[a]	L04A	P01B	A07E	B03B	N02A	L01B	H02A	D02B	M05B	D06A
Dementia[a]	N06D	Birthyear	N05A	A12A	C03C	N03A	M05B	Y	D03D	C09D
Congenital neurological anomaly	M03B	G04B	N03A	J01X	D07X	N05A	J01E	A12A	D01A	N05B
Multiple sclerosis[b]	L03A	M03B	G04B	N03A	N06A	N04B	B03B	S01A	J01X	C03C
Cancer[a]	Birthyear	L02B	H03A	Y	D06A	A04A	Gender	L02A	G03C	A12A
Chronic alcohol abuse[a]	N07B	N05A	Birthyear	N05B	G04C	A02B	N06A	M04A	A10B	N05C
Depressive disorder[a]	N06A	N05A	N05B	N06B	N05C	N07B	N03A	A03F	A11C	G04B
Stroke (including TIA)	B01A	V03A	C01D	C01B	C07A	Birthyear	C08C	C01A	S01C	S01E
Congenital cardiovascular anomaly	B01A	C07A	J01C	N03A	C09A	D06A	R03B	Y	D02A	S02C
Chronic back or neck disorder[b]	N02A	M01A	A02B	A06A	N02B	C05A	S02C	N03A	H02A	R05D
Osteoarthritis	Birthyear	B01A	Gender	M04A	C10A	N02B	C10B	C03A	S01E	N02A
Anxiety disorder, neurosis, PTSS	N06A	N05B	N05A	C07A	N01B	N05C	A03F	A06A	N03A	D05A
Mental retardation	N05A	N03A	D02A	D06A	A06A	N05B	Y	D10A	S01F	N01B
Hearing disorder	Birthyear	L02A	B02A	S01X	D05A	G04C	H02A	A10B	C08D	C07B
Anorexia	Gender	G03A	A06A	A01A	Y	J01X	G03H	R05D	G01A	A12B
Gastric or duodenal ulcer[b]	A02B	D05A	G04B	A07A	A03A	A03F	M01A	A11C	D06B	A06A
Tuberculosis[b]	J04A	D02A	D07X	C01B	J01A	C08C	C03C	S01C	S02C	C03E

Legend: First column gives name of chronic disease. The next 10 columns list the predictors used in the final RF-model, in order of decreasing importance. To facilitate comparison, diseases are presented in the same order as in table2.

a: For these diseases, comparison is possible with pharmaceutical groups listed in risk adjustment compulsory insurance. The shaded groups are also used for the detection of these diseases by Dutch insurers.[32]

b: These diseases have been compared with ATC-groups mentioned in the relevant Dutch treatment guidelines. The shaded groups are included in these guidelines.[33–37]

despite a low frequency (15 per 10 000 persons) seems to be very predictable from drugs utilization.

In table 3 predictors of all model output from the RF-analysis are ranked by importance. The shaded areas denote drugs that are also mentioned as indicator drugs for these diseases in the theoretical drug classifications we compared with. Only for cancer we found no similarities. For all other diseases, the ATC codes mentioned by insurers and guidelines are also strong predictors for the corresponding diseases in our RF-models. However, our models show a number of additional predictors for most disorders. Supplementary file S2 gives more information.

The actual prevalence in the training set and the calculated prevalence based on applying the final RF-models have been compared for the six diseases with a lower bound of the AUC 95% confidence interval >.7. Except Asthma, correlations are above .9. Asthma shows correlations of .43 for males and .66 for females, indicating poor performance. Looking at the graphs for osteoporosis, a large discrepancy exists between predicted and observed prevalence around the age of 70. Figure 1 gives an example (COPD, male). A full set of figures is found in Supplementary file S3.

Discussion

For a broad range of 29 diseases, RF was used to predict disease prevalence based on medication use. Predictive performance was acceptable for 6 out of 29 diseases and would result in reliable estimates of population prevalence. Furthermore, we find that theory-based indicator drugs were included in the range of diseases identified by the RF model. This seems to be independent from the performance of the models, which indicates that the RF algorithm can also be used to identify suitable predictors, even in those cases were the predictive performance is low. Especially for diabetes, heart failure and COPD we observe a high correlation between estimated and observed population.

