Mapping Chronic Disease Prevalence based on Medication Use and Socio-demographic variables: an Application of LASSO in healthcare in the Netherlands

Füssenich, Koen; Boshuizen, Hendriek; Nielen, Markus M J; Buskens, Erik; Feenstra, Talitha L

Published in:
Research Square

DOI:
10.21203/rs.2.21491/v1

Publication date:
2020

Mapping Chronic Disease Prevalence based on Medication Use and Socio-demographic variables: an Application of LASSO in healthcare in the Netherlands

Koen Füssenich (koen.fussenich@rivm.nl)
National Institute for Public Health and the Environment, Bilthoven; Department of Epidemiology, University Medical Center Groningen, Groningen University, Groningen https://orcid.org/0000-0003-0141-0004

Hendriek C. Boshuizen
Rijksinstituut voor Volksgezondheid en Milieu

Markus M.J. Nielen
Nederlands Instituut voor Onderzoek van de Gezondheidszorg

Erik Buskens
Universitair Medisch Centrum Groningen

Talitha L. Feenstra
Rijksuniversiteit Groningen

Research

Keywords: Disease prevalence, municipality, medication use

DOI: https://doi.org/10.21203/rs.2.21491/v1

Abstract

Objectives

Policymakers generally lack sufficiently detailed health information to develop localized health policy plans. Chronic disease prevalence mapping is difficult as accurate direct sources are often lacking. Improvement is possible by adding extra information such as medication use and demographic information to identify disease. The aim of the current study was to use a LASSO (Least Absolute Shrinkage and Selection) model on a wide set of variables including medication use to obtain small geographic area prevalence estimates for four common chronic diseases and investigate regional patterns of disease.

Methods

Administrative hospital records and general practitioner registry data were linked to medication use and socio-economic characteristics. The training set (n=707021) contained GP diagnosis and/or hospital admission diagnosis as the standard for disease prevalence. For the entire Dutch population (n = 16,777,888), all information except GP and hospital admission was available. A LASSO operator regression model for binary outcomes was used to select variables strongly associated with disease. Dutch municipality (non-)standardized prevalence estimates for stroke, CHD, COPD and diabetes were then based on the average of individual predicted probabilities.

Results

Adding medication use data as a predictor substantially improves model performance. Estimates at the municipality level are best for diabetes with a weighted percentage error (WPE) of 6.8%, and worst WPE for COPD, with 14.5%. Disease prevalence has clear regional patterns, also after standardization for age.

Conclusion

Adding medication use as an indicator of disease prevalence next to socio-economic variables substantially improved estimates at the municipality level. The resulting individual disease probabilities can be aggregated into any desired regional level and provide a useful tool to identify regional patterns and subsequently inform local policy.

Introduction

Chronic disease prevalence is an important indicator of public health. Large differences in disease prevalence have been observed between populations. These are influenced by demographic background, genetics, lifestyle, environmental factors and healthcare policy. As a result, disease prevalence rates strongly vary between small geographic regions. [1–3] Disease mapping may be used to visualize and analyse these differences, which allows for more efficient allocation of healthcare resources and specific local healthcare policies[4]. In the Netherlands, disease prevention has been delegated to municipalities, creating demand for disease maps at the municipal level or even at smaller geographic scale, such as neighbourhoods.

At the national level, disease prevalence data is often available from surveys, [5–7] hospitalization data, [8] GP registries, or insurance claims data. [9] Due to the high costs of collecting data and medical confidentiality, sample size will often be insufficient to create disease maps at a detailed geographic level.[10]

As sample sizes are low, researchers have to add extra information to arrive at good estimates for small area disease estimates [7]. Often, spatial dependencies are used, borrowing information from geographically proximate regions.[11] Alternatively, other disease related data available for those regions could be used. A frequently used indicator for disease is medication use.[12, 13]

Based on a theoretical link between disease and medication, usually medication use is applied as a direct indication of the disease being present. More recently, studies have explored medication use as a predictor in models using training sets with disease diagnosis and medication use data [14–16]. These studies use machine learning techniques to select medication groups with the highest predictive power. As not all persons that have a disease take the same medication, observing the link between diseases and medication use in data outperforms predictions based on medication found in literature.

