



Research article

Systematic population-wide ecological analysis of regional variability in disease prevalence

Donald Ruggiero Lo Sardo^{a,b,c,*}, Stefan Thurner^{a,b,d,e}, Johannes Sorger^b,
 Georgh Heiler^b, Michael Gyimesi^f, Alexander Kautzky^g, Michael Leutner^h,
 Alexandra Kautzky-Willer^{h,i}, Peter Klimek^{a,b}

^a Section for Science of Complex Systems, CeMSIIS, Medical University of Vienna, Spitalgasse 23, A-1090, Austria

^b Complexity Science Hub Vienna, Josefstädter Strasse 39, A-1080, Vienna, Austria

^c Physics Department, Sapienza University of Rome, Piazzale Aldo Moro 2, 00185, Rome, Italy

^d IIASA, Schlossplatz 1, A-2361, Laxenburg, Austria

^e Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM, 85701, USA

^f Austrian National Public Health Institute (GÖG), Stubenring 6, A-1010, Vienna, Austria

^g Department of Psychiatry and Psychotherapy, Medical University of Vienna, Waehringer Guertel 18–20, A-1090, Vienna, Austria

^h Department of Internal Medicine III, Clinical Division of Endocrinology and Metabolism, Medical University of Vienna, Waehringer Guertel 18–20, A-1090, Vienna, Austria

ⁱ Gender Institute, A-3571, Gars am Kamp, Austria



ARTICLE INFO

Keywords:

Social determinants of health
 Economical determinants of health
 Health inequalities
 Lifestyle

ABSTRACT

The prevalence of diseases often varies substantially from region to region. Besides basic demographic properties, the factors that drive the variability of each prevalence are to a large extent unknown. Here we show how regional prevalence variations in 115 different diseases relate to demographic, socio-economic, environmental factors and migratory background, as well as access to different types of health services such as primary, specialized and hospital healthcare. We have collected regional data for these risk factors at different levels of resolution; from large regions of care (*Versorgungsregion*) down to a 250 by 250 m square grid. Using multivariate regression analysis, we quantify the explanatory power of each independent variable in relation to the regional variation of the disease prevalence. We find that for certain diseases, such as acute heart conditions, diseases of the inner ear, mental and behavioral disorders due to substance abuse, up to 80% of the variance can be explained with these risk factors. For other diagnostic blocks, such as blood related diseases, injuries and poisoning however, the explanatory power is close to zero. We find that the time needed to travel from the inhabited center to the relevant hospital ward often contributes significantly to the disease risk, in particular for diabetes mellitus. Our results show that variations in disease burden across different regions can for many diseases be related to variations in demographic and socio-economic factors. Furthermore, our results highlight the relative importance of access to health care facilities in the treatment of chronic diseases like diabetes.

* Corresponding author. Section for Science of Complex Systems, CeMSIIS, Medical University of Vienna, Spitalgasse 23, A-1090, Austria.
 E-mail address: losardor@gmail.com (D.R. Lo Sardo).

<https://doi.org/10.1016/j.heliyon.2023.e15377>

Received 17 June 2022; Received in revised form 3 April 2023; Accepted 4 April 2023

Available online 11 April 2023

2405-8440/© 2023 Published by Elsevier Ltd.

This is an open access article under the CC BY-NC-ND license

(<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

1. Introduction

The Austrian health-care system (HCS) can be universally accessed by all Austrian citizens, European citizens and other population groups residing in Austria. The administration of the system is however not managed only at the national level, but results from a dialogue of the federal ministry of health with federal states and social insurance institutions. Whereas overall planning regarding the provision of health services happens on a national level via the Austrian structural health plan, the implementation of intramural care is the responsibility of the federal states, implementation of extra-mural care is the responsibility of the social insurance institutions. This is similar to many health-care systems such as those in Italy, Germany and North America. There is evidence that this leads to regional variations in healthcare quality [3,7,10,16,24,27]. Socio-economic, environmental and demographic determinants such as education, income, marital status, citizenship and temperature of the environment, have been shown to explain much of the variance in life expectancy [4,17]. Differences in health expenditures and their outcomes have been investigated by de Vries et al. [5], Rahman [21], Manning [14] and Giannoni [8], Lopez-Casnovas [13] discuss discrepancies in supply and demand for health services, Pappas et al. investigate the effects of socioeconomic determinants on avoidable hospitalizations [18], while Petrelli et al. investigate citizenship effects on hospitalizations [19]. However, no comprehensive study links a set of regional risk factors to a comprehensive set of pathology prevalence endpoints. The interplay between the regional and the national scale, in regulating the HCS, has proven to be a fragile yet vital element of the management of crises. During the COVID-19 outbreak for example, such coordination was a crucial discriminant in the success of the management in many countries including Italy, the US and China [23]. From this arises the challenge to monitor the how and why health conditions vary across different administrative regions. Previous studies making use of electronic health records (EHRs) have shown that their systematic use can greatly improve our awareness of the status of the health-care system and its users [12].