Our outcomes can be compared with a few other studies. Chaudhry[20] predicted the presence of diabetes with an AUC of .95 and dementia with an AUC of .875, higher than the .87 and .67 we found. However, for dementia he used dementia-coded doctor visits as predictors, while we use this as our definition of disease. Khalilia et al.[21] used data on hospital stays as predictors, on a very large set (8 million records). A training set was generated by bootstrapping. The average AUC he reports (.88) is much higher than those we found. For the two diseases which could be directly compared (diabetes and osteoporosis) he finds almost the same AUC (.879 and .870 respectively) as we found, .87 for both. Compared with these two previous studies, we included a relatively broad range of diseases and added the comparison with theory-based models.

While the method seems useful for some diseases, the predictive performance is still low for most diseases. This could have multiple causes. First, for some diseases, there is no standard pattern of drugs included in all treatment options. In addition, drugs might be

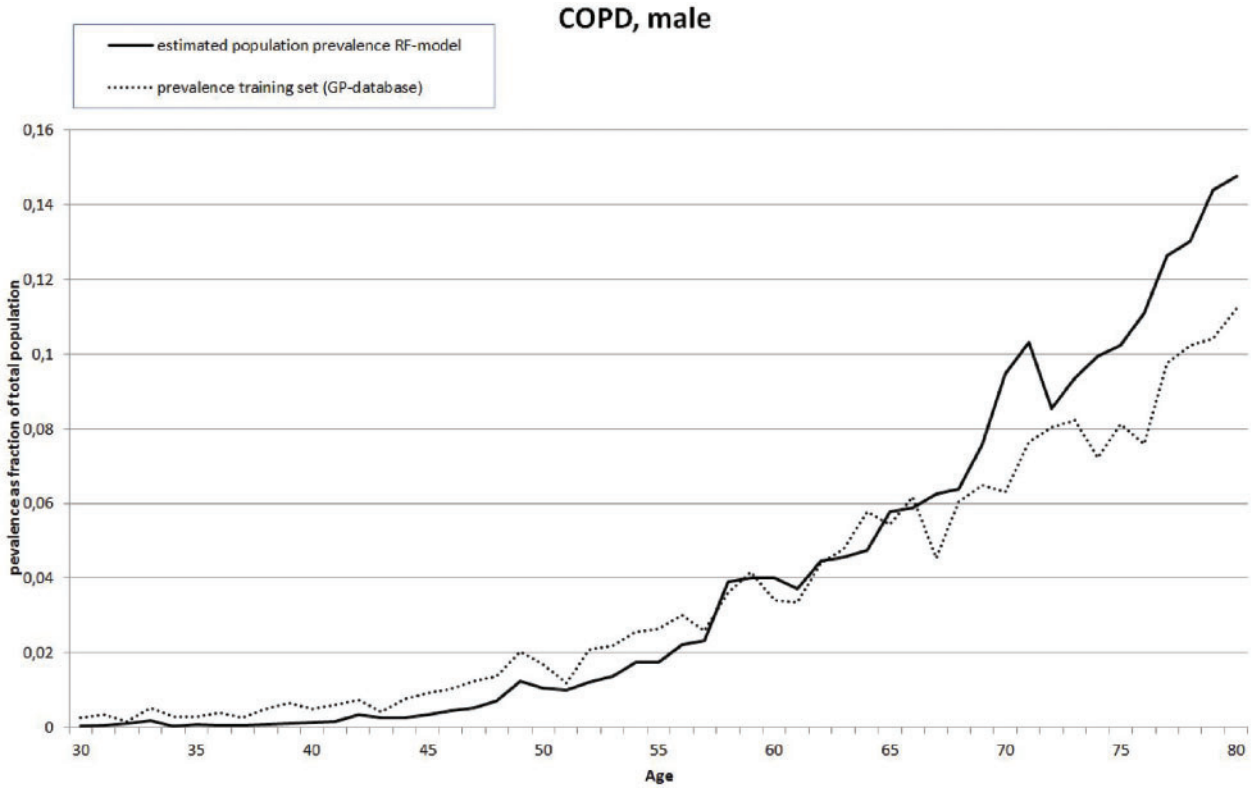

Figure 1 Example of comparison between Dutch population prevalence for ages 30–80 estimated from model applied to drug utilization data and estimation based on training set for COPD, male