While it has been shown that medication use can be a powerful indicator of disease, it has not been shown to what extend they can be applied to estimate regional disease prevalences. The current study investigates the added value of medication use and socio-economic variables compared to models using just age and gender to predict diabetes, chronic obstructive pulmonary disease (COPD), coronary heart disease (CHD) and stroke and it investigates the resulting regional patters in The Netherlands.

Methods

Data

All data used was accessed and analysed through the System of Social Statistical Datasets (SSD) of Statistics Netherlands. The SSD provides access to multiple administrative data sources, the ability to link pseudo-anonymised data at the individual level, and serves as a Trusted Third Party (TTP). Analyses took place in a secured environment and results can only be exported after control by SSD for privacy and security issues.[17] Dutch law allows the use of electronic health records for research purposes under strict conditions. According to this legislation, neither obtaining informed consent from patients nor approval by a medical ethics committee is obligatory for this type of observational studies containing no directly identifiable data (Dutch Civil Law, Article 7:458).

The population consisted of all those living in the Netherlands on December 31st 2012. Of the 16,779,412 persons recorded, for 16,777,888 persons (99.9%) data was available on date of birth, gender, marital status, municipality, ethnicity, being 1st or 2nd generation immigrant, percentile group of wealth, source of income, percentile group of household income and household composition.

Individual data on medication use were obtained from Medicijntab [18], ‘containing data on persons to whom medicines were dispensed and reimbursed under the statutory basic medical insurance in the year concerned.’ While all individuals have basic insurance, medications reimbursed differently or sold over the counter are not included. It was assumed that individuals with no record of a certain ATC3 code did not use this medication in the year of interest.

Diagnosis data was available from two sources, a primary care database and hospital records. When a person was registered in one of the practices participating in the primary care database, the person was included in what we will refer to as the ‘training set’. All Dutch inhabitants are registered in a primary care practice for insurance purposes. The NIVEL primary care database [21] comprises approximately 10% of the Dutch population, with most practices entering during 2002–2006. Diagnostic codes were given by general practitioners in ICPC-1 code [19], and covered all individuals registered to a GP practice as of date of entry of either the GP into the registry, or the individual into the GP practice.

Clinical and day admissions to hospitals were available from the National Medical Registry [‘Landelijke Medische Registratie’(LMR)] [20] from 2002–2012. For 2012 it was estimated that around 25% of admissions were missed by Statistics Netherlands, while there were fewer missing cases in the previous years [20]. Most hospitals reported in ICD9, while in 2012 several hospitals reported in ICD10.

If a person had been diagnosed with one of the codes available in Table 1, in either the hospital data (primary and secondary diagnosis) or the primary care data, we considered the person to have the disease/diagnosis category indicated. For stroke and myocardial infarction, having experienced the event in the period covered by the datasets was considered as a chronic disorder for the current study. When neither the hospital records, nor the GP registry indicated a diagnosis, the individual was considered disease free.

About 85% of patients in this database could be uniquely linked in the SSD environment to the full set of socio-demographic variables, resulting in a training set of 707,021 individuals, with full diagnostic information being present, as well as complete information on covariates.

Table 1
ICD10, ICD9 and ICPC codes [19] per disease

Disease	ICD10	ICD9	ICPC-1
Coronary Heart Disease	I20 – I25	410–414	K74-K76
Stroke	I60 – I69	430–434, 436–438	K90
Diabetes	E10 – E14	250, 648	T90
COPD	J40 – J44	490–492, 496	R91,R95

Table 2 shows the characteristics of the training set compared to the total Dutch population. Differences are very small, with a slightly elderly population, and slightly more pensions as source of income in the training set. The first and third quartiles are also similar for age, wealth- and income percentile.

