INFACT
Joint Action on Health Information

Deliverable 9.2 (Part B)
Inspiring Examples from European Countries

Implications for Innovative Use of Data Sources

October 23, 2020

Santé publique France

HIGIENOS INSTITUTAS

Submission Date: October 23, 2020

WP9 Lead: Department of Non-Communicable Disease and Injury, Santé Publique France
WP9 Co-Lead: Health Information Centre and Institute of Hygiene, Lithuania
Joint Action Coordination:
Sciensano | Rue Juliette Wytsmanstraat 14 |
1050 Brussels | Belgium| e-mail: infact.coordination@sciensano.be|
Website: www.inf-act.eu|Twitter: @JA_InfAct

This project is co-funded by the Health Programme of the European Union

Table of Contents

I. Summary

Background: The use of data linkage and artificial intelligence to estimating health indicators in public health have several advantages such as data linkage improves completeness and comprehensiveness of information to guide health policy process, whereas artificial intelligence allows to handle data with a large number of dimensions (features) and units (feature vectors) more efficiently with high precision. The capacity to use data linkage and/or the use of artificial intelligence to estimate and predict health indicators varies across EU-MSs (European Member States). However, the estimation of health indicators from linked administrative data is challenging due to several reasons such as variability in data sources and data collection methods, availability of a large number of variables, lack of skills and capacity to analyze big data. Therefore, if the MS or a group of MSs who have developed good practices or inspiring examples using data linkage and/or use of artificial intelligence, they could share and learn from each other. The main objectives of this study are to identify and share inspiring examples related to the innovative use of data sources (i.e., use of data linkage and/or artificial intelligence) and describing their study design, statistical method applied, study limitations, added value and implications of their results in public health in Member States.

Methods: We defined inspiring example as a study that takes into account the use of linked data and/or artificial intelligence (i.e., innovation aspect) to estimate health indicators and implied the related health indicators to target priority public health actions (i.e., surveillance, prevention, promotion, etc.), healthcare strategies or to guide/support public health policies according to their geographical regions, are considered as inspiring examples. We asked the InfAct project partners to share with us any study that corresponds to this definition and had been performed in the past or is ongoing at their institutes, their partner institutes or other research departments in their countries.

Results: We have identified 17 studies (13 studies related to data linkage, 2 studies applied machine learning and 2 studies used both data linkage and machine learning approaches) as inspiring examples from ten European countries. These studies covered 14 different domains of public health. Some of these studies applied classical statistical methods such as multilevel linear regression and some of these studies used artificial intelligence such as machine learning techniques. These studies highlighted that different data collection method, lacking completeness of information or inaccessibility to certain information challenge large linked datasets analysis. Using linked data and AI, the methodological and data analysis aspects can be improved. The results of these studies are used to improve public health surveillance, developing prevention strategies, evaluating health care services and guiding the health policy process.

Conclusions: These inspiring examples support countries to share different experiences and to learn from each other. Furthermore, these examples would help countries to develop, adopt and integrate innovative approaches (i.e., data linkage and artificial intelligence) improving the estimation of health indicators. These examples also allow comparing various approaches used for the innovative use of health information across MS. These inspiring examples support the development of methodological guidelines, potentially improving health indicators estimation using linked data and artificial intelligence. Eventually, the evidence produced by innovative techniques uses allows for better decisions in policymaking.

II. Background

In the past, some initiatives such as BRIDGE-Health (Bridging Information and Data Generation for Evidence-based Health Policy and Research) have been taken by the European commission[1]. The main objectives of these initiatives were to improve the health information system by providing the European Core Health Indicators (ECHI), for comparable health information and knowledge system to monitor health at EU level. The InfAct (Information for Action) project is the continuation of those projects and a joint action of MSs towards a sustainable health information system that supports country knowledge, health research and policy-making process[2]. InfAct gathers 40 national health authorities from 28 Member States (MSs). Innovation in health information for public health policy development is part of this joint action.

Health information is essential to building up country-specific and cross-country knowledge and provides a basis for national health policies. There are large differences in terms of quality and comparability of health information between and within Member States (MSs). The availability of administrative data generated from different sources is increasing and the possibility to link these data sources with other databases[3]. Many countries have already invested in data linkage of their traditional health data systems and increased interoperability[4]. More efficient ways for data linkage of different sources and/or to analyze big datasets by applying artificial intelligence, are required to generate comparable and timely health information across EU/EEA MSs.

The **innovation in health information** is described as the use of 1. linkage of different data sources (health surveys and/or disease-specific and/or population-based registries and/or national cohort and/or clinical research datasets and/or administrative data and/or electronic health records and/or X-data sources) with each other using linkage technology and/or 2. applying artificial intelligence either to linked data or to an individual large dataset, allowing a better understanding of what determines population health or the efficiency of the health system and decision making at different geographical levels or other categorization parameter level. Machine learning (ML) is an application of AI that provides systems the ability to learn automatically and improve from experience without it being explicitly programmed[5].

Using these innovative techniques have several advantages such as data linkage improves completeness and comprehensiveness of information to guide health policy process[6], and artificial intelligence allows to handle data with a large number of dimensions (features) and units (feature vectors) more efficiently with high precision. The capacity to use data linkage and/or the use of artificial intelligence to estimate and predict health indicators varies across EU-MSs (European Member States)[7]. However, the estimation of health indicators from linked administrative data is challenging due to several reasons such as variability in data sources and data collection methods, availability of a large number of variables, lack of skills and capacity to link and analyze big data[8]. Due to varying health information system across MSs, it makes challenging to learn from each other experiences. Therefore, if the MS or a group of MSs who have developed good practices or inspiring examples using data linkage and/or use of artificial intelligence, they could share with others. This could help them to develop, adopt and integrate those approaches according to their country context to improve the quality and comparability of the health information system. This study will provide a mechanism to disseminate those inspiring examples/practices into actions.

III. Objective

The main objectives of this study were to identify and share inspiring examples related to the innovative use of data sources (i.e., use of data linkage and/or artificial intelligence) and describing their study design, statistical method used, study limitations, added value and implications of their results in public health (i.e., health status monitoring or health system performance or health policy process) in Member States.

IV. Methodology

As a first step, we defined the inspiring example in the context of innovation of health information.

Definition of Inspiring examples:

"A study that takes into account the use of linked data and/or artificial intelligence (i.e., innovation aspect) to estimate health indicators and implied the related health indicators to target priority public health actions (i.e., surveillance, prevention, promotion, etc.), healthcare strategies or to guide/support public health policies according to their geographical regions, is considered as an inspiring example."

We asked the InfAct project partners to share with us any study that corresponds to this definition and had been performed in the past or is ongoing at their institutes or partner institutes or other research departments in their country.

We have developed one-page document to report the main contents of inspiring example systematically *(Annex 1)* and shared with InfAct partners. We reported inspiring examples under three main categories based on the use of data linkage and/or artificial intelligence: first related to the data linkage, second machine learning method and third related to both data linkage and machine learning method.

V. Results

We have identified 17 studies as inspiring examples from ten European countries that have been performed and some studies are ongoing. These studies estimated health indicators, either by using data linkage (13 studies), or machine learning methods (2 studies) or both data linkage and machine learning approaches (2 studies). These examples cover the following domains of public health: obesity, injury, psychotic illness, disability and chronic health conditions, industrial pollution and cancer, suicidal prevention, health inequalities, cardiovascular diseases, occupational health of cancer patients, mental health, primary care, epilepsy: neurological disease, vaccine hesitancy and diabetes. A summary of these examples is reported in *table 1*.

Table 1: Summary of inspiring examples from European countries

S/No	European countries	Authors	Domain	Title	Electronic link to the published studies
A	**Data linkage studies**				
1	UK	Kate E Mason et al, 2017	Obesity	Association between fast food and physical activity environments and adiposity in mid-life: cross-sectional, observational evidence from UK Biobank	bit.ly/2WS0Zr3
2	FR	Ludivine Orriols et al, 2014	Injury	Long-term chronic diseases and crash responsibility: A record linkage study	bit.ly/3atYf77
3	UK	Keith Lloyd et al, 2015	Psychotic illness	A national population-based e-cohort of people with psychosis (PsyCymru) linking prospectively ascertained phenotypically rich and genetic data to routinely collected records: Overview, recruitments and linkage	bit.ly/2wMbVvN
4	BE	Johan Van der Heyden et al, 2015	Disability and chronic health conditions	Activity limitations predict health care expenditures in the general population in Belgium	bit.ly/2QWjTcv
5	ES	Pablo Fernández-Navarro et al, 2019	Industrial pollution and cancer	Use of non-health databases for health surveillance: En-risk application	Unpublished study
6	LT	Ausra Zelviene et al, 2019	Suicidal prevention	Innovative use of health information on suicide prevention	Unpublished study
7	SI	Tina Lesnik et al, 2019	Health inequalities	Using individual linked-datasets to monitor health inequalities	Unpublished study
8	IT	Luigi Palmieri et al, 2019	Cardiovascular diseases	Using record linkage to estimating periodically the occurrence and case fatality rate of CVD in different region of Italy	https://bit.ly/3cNPB4R
9	BE	Régine L Kiasuwa-Mbengi et al, 2019	Return to work of patients with cancer/Occupational health of cancer patients	The EMPCAN study: protocol of a population-based cohort study on the evolution of the socio-economic position of workers with cancer *(ongoing)*	bit.ly/2vYNxGZ
10	UK	Amy Mizen et al, 2018	Mental health	Longitudinal access and exposure to green-blue spaces and individual-level mental health and well-being: protocol for a longitudinal, population-wide record-linked natural experiment *(ongoing)*	bit.ly/39sjtkJ
11	AT	Stefan Mathis-Edenhofer et al, 2019	Primary care	Regional health care profiles: Case studies on catchment areas of envisaged primary care units *(ongoing)*	Unpublished study
12	SW	InfAct focal person: Rosita Claesson Wigand et al, 2019	Public health reporting	Web-based public health reporting: The National Public Health Reporting System of Sweden	Description of Swedish national public health monitoring system
13	UK	Samantha Turner et al, 2020	Burden of Injury estimates	Improving Burden of Injury Estimates: Case study based on injuries in Wales, 2012 - 2017	Unpublished study
B	**Machine learning method study**				
14	UK	Brian Cleland et al, 2018	Mental health	Insights into antidepressant prescribing using open health data	bit.ly/2UqQXM3

15	UK	Beata Fonferko-Shadrach et al, 2019	Epilepsy: neurological disease	Use of natural language processing to extract structured epilepsy data from unstructured clinic letters: Development and validation of the ExECT (Extraction of epilepsy clinical text) system	bit.ly/3byZtOK
C	**Both data linkage and machine learning method studies**				
16	HR	A Bell et al, 2019	Vaccine hesitancy	Proactive advising: A machine learning driven approach to vaccine hesitancy	https://bit.ly/3dmLe0e
17	FR	Sonsoles Fuentes et al, 2020	Diabetes	Artificial intelligence for diabetes research: Development of type 1 / 2 diabetes classification algorithm and its application to surveillance using a nationwide population-based medico-administrative database in France	*(Manuscript under revisions)*

Here we describe the summary of important information of background, methods, main results, study limitations, conclusions, added-value and implications of available evidence in public health. We reported these examples under three categories: A. data linkage, B. machine learning and C. both data linkage and machine learning.

A. Examples related to data linkage

1. Association between fast food and physical activity environments and adiposity in mid-life: cross-sectional, observational evidence from UK Biobank

Link to published article: bit.ly/2WS0Zr3

Summary[9]

Background

Obesity is strongly linked to a range of chronic diseases, including type 2 diabetes and cardiovascular disease, and contributes substantially to excess morbidity, mortality, and rising health-care costs globally. Using observational data from UK Biobank, a large sample of adults in a crucial period of the life course for the development of chronic disease, we assessed whether the number of formal physical activity facilities near an individual's place of residence and proximity to fast-food outlets is independently associated with objectively measured adiposity in UK adults. We also explored whether these associations differ by sex or income and whether findings might be affected by residual confounding.