prescribed for multiple diseases. For instance, the two strongest predictors for asthma and COPD are the same (R03A and R03B, table 3). As a result, misclassification of asthma and COPD patients is likely to occur, which has not been further investigated in this study. Furthermore, patients and GPs may deal with diseases in different ways. Based on patient characteristics a GP will sometimes advise lifestyle changes instead of drugs, but will treat similar cases in other instances immediately with drugs. In addition, the patient may have treatment preferences. The relationship between diagnosis and drugs can also change over time. Innovation or policy changes can strongly influence prescription behaviour, making regular calibration of the algorithms necessary.

Second, the predictive power is likely limited due to weaknesses of the current data. In the current training set, only 3 years of diagnoses are used. While many patients with a chronic disease are visiting a GP more than once every 3 years, some patients who visit less frequently will not occur as diseased in the training set. Furthermore, some diseases might not be treated primarily by a GP, but directly in the hospital, also resulting in missing diagnoses in the training set. As the training set serves as a 'golden standard', any diagnosis errors in the training set will translate into the final predictions. Investing in a smaller set of persons for which disease diagnosis is even more reliable, e.g. through the use of cohort studies may provide a training set with better performance. The disadvantage of such a cohort, and the advantage of our current approach is that for rare diseases, relationships between disease and drugs would have to be derived from only a very limited number of disease cases.

Next to errors in disease diagnosis, drugs use measures could also be improved. Drug use often varies between years. Grouping multiple years of drug use could improve results. Also, more complete drug utilization data could be obtained by including inpatient drugs. For some diseases, utilizing more detailed pharmaceutical predictors, such as ATC4 or ATC5 groups, would improve results.

Even though improvements might be needed to obtain reliable prevalence estimates for most diseases, for 16 out of 17 diseases for which theory-based predictors were found within existing guidelines, important similarities were found. This means that even though the predictive power of the algorithm on the current data is insufficient, it is still possible to identify relevant drug groups. Compared with purely theory-based models, the RF algorithms have the important advantage of coming with confidence intervals and information about model performance. From this similarity we also infer that Dutch general practioners broadly follow existing pharmaceutical guidelines. Cancer was the only disease for which the drugs found using the RF algorithm differ from theory. This could be the result of grouping all cancers together, and many drugs used in cancer treatment were not covered by the dataset as they were prescribed in a hospital setting.

We do not want to suggest that prediction models can entirely replace current GP registers or population surveys. On the contrary, since without these registries the models cannot be built or validated. However, even in countries like the Netherlands which are covered by both population surveys and GP networks the method is of practical value, as it allows for analysis on subgroups, such as regions or stratifications by socio-economic status. The primary care database used as training set has been enlarged in recent years, but still covers at this moment only 10% of the population and the Dutch GPs. Using drug use will allow for better prevalence estimates for the 90% not covered.

Because the full population is covered in the prescription data we use, and the model provides estimates of the probability of having a disease at the individual level, other useful applications would be pre-selecting subjects for medical trials, or making case-mix corrections, e.g. for comparing hospital performance.

To conclude, combining diagnosis data and drug use by the RF algorithm provides can be a useful tool to predict population prevalence. Applications include situations where the diagnosis data is not necessarily representative for the population of interest,

but the relation found between diagnosis and drug use is representative. Furthermore, it can be used to select relevant drug use groups in almost all cases.

Supplementary data

Supplementary data are available at *EURPUB* online.

Acknowledgements

This study has been approved by both the Central Commission for Statistics of Statistics Netherlands (CBS) and the Steering Committee of the NIVEL Primary Care Database. We kindly thank them for their permission to use their data. CBS is also thanked for providing a secure environment for the analysis of the data. Wien Limburg from the Dutch National Institute of Public Health and the Environment (RIVM) has been of enormous help in drafting the final version of the manuscript, and solved many English language issues.