Table 2
Descriptive statistics in percentages

Variable	Training set	Dutch Population
Mean Age	40.6	40.3
Mean Wealth Percentile	50.3	50.5
Mean Income Percentile	60.7	59.9
Percentage Females	51.1	50.5
Marital Status		
Unmarried	46.5	47.0
Divorced	7.3	7.1
Widowed	5.4	5.2
Married	40.8	40.7
Source of Income		
Labor	57.2	57.1
Owned company	14.8	14.7
Wealth	0.4	0.4
Social benefits	8.2	8.1
Pension	18.3	17.8
Study Financing	0.6	0.8
Other	0.1	0.1
No Income	0.4	1.0
Ethnic Group		
Moroccan	2.0	2.2
Turkish	2.2	2.4
Surinam	2.1	2.1
Netherlands Antilles and Aruba	0.9	0.9
Native	80.2	78.9
Other western	4.0	4.2
Other non-western	8.5	9.4

Variable	Training set	Dutch Population
Immigrant generation		
Native	80.2	78.9
1st generation	9.3	10.7
2nd generation	10.5	10.4
Type of household		
1 person	15.8	16.5
Married couple with children	39.0	39.2
Married couple without children	20.0	19.8
Non-married couple with children	9.1	8.3
Non-married couple without children	6.2	6.3
1 parent with children	8.1	7.9
Institutional	1.2	1.4
Other	0.5	1.4
Source of Income		
Labor	57.2	57.1
Owned company	14.8	14.7
Wealth	0.4	0.4
Social benefits	8.2	8.1
Pension	18.3	17.8
Study Financing	0.6	0.8
Other	0.1	0.1
No Income	0.4	1.0

Data analysis

First, we estimated disease probabilities on the individual level. Then, we aggregated these probabilities into prevalence at the municipality level. All analyses were done separately for all diseases.

For our prediction model, next to ATC3 medication codes, a range of socio-economic variables was available as potential predictors. Table 2 lists the variables included and their factor levels where appropriate. Adding all interaction terms with age and age^2, this amounted to 699 potential predictors. Percentile scores for income and wealth were added next to their second and third degree polynomials. Three models were distinguished and estimated separately for each disease: The complete model with all 699 predictors, the medication only model, with 182 predictors reflecting ATC3 codes, and the socio-demographics only model with 146 predictors, excluding medication use information.

In order to reduce the number of predictors, a Least Absolute Shrinkage and Selection operator (LASSO) model, with a logit link was fitted using the R package 'glmnet'[21], with the four diseases separately as dependent variables. The shrinkage parameter was chosen that minimizes the misclassification error based on tenfold cross-validation plus one standard error[21], or such that at least 10 predictors were included, whichever of the two included the most variables. Levels of a categorical predictor were considered as separate variables.

Finally based on the total Dutch population, for each municipality, the disease prevalence was computed as the average of the predicted individual disease probabilities.

To assess the internal validity of the resulting prevalence estimates at the municipality level, 5-fold cross validation was used for the LASSO procedure.

Based on the cross-validation, the weighted percentage error (WPE) was computed at the municipality level,

$$\sum\nolimits_{m\in M} w_m\left((P_m - O_m)/O_m\right),$$

where M is the set of municipalities, 0_m is the observed prevalence (percentage) for municipalities in the training set, directly based on the registry data. P_m is the estimated prevalence using either the complete, the medication only or the socio-demographics only model, and w_m is the weight, computed as subpopulation size in the training set compared to the size of the training set, such that the sum of the weights is 1. For municipalities with few persons in the training set, 0_m is zero for several diseases. Hence, only municipalities with more than 500 persons in the training set were included in the WPE.