We build on this literature by analyzing data gathered from publicly available sources: the European Commission, the Austrian statistical office (Statistik Austria), the Austrian noise-level monitoring branch of the ministry of interior affairs (Bundesministerium), Open Street Maps to identify points of interest (POI) like fast food restaurants and marketing data (WIGeoGIS). A novelty in our analysis is the use of geolocalized information on points of care using the Google Maps API to calculate the distance from hospitals to urban centers. Also, the use of geographical information, means we can investigate certain potential indicators of lifestyle such as the fast food density, that could be informative for diagnose blocks associated with chronic diseases among others. To estimate disease prevalence, we use a nationwide dataset on hospital stays in each region of care (*Versorgungsregion*) for which we have a primary diagnosis in the form of a level-3 ICD-10 code that we aggregate into blocks following the guidelines of the World Health Organization

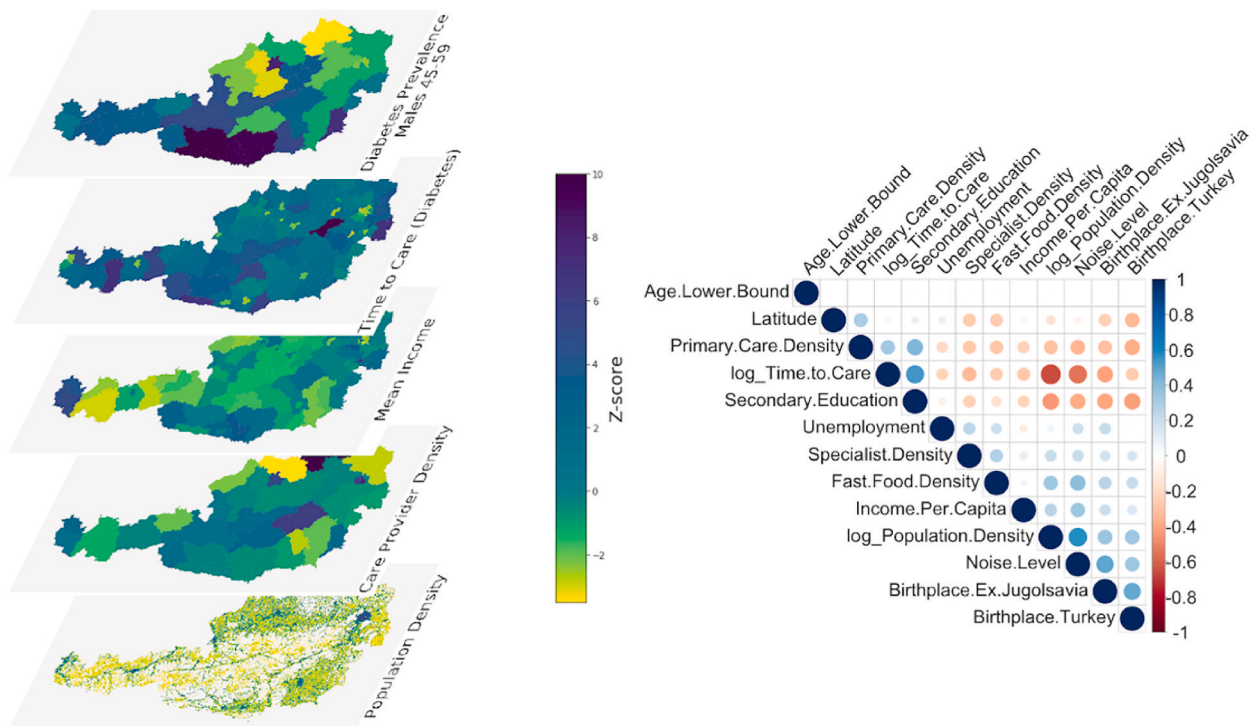


Fig. 1. Depiction of the analysis performed on diseases of the diabetes mellitus block (ICD-10: E10 to E14). In the left panel: a sample of the input data for the model. The five layers describe the regional distribution of standard scores (or Z-scores) at the highest resolution available for the measure. From top to bottom the values are: the prevalence of hospital diagnosis of diabetes for males aged 45 to 59, the logarithm of time to care (specific to diabetes), the mean income, the care provider density, the population density. In the right panel: the Kendall correlation coefficient between predictors selected for the model, excluding sex, being the only categorical variable.

(WHO). For each diagnostic block, we measure the correlation of each risk factor with the disease prevalence using multivariable regression analysis. We find that up to 80% of the variance of certain diagnostic blocks, such as hernia, diseases of the inner ear and disorders due to the use of psychoactive substances, can be predicted using this simple model. While for other diagnostic blocks, such as radiation-related cutaneous disorder and hemolytic anemias, the variation cannot be explained using this method (adjusted-R2 < 0.15). Overall, the strongest risk contributions, after adjusting for age and sex, come from population density, the distance to the healthcare facilities and migratory backgrounds.