Funding

Research was conducted from the research positions of the authors within their academic institutions. No external funding was received.

Conflicts of interest: None declared.

Key points

- Disease prevalences can be estimated from drug use data by Random Forest (RF), a machine learning tool.
- No prior knowledge about the relationship between drug use data predictors and disease is needed.
- Survey-based prevalence estimates can easily be elaborated with indepth subgroup analyses
- Routine application in public health planning and monitoring is possible.

References

1 Williams R, Wright J. Epidemiological issues in health needs assessment. *BMJ* 1998;316:1379–82.

2 Ward BW, Nugent CN, Blumberg SJ, Vahratian A. Measuring the prevalence of diagnosed chronic obstructive pulmonary disease in the United States using data from the 2012-2014 National Health Interview Survey. *Public Health Rep* 2017;132:149–56.

3 Shin HY, Kang HT. Recent trends in the prevalence of chronic kidney disease in Korean adults: Korean National Health and Nutrition Examination Survey from 1998 to 2013. *J Nephrol* 2016;29:799–807.

4 Du Y, Heidemann C, Gosswald A, et al. Prevalence and comorbidity of diabetes mellitus among non-institutionalized older adults in Germany—results of the national telephone health interview survey 'German Health Update (GEDA)' 2009. *BMC Public Health* 2013;13:166.

5 Niiranen TJ, Lyass A, Larson MG, et al. Prevalence, correlates, and prognosis of healthy vascular aging in a Western community-dwelling cohort: the Framingham Heart Study. *Hypertension* 2017;70:267–74.

6 Darweesh SK, Koudstaal PJ, Stricker BH, et al. Trends in the incidence of Parkinson disease in the general population: the Rotterdam Study. *Am J Epidemiol* 2016;183:1018–26.

7 Caspersen CJ, Bloemberg BP, Saris WH, et al. The prevalence of selected physical activities and their relation with coronary heart disease risk factors in elderly men: the Zutphen Study, 1985. *Am J Epidemiol* 1991;133:1078–92.

8 Filipovic-Pierucci A, Samson S, Fagot JP, Fagot-Campagna A. Estimating the prevalence of depression associated with healthcare use in France using administrative databases. *BMC Psychiatry* 2017;17:1.

9 Koster I, Huppertz E, Hauner H, Schubert I. Costs of Diabetes Mellitus (CoDiM) in Germany, direct per-capita costs of managing hyperglycaemia and diabetes complications in 2010 compared to 2001. *Exp Clin Endocrinol Diabetes* 2014; 122:510–6.

10 Winnard D, Wright C, Taylor WJ, et al. National prevalence of gout derived from administrative health data in Aotearoa New Zealand. *Rheumatology (Oxford)* 2012;51:901–9.

11 Wirehn AB, Karlsson HM, Carstensen JM. Estimating disease prevalence using a population-based administrative healthcare database. *Scand J Public Health* 2007; 35:424–31.

12 van Oostrom SH, Picavet HS, van Gelder BM, et al. Multimorbidity and comorbidity in the Dutch population—data from general practices. *BMC Public Health* 2012;12:715.

13 Von Korff M, Wagner EH, Saunders K. A chronic disease score from automated pharmacy data. *J Clin Epidemiol* 1992;45:197–203.

14 Chini F, Pezzotti P, Orzella L, et al. Can we use the pharmacy data to estimate the prevalence of chronic conditions? A comparison of multiple data sources. *BMC Public Health* 2011;11:688.

15 Carral F, Olveira G, Aguilar M, et al. Hospital discharge records under-report the prevalence of diabetes in inpatients. *Diabetes Res Clin Pract* 2003;59:145–51.

16 Herrett E, Shah AD, Boggon R, et al. Completeness and diagnostic validity of recording acute myocardial infarction events in primary care, hospital care, disease registry, and national mortality records: cohort study. *BMJ* 2013;346:f2350.

17 van den Dungen C, Hoeymans N, van den Akker M, et al. Do practice characteristics explain differences in morbidity estimates between electronic health record based general practice registration networks? *BMC Fam Pract* 2014;15:176.