Next to the unstandardized results, standardized results for age were calculated by applying weights to each individual, before averaging to the municipality level. This estimate allowed to investigate regional differences that remain after correcting for differences in the age of the population. Weights were computed by comparing the age distribution of the municipality to the total Dutch population. Five-year age categories were applied for ages 20–85, while all persons aged below 20 years of age were combined in a single category and also all persons aged 85 years and over were combined in a single category.

Results

Figure 1 shows the AUC for the four diseases and models. As an AUC closer to 1 indicates a better fit, we see that a model with only age and gender already fits well, especially for stroke and CHD. Adding socio-economic variables barely improves the AUC further. Adding medication use, however, does improve the AUC for all four diseases. This improvement is largest for diabetes.

Figure 2 shows the fit at the municipality level in the training set. A lower WPE indicates a better fit. As to be expected, we see that adding more information generally improves the model, and that only age and gender always perform the worst. However, we observe that medication use is very predictive for CHD and diabetes, where socio-economic variables do not further improve the model. For COPD and stroke, there is a more gradual improvement. Overall, the error made for COPD is relatively large, even though adding medication and socio-economic variables does decrease the error by several percentage points.

Figure 2 shows the age-standardized maps. Clear regional patterns were observed, which also differ per disease. Especially the different pattern for stroke is clear and important information for capacity building and prevention policy. Appendix 1 shows the unstandardized results, which show a slightly different pattern and larger differences. The northern province of Groningen and the south of Limburg show the highest prevalence.

Discussion

In this study we assessed the role of medication use data, demographic information (age and gender) and socio-economic predictors in creating models to estimate disease prevalence at the individual level. Using these models allows the creation of maps at any desired level of regional granularity. Maps at the municipality level indeed revealed clear regional patterns that differed by disease.

Looking at cross-validation results in the training-set, we found that the weighted percentage error at the municipality level when comparing the models including both medication use and socio-economic variables was least for diabetes at 6.2%, while it was highest for COPD, with 14.4%.

Adding medication use as predictor improved estimates substantially compared to models that only included socio-economic variables or age and gender. This effect was strongest for diabetes, and weakest for stroke. Other researchers estimating disease prevalences at a small-area level have used mainly age, gender, ethnicity, education or income as predictors, and frequently relied on spatial dependencies to attain estimates for small regions. [6, 7, 22, 23] Adding medication use substantially improves these estimates.

The current method has several limitations. First, it requires more variables than survey based methods, at least for a training set, while all relevant predictors also have to be available for the entire population for whom estimates are to be obtained. Access to information on medication use, GP and hospital records maybe restricted or difficult to link at the individual level. However, the training set could also be based on alternative sources if these would be more easily available, as long as data on diagnosis as well as medication use and other predictors are available, and the set is representative for the population at large. The main message is that, once a registry is envisioned to be used for prevalence estimates, it is worthwhile considering it as a training set rather than directly extrapolating from the registry diagnoses to the entire population. Indeed, applying predictors that are also easily available for the entire population to enlarge the precision of regional prevalence data, over what can be obtained by simple age and gender based adjustments appears worthwhile.

In the current study, while diagnosis and medication use data were available, the data sources at hand have their limitations. We had diagnosis data available from GP and hospital sources. However from the GP records, 85% can be linked individually, while 25% of the hospital records in 2012 are missing. We did include multiple years of data to capture as much information as possible. Furthermore, we only observed diagnosed cases. Persons who may have had the disease but never went to see a medical professional will not be included in any administrative data source. As such, the prevalence estimates reflect estimates of formally diagnosed disease.

While most of the available data are indicator functions, age, income and wealth are count and percentile scores. The applications of LASSO forced making assumptions with respect to linearity, while we were only able to add polynomials of age, income and wealth. Furthermore, we only added interactions with age and age^2, while interactions with socio-economic variables or between ATC groups could be predictive of disease as well.