2. Results

2.1. Characteristics of the dataset

Of the 131 disease blocks defined by the WHO, we have excluded 16 from our analysis due to insufficient statistics or age and sex specificities of the disease (we refer to <https://www.icd10data.com/ICD10CM/Codes> for ease of consultation; see SI for a list of excluded blocks). Our data, aggregated at the level of the regions of care, comprises 29 regions divided into 12 demographic groups: 6 age groups of 15 years each and 2 sexes. For each of these regions, we have collected publicly available data and added 2 indicators: time to care for the treatment of the diseases in a diagnostic block and the number of fast-food restaurants per 100.000 individuals. Time to care is specific to each diagnostic block and ranges from 100 s to 5000 s with a mean of 1000. The fast-food density is between 1 and 63 establishments per 100.000 individuals (see SI for further information on these indicators). Data for each layer is assigned to its specific location (an example for a few layers can be seen in Fig. 1) and combined to the region of care level.

2.2. Multivariate ecological modelling

For each diagnostic block, we calibrate the same linear model (as defined in the SI). We find that the explained variance of the model varies significantly across disease blocks, going from an adjusted R2 of 0.7 for hernia to 0.02 for hemolytic anemias. Well-predicted diagnoses include “diseases of the ear and mastoid process” (R2 = 0.46 ± 0, 16) and “diseases of the circulatory system” (R2 = 0.39 ± 0, 13). The chapter that the model explains the least are “diseases of the blood and blood-forming organs and certain disorders involving the immune system” (R2 = 0.12 ± 0, 07) and “injury, poisoning and certain other consequences of external causes” (R2 = 0.14 ± 0, 14). We also controlled for variance inflation due to multicollinearity of predictors. 11% of models have variance inflation above 10 and therefore have inflated error estimates of the coefficients. These are therefore not among the most significant results (for further details see SI Fig. 1 where we report the relation between R2 and VIF). In Fig. 2, we summarize the results of the model calibration across the first 40 most diagnosed diseases in Austria. Each row in the lower half corresponds to one indicator, while each column corresponds to a certain diagnostic block. The circles represent the contribution of each indicator to the prevalence of the disease as estimated by the model in terms of average marginal effect (AME). A red circle identifies a risk factor, while a blue circle

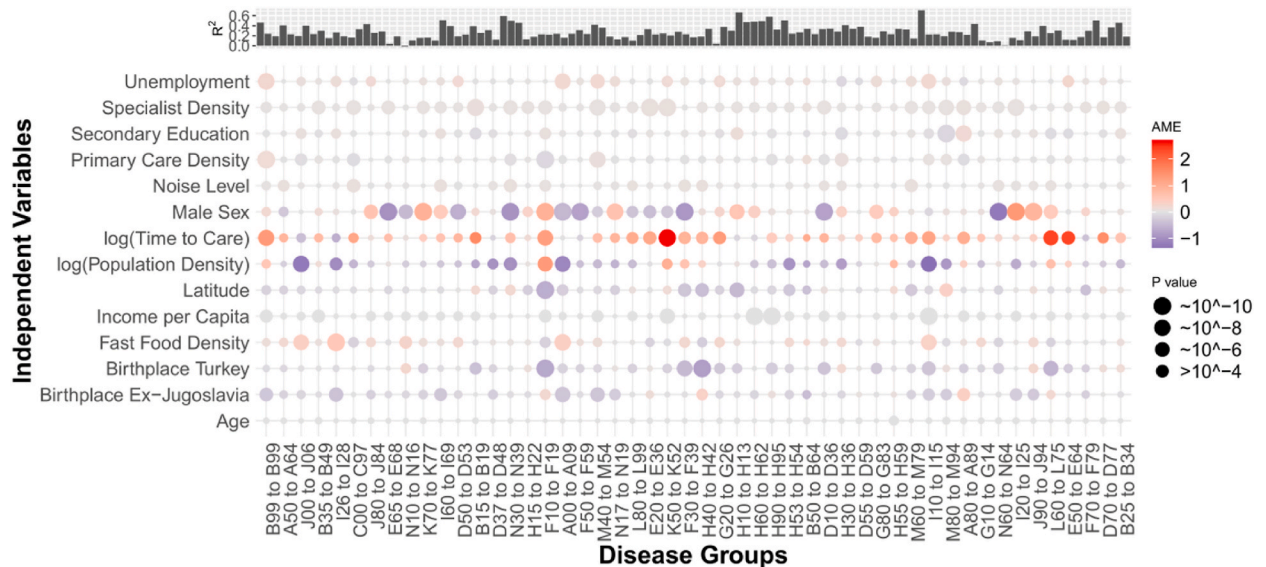


Fig. 2. A graphic representation of the linear models for certain disease blocks. Only models of the first 40 most common diagnostic blocks in Austria are shown. In the top panel, the R² of the model is the amount of variance the model can explain. In the bottom panel, the average marginal effect (AME) each independent variable has on the age-standardised prevalence of the disease is shown. The color represents the direction of the effect: red means higher disease prevalence in regions where the independent variable assumes higher values, blue means lower prevalence. The size of the circle represents the significance of the effect: larger dots are more significant, smaller dots are less significant. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