18 Koster I, von Ferber L, Ihle P, et al. The cost burden of diabetes mellitus: the evidence from Germany—the CoDiM study. *Diabetologia* 2006;49:1498–504.

19 Renard LM, Bocquet V, Vidal-Trecan G, et al. An algorithm to identify patients with treated type 2 diabetes using medico-administrative data. *BMC Med Inform Decis Mak* 2011;11:23.

20 Chaudhry MR. *Predicting Individual-level Probabilities of Dementia and Diabetes Using Health Services Administrative Data*. Toronto: University of Toronto, 2015.

21 Khalilia M, Chakraborty S, Popescu M. Predicting disease risks from highly imbalanced data using Random Forest. *BMC Med Inform Decis Mak* 2011;11:51.

22 National Health Care Institute (ZiN). *Pharmacotherapeutical Compass*. 2017. Available at: https://www.farmacotherapeutischkompas.nl/

23 Breiman L. Random Forests. *Mach Learn* 2001;45:5–32.

24 Liaw A, Wiener M. Classification and regression by randomForest. *R News* 2002;2:18–22.

25 National Health Care Institute (ZiN). 2017. Available at: https://english.zorginstituutnederland.nl/

26 Netherlands Institute for Health Services Research (NIVEL). *NIVEL Primary Care Database*. 2018. Available at: https://www.nivel.nl/en/nivel-primary-care-database.

27 Statistics Netherlands (CBS). Share of 80+ population living in institutional households for 2010. 2018. Available at http://statline.cbs.nl/Statweb/publication/?DM=SLNL&PA=37620&D1=0,11&D2=0&D3=112-115&D4=15&VW=T.

28 World Health Organization. International Classification of Primary Care (ICPC-2), 2018. Available at: http://www.who.int/classifications/icd/adaptations/icpc2/en/.

29 O'Halloran J, Miller GC, Britt H. Defining chronic conditions for primary care with ICPC-2. *Fam Pract* 2004;21.

30 Bakker B, Rooijen J, Van Toor L. The system of social statistical datasets of Statistics Netherlands: an integral approach to the production of register-based social statistics 2014;30:411–24.

31 Statistics Netherlands (CBS). Microdata: conducting your own research. 20 December 2018. Available at: https://www.cbs.nl/en-gb/our-services/customised-services-microdata/microdata-conducting-your-own-research.

32 National Health Care Institute (ZiN). FKG ATC-referentiebestand somatische zorg vereveningsmodel 2016. 13 October 2015.

33 National Health Care Institute (ZiN). Pharmaco-therapeutic information tuberculosis. 2018. Available at: https://www.farmacotherapeutischkompas.nl/bladeren/groepsteksten/tuberculosemiddelen

34 National Health Care Institute (ZiN). Pharmaco-therapeutic information multiple sclerosis. 2018. Available at: https://www.farmacotherapeutischkompas.nl/bladeren/indicatieteksten/multipele_sclerose.

35 De Jong L, Janssen P, Keizer D, et al. NHG-Standaard Pijn. *Huisarts Wet* 2015;58:472–85.

36 Winters JC, Van der Windt DAWM, Spinnewijn WEM, et al. NHG-standaard schouderklachten. In: Wiersma T, Boukes FS, Geijer RMM, Goudswaard AN, editors. *NHG-Standaarden 2009*. Houten: Bohn Stafleu van Loghum, 2009: 1213–29.

37 National Health Care Institute (ZiN). Pharmaco-therapeutic information peptic ulcer. 20 December 2018. Available at: https://www.farmacotherapeutischkompas.nl/bladeren/indicatieteksten/peptische_ulcera_met_positieve_helicobacter_pylori_test.

38 Steyerberg E. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. New York: Springer-Verlag, 2009.

39 LeDell E, Petersen M, van der Laan M. Computationally efficient confidence intervals for cross-validated area under the ROC curve estimates. *Electron J Stat* 2015;9:1583–607.