Also, the current method assumes consistency in prescribing behaviour among medical professionals, and especially GPs among the population of interest. While the Netherlands has centralized prescription guidelines, medical professionals may still treat patients differently. With multiple GPs working in one municipality, this partially averages out. Still, for any estimated difference, the question remains whether this is entirely due to differences in underlying health status or partly attributable to differences in prescription pattern across municipalities. Further research separating the two would add to the interpretation of regional differences observed.

Interestingly, applying the method to the Netherlands, we observe clear regional patterns in disease that surpass random noise. We therefore believe this method recommend our approach as a useful tool to monitor and observe regional trends, and identify areas that may require extra attention. For instance, the high prevalence of stroke in the Southern part of the Netherlands may indicate that policy makers should make available sufficient emergency care as well as develop preventive policies in these municipalities.

Regional patterns for the four diseases are also different, indicating that dedicated local policy would be beneficial. Relating such patterns to e.g. lifestyle risk factor prevalence and/or socio-demographics could support policy choices in prevention and capacity planning.

Conclusion

In this manuscript, we assessed whether medication use and demographic variables can be used to reliably estimate municipality disease prevalence for stroke, coronary heart disease, diabetes and COPD in the Netherlands. Adding medication use next to socio-economic variables substantially improved estimates at the municipality level.

The resulting individual disease probabilities can be aggregated into any desired regional level and provide a useful tool to explore regional patterns and develop a specific local policy.

Abbreviations

LASSO
Least Absolute Shrinkage and Selection
GP
General Practitioner
CHD
Coronary Heart Disease
COPD
Chronic Obstructive Pulmonary Disease
WPE
Weighted percentage error
SSD
System of Social Statistical Datasets
TTP
Trusted Third Party
ATC
Anatomical Therapeutic Chemical Classification System
AUC
Area under the receiver operating characteristic curve

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

All data used was accessed and analysed through the System of Social Statistical Datasets (SSD) of Statistics Netherlands. The SSD provides access to multiple administrative data sources, the ability to link pseudo-anonymised data at the individual level, and serves as a Trusted Third Party (TTP). All data is only available after authorization of Statistics Netherlands.

Competing interests

The authors declare that they have no competing interests.

Funding

Unrestricted funding by RIVM and UMCG

Authors' contributions

KF, HB, EB and TF designed the article. HB, EB and TF acquired the funding. KF analyzed the data. MN provided interpretation of the NIVEL Primary Care database. KF drafted the manuscript. All authors discussed outcomes and their interpretation. All authors read and approved the final version of the manuscript.

Acknowledgements

We would like to thank Erik Koffijberg en Jan van de Kassteele for critically reviewing the paper and their suggestions.

References

1. Mackenbach, J.P., *Socio-economic health differences in The Netherlands: a review of recent empirical findings.* Soc Sci Med, 1992. **34**(3): p. 213-26.
2. Rijksinstituut voor Volksgezondheid en Milieu, *Atlas VZInfo.* 2018.
3. Centraal Bureau voor de Statistiek and Planbureau voor de Leefomgeving, *Regionale Verschillen in Sterfte Verklaard*. 2013.
4. Lawson, A.B. and F.L.R. Williams, *An Introductory Guide to Disease Mapping*. 2001: John Wiley & Sons Ltd. .
5. Terashima, M., D.G.C. Rainham, and A.R. Levy, *A small-area analysis of inequalities in chronic disease prevalence across urban and non-urban communities in the Province of Nova Scotia, Canada, 2007–2011.* BMJ Open, 2014. **4**.
6. Wang, Y., et al., *Comparison of Methods for Estimating Prevalence of Chronic Diseases and Health Behaviors for Small Geographic Areas: Boston Validation Study, 2013.* Preventing Chronic Disease 2017. **14**.