Methods

We used cross-sectional observational data from UK Biobank (i.e., a large population-based cohort). Participants were aged 40–70 years and attended 21 assessment centers between 2006 and 2010.

Exposures: The physical activity environment was defined as the density (count) of formal physical activity facilities (i.e., any land use classified in the commercial-leisure such as leisure centers, gym, swimming pools, etc.) within a 1000 m street-network buffer around each individual's place of residence. For each individual, the street-network distance (in meters) from residential address to the nearest fast-food outlet, classified as "hot/cold fast-food

outlet/takeaway" in the UK Ordnance Survey Address Base Premium database, was available. We used these distances and the distribution of data to categorize individuals as living closer than 500 m, 500–999 m, 1000–1999 m, or at least 2000 m from their nearest fast-food outlet.
Outcomes: Three adiposity outcomes were used to measure the association with exposure indicators (i.e., physical activity facilities and fast-food outlets): waist circumference, body-mass index (BMI), and body fat percentage. Using linked data on environments around each participant's residential address, we examined whether the density of physical activity facilities and proximity to fast-food outlets were associated with waist circumference, body-mass index (BMI), and body fat percentage (i.e., adiposity measures).
Statistical analysis: We used multilevel linear regression models with random intercepts and random coefficients for the main exposure to estimate independent associations between each environmental exposure and each adiposity outcome, accounting for the nesting of individuals within assessment centres. The estimates were adjusted for potential confounders and conducted several sensitivity analyses.

Results
Complete case sample sizes were 401 917 (waist circumference models), 401435 (BMI), and 395 640 (body fat percentage). The greater density of physical activity facilities within 1000 m from home was independently associated with smaller waist circumference and lower BMI and body fat percentage. Compared with people with no nearby physical activity facilities, those with at least six facilities close to home had 1·22 cm smaller waist circumference (95% CI: −1·64 to −0·80), 0·57 kg/m² lower BMI (−0·74 to –0·39), and 0·81 percentage points lower body fat (−1·03 to −0·59) *(table 2)*. Living further from a fast-food outlet was weakly associated with waist circumference and BMI, mostly among women. Compared with people living fewer than 500 m from a fast-food outlet, those living at least 2000 m away had 0·26 cm smaller waist circumference (95% CI: −0·52 to 0·01), 0.10 kg/m² lower BMI (−0.24 to 0.04), and 0.10 percentage points lower body fat (−0.26 to 0.05) *(table 3)*.

Study limitations
The weaker associations for access to fast foods are more likely to be underestimated the available food environment measure, some fast-food outlets as restaurants and the inability to account simultaneously for both healthy and unhealthy food outlets. Although the UK Biobank sample is very large, the response rate was low (5.5%) and the sample showed evidence of healthy volunteer bias. The non-null associations were observed between access to physical activity facilities and height in a negative control analysis suggest that some unmeasured confounding of the physical activity environment associations with adiposity remained. Studies of neighbourhood effects are particularly susceptible to bias arising from residential mobility, where movement between neighbourhoods over time increases the risk of exposure misclassification, and leaves open the potential for reverse causation in cross-sectional analyses. Current adiposity might also reflect exposure to neighbourhood environment earlier in life, posing a further challenge for causal inference. Another limitation was the unavailability of equivalent metric for the physical activity and food environment that might allow a direct comparison of their influence.

Table 2: Association between the density of physical activity facilities and adiposity outcomes: multilevel regression results

	Model 0	Model 1	Model 2	Model 3
Waist circumference, cm (n=401 917)				
Number of facilities				
0	0 (ref)	0 (ref)	0 (ref)	0 (ref)
1	0·15 (−0·02 to 0·32)	−0·15 (−0·30 to −0·01)	−0·13 (−0·26 to 0·00)	−0·19 (−0·32 to −0·06)
2–3	0·08 (−0·20 to 0·35)	−0·43 (−0·70 to −0·17)	−0·29 (−0·51 to −0·08)	−0·40 (−0·62 to −0·17)
4–5	−0·14 (−0·55 to 0·27)	−0·80 (−1·19 to −0·42)	−0·51 (−0·82 to −0·19)	−0·65 (−0·96 to −0·33)
≥6	−0·67 (−1·25 to −0·09)	−1·51 (−2·04 to −0·98)	−1·03 (−1·45 to −0·62)	−1·22 (−1·64 to −0·80)
Body-mass index, kg/m² (n=401 435)				
Number of facilities				
0	0 (ref)	0 (ref)	0 (ref)	0 (ref)
1	0·04 (−0·04 to 0·12)	−0·08 (−0·15 to 0·00)	−0·06 (−0·13 to 0·01)	−0·07 (−0·14 to 0·00)
2–3	−0·03 (−0·15 to 0·10)	−0·22 (−0·34 to −0·09)	−0·16 (−0·26 to −0·05)	−0·18 (−0·28 to −0·08)
4–5	−0·18 (−0·36 to 0·00)	−0·43 (−0·59 to −0·26)	−0·30 (−0·43 to −0·17)	−0·33 (−0·46 to −0·20)
≥6	−0·42 (−0·67 to −0·17)	−0·73 (−0·96 to −0·50)	−0·52 (−0·69 to −0·34)	−0·57 (−0·74 to −0·39)
Body fat, % (n=395 640)				
Number of facilities				
0	0 (ref)	0 (ref)	0 (ref)	0 (ref)
1	0·03 (−0·07 to 0·14)	−0·11 (−0·20 to −0·01)	−0·08 (−0·16 to 0·00)	−0·11 (−0·20 to −0·03)
2–3	−0·07 (−0·24 to 0·11)	−0·30 (−0·46 to −0·13)	−0·21 (−0·34 to −0·07)	−0·27 (−0·40 to −0·13)
4–5	−0·28 (−0·53 to −0·02)	−0·58 (−0·81 to −0·35)	−0·40 (−0·58 to −0·22)	−0·48 (−0·67 to −0·29)
≥6	−0·63 (−0·96 to −0·30)	−1·00 (−1·29 to −0·70)	−0·71 (−0·92 to −0·49)	−0·81 (−1·03 to −0·59)

Density is defined as number of physical activity facilities in a 1000 m street-network buffer. Data are mean difference (95% CI). Model 0: adjusted for age and sex. Model 1: model 0 plus adjustment for ethnicity, urban or non-urban status, and area deprivation. Model 2: model 1 plus adjustment for individual socioeconomic characteristics (income, education, and employment status). Model 3: model 2 plus adjustment for residential density and distance to nearest fast-food outlet.

Table 3: Associations between proximity to fast-food outlets and adiposity outcomes: multilevel regression levels

	Model 0	Model 1	Model 2	Model 3
Waist circumference, cm (n=401 917)				
Distance				
<500 m	0 (ref)	0 (ref)	0 (ref)	0 (ref)
500–999 m	−0·48 (−0·67 to −0·29)	−0·08 (−0·24 to 0·07)	−0·07 (−0·22 to 0·07)	−0·15 (−0·30 to −0·01)
1000–1999 m	−0·69 (−1·01 to −0·38)	−0·04 (−0·29 to 0·22)	−0·08 (−0·27 to 0·12)	−0·22 (−0·44 to 0·00)
≥2000 m	−1·20 (−1·59 to −0·81)	−0·05 (−0·40 to 0·30)	−0·11 (−0·36 to 0·14)	−0·26 (−0·52 to 0·01)
Body-mass index, kg/m² (n=401 435)				
Distance				
<500 m	0 (ref)	0 (ref)	0 (ref)	0 (ref)
500–999 m	−0·20 (−0·29 to −0·11)	−0·04 (−0·10 to 0·03)	−0·03 (−0·09 to 0·03)	−0·08 (−0·14 to −0·02)
1000–1999 m	−0·24 (−0·39 to −0·09)	0·01 (−0·11 to 0·12)	−0·01 (−0·10 to 0·08)	−0·10 (−0·20 to −0·01)
≥2000 m	−0·40 (−0·60 to −0·21)	0·03 (−0·15 to 0·22)	0·01 (−0·13 to 0·14)	−0·10 (−0·24 to 0·04)
Body fat, % (n=395 640)				
Distance				
<500 m	0 (ref)	0 (ref)	0 (ref)	0 (ref)
500–999 m	−0·18 (−0·27 to −0·08)	−0·02 (−0·10 to 0·07)	−0·02 (−0·10 to 0·06)	−0·08 (−0·16 to 0·00)
1000–1999 m	−0·19 (−0·35 to −0·03)	0·07 (−0·07 to 0·21)	0·04 (−0·07 to 0·15)	−0·07 (−0·19 to 0·05)
≥2000 m	−0·46 (−0·67 to −0·24)	0·05 (−0·15 to 0·26)	0·00 (−0·14 to 0·15)	−0·10 (−0·26 to 0·05)

Distances given are distance to nearest fast-food outlet. Data are mean difference (95% CI). Model 0: adjusted for age and sex. Model 1: model 0 plus adjustment for ethnicity, urban or non-urban status, and area deprivation. Model 2: model 1 plus adjustment for individual socioeconomic characteristics (income, education, and employment status). Model 3: model 2 plus adjustment for residential density and density of local physical activity facilities.

Conclusions

This study shows strong associations between high densities of physical activity facilities and lower adiposity for adults in mid-life. We observed weaker associations for access to fast food, but these are likely to be underestimated owing to limitations of the food environment measure. Policy makers should consider interventions aimed at tackling the obesogenic built environment

Added value of this study

This study is one of the first published to use UK Biobank to examine associations between features of the neighbourhood built environment and adiposity. The sample is made up of adults in mid-life—a crucial period of the life course for the development of chronic disease. By using a very large dataset that covers much of the UK, enabled to provide evidence that relates to a wider geographical area than do most UK-based studies, and to examine sex and income differences. Making use of extensive covariate data available in UK Biobank, we were able to more comprehensively adjust for sources of confounding than have many other studies, and with additional sensitivity analyses, we were able to examine the robustness of our findings to residual confounding and different model specifications, further strengthening our findings.

Implications of available evidence

The results of this study provide evidence to support the hypothesis that increasing access to local physical activity facilities and, possibly, reducing access to fast-food close to residential areas has the potential to reduce overweight and obesity at the population level. Policy-makers

should consider interventions aimed at modifying residential environments to better facilitate healthy lifestyles but recognizing that such an approach might be more effective in some groups than in others.

2. Long-term chronic diseases and crash responsibility: a record linkage study

Link to published article: bit.ly/3atYf77

Summary[10]

Background
Driving has become a basic activity of daily living. Available data suggest that the impact of health is an important issue regarding driving abilities and road safety. In France, chronic diseases are registered in the healthcare insurance database. We took advantage of a nationwide record linkage study to investigate the risk of being responsible for a road traffic crash associated with the presence of long-term disease.

Methods
A case-control analysis comparing responsible versus non-responsible drivers was conducted. Data from three French national databases were extracted and matched: the national health care insurance database, police reports and the national police database of injurious crashes. Cases were defined as drivers who were deemed responsible for the crash, while controls were drivers who were not responsible. Responsibility levels in the crash were determined by a standardized method adapted from Robertson and Drummer 1994. This method, previously used in France, takes into consideration six factors likely to reduce driver responsibility for the crash: road environment, vehicle-related factors, traffic conditions, type of accident, traffic rule obedience and difficulty of the driving task. Exposure to chronic conditions were compared between responsible and non-responsible drivers. In France, patients are fully reimbursed for health care expenses related to recognized long-term diseases. In the analysis, diseases were classified according to the International Classification of Diseases 10th edition (ICD-10) code that is available in the HCI database (299 ICD-10 codes).
Statistical analysis: Logistic regression was used to compare the excluded and included subjects by taking into account six factors (as described above). Lasso (least absolute shrinkage and selection operator) method was used to fit a single model adjusted for crash-related and sociodemographic factors, including all the 299 long-term diseases as covariates. Adjustment variables (age, gender, socioeconomic category, year, season, day, time and location of crash, vehicle type, injury severity, blood alcohol concentration and exposure to level 2 and 3 medicines) were forced into the model; the proper amount of shrinkage of the long-term disease covariates was estimated using the Akaike information criterion (AIC) and was corrected for bias.