The European Journal of Public Health, Vol. 29, No. 4, 621–625

doi:10.1093/eurpub/cky272 Advance Access published on 18 January 2019

Increase in emergency department visits related to cannabis reported using syndromic surveillance system

G.N. Noel[1,2,3], A.M. Maghoo[3], F.F. Franke[4], G.V. Viudes[1], P.M. Minodier[1]

1 PACA Regional Emergency Department Observatory (ORUPACA), Hyères, France
2 Pediatric Emergency Department, APHM, Marseille, France
3 Public Health Department, EA 3279, Chronic Diseases and Quality of Life, Aix-Marseille University, Marseille, France
4 Santé Publique France, French National Public Health Agency, Regional Unit (CIRE Provence-Alpes-Côte d'Azur and Corsica), Marseille, France

Correspondence: Guilhem Noel, Observatoire Régional des Urgences PACA (ORU PACA), 145 Chemin du Palyvestre, 83400 Hyères, France, Tel: +33 04 98 080 080, Fax: +33 04 94 57 09 09, e-mail:gnoel@orupaca.fr

Background: Cannabis is illegal in France but, as in many countries, legalization is under debate. In the United States, an increase of emergency department (ED) visits related to cannabis exposure (CE) in infants and adults was reported. In France, a retrospective observational study also suggested an increase of CE in children under 6 years old. This study only included toddlers and the data sources used did not allow repeated analysis for monitoring. **Methods:** Our study aimed to evaluate the trend in visits for CE in ED in patients younger than 27 years old in Southern France. A cross-sectional study using the Electronic Emergency Department Abstracts (EEDA) included in the national Syndromic Surveillance System. CE visits were defined using International Classification of Disease (ICD-10). **Results:** From 2009 to 2014, 16 EDs consistently reported EEDA with <5% missing diagnosis code. Seven hundred and ninety seven patients were admitted for CE including 49 (4.1%) children under 8 years old. From 2009–11 to 2012–14, the rate of CE visits increased significantly across all age groups. The highest increase was in the 8–14 years old (+144%; 1.85–4.51, $P < 0.001$) and was also significant in children under 8 (0.53–1.06; $P = 0.02$). Among children under 8, hospitalization rate (75.5% vs. 16.8%; $P < 0.001$) and intensive care unit admissions (4.1% vs. 0.1%; $P < 0.001$) were higher compared with patients older than 8 years. **Conclusion:** These trends occurred despite cannabis remaining illegal. EEDA could be useful for monitoring CE in EDs.

Introduction

Cannabis is illegal in France but, as in many countries, legalization is under debate. Survey data could be used to evaluate the prevalence of cannabis use in the general population whereas emergency department (ED) visits for cannabis exposure (CE) are symptomatic of pathologic situations associated with cannabis. Zhu used ED visits in the United States to report an increase in CE visits in adults.[1] In the USA, various studies[2] using Poison Control Center (PCC) data[3–6] or hospital data[7,8] have reported that legalization of cannabis was associated with an increase in un-intentional cannabis ingestions by young children. In France, without any change in cannabis law, an increase in phone calls to PCC was first reported in 2009.[9] In 2017, Claudet also reported an increase in hospital admissions for CE in a retrospective observational study (2004–14) using hospital data and retrospective reading of patients' medical file.[10] However, this study could have over-estimated the increase, since during the first years of the study, only inpatients' medical file was computerized. Children who had not been admitted to hospital following the ED visit were thus not caught in the study in many EDs. Moreover, the methodology used did not allow for an easy analysis repetition to survey the trend. In France, each ED admission has to be reported daily through Electronic Emergency Department Abstracts (EEDA). These EEDA are transmitted to OSCOUR® network included in the French Syndromic Surveillance (SSS) System SurSaUD® coordinated by Public Health France.[11] EEDA have previously been used for various epidemiological studies in EDs.[12–15] Our study aimed to measure, in southern France, the trend of CE visits in EDs, in children and young adults, using daily available data included in the French national SSS.

Methods

Data source

Our cross-sectional study analysed EEDA related to patients younger than 27 years old. EEDA have been included in the French SSS since 2004. They are directly collected from patients' computerized medical file filled in during medical consultations. Details of this network have been published elsewhere.[3,4] The Provence-Alpes-Côte d'Azur