7. van de Kassteele, J., et al., *Estimating the prevalence of 26 health-related indicators at neighbourhood level in the Netherlands using structured additive regression.* International Journal of Health Geographics, 2017. **16**(23).
8. Lee, D.C., et al., *Determining Chronic Disease Prevalence in Local Populations Using Emergency Department Surveillance.* American Journal of Public Health, 2015. **105**(9): p. 67-74.
9. Kappelman, M.D., et al., *The Prevalence and Geographic Distribution of Crohn's Disease and Ulcerative Colitis in the United States.* Gastroenterology, 2007. **5**(12): p. 1424-1429.
10. Waller, L.A. and B.P. Carlin, *Disease mapping.* Chapman Hall CRC Handb Mod Stat Methods, 2010.
11. Wakefield, J., *Disease mapping and spatial regression with count data.* Biostatistics, 2007. **8**(2): p. 158-183.
12. Von Korff, M., E.H. Wagner, and K. Saunders, *A chronic disease score from automated pharmacy data.* J Clin Epidemiol, 1992. **45**(2): p. 197-203.
13. Cossman, R.E., et al., *Correlating pharmaceutical data with a national health survey as a proxy for estimating rural population health.* Popul Health Metr, 2010. **8**: p. 25.
14. Slobbe, L.C.J., et al., *Estimating disease prevalence from drug utilization data using the Random Forest algorithm.* Eur J Public Health, 2019.
15. Khalilia, M., S. Chakraborty, and M. Popescu, *Predicting disease risks from highly imbalanced data using random forest.* BMC Med Inform Decis Mak, 2011. **11**: p. 51.
16. Chaudhry, M.R., *Predicting Individual-level Probabilities of Dementia and Diabetes using Health Services Administrative Data*, in *Health Policy, Management and Evaluation*. 2015, University of Toronto.
17. Bakker, B.F.M., J. van Rooijen, and L. van Toor, *The System of social statistical datasets of Statistics Netherlands: An integral approach to the production of register-based social statistics.* Statistical Journal of the IAOS, 2014. **30**(4): p. 411-424.
18. College voor Zorgverzekeringen, *Documentatierapport Verstrekkingen van geneesmiddelen aan personen (MEDICIJNTAB)*. 2012.
19. Bentsen, B.G., *International classification of primary care.* Scand J Prim Health Care, 1986. **4**(1): p. 43-50.
20. Centraal Bureau voor de Statistiek. *Documentatierapport Landelijke Medische Registratie (LMR) 2012*. Available from: https://www.cbs.nl/nl-nl/onze-diensten/maatwerk-en-microdata/microdata-zelf-onderzoek-doen/microdatabestanden/lmr-landelijke-medische-registratie.
21. Friedman, J., T. Hastie, and R. Tibshirani, *Regularization Paths for Generalized Linear Models via Coordinate Descent.* J Stat Softw, 2010. **33**(1): p. 1-22.
22. de Graaf-Ruizendaal, W.A. and D.H. de Bakker, *The construction of a decision tool to analyse local demand and local supply for GP care using a synthetic estimation model.* Hum Resour Health, 2013. **11**: p. 55.

23. Yasaitis, L.C., M.C. Arcaya, and S.V. Subramanian, *Comparison of estimation methods for creating small area rates of acute myocardial infarction among Medicare beneficiaries in California.* Health Place, 2015. **35**: p. 95-104.

Supplementary Materials Legends

Supplementary Figure: Estimated unstandardized disease prevalence (%) for all Dutch municipalities.