Results
69,630 drivers involved in an injurious crash in France between 2005 and 2008, were included. 6210 (8.9%) were suffering from at least one long-term disease. Among them, 866 were suffering from two long-term diseases, 125 from three long-term diseases and 24 from more than three. Some disorders were more frequent among responsible drivers than among the French population, in particular, long-term psychiatric disorders (21.1% versus 14.4%) and chronic active liver disease and cirrhosis (4.2% versus 2.4%). When adjusted for prescription of medicines, blood

alcohol, demographic driver characteristics and crash characteristics, the increased risk of being responsible for a crash, was found in drivers registered in the French healthcare database with five following long-term diseases: epilepsy (odds ratio [OR]=2.53 [95% CI: 1.53-4.20]), type 1 diabetes (OR=1.47) [1.12-1.92], alcoholic liver disease (OR=3.37 [1.40-8.13]), asthma (OR=1.72 [1.13-2.60]) and specific personality disorders (OR=1.35 [1.05-1.74]) *(table 1)*. No association was found for cardiovascular diseases or Alzheimer's disease. The combined prevalence of these diseases in drivers was 1.1%.

Table 1: Long-term diseases selected with the different modeling strategies: bootstrap-enhanced Lasso, separate logistic regression with and without Bonferroni correction

	n	Mean age (SD)	Bias corrected OR [IC 95%][a]	p-Value[b]
Epilepsy	80	38.7 (10.3)	2.53 [1.53–4.20]	3 × 10−4[c]
Insulin-dependent diabetes	238	46.8 (16.1)	1.47 [1.12–1.92]	0.0047
Alcohol liver disease	37	54.7 (10.2)	3.37 [1.40–8.13]	0.0061
Asthma	105	46.5 (17.1)	1.72 [1.13–2.60]	0.0084
Specific personality disorders	298	41.9 (12.0)	1.35 [1.05–1.74]	0.0190

a. *Lasso logistic regression, adjusted for age, gender, socioeconomic category, month, time of day, vehicle type, alcohol level, injury severity, exposure to medicines affecting driving abilities and other long-term diseases.*
b. *p-Value obtained in the separate logistic regression model, adjusted for age, gender, socioeconomic category, month, time of day, vehicle type, alcohol level, injury severity, exposure to medicines affecting driving abilities and other long-term diseases.*
c. *Significant after Bonferroni correction (p < (0.05/299).*

Study limitations

The method used to analyze crash responsibility level does not capture the risk for non-responsible drivers, of being unable to avoid a crash, which may be linked to the presence of a disease. This would underestimate the results. The limited association found between responsibility crash and long-term disease could be due to a 'risk compensation phenomenon'. Drivers with a well-identified chronic disease may try to alleviate their risk with lower speed, shorter distance, avoidance of roads perceived as hazardous and some of them may drive less during that study period. The analyses were not adjusted for the amount of driving, as this information was not available. It was not possible to account for crashes that occurred before the study period. We did not have information on all diseases and only long-term diseases that imply chronic treatment and expensive medical care are registered in the database. The request for registration leading to full reimbursement of health care expenses related to the disease must be made by the practitioner and this initiative is not systematic. As the request has to be approved by the healthcare insurance authorities, it can take some time for a disease to be registered, so the actual date of symptom occurrence might be uncertain. One limitation of the Lasso method is that with a proper amount of shrinkage relevant covariates are retained and to address this issue, the bootstrap-enhanced Lasso approach was applied.

Conclusions

The crash responsibility study of 69,630 drivers from a French national record linkage study found an increased risk of being responsible for an injurious road traffic crash in drivers registered in

the French long-term disease database with the five following long-term diseases: epilepsy, type 1 diabetes, alcoholic liver disease, asthma and specific personality disorders. However, results should be considered cautiously with regards to potential regulatory driving judgments that could have a negative impact on patients' social life.

Added value of this study
Using the French healthcare insurance database has several advantages: data collection is prospective and medicine dispensation is registered. The high statistical power achieved by a record linkage method, combined with the Lasso analysis, allowed us to study some long-term diseases, including the rarest ones.

Implications of available evidence
These results showed an impact of several chronic diseases on the risk of road traffic crashes and may update the list of medical conditions that may impair driving skills. Providing patients with proper information on the effect of these diseases on their abilities to drive may improve their driving behaviours.

3. A national population-based e-cohort of people with psychosis (PsyCymru) linking prospectively ascertained phenotypically rich and genetic data to routinely collected records: Overview, recruitments and linkage

Link to published article: bit.ly/2wMbVvN

Summary[6]

Background
Schizophrenia, bipolar disorder and other psychoses are complex inter-related disorders, both genetically and neuro-developmental aspects. Gene-environment interactions can run in both directions with environmental and psycho-social factors contributing to explanations of variance. There is an urgent need to understand the exact pathogenesis and underlying biological mechanisms resulting from gene-environment interactions which give rise to specific symptoms. The overarching purpose of creating this unique research platform is to develop our understanding of the causes, course and outcomes of psychosis. If more is understood about the pathology of these diseases, a contribution can be made to the development of better diagnostic, predictive, preventative and therapeutic approaches. This paper describes the recruitment process and data collected for the creation of a well-characterised e-cohort. It also describes the linkage of this data to routinely collected health and administrative data and the creation of a much larger e-cohort from routinely collected primary care sources.

Methods
PsyCymru was initially established as a proof of concept to investigate the feasibility of linking a prospectively ascertained, well-characterised (linked clinical cohort) of people with psychosis in Wales, UK with large amounts of anonymised routinely collected health record data. We are now additionally linking genetic data. PsyCymru aims to create a research platform and infrastructure for psychosis research in Wales by the establishment of two cohorts. The first is a well-characterised clinically assessed cohort of 490 individuals aged 16 and over, including

genetic data. Consented individuals underwent a structured interview using a series of well-validated questionnaires and gave blood samples for DNA extraction for sequencing and candidate gene identification. This data was linked to routinely collected health and social datasets with identity encryption used to protect privacy. The second is a much larger (12,097 individuals) but less well characterized population-based e-cohort of prevalent psychosis cases created using a previously validated algorithm applied to anonymised routine data. Both cohorts can be tracked prospectively and retrospectively using anonymised routinely collected electronic health and administrative data in the Secure Anonymised Information Linkage (SAIL) databank.

Results

We have successfully imported the results from the questionnaires into SAIL and allocated an Anonymous Linking Field (ALF) for the majority (92.4%) of the participants in the study. The ALF is the unique anonymous identifier for an individual and remains the same irrespective of the data source. To create a population-based electronic cohort, the algorithm against the "gold standard" was already established using operational criteria checklist for psychotic and affective illness (OPCRIT). The algorithm had favourable test characteristics, with a very good ability to detect patients with psychotic disorders (sensitivity > 0.7) and an excellent ability not to falsely identify patients with psychotic disorders (specificity > 0.9) in primary care. This exercise has allowed us to create a validated e-cohort for research purposes. The e-cohort (n=12,097) is consistent with the literature in terms of gender distribution, deprivation and urbanicity. There were more males than females with schizophrenia and other psychotic disorder groups, whilst a greater proportion of females had bipolar disorders. Over half of the cohort 8204/12,097 (67.82%) were issued with a prescription for a psychiatric-related drug with polypharmacy being common and 3893/12,097 (32.18%) were untreated during 2010.

Study limitations

The limitations of this research platform include the standard issues inherent in any routinely collected data source such as data quality and completeness. For example, completeness varies between datasets and individual variables, prescriptions are very well recorded in primary care, whilst data relating to employment, occupation and socioeconomic status is not readily available. A further limitation is that the amount of historical data available varies between individuals and between datasets. Re-use of routine data requires a detailed understanding of the nuances of each of the datasets.

Conclusions

This unique platform pools data together from multiple sources; linking clinical, psychological, biological, genetic and health care factors to address a wide variety of research questions. This resource will continue to expand over the coming years in size, breadth and depth of data, with continued recruitment and additional measures planned.

Added value of this study

One of the unique features of this study is the ability to link study participants' clinical and genetic data to routinely collected health and administrative data in the SAIL databank and then follow these individuals through time, all within a framework that ensures privacy protection.
Retrospective data available can be used to understand disease progression and evaluate potential risk factors and the ongoing routine updates provide us with the opportunity to track

and follow-up these individuals across multiple health care settings in a cost-effective and in-obtrusive manner.

Implications of available evidence

Using data from two linked clinical cohorts, long-term outcomes, surveillance from ongoing and novel treatment interventions and how health services are utilized and benefited from, can be evaluated offering the potential to improve health care delivery.

4. Activity limitation predict health care expenditures in the general population in Belgium

Link to published article: bit.ly/2QWjTcv

Summary[11]

Background

Health care expenditures represent an increasing part of the GDP in all OECD countries. In Belgium, the proportion of the GDP devoted to health care rose from 8.1% to 10.5% between 2000 and 2011. There is no doubt that - apart from the development of new technologies and drugs - population ageing and the associated higher burden of ill-health have contributed to this trend.

Disability and chronic conditions both have an impact on health expenditures and although they are conceptually related, they present different dimensions of ill-health. Recent concepts of disability combine a biological understanding of impairment with the social dimension of activity limitation and resulted in the development of the Global Activity Limitation Indicator (GALI). This study investigated the relationship between activity limitations on health care expenditures in Belgium concerning chronic conditions. This was assessed for the total health expenses and reimbursed and out-of-pocket expenses separately. Furthermore, it was assessed to which extent differences in health care expenditures by activity limitation can be explained or differ by socio-demographic characteristics. It is reported by the predictive value of the GALI on health care expenditures in relation to the presence of chronic conditions.

Methods

Data from the Belgian Health Interview Survey 2008 were linked with data from the compulsory national health insurance to obtain health care expenditures of the final study sample (n = 7,286).

Health care expenditures were obtained for the years 2008, 2009 and 2010. A distinction was made between health care expenditures for (1) ambulatory care (excluding the cost of pharmaceuticals), (2) hospital care and (3) reimbursed medicines obtained in pharmacies. Activity limitations were measured with the GALI. The initial version of the question was used: “For the past 6 months or more have you been limited in activities people usually do because of a health problem? (Yes, strongly limited/Yes, limited/No, not limited)”. Information was available on the one-year prevalence of 30 specific chronic conditions and health problems. Chronic conditions were included in the multivariate model if there appeared to be an independent association with health expenditure after adjustment for age, gender and the other chronic conditions. The effect of activity limitation on health care expenditures was assessed via cost ratios from multivariate linear regression models. To study the factors contributing to the difference in health expenditure between persons with and without activity limitations, the

Blinder-Oaxaca decomposition method was used. The Blinder-Oaxaca technique demonstrates the relative importance of each predictor. The decomposition illustrates the fraction of the gap in health care expenditures that is attributable to group differences in the magnitude of the determinants (the explained or prevalence component) and group differences in the effects of these determinants (the unexplained or impact component). The Blinder-Oaxaca decomposition method is particularly useful to study differences in health care expenditures between two groups, but it has also been used in studies in which the contribution of both the prevalence and the impact of determinants to explain differences between groups, was investigated for other health outcomes.

Results

Activity limitations are a strong determinant of health care expenditures. People with severe activity limitations 461 (5.1%) and moderate activity limitations 1364 (14.9%) accounted for 16.9% and 31.6% of the total health expenditure respectively, whereas those without activity limitations 5461 (79.0%) were responsible for 51.5% of the total health expenditure. On average 57.0% of the expenses were ambulatory costs, 21.8% hospital costs (excluding fixed costs, as explained in the methods section), 18.2% costs for reimbursed medicines obtained in pharmacies and 3.0% not specified. The large majority of expenses (84.0%) were covered by the health insurance, 10.9% were out-of-pocket payments and 5.1% supplements. These observed differences in health care expenditures can to some extent be explained by chronic conditions, but activity limitations also contribute substantially to higher health care expenditures in the absence of chronic conditions (cost ratio 2.46; 95% CI 1.74-3.48 for moderate and 4.45; 95% CI 2.47-8.02 for severe activity limitations). The association between activity limitation and health care expenditures is stronger for reimbursed health care costs than for out-of-pocket payments.