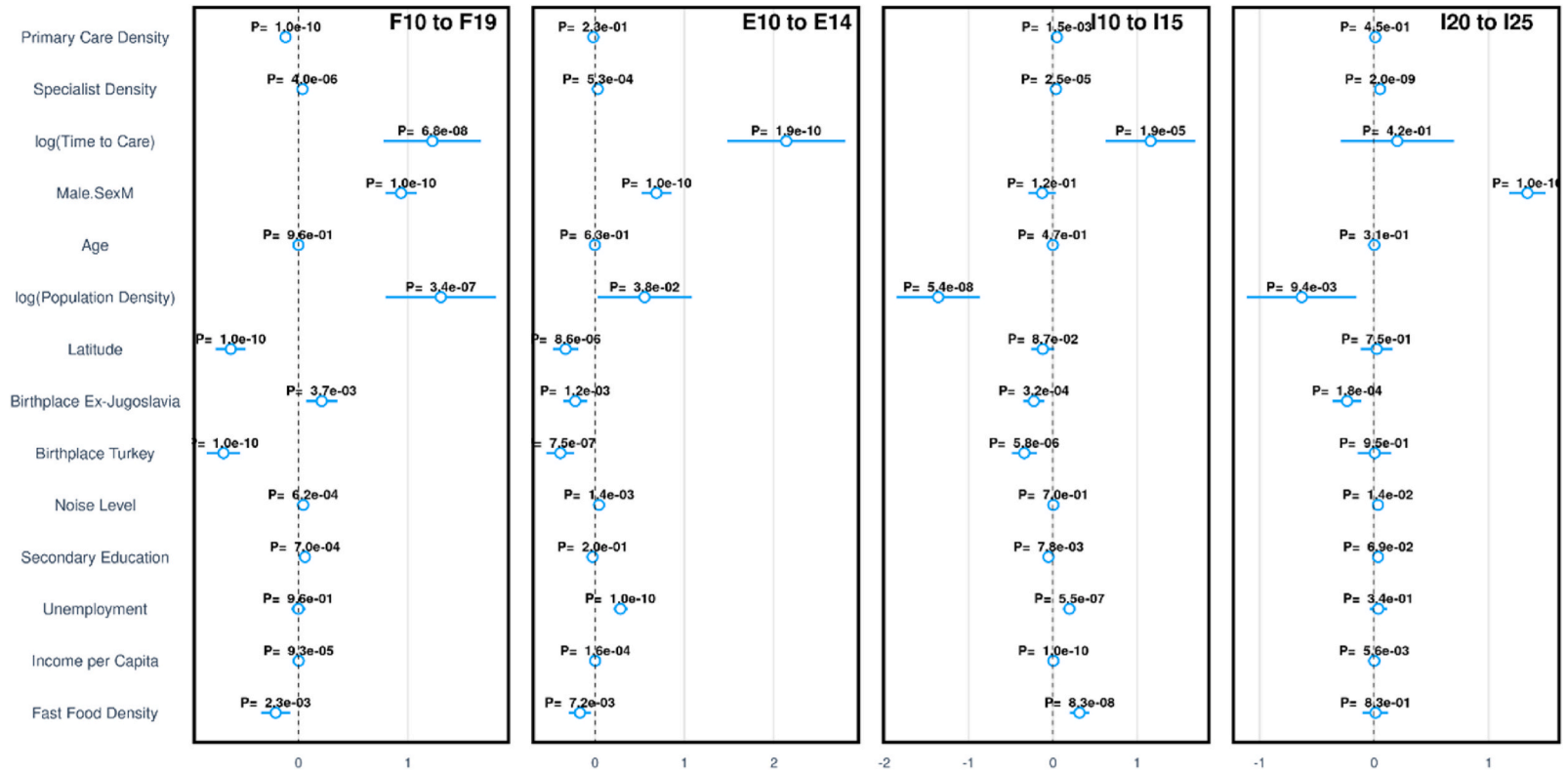


Fig. 3. A detailed view of the regression coefficients of four disease blocks: **F10 to F19** mental and behavioural disorders due to psychoactive substance use, **E10 to E14** diabetes mellitus, **I10 to I15** hypertensive diseases and **I20 to I25** ischaemic heart diseases. The blue circles show the average marginal effect and the bars correspond to the 95% confidence interval. A positive value (>0) indicates a concurrence between the predictor and the diagnose prevalence. A negative effect, conversely, indicates that areas where the predictor is high have a lower diagnose prevalence. The value **P** above each line indicates the p-value (or statistical significance) of each marginal effect. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

indicates a protective effect. The size of the circle specifies the significance of the indicator with large circles being more significant than the small ones. Given that we are testing the effect of 14 indicators across 115 diagnostic blocks, we will treat as significant results with a p-value lower than 10^{-6} using a Benjamini-Hochberg correction as a false discovery rate of 10^{-3} . Examples of significant contributions to the prevalence model are: disorder of the thyroid gland is less prevalent in densely populated areas (codes E00 to E07, $\rho = -1.7$, p-value = 1.1×10^{-10}); diseases of the middle ear and mastoid are more prevalent in areas with lower access to primary care (codes H65 to H75, $\rho = -0.08$, p-value = 5.6×10^{-7}); diseases of esophagus, stomach and duodenum are more prevalent in areas of higher income (codes K20 to K34, $\rho = 4.7 \times 10^{-3}$, p-value = 1.4×10^{-10}) and diseases of hair and nails are more prevalent in areas where the average distance to a relevant hospital ward is greater (codes L60 to L75, $\rho = 2.4$, p-value = 3.4×10^{-7}). At the top of Fig. 2, we display the predictive power of the calibrated model in terms of its R2. To correct for age being the strongest driver of disease prevalence across most diagnostic blocks, we have normalized diagnosis prevalence within each age group. Excluding age, the strongest predictors over all are sex, population density, time to care and migratory backgrounds. Time to care is strongly correlated with population density (see Fig. 1). However, the increased time to care is a risk factor for most diseases, while population density has an opposite effect in many important cases, such as for hypertensive diseases [I10 to I15] and ischemic heart diseases [I10 to I25] (see Fig. 3 for a detailed view of the average marginal effect on these diseases). Among other effects, we notice also a protective effect of Turkish migratory background with respect to “mental and behavioral disorders due to psychoactive substances” [F10 – F19]. Conversely, we see that high noise levels (induced by street traffic) are not strongly related to any diagnostic block.