Figures

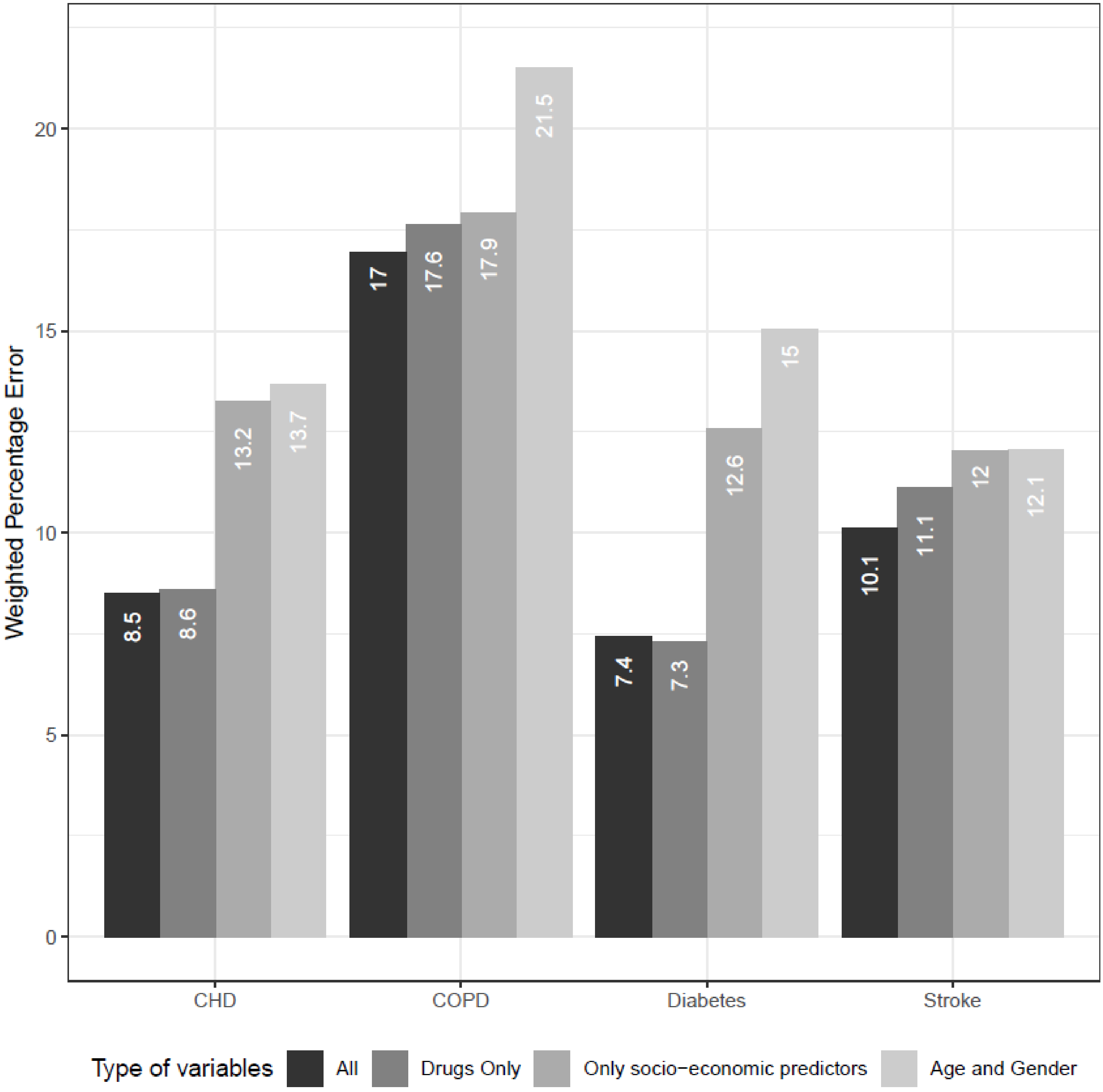

Figure 1

Y-axis: Deviation (%) between the estimated prevalence (%) aggregated by municipality and observed prevalence (%) in the training set, weighed by municipality size. X-axis: All: both ATC3 codes and socio-economic predictors, Drugs only : only ATC3 codes, only socio-economic predictors: only socio-economic predictors, or Age and Gender : only age and gender.