Study limitations

The initial study sample was representative of the total population, some groups were excluded due to people for whom no linkage could be done (4.4%), people for whom no self-administered questionnaire was available, mostly because of a proxy interview (18.6%), people who did not answer the GALI question (1.2%) and respondents who died in the 12 months following participation in the survey (less than 0.5% of the total). Even though it is known that health care expenditures increase drastically during the last year of life, it is assumed that the exclusion of the latter group did not affect the results substantially, because of its small size. In contrast, the exclusion of proxy interviews, in which information on the GALI is lacking, may have had a bigger impact on the results. Respondents interviewed by proxy are probably in worse health, and therefore have higher health care expenditures than self-respondents. Whereas this may have influenced the level of the health care expenditures, it is not sure that this had an impact on the associations that were investigated. Some health care expenditures were not included, e.g. fixed day fees for a hospital stay. The same applies for health care expenditures that were not reported within the compulsory health insurance system, although these represent only a marginal fraction in comparison with the reported health care expenditures. Information on chronic conditions was based on self-reports. The validity of self-reported specific chronic diseases is limited and strongly depends on the type of disease. Furthermore, chronic conditions included in the study were restricted to the list of diseases available in the Belgian HIS 2008. Measuring disability is challenging. Although the GALI has been validated, it remains a self-reported item, with several crucial conceptual elements included in one question.

Conclusions
Activity limitations, both moderate and severe, are a major driver for health care expenditures. This is particularly the case for reimbursed health care expenditures. Chronic conditions explain to a certain extent, differences in health care expenditures by the level of activity limitation. However, in the absence of chronic conditions, activity limitations appear to be an important determinant of health care expenditures. To make projections on health care expenditures, routine data on activity limitations are essential and complementary to data on chronic, conditions.

Added value of this study
This is the first published study to investigate differences in health care expenditures by disability status in a representative sample of a European country. The Belgian health system is based on a Bismarckian model, which is essentially characterized by a premium financed social insurance system with a mixture of public and private providers. This is opposed to the Beveridge model in the UK and the Scandinavian countries, based on taxation with many public providers, and the private insurance model of the US.

Implications of available evidence
Our findings are also important from a health policy perspective. In the planning of strategies for health care cost containment, policymakers should not only consider the impact of chronic diseases on health care expenditures but also be aware of the role of disability and its consequences. Reducing activity limitations in the population, e.g. by measures that facilitate the participation of people with functional limitations in the society, is cost-effective, not only because it increases the quality of life and the productivity of people, but also because of its direct impact on the health care costs.

5. Use of non-health EU-databases for health surveillance: En-risk Application

Unpublished study

Summary

Background
Being able to combine health information with environmental health determinants is very important, both for surveillance or epidemiological monitoring and for risk studies in health. Within the European Union, many non-health data can be used. Its integration with health data is difficult, represents an important challenge, and requires specific expertise. A good example of a potentially useful source of significant environmental data relevant for health is the European Pollutant Release and Transfer Register (E-PRTR), which allows estimating exposure to industrial pollution. We aimed to develop an easy-to-use tool that, without requiring advanced statistical knowledge, allows performing an initial screening suggesting the presence/absence of excess risk of a disease linked to residential proximity to industrial pollution.

Methods
En-risk: A java interactive tool was developed to merge the information of E-PRTR and the municipal mortality/morbidity data and to perform an exploratory spatial analysis of the association between them by type of industrial facility using distance as a proxy of exposure.

The application needs cartography of the country and a database of the annual observed deaths (mortality) or cases (morbidity) and population broken down by age groups (18) and by sex per municipality. With this information, the application directly calculates the expected number of deaths or cases of the selected disease, using as reference the rates by age group and sex for the whole country; b) the distance from the municipal centroids (information obtained from the shapefile) to the location of all the industrial facilities included in the E-PRTR. These distances allow classifying municipalities as exposed or not exposed to industrial pollution, according to the already published methodological criteria[1]. With these elements, it performs a spatial association analysis to evaluate whether there is any excess of mortality/morbidity in those municipalities exposed to industrial pollution compared to those not exposed, globally and by industrial sectors. In summary, relative risks (RRs) of dying from cancer between exposed and non-exposed municipalities are estimated using Bayesian conditional autoregressive models proposed by Besag et al.[2] with explanatory variables. The industrial pollution exposure was defined as the proximity of population centroids to pollutant sources, considering towns without any nearby pollutant industry as the reference for comparison purposes. The random-effects terms included two components: a spatial term containing municipal contiguities and the municipal heterogeneity term to control for the possible spatial effects of dependence and heterogeneity.
If the user has loaded additional information (social and economic environment information), the analysis could be also performed considering them as possible confounding factors. Finally, municipal lung cancer deaths (2005-2009) in Spain provided by the National Institute of Statistics were analyzed with this application as an example.

Results
En-risk gives a table of Relative Risks of mortality/morbidity due to exposure to industrial pollution by industrial sector and sex. The analysis using the municipal lung cancer deaths in Spain showed an excess of mortality associated with the proximity to several industrial sectors.

Study limitations
The ecological study that is performed with the application may display biases deriving from errors of classification of what is deemed to be exposed. In the statistical analysis, the estimation of population exposure to industrial pollution is based on the distance from town centroids to industrial facilities. We assumed an isotropic model of exposure. This could introduce a problem of misclassification, because real exposure is critically dependent on other variables, such as prevailing winds or geographic landforms. However, unless the errors of classification of exposure are different for the groups being compared, these types of biases tend to mask the associations by shifting the relative risks towards unity.
A further possible bias lies in the use of centroids as coordinates for pinpointing the entire population of a town, when, in reality, the population may be widely dispersed. We assumed that the whole municipal population was exposed to the same type and amount of pollutant substances. Nevertheless, the use of small areas as units reduces the risks of ecologic bias and misclassification stemming from these assumptions.
Other limitations are related to the impossibility of estimating intensity, duration and variability of exposure. Populations residing in the proximity of pollutant industries could potentially be exposed to large amounts of toxic substances. However, it was not possible to estimate the intensity, duration or even the variability of exposure, due to a lack of knowledge about the dates when emissions began, the annual amounts involved, and the influence of meteorology on the dispersion and spread of pollutants. Moreover, the heterogeneity of the industrial facilities

within the same industrial group or other non-industrial sources of carcinogenic pollutants were not taken into account, which are other possible sources of bias.
Due to the methodological shortcomings and limitations mentioned, these studies are considered exploratory and do not allow causal associations to be established. Moreover, the mechanisms for disparities in disease survival or other parameters are multidimensional and vary according to the type of disease or the specific health care system involved. These mechanisms may pertain to screening, treatment, diagnostic conditions, access to specialized care, or follow-up modalities, possibly inducing spatial heterogeneities in mortality or morbidity.

Conclusions
En-risk facilitates the study of the relationship between industrial pollution and health all around Europe. It can be used by public health services to identify health problems.

Added value of this study
En-risk performs an initial screening, suggesting the presence/absence of excess risk of a disease linked to residential proximity to industrial pollution. The results, whose interpretation clearly needs public health expertise, can generate useful hypotheses to carry out ad-hoc studies to deepen into it.

Implications of available evidence
The application can be used by public health services to identify health problems and to point to key policy interventions to reduce the impact of industrial pollution on health. Finally, the same approach, handy and cheap, can be applied to other geographically based European environmental databases and its sustainability is clear, because it is a normative tool and might improve the interoperability of health information systems with non-health data that would be included in machine learning algorithms in the future.

[1] *Fernández-Navarro, P. et al, G. Industrial pollution and cancer in Spain: An important public health issue. Environ. Res. 2017*
[2] *Besag, J., York, J., Mollié, A., 1991. Bayesian image restoration, with two applications in spatial statistics. Ann. Inst. Stat. Math.*

6. Innovative use of health information on suicide prevention in Lithuania

Unpublished study

Summary

Background
According to the World Health Organization, Lithuania has the highest suicide rate in Europe and one of the highest in the world. Lithuania states this problem as a priority in the strategic documents puts efforts in trying to solve the suicide problem, but the results remain failing. Approximately 1000 people in Lithuania die by suicide, and about 5000 people receive treatment in the emergency department each year due to deliberate self-harm[1]. The high suicide rates in Lithuania have been associated with a range of factors, including rapid socioeconomic transition, increasing psychological and social insecurity, and the absence of a national suicide prevention strategy[2].

[1] *Institute of Hygiene, 2015*
[2] *OECD, 2012*

The general practitioner (GP) does consultations of a large number of patients who subsequently commit suicide. This suggests an approach to case finding based upon risk factors, sensitivity to high-risk situations in depressed patients, and assessment of suicidality in patients being treated for depression are appropriate in the primary care setting. This may uncover occasional patients who make their intentions known and are amenable to intervention.
The main objectives of this study were to determine the use of health services before death by suicide and the registered mental or behavioural disorder in the year before death by suicide.

Methods
We performed a case study at the national level. Record linkage was employed to investigate the health service use and demographics of Lithuanians who died by suicide. The death records were obtained from Causes of death register that was linked to health service records from the National Health Insurance Fund using demographic personal data –sex, date of birth, date of death and registration municipality. The demographic data available included sex, age and residence location. Using a mortality database maintained by the national institute of Hygiene records with suicide coded as the underlying cause of death (ICD-10-AM codes X60-X84) for two-years from January 1, 2013, to December 31, 2014, were selected. 99.1% (1997/2015) of persons who committed suicide were linked using demographic data.
We estimated the following health indicators: use of health services by sex, age, place of residence;
use of GP services in a year before suicide in less than 90, 180, 365 days before suicide by sex, age, place of residence; use of mental health services in a year before suicide in less than 90, 180, 365 days before suicide by sex, age, place of residence; use of emergency care services in a year before suicide in less than 90, 180, 365 days before suicide by sex, age, place of residence; and presence of mental illness (recorded ICD F code) by sex, age, place of residence.

Results
The 80% (1602/1997) of people who later committed suicide had had at least one patient record in the previous year. A higher percentage of women than men had accessed health services. The use of health services was also higher for the residents of urban areas than for their rural counterparts. Almost all suicides in the 19 years and under age group had had at least one patient record in the year before death (table 1).

Table 1. Use of health services in the year before death by suicide

Variable	User (n=1602)		Non user (n=395)		Total (n=1997)	
	n	%	n	%	n	%
Gender						
Male	1283	78.3	356	21.7	1639	100
Female	319	89.1	39	10.9	358	100
Age						
19 and under	71	93.4	5	6.6	76	100
20-34	297	76.2	93	23.8	390	100
35-49	400	75.5	130	24.5	530	100
50-64	460	78.5	126	21.5	586	100
65+	374	90.1	41	9.9	415	100
Place of residence						
Urban	903	83.8	175	16.2	1078	100
Rural	699	76.1	220	23.9	919	100

Two-thirds of suicides had had at least one GP visit in the previous year. The use of psychiatric services was not that usual – less than one-third of persons who committed a suicided visited a psychiatrist in a year before a suicide.

Mental illness is consistently cited as one of the key factors contributing to suicide risk. The 28% (556/1997) of the study population had had a registered mental or behavioural disorder within one year before death by suicide. Registered mental health illnesses were more prevalent among females and people living in urban areas. Mental disorders were more prevalent among older patients (Table 2).