3. Discussion

Here we performed an ecological analysis that seeks to relate changes in the regional prevalence of 115 different groups of diagnoses to socio-economic variables, demographic factors, access to certain POIs like fast food restaurants, environmental factors such as noise, migration status, as well as access to intra- and extra-mural healthcare facilities. While prevalence of some diagnoses shows little correlation with any of these factors, for diseases like diabetes mellitus, ischemic heart diseases, or substance abuse we can indeed explain a substantial amount of their regional variation. For many chronic disorders like diabetes or cardiovascular diseases their prevalence is strongly driven by distance to hospital wards in which such disorders are typically treated, the larger this distance, the higher the corresponding disease prevalence.

We find that population density, sex and time to care contribute strongly to the predictive power of most disease-specific models. Time required to reach hospital care is always positively related to the disease prevalence. On the other hand we find that the accessibility of non-hospital health-care systems does not contribute substantially to the predictive power of the model. This may be due either to insufficient statistics or the fact that the data we used did not enable us to distinguish between specialists, as we were able to do in the case of time to care. Note that in Austria diagnoses information is only available for inpatient care, which might also explain why distance to inpatient care shows a stronger correlation than distance to outpatient care facilities. A more detailed analysis needs to be carried out in order to verify these hypotheses. Socio-economic predictors, such as income and education, did not contribute strongly, with the exception of unemployment, that shows an increased risk of the prevalence of diabetes.

We find stronger predictive power for specific chronic diseases, especially related to heart conditions and the thyroid gland, some psychological disorders and some pathologies of the ear. For many of these, sex plays an important role: there is a higher risk of heart diseases in men and a higher risk of thyroid-related diseases in women. The close relationship between male sex and cardiovascular diseases (cvd) is an important issue. Hence the effect of sex (male) is strongly related to atherosclerosis. Protective effects of female sex hormones on cardiovascular wall properties have been described earlier. Also, a more favorable body fat distribution, adipokine secretion, better insulin sensitivity and secretion especially in the younger ages of females contribute to the lower cvd and diabetes risk compared to males [9]. Regarding cvd prevention, also a lower prevalence rate of diabetes, arterial hypertension and dyslipidemia in females in the premenopausal state has to be mentioned here ([22]. On the other hand the presence of these comorbidities also predict higher relative risk of cvd in women than men, especially at younger age. For well-predicted psychological disorders, certain migratory backgrounds show a negative correlation, while the population density is a risk factor. Given the correlation between population density and migration status, this might also indicate inequalities in health care utilization.

We observe substantially lower predictive power for oncological disease groups, infectious diseases, eye, ear and cutaneous infections. Our results might directly be useful for a more evidence-based approach to regional health planning. Health planning in Austria often uses a regionalized system based on socio-economic and demographic indicators [15]. For instance, decisions on establishing primary care units at a given place are supported by constructing regional care profiles for the corresponding catchment area. These profiles include indicators such as population size in different age groups, regional prevalence of diseases like obesity or diabetes mellitus type 2, or life expectancy. Our results suggest that adding an indicator quantifying the travel distance to specific hospital wards (“time to care”) in such a system might further contribute to tailoring the provision of health services to the needs of the population.

Regarding the relevance of our work beyond the Austrian health-care system, our claims are three-fold: we present a methodological improvement through the use of geolocalized data, we discuss a novel indicator (time to care) relevant both to the improvement of current and future models of health-care provision. The methodology implemented can indicate regional criticalities and support the planning process for regional care. For example, in the discussion on the creation of rural health centers in the United States similar analysis might provide important information. Further, the importance of time to care, while shown only in the context of Austrian healthcare, might prove to be even more relevant in health-care systems where this indicator has greater variations, such as nations with significantly larger surfaces and greater diversity in health-care-provision density. Finally, the results of our work provide relevant information for the greater discussion on the design of novel models for health-care systems, specifically for the organization

of regionalized care.

Our work's main constraint is the lack of geolocalized data on diagnosis, specifically in the out-patient sector. The level of granularity for the diagnosis data is approximately that of the federal district, while other data sets were available on a much finer scale. Furthermore, the lack of out-patient data is cause for concern in the case of some diseases, such as diabetes, for which hospitalization should not be necessary. Moreover, data on the distribution of different care-provider specializations could prove more insightful if matched to the specific disease typology. Finally, more work needs to be done to correctly understand whether the effects we see are the result of biological phenomena, such as distance to care being critical for certain disease types, or the result of socio-economic differences such as the divide between urban and rural communities.