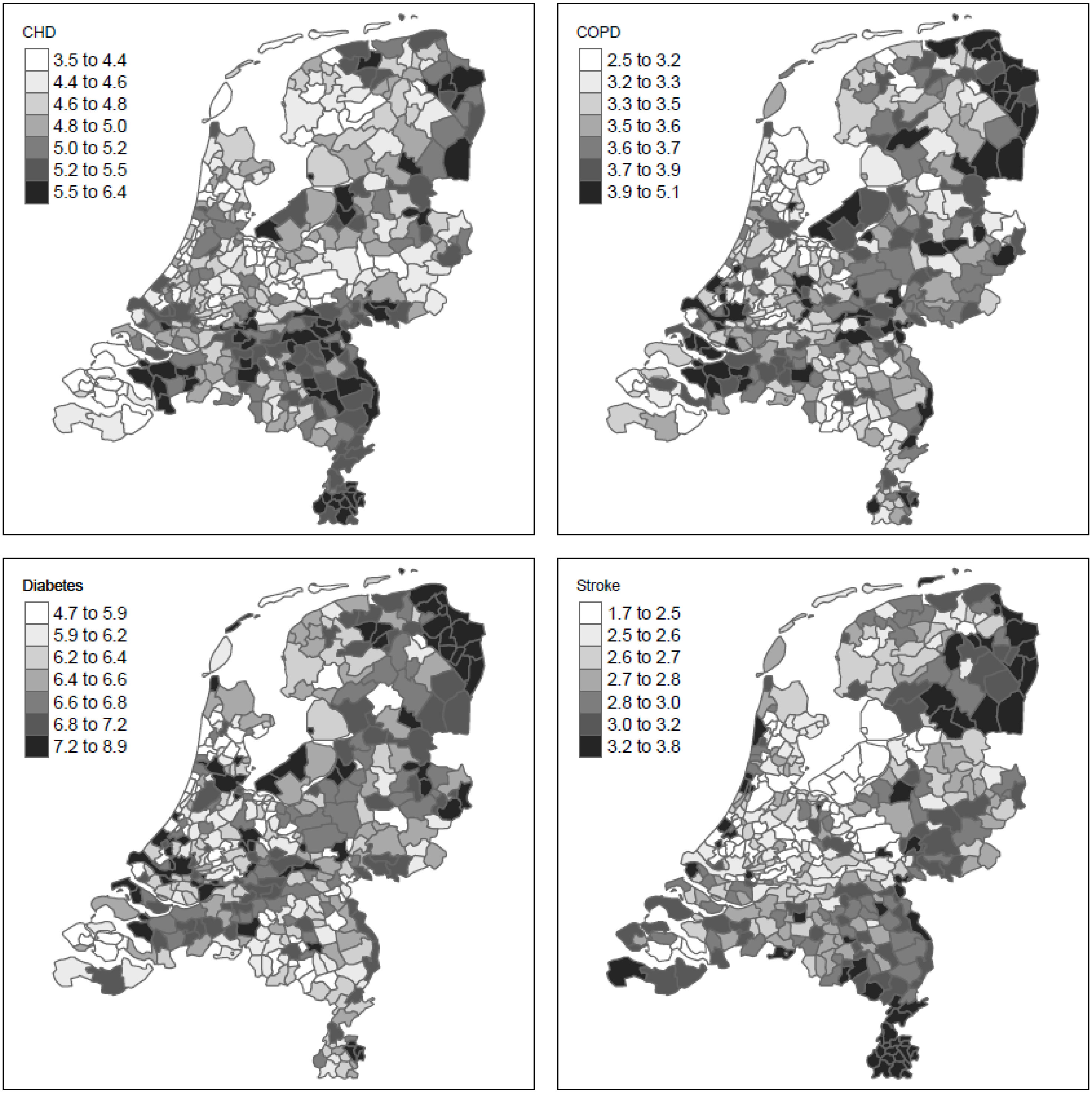

Figure 2

Estimated standardized disease prevalence (%) for all Dutch municipalities. Standardized for age using direct standardization.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- suppfig.png