Table 2. Registered mental or behavioural disorder in the year before death by suicide

Variable	F code in ICD-10-AM (n=556)		No F code in ICD-10-AM (n=1046)		Total (n=1602)	
	n	%	n	%	n	%
Gender						
Male	406	31.6	877	68.4	1283	100
Female	150	47.0	169	53.0	319	100
Age						
19 and under	11	15.5	60	84.5	71	100
20-34	84	28.3	213	71.7	297	100
35-49	132	33.0	268	67.0	400	100
50-64	172	37.4	288	62.6	460	100
65+	157	42.0	217	58.0	374	100
Place of residence						
Urban	348	38.5	555	61.5	903	100
Rural	208	29.8	491	70.2	699	100

Discussion

Our results highlighted that almost 80% of people who later committed suicide had had at least one patient record in the previous year. Among these people, 28% had a registered mental or behavioural disorder within one year before death by suicide. With the current degree of implementation of existing policies and clinical practice guidelines, GPs are the first health professionals encountered by many patients, it is critical for them to identify and to manage their mental health problems. In Lithuania, a study[3] has demonstrated a huge discrepancy between the sense of responsibility that many GPs felt about managing patients' mental health problems and their self-perceived competencies. Only 8.8% of GPs mentioned that their knowledge of mental healthcare is sufficient. This could be a possible explanation for the large unmet mental healthcare need in primary healthcare. Study participants suggested that the capacity building of GPs in mental healthcare could be a promising intervention to increase their involvement in the field. Some studies have been performed in other countries that also emphasised the need to improve GPs' knowledge and skills in diagnosing and treating mental health disorders. Our study did not find differences in self-perceived competencies according to the GPs' age. This suggests that capacity-building interventions should target all age groups - the further education of active GPs and the curriculum of undergraduate training as well as the

[3] *Lina Jaruseviciene et al, Preparedness of Lithuanian general practitioners to provide mental healthcare services: a cross-sectional survey. Int J Ment Health Syst. 2014*

residency in family medicine to ensure that graduates join the workforce with adequate knowledge of mental illness and will improve their competencies in diagnosis and treatment.

Study limitations
The study has some limitations. The data was extracted from the administrative sources and the purpose of their collection is not for epidemiological research. The data from the private sector, which has no contract with the National insurance fund, is not available in the national database, for example, some psychiatric services are provided by the private healthcare sector. One potential limitation is about stigma in writing psychiatric diagnosis and trying to avoid them even if they are present.

Conclusions
In this study, linked data was used to evaluate the implementation of suicide prevention actions. GPs are the first health professionals encountered by many patients, they must be able to identify and to manage their mental health problems. Mentally ill patients are vulnerable to suicide. 28% of the study population had had a registered mental or behavioural disorder within one year before death by suicide.

Added value of this study
The study was conducted to identify the scale of people entering the health system who can potentially commit suicide and to provide evidence on feasible actions to prevent potential suicide cases. Moreover, these results would provide mental health services. Data linkage of various data sources such as death registry with primary health care, mental health, emergency care services and hospital discharge data, allowed identifying the potential suicide cases.

Implications of available evidence
The results of this study were used to develop suicide prevention strategies and were implemented in Lithuania. The results of this study are in line with research findings elsewhere as Lithuanian GPs try to provide healthcare services for mentally ill patients and suggestions for how to improve the situation. As it has become more evident that different health systems face similar difficulties in integrating mental healthcare into primary care. Therefore, it is an increasing need for generalizable solutions to these problems and the transferability of solutions should be assessed in future research.

7. Using individual linked datasets to monitor health inequalities in Slovenia

Unpublished study

Summary

Background
Health inequalities are the main contributor to morbidity and mortality in all over the world. Monitoring and reporting of health inequalities is fundamental to encourage public discussion and political actions on strategies and measures to decrease health inequalities. Systematic monitoring of health status in relation to socioeconomic determinants is among the objectives of the Resolution on the National Health Care Plan 2016-2025. By now, monitoring of socioeconomic determinants of health has not become routine and does not feature in periodic

reports except for a few ones, which are dedicated to health inequalities. For that purpose, we have developed and tested several pathways for retrieving and linking data on health status with socioeconomic data from different sources and stratifying data at various levels. These newly established pathways of data linkage can be further developed and upgraded to the routine monitoring system of health inequalities. Our second national report of 2018, foreseen by national health care strategy aimed to present health inequalities in Slovenia. One of the main research questions was "what was the impact of the last economic crisis on health inequalities?" Monitoring health inequalities with the sound data linked at an individual-based dataset in Slovenia was the most desirable scenario. To disaggregate lifestyle and health outcome indicators according to various socio-economic distribution/strata, we linked different health databases with census data on educational attainment. The linkage was made using a personal identification number (PIN). Other determinants include such as profession, income, work position was also assessed but were not of enough good quality to be used for linkage.

Methods

We performed a study using data from the national health report on health inequalities in Slovenia. We compared different health outcomes incidence or prevalence by socio-economic determinants and time trends at the national level. We used the individual-level data, registered in the population-based registries/databases such as mortality data, perinatal information system, hospital discharge database and drug prescription database. These data sources are linked with national register-based census data that is managed annually. We use a deterministic type of linkage.

We made some assumptions at the beginning of this study such as:

- The deterministic linkage will be used due to several reasons such as PIN number is of excellent quality in all data sources, with probabilistic linkage some marginal but for this analysis, important groups might be neglected or overlooked.
- The highest level of achieved education will be used as a proxy measure of socioeconomic status; the income of the household was not available at that time. The information about educational attainment was available for the whole population since 2011 (the first register-based census).
- Our report will not be focused on regional inequalities due to another on-going parallel project. Nevertheless, the linked data would also enable the estimation of health inequality indicators at the regional level.

The level of highest educational achievement was presented as the socioeconomic covariates. This dataset is held by the National Statistical Office and is built from different administrative data sources, and comprises the total Slovenian population. Statistical Office translates the data on education from a national classification (Klasius) to the "International Standard Classification of Education (ISCED)" that was used to present different educational levels. For the reporting purpose, we classified the level of education into three main groups (low, medium and high education). We selected the following indicators to compare them by educational gradient and their change over time: life expectancy at the age of 30 years; health expectancy at 30 years of age; beginning of life; youth health; self-assessment of health; smoking and lung cancer; vaccination again tick-borne meningoencephalitis; alcohol and mortality; nutrition, physical activity, obesity and cardiovascular diseases; mental health; mortality due to unintentional injuries; and elderly's people health.

Results

The results are presented in the national health inequality report of 2011. It is available in Slovenian language available at the following link and we plan to translate that into the English language: http://www.nijz.si/sl/publikacije/neenakosti-v-zdravju-v-sloveniji-v-casu-ekonomske-krize.

The linkage of different health databases and sociodemographic database resulted in analyses of health inequalities based on educational attainment differences. Our results showed as stated in the report published in 2011 "the health inequalities in Slovenia still exist in most of the selected indicators. The main difference is usually found with persons of lower socio-economic situations are also related to the level of achieved education."

The main results of this study showed that the educational gap in most selected indicators was unchanged after the last economic crisis (Table 1). However, the gap was reduced for the following indicators: life expectancy at 30 years of age, health expectancy at 30 years of age, mortality from unintentional injuries in adults and mortality due to falls among elderly people (Table 1). The inequalities in smoking among women were reduced since the share of educated women among smokers was reduced. On the contrary, educational inequality related to high-risk intoxication in women has increased due to the increase in highly-educated women. Despite the indicator of mortality from alcohol, prescribed reasons for death have increased in the observed period, while the trend of decrease of mortality from suicide in men has halted, with educational inequality increasing in both"[4].

Table 1: Change in indicators of health inequalities by educational attainment in Slovenia

Indicator		Educational gradient[1]	Inequality in time[2]
Life expectancy at 30 years of age		Detected	Reduced
Health expectancy at 30 years of age		Detected	Reduced
Beginnings of life	Smoking during pregnancy	Detected	Unchanged
	Examinations during pregnancy and birth preparations	Detected	Unchanged
	Pregnancy results	Detected	Unchanged
Youth health	Self-assessment of health	Detected*	Unchanged
	School strain	Detected*	Unchanged
	Obesity	Detected*	Unchanged
Self-assessment of health		Detected	Unchanged
Smoking and lung cancer	Share of smokers	Detected	Unchanged with men Reduced with women (highly educated women smoke less)
	Mortality from lung cancer	Detected	Unchanged
Vaccination against tick-borne meningoencephalitis		Detected	/[3]
Alcohol and mortality	High-risk intoxication	Detected (highly-educated persons are more exposed)	Unchanged with men Increased with women (highly educated women are riskier with intoxication)
	Mortality from alcohol-related causes	Detected	Unchanged
Nourishment, activity,	Eating vegetables	Not detected	/
	Physical activity	Detected	/
	Obesity	Detected	Unchanged
	Cardiovascular diseases	Detected	/

[4] *Report 2018: Neenakosti v zdravju v Sloveniji v času ekonomske krize (eng: Health inequalities in Slovenia during the economic crisis, in translation)*

obesity and cardiovascular diseases	Hospitalization due to cardiovascular diseases	Detected	/
	Receiving medicine due to arterial hypertension	Detected	/
Mental health	Suspicion of major depression	Detected	/
	Experiencing anxiety	Detected	/
	Use of antidepressants	Detected	/
	Use of anxiolytics	Detected	/
	Suicide	Detected	Unchanged
Mortality due to unintentional injuries		Detected	Reduced
Health with elderly	Self-assessment of good health	Detected	/
	Functionality	Detected	/
	Visit to the dentist/orthodontist	Detected	/
	Mortality from falling	Detected	Reduced

[1] When the educational gradient is detected the value of the indicator is usually the lowest among the people with lower education.

[2] The inequality stands for the difference in the health results between the persons with low or high education, and the change of inequality stands for the long-term period – the comparison between 2007 and 2014 for inequalities using EHIS data and for mortality data, the same period and made a three years average.

[3] No data.

* It is not an educational gradient, but a gradient of subjective assessment of the family well-being.

Discussion

We performed a data linkage of the socio-economic database with the prescription database. Data linkage allowed us to estimate different indicators related to mortality, life expectancy, healthy life years, hospitalization, etc. Each of the yearly databases of health outcome was linked separately with the corresponding yearly database of educational attainment. Later than three years, the average value of indicators was calculated. Regarding data linkage, as already mentioned, due to the excellent quality of PIN numbers, we have not experienced any problems. Various institutions in the Slovenian health and statistical system perform data linkage of different sources. The sound data and relevant socioeconomic information can be obtained from the National Statistical Office from administrative or statistical data sources and registers.

We propose two recommendations: first is to develop a detailed inventory of national databases to identify/detect useful information/data and second to adopt multiple approaches to use different data sources stored in different organizations in a country.

Study limitations

During this study, we encountered some challenges and obstacles. Data linkage of health and non-health data sources require the exchange of sensitive personal data between institutions. Therefore, some legal obstacles might have appeared and we used some legal tools that enable to exchange the datasets, which made possible to perform this study. A small sample of people for whom no linkage could be done due to incorrect PIN number or missing data on educational attainment. There are data available on health inequalities at regional levels as called 'health profiles' but not at local levels.

Conclusions

Data linking from different institutions has resulted in comprehensive, sound and valid datasets, which enable us to set the relevant indicators, analyze data to monitoring health inequalities and to answer various research questions. This study has also highlighted the new improvements and analytical possibilities to analyze linked datasets. Broadening the socioeconomic distribution on various strata (for example mortality data based on income or medical

prescription on income level), the new channels are established where the data on income will be available. One of the future challenges is inter-regional inequalities where in-depth analyses would help us to understand and support the policymaking process and actors. The existing different relevant data sources can give us information on health status as well as information on the socio-demographic and socio-economic status of the population, which is needed to monitor health inequalities in the country.

Added value of this study
Due to data linkage, the health inequalities in relation to educational attainment and changes over time were highlighted in Slovenia. Data linkage can importantly enrich the initial data source, which can create more informative results, allow having a larger sample size (or full-population coverage) and enable to calculate regional estimates in some cases.

Implications of available evidence
The report was commissioned by the Ministry of Health and is aimed to give evidence of health inequalities for future policy development in Slovenia.

8. Using record linkage for estimating periodically the occurrence and case fatality rate of CVD in different regions of Italy

Link to published article: https://bit.ly/3cNPB4R

Summary[12]

Context
The Italian register of cardiovascular diseases is a surveillance system of fatal and nonfatal cardiovascular events in the general population aged 35–74 years. It was launched in Italy at the end of the 1990s to estimate periodically the occurrence and case fatality rates of coronary and cerebrovascular events in different geographical areas of the country. The main objective of this study was to illustrate the adopted methods, to report estimates of attack rates and 28-day case-fatality rates for coronary and cerebrovascular events.