4. Conclusion

We set out to explore the regional variability in the structure of demands made of the health-care system. We have shown that the use of electronic records enables a comparison between different regions in the prevalence of disease diagnoses. Using a combination of data on hospital diagnoses and sources for demographic, socio-economic, environmental data and the accessibility of the health-care system, we have developed a linear model for each diagnostic block. Each of these models sets a lower boundary for the explanatory power of these resources on the disease prevalence. We found that for certain disease groups, some psychological and heart-related conditions and thyroid disorders, the explanatory power of our model is up to about 80% of the observed variance. We observed that the distance to a relevant care point positively correlates with the prevalence of most disease types. This raises questions about how best to structure and distribute care centers. Our findings highlight the need for awareness of the geographic distribution of resources, such as hospital departments and out-patient centers, based on predictable healthcare demands.

Author contribution statement

Donald Ruggiero Lo Sardo: Conceived and designed the experiments; Contributed reagents, materials, analysis tools or data; Analyzed and interpreted the data; Wrote the paper.

Stefan Thurner: Conceived and designed the experiments; Wrote the paper.

Johannes Sorger: Contributed reagents, materials, analysis tools or data; Analyzed and interpreted the data; Wrote the paper.

Georgh Heiler: Contributed reagents, materials, analysis tools or data; Wrote the paper.

Michael Gyimesi: Contributed reagents, materials, analysis tools; Analyzed and interpreted the data; Wrote the paper.

Alexander Kautzky: Analyzed and interpreted the data.

Michael Leutner, Alexandra Kautzky-Willer: Analyzed and interpreted the data; Wrote the paper.

Peter Klimek: Conceived and designed the experiments; Analyzed and interpreted the data; Wrote the paper.

Data availability statement

Data will be made available on request.

Acknowledgements

We thank Vito D. P. Servedio for help with the online visualization. We acknowledge support from the GENDER-NET Plus ERA-NET Cofund project GOING-FWD.

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.heliyon.2023.e15377>.

Data

In this section we detail how the data was gathered and prepared for the regression analysis.

Dataset 1: Global human settlement

This spatial raster dataset depicts the distribution and density of population, expressed as the number of people per cell. Residential population estimates for the year 2015 provided by CIESIN GPWv4.10 were disaggregated from census or administrative units to grid cells, informed by the distribution and density of built-up areas using the Global Human Settlement Layer (GHSL). For more information see [6].

Dataset 2: Demographics from Statistik Austria

This dataset on demographic information is comprised of different layers we obtained from the Austrian statistical office, Statistik Austria. We use the layers for age distribution, and sex at the political district (*bezirk*) level; and employment, education, households,

families and companies at the municipality level (*gemeinde*). The resident population consists of people staying at least 90 days (including key day) in one location (registration of main residence). For this work we use data from 2017. Since each federal region (*versorgungsregion*) is comprised of, and covers completely, a set of political districts, we define the series $\{vk\}_k$ such that the k -th element is the federal-region index corresponding to political district k . We also define f_{askv} : the fraction of inhabitants of a given age group, and sex in *bezirk* k . For each of the layers at the *gemeinde* level of aggregation, we define a similar series $\{kg\}$ and an independent variable L_{kg} , where L stands for a generic layer.

Dataset 3: Diagnosis dataset

This data was provided by Gesundheit Österreich GmbH (GÖG), the Austrian National Public Health Institute and covers approximately 2,7 M hospitalizations of circa 9 M individuals for the year 2014. The dataset includes long-term care facilities such as rehabilitation centers and psychiatric hospitals. For each entry, the dataset documents the patient anonymized ID, age (in 15 y groups), sex, the federal region of origin, the admission and release date, region of care, the release type (release, transfer, death), the department of care, and most relevant for our study, the main and side diagnoses in the form of level-3 ICS-10 codes. Since we want to run a comprehensive assessment of the health status of the Austrian population, we aggregated the diagnoses in diagnostic blocks as defined by the World Health Organization (WHO). There are 115 diagnostic blocks. For a full taxonomy of the ICD-10 codes see Ref. [26]. For the purpose of this work, we indicate variables that correspond to individuals using the subscript μ . We thus define d_μ the diagnoses of patient μ and the set $D_{asvd} = \{\mu | v_\mu = v \wedge d_\mu \ni d \wedge a_\mu = a \wedge s_\mu = s\}$ of all patients of a given age, sex, federal region that have a certain diagnosis in their history. We also denote $H_{d\Delta} = \{\mu | d_\mu = d \wedge \Delta_\mu = \Delta\}$ the set of patients with diagnosis d who received care in department Δ . Finally, we need to define $B_{asvb} = \sum_{d \in b} |D_{asvd}|$ the number of patients of age a , sex s , from region v , with a diagnosis from block b .

Dataset 4: Distance to care

For each municipality (*gemeinde*) and for each diagnosis, we used the Google Maps API to query the time necessary to reach the relevant hospital wards from the center of the inhabited area or the official location for the municipal center (*gemeindeamt*). District center locations were geocoded via Google Maps. We manually verified the locations of each *gemeindeamt* and repositioned the location to a densely populated area (using a visual overlay of population densities on the map), for cases where the Google Maps algorithm could not find a suitable location. Once the municipal center locations were verified, we used the Google Maps API to issue distance queries between hospital and *gemeindeamt* locations, calculating the distance and travel time (based on average traffic conditions) between each point pair. For further exploration of this data visit: https://www.csh.ac.at/vis/med_public/hospital_explorer/ [25]. Since a diagnosis is not necessarily treated in only one kind of ward, we perform a weighted average of distances to wards offering the required type of care, i.e., $T_{gd} = \sum_{\Delta} \overline{H_{d\Delta}} T_{g\Delta}$, with $\overline{H_{d\Delta}} = H_{d\Delta} \vee \sum_{\Delta} \overline{H_{d\Delta}}$.

Dataset 5: Marktdaten Österreich.

This dataset consists of demographic and population level statistics collected by WIGeoGIS. For each postal code, we have information on: inhabitants by main and secondary residence, population structure (age groups, gender, nationality, origin), the average income and household size among others. We aggregate the postal codes to the *gemeinde*.

Dataset 6: Fast food density

Using publicly available data from open street maps, we gathered information on the location of fast foods. We assign each fast food restaurant to a municipality and compute the number of fast food restaurants per capita in the area.

Dataset 7: Noise

Data on noise levels was taken from the Bundesministerium of Austria. For information on the data collection methods see Ref. [11]. From the available data, we took 24 h averages of noise caused by road traffic. This data is organized in shapefiles such as polygons and points. It was spatially aggregated to the population density layer and a weighted average was performed, using population as a weight, to obtain federal state averages.

Data preparation

Since the data at hand were stored in different geographical projections, we harmonized them using Mollweide espg:54009. We assign each point in the raster from dataset 1 to a *gemeinde*. Some areas along the border are not uniquely assigned and were discarded. The total population after this step is: 8.52 M. To compute diagnostic-block prevalences, we must first compute the size of each unique age, sex and region specified group. We define $M_{asv} = N_v f_{asv}$, where N_v is the total population of the region. Since age is often the main driver of disease prevalences, we also standardized each prevalence within the age groups. Each disease is assigned a prevalence as follows

$$P_{asvb} = \frac{B_{asvb}}{M_{asv}}$$

$$P_{svb} = \frac{P_{asvd} - P_{asvba}}{\sigma(P_{asvb})_a}$$

where the symbols $\langle \cdot \rangle_a$ and $\sigma(\cdot)_a$ indicate mean and standard deviation within a certain age group. The data preparation methods described were implemented in python, in particular using the GeoPandas library. The code for the preparation is available at <https://github.com/complexity-science-hub/GeoDisPrev>.

Dependent and independent variables in the regression analysis

In the previous sections, we have presented the sources, both publicly accessible and sensitive data sets, used for this study. Special attention was paid to the collection of data measuring the distance from the municipality (*gemeindeamt*) to the centers of care, selecting departments relevant to the disease being analyzed.

Prevalences from each diagnostic block P_{asvb} are used as outcome of a series of layered models in which a group of predictors of a certain category are added at each step: a purely demographic model, a model also informed by data on the health-care system, a model that includes migratory background distributions in the population and finally a complete model with the addition of socio-economic predictors and noise levels. The predictors are:

- Age Group: A growing index corresponding to the age group for which the prevalence is computed. Groups cover 15 y of age with the last being from 75 onward.
- Birthplace ex Yugoslavia: The percentage of people in region v with a migratory background originating in ex Yugoslavia.
- Birthplace Turkey: The percentage of people in region v with a migratory background originating in Turkey.
- Fast Food Density: The number of fast food vendors per capita in region v.
- Income Per Capita: The average net income of region v's inhabitants.
- Latitude: The weighted average of region v, with the population of each subregion being used as weight.
- log_Population.Density: The natural logarithm of the number of people living in a $250 \times 250\text{m}^2$ area. This is an average over all inhabited areas of the region.
- log_Time to Care: the natural logarithm of the time from the *gemeindeamt* to the nearest hospital department, averaged over all possible departments that treat diagnostic block b and over all municipalities in region v.
- Noise Level: The average noise level due to road traffic in the inhabited areas of the region.
- Primary Care Density: The number of primary care providers per inhabitant in the region.
- Secondary Education: The percentage of people in region v who have achieved secondary education but not tertiary.
- SexM: The percentage of people of male sex in the region.
- Specialist. Density: The number of specialized care providers per inhabitant in the region.
- unemployment: The percentage of people older than 15 years of age who are unemployed in the region.