Methods
We performed a case study at the sub-national level to report attack rates and case fatality rates for coronary and cerebrovascular events. We used the mortality database and hospital discharge record (HDR) database linked at the individual level. Both deterministic and probabilistic data linkage techniques were applied. We used the following input variables from two mentioned databases: 1. mortality database: demographic and ID information (e.g., name, surname, gender, date of birth, place of birth, residence), date and place of death, and 4 causes of death according to ICD 9 and 10; and 2. HDR database: demographic and ID information (e.g., name, surname, gender, date of birth, place of birth, residence), date of admission and discharge, hospital of admission and discharge, type and mode of admission, and 4 clinical diagnoses of discharge according to ICD 9 and 10. We applied MONICA (MONItoring of CArdiovascular diseases diagnostic) criteria (information collected from clinical charts) to validate the potential coronary and cerebrovascular events. We estimated the following health indicators: attack rate, which includes both first and recurrent events (separately for fatal and nonfatal cases); case fatality (as the ratio between fatal and all events) rate for both coronary

and cerebrovascular events. The main health outcome was the occurrence of coronary and cerebrovascular events among the resident adult population.

Results
We estimated the attack rate for both fatal and non-fatal coronary and cerebrovascular events and case fatality rate for 2-year period (1998-99) among the resident adult population in 8 geographical areas (2 Regions, 4 municipalities, 1 sub-metropolitan area, 1 multi-municipality area) of the country. These estimates of fatal and non-fatal attack rates and case fatality rates were provided by the geographical distribution, age and sex[1,2].

Study limitations
There are some limitations to this study and would require an in-depth analysis to further improve the surveillance system: first, the age range (i.e., 35-44, 45-54. 55-64. 65-74 age groups) was limited, therefore the representativeness of the general population is not assured, especially for the pathologies of coronary and cerebrovascular diseases that particularly affect the elderly people. Second, the choice of the geographical area to be included for surveillance depends on the occurrence of the event. According to EUROCISS recommendations, a minimal number of 300 total events per year in the age range of 45–74 years, is needed to produce reliable rates. Hence, if the number of events per year is less than the threshold of 300, a wider population and geographical area are required. Third, the implementation of a surveillance system with the systematic validation of all suspected events is expensive; thus, a periodical collection (e.g. once every five years) could be performed to evaluate possible trends over time. Fourth, the collection and coding of secondary causes of deaths are necessary but challenging since, in Italy, only the main cause of death is codified according to the ICD-9 or 10 revision and available in the electronic administrative database; the secondary causes of death are included in the database just as a description; therefore, they need to be codified. Finally, the fatal and nonfatal 500 events should be validated by considering several days during the year to take into account possible seasonal variability.

Conclusion
This study supports to implement of population-based coronary and stroke event registers following standardized methodologies to produce comparable data at the European level.

Added value of this study
The study is an example of the Italian register of cardiovascular diseases that perform record linkage of different data sources including mortality register and HDR. This register can provide health-monitoring indicators in some areas of Italy.

Implications of available evidence
This study provides important information on the occurrence of CVD, survival and case fatality in the general population, which can be integrated with other surveillance systems in the country. Moreover, in terms of policy planning, this study provides essential information for planning health care services and prevention programmes.

[1] *Giampaoli S et al, Populaton-based register of stroke: manual of operations. Eur J Cardiovasc Prev and Rehab 2007*
[2] *Madsen M et al, Population-based register of acute myocardial infarction: manual of operations. Eur J Cardiovasc Prev Rehabil. 2007 Dec*

9. The EMPCAN study: protocol of a population-based cohort study on the evolution of the socio-economic position of workers with cancer *(ongoing)*

Link to published article: bit.ly/2vYNxGZ

Summary[13]

Background

The improvements in cancer control led to an increase in the number of cancer survivors, notably, in the working-age population (16–64 years). For this group, the ability to keep or resume work means maintaining the level of the household income, self-esteem, back to "normal", to feel being cured, preservation of the quality of life, etc. However, the return-to-work (RTW) pathway is not obvious and encounters many challenges. It can imply temporary adjustments, different working capacities or new occupational aspirations. There is a strong need to assess and understand their reintegration into the labour market, which underlines and ensures their social integration and quality of life. The objectives of the EMPCAN study are therefore to measure the scale of return-to-work after cancer and to identify the determining factors, allowing for the implementation of an adequate socio-professional support.

Methods

The EMPCAN study is a retrospective population-based cohort study, using data coming from three Belgian administrative registers. It includes workers aged 16–64 years, diagnosed in Belgium with breast, colorectal, lung, corpus uteri, prostate, head and neck or testis cancer between 1st January 2004 and 31st December 2011. We requested data from the Belgian Cancer Registry and the Crossroad Bank for Social Security. We included all socially insured Belgian workers diagnosed between 2004 and 2011 with colorectal, breast, head & neck, prostate, testis, lung and corpus uteri cancer. We excluded those patients who were long-term unemployed, disabled, handicapped or on sick leave at these dates. The end of (administrative) follow-up was 31st December 2012. We include demographic, health-related and work-related factors in the analysis and observed how these factors interplay to determine the working status. After having solved legal, ethical and technical issues for the coupling/linkage, we will perform three types of data analysis: 1. exploratory data analysis, 2. logistic regression and 3. multistate modelling. For the exploratory analysis, we will clean and describe the main characteristics of the individuals in the EMPCAN study and will explore possible effect modification of variables (e.g., between cancer, stage and treatment) on the outcome. According to the main objectives of the EMPCAN study, we will use logistic regression considering the (return to the) professional activity as the main event of interest, whether it is part-time or full-time. The determining factors are demographic, cancer-related and work-related variables. To ensure proper estimation of this event, we will have to take into account the probability of occurrence of other events, i.e. death and retirement. We will perform survival analysis with competing risks analysis and provision of cause-specific hazards using the Fine and Gray model. An important aspect of our data and event of interest is that workers can enter and exit the different socio-economic positions, over time. To capture this reality, the best method is to use a multistate model using transitions probabilities among different socio-economic positions and finally, group-based modelling for longitudinal data using the 'proc traj' package in SAS.

Discussion

The EMPCAN will bring important insights into the RTW after cancer and will allow the identification of the determining factors to be considered for the development of occupational rehabilitation interventions. The combination of the results from both cross-sectional and longitudinal approaches will allow a better understanding and explanation of the changes and stability of the socioeconomic positions in the years after the cancer diagnosis. This should provide elements for the identification of impediments or facilitators that relate to the social security scheme. Eventually, the more specific needs and remaining gaps in the knowledge regarding the RTW after a long sickness absence will be identified and suggested as future research perspectives.

Study limitations

The main limit of EMPCAN is that it doesn't include three important aspects to completely (comprehensively) understand and explain the RTW after cancer. ***First***, the patient's experience and self-assessment of their health status and (new) occupational aspirations, are not captured in administrative data. ***Second***, we do not include the employer's perspective with their ability and willingness to reintegrate workers with cancer. ***Thirdly***, cancer-related information is provided at baseline as the treatments are those provided during the first twelve months following the month of incidence, which limits the possibility to appreciate the health status of the workers in the following years on which he is observed. Similarly, cancer treatment-related or pre-existing comorbidities could be of high importance in the ability to work but are currently missing in our data. However, in a previous study, we build a proxy for the health status, using the relative survival corresponding to the cancer site. This exercise could also be performed in EMPCAN, using additional information on the relative survival, the stage at diagnosis and the treatments received.

Added value of this study

The main strength of the EMPCAN study is its representativeness. A recent review on quantitative studies on RTW after cancer reports that only one of the twelve included studies was population-based. The main reason is probably that the linkage/coupling of health, socio-economic and administrative data requires many legal and ethical steps. Another strength of this study is that most of these data are not available neither accessible at the national level.

Implications of available evidence

The results of the EMPCAN study will not only raise the awareness among health professionals and policymakers but will also allow the provision of evidence-based support to professional reintegration policies and better planning and organization of vocational rehabilitation programs.

10. Longitudinal access and exposure to green-blue spaces and individual-level mental health and well-being: protocol for a longitudinal, population-wide record-linked natural experiment *(ongoing)*

Link to published article: bit.ly/39sjtkJ

Summary[14]

Background

Globally, 686 million people suffer from common mental health disorders (CMDs) such as depression or anxiety. In the UK, CMDs are experienced by around one in four of the population, and mental ill-health costs the economy over £100 billion per annum in health, social care and quality of life loss costs. Subjective well-being (SWB) is also related to mental and physical health outcomes, including life expectancy, and is a key marker of quality of life. With increasing impacts on wider societal costs, CMDs and promoting SWB are growing in importance. Access to natural environments considered here as 'green-blue spaces' (GBS) such as parks and beaches, may provide opportunities to support and promote good public mental health and well-being. The evidence based on the impacts of GBS, on mental health and well-being is growing rapidly. Current research suggests that the benefits may differ by population group, context and health outcome. Studies suggest that access and exposure to green-blue spaces (GBS) have beneficial impacts on mental health. However, the evidence base is limited concerning longitudinal studies. The main aim of this longitudinal, population-wide, record-linked natural experiment, is to model the daily lived experience by linking GBS accessibility indices, residential GBS exposure and health data; to enable quantification of the impact of GBS on well-being and common mental health disorders, for a national population.

Methods

Study design: The GBS project is a retrospective and controlled population-wide study. This research will estimate the impact of neighbourhood GBS access, GBS exposure and visits to GBS on the risk of common mental health conditions (CMD) and the opportunity for promoting subjective well-being (SWB); both key priorities for public health.

Data sources: We will use a Geographic Information System (GIS) to create quarterly household GBS accessibility indices and GBS exposure using a digital map and satellite data for 1.4 million homes in Wales, UK (2008–2018). We will link the GBS accessibility indices and GBS exposures to individual-level mental health outcomes for 1.7 million people with demographic data, general practitioner (GP) data and data from the National Survey for Wales [NSW] (n=~12 000) on well-being in the Secure Anonymized Information Linkage (SAIL) Databank.

We will examine the risk of CMD using longitudinal changes in access to neighbourhood GBS and if these associations are modified by multiple socio-physical variables, migration and socioeconomic disadvantage. Subgroup analyses will examine associations by different types of GBS. This longitudinal study will be augmented by cross-sectional research using survey data on self-reported visits to GBS and SWB to investigate whether visits to GBS improve SWB.

Study participants: Our study population contains people aged 16 years and older living in Wales, UK. To evaluate the association between changes in access and exposure to GBS, on the risk of CMD, study sample includes the total adult population registered with a general practitioner (GP), providing GP records to SAIL. This is expected to be about 1.7 million adults in Wales. To investigate whether visits to GBS improve SWB, study population includes a representative sample of the adult population in Wales based on the NSW for 2 years (cross-

sectional samples in 2016/2017 and 2017/2018). The NSW has an annual sample of approximately 12 000 responders and GBS visit questions are asked of 50% of that sample. This provides a total cross-sectional sample of 12 000 (over the 2 years) for this part of the study.
Outcomes and analysis: There are one primary and two secondary outcomes of this study. The primary outcome is the change in counts of CMD treatments for adults identified as CMD cases within the corresponding periods for the 70% of adults in Wales for whom we have GP data records in SAIL (1.7M adults). Prevalence algorithms will be applied to detect cases of CMD (anxiety and depression) from routinely collected GP data. The first secondary outcome is SWB, measured by the Warwick-Edinburgh Mental Wellbeing Scale (WEMWBS) in NSW for two survey years (2016/2017 and 2017/2018) for a representative population. We have the total number of events recorded for each person in a WLGP dataset. The second secondary outcome is the GP event and we will calculate that by converting the GP event from a binary daily activity or no activity aggregated to quarterly counts. We have two distinct types of outcomes: continuous (SWB) and counts (CMD treatments and GP event days). We intend to analyze the continuous outcome using a linear model. Poisson models will be used to analyze the longitudinal count data.