In the purely demographic model, we use: sex, age, population density and latitude as predictors. In the health-care system informed model, we add: primary care density, specialist density and time to care. The third model includes the percentages of people born in Turkey and ex Yugoslavia. Finally, noise level, secondary education percentage, unemployment rate and income are included. At each step in the modeling procedure, variance inflation was computed to make sure multicollinearity was acceptable. For the diagnoses discussed in greater detail, such as diabetes mellitus, the maximum value for the final model is 6.2 for time to care (for further details see the SI). The results of the complete model are displayed in Fig. 2, while tables with detailed results are available in the SI. For each diagnostic block, we measure the predictive power of the model in the form of an adjusted r-squared. We then use permutation importance tests, Bonferroni adjusted significance and the model coefficients to uncover the contribution of each predictor. In order to validate the results of the analysis we use a battery of tests: r-squared, QQ plots and VIF monitoring (See SI for QQ-plots of each diagnose block). The implementation for the regression analysis was performed in R. Code for this implementation can be accessed at <https://github.com/complexity-science-hub/GeoDisPrev>.

References

- [3] I. Borboudaki, M. Linardakis, A. Markaki, A. Papadaki, A. Trichopoulou, A. Philalithis, Health service utilization among adults aged 50+ across eleven European countries (the SHARE study 2004/5), *J. Publ. Health* (2020) 1–11.
- [4] M.R. Cullen, C. Cummins, V.R. Fuchs, Geographic and racial variation in premature mortality in the US: analyzing the disparities, *PLoS One* 7 (4) (2012), e32930.
- [5] E.F. de Vries, Unraveling the drivers of regional variation in healthcare spending by analyzing prevalent chronic diseases, *BMC Health Serv. Res.* 18 (1) (2018) 323.
- [6] A.a. Florczyk, GHSL Data Package 2019, European Commission Joint Research Center, 2019, p. 29788.
- [7] V.R. Fuchs, More variation in use of care, more flato-of-the-curve medicine, *Health Aff.* 23 (2004) VAR 104–Var107.
- [8] M.a. Giannoni, The regional impact of health care expenditure: the case of Italy, *Appl. Econ.* 34 (14) (2002) 1829–1836.

- [9] A.a. Kautzky-Willer, Sex and gender differences in risk, pathophysiology and complications of type 2 diabetes mellitus, *Endocr. Rev.* 37 (3) (2016) 278–316.
- [10] T. Kopetsch, H. Schmitz, Regional variation in the utilization of ambulatory services in Germany, *Health Econ.* 23 (12) (2014) 1481–1492.
- [11] laerminfo.at, *Current Noise Maps 2017, 2020*. Retrieved from, <https://www.laerminfo.at/laermkarten.html>.
- [12] L. Laranjo, D. Rodrigues, A. Pereira, R.T. Ribeiro, J. Boavida, Use of electronic health records and geographic information systems in public health surveillance of type 2 diabetes: a feasibility study, *JMIR Public Health Surveill.* 2 (1) (2016) e12.
- [13] G.a.-F. Lopez-Casasnovas, Diversity and regional inequalities in the Spanish 'system of health care services, *Health Econ.* 14 (S1) (2005) S221–S235.
- [14] W.G. Manning, *Explaining Geographic Variation. Geographic Variation in Health Care Spending and Promotion of High-Value Care Interim Report*, Institute of Medicine, 2012.
- [15] S.a. Mathis-Edenhofer, Regional health care profiles—an improved method for generating case studies on the catchment areas of envisaged primary health care units in Austria: a report to the InfAct Joint Action, *Arch. Publ. Health* 80 (1) (2022) 1–8.
- [16] C. Mu, J. Hall, What explains the regional variation in the use of general practitioners in Australia? *BMC Health Serv. Res.* 20 (2020) 1–11.
- [17] S.a.-E. Muratov, Regional variation in healthcare spending and mortality among senior high-cost healthcare users in Ontario, Canada: a retrospective matched cohort study, *BMC Geriatr.* 18 (1) (2018) 262.
- [18] G.a. Pappas, Potentially avoidable hospitalizations: inequalities in rates between US socioeconomic groups, *Am. J. Publ. Health* 87 (5) (1997) 811–816.
- [19] A.a. Petrelli, Socioeconomic and citizenship inequalities in hospitalisation of the adult population in Italy, *PLoS One* 15 (4) (2020), e0231564.
- [21] T. Rahman, Determinants of public health expenditure: some evidence from Indian states, *Appl. Econ. Lett.* 15 (11) (2008) 853–857.
- [22] J.G.-C. Regensteiner, Sex differences in the cardiovascular consequences of diabetes mellitus: a scientific statement from the American Heart Association, *Circulation* 132 (25) (2015) 2424–2447.
- [23] G.a. Sartor, COVID-19 in Italy: considerations on official data, *Int. J. Infect. Dis.* 98 (2020).
- [24] H. Schmitz, More health care utilization with more insurance coverage? Evidence from a latent class model with German data, *Appl. Econ.* 44 (34) (2012) 4455–4468.
- [25] J. Sorger, *Hospital Explorer (Austria)*, 2018. Retrieved from, https://www.csh.ac.at/vis/med_public/hospital_explorer/.
- [26] WHO, *ICD-10 Version:2019*, 2019. Retrieved from, <https://icd.who.int/browse10/2019/en>.
- [27] X. Zhang, B. Yu, T. He, P. Wang, Status and determinants of health services utilization among elderly migrants in China, *Global Health Res. Policy* 3 (1) (2018) 8.