Discussion
This study will be the first of its kind to link GIS-generated GBS accessibility and exposure measures with routinely collected, longitudinal health data and cross-sectional survey data for a whole population. We plan to create a longitudinal GBS dataset for Wales. Using multisource data, we will build a dataset that records local-level changes in GBS for Wales, for 11 years. Longitudinal studies that have previously examined GBS exposure have used cross-sectional environment data to calculate GBS accessibility and exposure indices. This may be because it is a resource and time-intensive task to create a longitudinal GBS dataset, and longitudinal data are not always available. In addition, bringing together data from different sources, harmonizing the data and deriving GBS indices from the data requires expertise and specialized skills. A national study using routinely collected data on a national scale is timely following recommendations from a report on a prospective quasi-experimental study.

Study limitations
The main limitation of this study is a large number of potentially confounding variables. Despite this is a non-randomized study using routinely recorded data may omit some unknown confounders, thereby introducing a moderate level of bias due to confounding..

Added value of this study
This study will create novel GBS accessibility indices to use network routes from a variety of sources, and access points, to model how people may access GBS. The way that studies measure access to GBS is methodologically diverse and there is no consensus on that which is the best measure to use.

Implications of available evidence
Our study will work with stakeholders and policy-makers to develop a GBS typology that can be used to provide evidence that can be translated to help policy and practice. Improved evidence on the impacts of GBS on mental health is required to inform decisions relating to planning, area regeneration and environmental management.

11. Regional health care profiles- case studies on catchment areas of envisaged primary care units in Austria *(ongoing)*

Unpublished study

Summary

Background

In Primary Care Act[1,] the primary care units are described as the Austrian's novel form of primary care institutions in response to the health policy to ensuring a solidarity-based health system with a high level of quality. The novel primary care units are a turning point in the traditional structure of the Austrian health care landscape as they represent a comprehensive inter-sectoral approach. Additionally, the novel primary care units pursue a team-based approach: general practitioners and other employees with different health and social professions work together cooperatively to contribute equally to the care provided.

The requirements of the primary care unit are described in "PVG 2017"[2] and this law requires to meet the local needs. Primary care units have to provide health care close to the patient's home, needs-based opening hours and accessibility for acute cases outside the opening hours. The scope of services covered by the primary care units includes a broad spectrum of diagnostic, therapeutic and nursing services with several additional tasks. As far as possible and practical from a medical point of view, acute treatment should be provided in the primary care unit to reduce the load on secondary levels of care. In addition to acute care, the range of tasks also includes chronic care, health promotion and prevention. The individual primary care unit is based on a specific care concept that specifies the care objectives and care tasks about the respective population in the catchment area with its local health and epidemiological profile. The care concept also describes the specific opening hours for service provision and the cooperation within the primary care unit team and with other caregivers

Here we introduce the regional health care profiles representing a comprehensive description of selected regions. The regional health care profiles include five domains within the catchment area: 1. Demography and socio-economic status, 2. Prevention and risk factors, 3. Epidemiology and mortality, 4. Service supply and 5. Outpatient utilisation.

The main objective of specific regional case studies using regional health care profiles for primary care (RVP/PV) is to support the heads (i.e., usually general physicians) or managers of primary care units to define the particular care concept for their envisaged primary care units. The regional health care profiles for primary care represents a comprehensive compilation of region-specific figures and facts about the location of a primary care unit by considering the municipalities in the catchment area.

Methods

We used a dynamic, highly interactive software tool, where different locations can be selected and compared to generate regional health care profiles. Additionally, the broadness of the catchment area can be set variably. The tool is based on the data and functionalities of the geographical information system already available in the Austrian Health Information System[3] (ÖGIS) and makes use of many skills and experiences associated with ÖGIS. ÖGIS provides numerous already prepared and quality-assured data sources and indicators, which feed into the tool for the regional health care profiles. On the other hand, new data sources have been integrated and linked with already existing data sources. New functionalities of ÖGIS have been achieved. For example, a new export option was implemented for the complex calculation of indicators differing according to various isochrones and related catchment areas. This allows an

automated output of about 6,500 individual values (needed as input for the indicators) in one single output process. A function for automatic cleaning of outliers has also been implemented. Regional health care profiles combine data at a high level of spatial resolution (i.e. a total of 2,122 Austrian municipalities) and identified 35 indicators. Indicators for each of the 2,122 municipalities are calculated as averages of all municipalities within a 10-, 15- and 20-minute car accessibility radius to the primary care units. Complex information technologies were used to handle the large amounts of data and extensive calculations as well as to meet the requirement for automated report generation. We used record linkage at the level of anonymised individual personal identifiers for four indicators (rate of hospitalisations of residents with heart diseases, femoral neck fracture, cerebrovascular disease and cancer within 2 years).
The visualization of 35 indicators related to five domains already mentioned, takes place by generating box plots and circles/squares laid over them (cf. pages 2-3 in appendix 2). The underlying visualisation technique was originally designed within the framework of the EU project "I2SARE" (cf. e.g. here[4]). Currently, I2SARE presentation technique is routinely used by several EU states and has been adapted for the current project in a more detailed local application. After multiple discussions with experts in the field of primary care, the indicators and their visualisation were optimised to be intuitively understandable and comprehensible. The primary testing was carried out together with physicians and non-physician PHC (Primary Health Care) experts in the initial phase of the project. In addition to the graphical representation, numerical values of the indicators and absolute values for the catchment area are shown in tabular form (cf. page 3 in appendix 2), which makes it possible to estimate the size of the population to be covered in the catchment area. Finally, the catchment area (ochre-coloured) used to calculate the indicator values and the location (black dot) of the envisaged primary care unit are displayed on a map (cf. map on the left side on page 1 in appendix 2). Optionally, a second area (green) can be defined and used for a head-to-head comparison.

Results

We have identified 35 indicators from following five domains, which provide an overview of health (care) related characteristics in the catchment area as described below:

1) Demography and socio-economic status

The age distribution of inhabitants in the catchment area shows that more services for younger or older people should be provided or referred to, according to the age distribution. Under the basic principle of health care, target groups should be focussed. The 'average income per income-receiver' represents the income situation of the regional population.

2) Prevention and risk factors

The proportion of people with self-assessed '(very) poor' state of health indicates the need for care due to difficulties in coping with everyday tasks/activities. The health behaviour indicators include the proportion of people who smoke, who do not exercise enough and the proportion of obese people. These determinants indicate that more preventive and health-promoting measures should be offered.

3) Epidemiology and mortality

Life expectancy at birth is often linked with structural factors. The prevalence of diabetes mellitus type 2, mental disorders, disorders of the musculoskeletal system or chronic head/cross/neck pain indicates the occurrence of typical diseases in the catchment area. The proportion of people living in single-person households aged 65+ serves as an indicator for single persons with possible restrictions in social participation or a higher need for support. The proportion of long-term care allowance recipients and the proximity and number of nursing homes in the catchment area indicate a possible increase in the need for outreach services.

4) Service supply

The presence of existing services in the fields of general medicine and pediatrics or other health care structures (hospitals, pharmacies, nursing facilities) can answer questions about the necessary structure and sizing of the personnel pool within the envisaged primary care unit and point out possible cooperation within the region-specific primary care tasks.

5) Outpatient utilisation

There are often regional differences in the utilisation of health services. Below-average values of the indicators (average values are displayed as the red line in the figure on page 2 in the appendix 2) may indicate access problems or a 'relatively healthy population'. Above-average values may indicate special regional care needs that could be covered by the envisaged primary care unit.

Study limitations

There are some limitations, which should be considered in this context: ***First*****,** there are limitations of data sources: As some data were not available at the low aggregation level of municipalities, calculations had to be interpolated for some indicators. This leads to a blurring of regional outputs in some cases. ***Second*****,** the size of the catchment areas and the respective number of inhabitants based on accessibility by car, vary greatly depending on the degree of urbanization, therefore comparability of different catchment areas is limited. ***Third***, due to legal restrictions on the use of data sources, the dissemination of the regional health care profile is restricted. ***Finally*****,** the interpretation of the indicators may be challenging or even impossible due to still lacking region-specific information (e.g. the number of health care providers in the fields of physiotherapy, occupational therapy or logopaedics). We recommend that the accessibility by public transport should be considered in the future.

Conclusions

Regional health care profiles play an important role in determining and describing the possible aims of envisaged primary health care institutions by providing comprehensive information on population health, utilization of health services and health structures. In addition to assessing the scope and nature of health care, they also provide information on public health interventions that are needed. To address new requirements of regional health care profiles (as they might shift dynamically), a yearly update is planned based on the feedback of the users.

Added value of this study

We used the geographical information system (GIS) and high level of spatial resolution to describe the regional health care profiles by integrating data from 2,122 Austrian municipalities. This is an innovative approach using GIS to improve the primary care services at municipal level.

Implications of available evidence

The results of this case study have already been shared with various stakeholders in Austria to further improve the planning of new interventions at primary care units. This represents the provision of locally targeted evidence-based support to health care professionals. The profiles have already successfully supported the process to establish several primary care units.

Appendix 1: List of Indicators (N = 35)

Demography and socio-economic status

1. Inhabitants in the catchment area
2. Proportion of children under 14 years (<15a)
3. Percentage of population aged 65 and over
4. Percentage of population aged 75 and over
5. Percentage of people aged 65 and over in one-person households
6. Average income per income-receiver

Prevention and risk factors

1. Percentage with self-rated health as "very bad" or "bad" (self-reported), EW >=15a
2. Percentage of smokers (daily and occasional), inhabitants >=15a
3. Percentage with too little exercise, inhabitants >=15a
4. Percentage with obesity, inhabitants >=15a

Epidemiology and mortality

1. Life expectancy at birth (men)
2. Life expectancy at birth (women)
3. Prevalence diabetes mellitus Type 2
4. Prevalence of mental disorders
5. Prevalence disorders of the musculoskeletal system
6. Prevalence of chronic head/cross/neck pain, inhabitants >=15a
7. Percentage of long-term care benefit recipients / level 1-3
8. Percentage of long-term care benefit recipients / level 4-7
9. Rate of hospitalized patients with heart disease in 2 years
10. Rate of hospitalized patients from 65 years of age with femoral neck fracture in 2 years
11. Rate of hospitalized patients with cerebrovascular disease within 2 years
12. Rate of inpatients with cancer in 2 years

Service supply

1. Inhabitants per general practitioner
2. Percentage of general practitioners with a statutory health insurance contract aged 55 +
3. Inhabitants per private general practitioner
4. Children per specialist for paediatrics
5. Inhabitants per specialist for internal medicine financed via health insurance funds
6. Distance to the nearest acute care hospital (incl. outposts), minutes by car
7. Number of pharmacies in the catchment area (excl. in-house pharmacies)
8. Distance to the nearest nursing home with nursing places (minutes by car)
9. Residents aged 65 and over per old people's home or nursing home in the catchment area

Outpatient utilisation

1. Percentage of inhabitants visiting general practitioners of statutory health insurance
2. General practitioner equivalents per 100,000 inhabitants (incl. outpatient departments of hospitals)
3. Paediatrician equivalents per 100,000 children (<15a; incl. outpatient departments of hospitals)
4. Internal medicine equivalents per 100,000 inhabitants (incl. outpatient departments of hospitals)

Appendix 2: Pages 1-3 from the regional care profile (Page 4 and 5 are indicator definitions)

Page 1 from the regional care profile: map with location and catchment area; introductory text

Regionales Versorgungsprofil Primärversorgung für den Einzugsbereich von Vorau

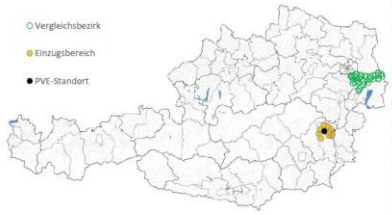
PVE
Primär-Versorgungs-Einheiten für Ihre Gesundheit

Dieses **regionale Versorgungsprofil Primärversorgung (RVP/PV)** wird vom Bundesministerium für Arbeit, Soziales, Gesundheit und Konsumentenschutz (BMASGK) für Gründerinnen und Gründer von Primärversorgungseinheiten (PVE) zur Verfügung gestellt. Die RVP/PV wurden im Rahmen der aktuellen Gesundheitsreform ("Zielsteuerung-Gesundheit") erstellt. Das RVP/PV ist eine umfassende Zusammenstellung regionsspezifischer Zahlen und Fakten für einen potenziellen PVE-Standort unter Berücksichtigung der Gemeinden im Umkreis (=Einzugsbereich [EZB]). Dieser **EZB** wird über die Entfernung umliegender Gemeinden zur geplanten PVE-Standortgemeinde in Pkw-Minuten definiert.

Das RVP/PV enthält **Indikatoren** zum demografischen und sozioökonomischen Status der Bevölkerung im Einzugsbereich, zum Thema Prävention und Risikofaktoren passend zu typischen Aufgabenbereichen einer PVE (wie z.B. Krankenbehandlung, Prävention, Gesundheitsförderung) sowie zur Epidemiologie und zur Lebenserwartung. Weiters sind auch Informationen zum bestehenden Versorgungsangebot (Anzahl praktizierender Ärztinnen und Ärzte, nahegelegene Spitäler, Pflegeheime u.a.) sowie zur Inanspruchnahme von Gesundheitseinrichtungen enthalten.

Auf **Seite zwei** werden epidemiologische Kennzahlen des Einzugsbereichs **grafisch** dargestellt (jeweils im Vergleich zum entsprechenden Bundes- oder Bundeslanddurchschnitt, optional auch für einen Vergleichsbezirk). Auf **Seite drei** sind die Zahlenwerte der Indikatoren bzw. die Absolutwerte für den Einzugsbereich **tabellarisch** dargestellt, womit sich die Größenordnung der zu versorgenden Bevölkerung im EZB abschätzen lässt. Die **Definitionen** der Indikatoren sind **auf den letzten beiden Seiten** dargestellt.

Hinweis für die Betrachtung am Bildschirm: Ziehen Sie auf Seite 2 und 3 mit der Maus über den Indikator-Text für weitere Details zum Indikator (**„Mouseover"**).



Einzugsbereich von Vorau (20 Pkw-Minuten)

Dechantskirchen, Grafendorf bei Hartberg, Lafnitz, Miesenbach bei Birkfeld, Neustift an der Lafnitz, Pöllau, Rohrbach an der Lafnitz, Sankt Lorenzen am Wechsel, Vorau, Waldbach-Mönichwald, Wenigzell
(Sortierung in alphabetischer Reihenfolge)

Vergleichsbezirk Bruck an der Leitha

Inhalte

Auf den Seiten 2 und 3 finden Sie Indikatoren zu den folgenden fünf Bereichen:

1) Demografie und sozioökonomischer Status

Die Altersstruktur der Bevölkerung im Einzugsgebiet zeigt, ob eher ältere oder eher jüngere Menschen im Einzugsgebiet leben. Entsprechend dem Grundprinzip der Zielgruppenorientierung sind je nach Altersstruktur vermehrt Angebote für jüngere oder ältere Menschen vorzusehen bzw. ist auf diese zu verweisen. Das „Durchschnittseinkommen pro Einkommensbezieher/-in" stellt die Einkommenssituation der regionalen Bevölkerung dar.

2) Prävention und Risikofaktoren

Der Anteil an Personen mit selbst eingeschätzter „(sehr) schlechter" Gesundheit verweist auf Versorgungsbedarf hinsichtlich Einschränkungen in der Alltagsbewältigung. Dargestellte Indikatoren zum Gesundheitsverhalten sind der Anteil an Personen, die rauchen, an Personen mit zu wenig Bewegung und der Anteil adipöser Menschen; sie geben Hinweise auf verstärkt anzubietende präventive und gesundheitsförderliche Maßnahmen.

3) Epidemiologie und Mortalität

Die Lebenserwartung bei Geburt steht häufig in einem Zusammenhang mit strukturellen Einflussfaktoren. Die Prävalenzen von Diabetes Mellitus Typ 2, psychischen Störungen, Störungen des Bewegungs-/Stützapparats oder chronischen Kopf-/Kreuz-/Nackenschmerzen geben Hinweise auf das Vorkommen von typischen Erkrankungen im Einzugsbereich.

Der Anteil an in Einpersonenhaushalten lebenden Menschen 65+ dient als Indikator für "ältere" alleinstehende Personen mit möglichen Einschränkungen in der sozialen Teilhabe bzw. höherem Unterstützungsbedarf. Der Anteil an Pflegegeldbezieherinnen/-beziehern sowie die Nähe zu und die Anzahl an Pflegeheimen im Einzugsgebiet geben Hinweise auf einen eventuell erhöhten Bedarf an aufsuchender Betreuung.

4) Versorgungsangebot

Das bestehende Angebot in den Fachbereichen Allgemeinmedizin bzw. Kinderheilkunde sowie an weiteren Gesundheitsversorgungsstrukturen (Spitäler, Apotheken, Pflegeeinrichtungen) kann Fragen der erforderlichen Struktur und Dimensionierung des Personalangebots im Rahmen einer geplanten PVE beantworten und mögliche Kooperationen im Rahmen der regionsspezifischen Primärversorgungsaufgaben aufzeigen.

5) Ambulante Inanspruchnahme

Regional bestehen oft Unterschiede in der Inanspruchnahme von Gesundheitsangeboten. Unterdurchschnittliche Werte können auf Zugangsprobleme oder auf eine "relativ gesunde Bevölkerung" hinweisen, überdurchschnittliche Werte können besondere regionale Versorgungsbedürfnisse anzeigen, die durch verstärkte PVE-Versorgung abgedeckt werden könnten.

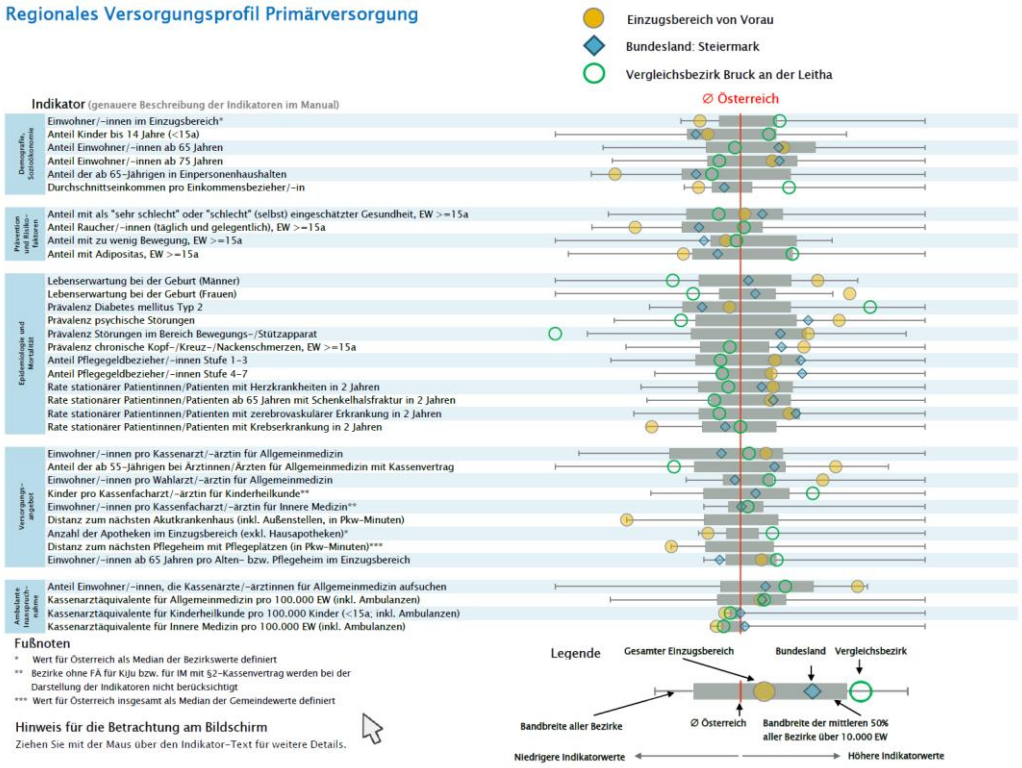
Gesundheit Österreich GmbH

Bundesministerium Soziales, Gesundheit, Pflege und Konsumentenschutz

Hauptverband der österreichischen Sozialversicherungsträger

Seite 1 von 5

Page 2 from the regional care profile: visualisation of the indicators



Page 3 from the regional care profile: relative and absolute values of the indicators

Regionales Versorgungsprofil Primärversorgung

Indikator (genauere Beschreibung der Indikatoren im Manual)	Indikatorwerte: Einzugsbereich	Vergleichsbezirk	Bundesland	Österreich	Einheit Indikator-Wert	Erwartete Absolutwerte für das Jahr 2019[1]: im Einzugsbereich	im Vergleichsbezirk	Einheit
Einwohner/-innen im Einzugsbereich*	25.778	102.010	1.243.052	8.858.775	Anzahl	25.778	102.010	Personen
Anteil Kinder bis 14 Jahre (<15a)	13,7%	15,1%	13,4%	14,4%	Anteil (in %)	3.528	15.400	Personen
Anteil Einwohner/-innen ab 65 Jahren	20,4%	18,6%	20,3%	18,8%	Anteil (in %)	5.268	19.011	Personen
Anteil Einwohner/-innen ab 75 Jahren	10,2%	8,9%	10,4%	9,4%	Anteil (in %)	2.637	9.085	Personen
Anteil der ab 65-Jährigen in Einpersonenhaushalten	23,2%	30,4%	29,2%	32,4%	Anteil (in %)	1.224	5.772	Personen
Durchschnittseinkommen pro Einkommensbezieher/-in	€ 24.287	€ 32.592	€ 26.649	€ 28.111	Euro	€ 24.287	€ 32.592	Euro
Anteil mit als "sehr schlecht" oder "schlecht" (selbst) eingeschätzter Gesundheit, EW >=15a	4,7%	4,1%	5,0%	4,6%	Anteil (in %)	1.035	3.531	Personen
Anteil Raucher/-innen (täglich und gelegentlich), EW >=15a	22,5%	30,3%	27,1%	30,0%	Anteil (in %)	5.015	26.210	Personen
Anteil mit zu wenig Bewegung, EW >=15a	74,1%	74,9%	72,6%	75,1%	Anteil (in %)	16.487	64.832	Personen
Anteil mit Adipositas, EW >=15a	11,9%	16,5%	13,4%	14,3%	Anteil (in %)	2.658	14.309	Personen
Lebenserwartung bei der Geburt (Männer)	80	78	79	79	Lebensjahre	80	78	Lebensjahre
Lebenserwartung bei der Geburt (Frauen)	86	83	84	84	Lebensjahre	86	83	Lebensjahre
Prävalenz Diabetes mellitus Typ 2	5,8%	8,4%	5,3%	6,0%	Anteil (in %)	1.506	8.589	Personen
Prävalenz psychische Störungen	15,4%	12,0%	14,8%	13,3%	Anteil (in %)	3.979	12.217	Personen
Prävalenz Störungen im Bereich Bewegungs-/Stützapparat	28,7%	14,2%	27,1%	24,8%	Anteil (in %)	7.397	14.453	Personen
Prävalenz chronische Kopf-/Kreuz-/Nackenschmerzen, EW >=15a	36,6%	32,3%	35,3%	32,9%	Anteil (in %)	8.144	27.988	Personen
Anteil Pflegegeldbezieher/-innen Stufe 1-3	3,9%	3,3%	4,2%	3,5%	Anteil (in %)	1.009	3.374	Personen
Anteil Pflegegeldbezieher/-innen Stufe 4-7	2,0%	1,5%	2,3%	1,7%	Anteil (in %)	516	1.511	Personen
Rate stationärer Patientinnen/Patienten mit Herzkrankheiten in 2 Jahren	2.548	2.239	2.467	2.320	Pat/100.000 EW	657	2.284	Personen
Rate stationärer Patientinnen/Patienten ab 65 Jahren mit Schenkelhalsfraktur in 2 Jahren	295	247	298	269	Pat/100.000 EW(>=65a)	61	205	Personen
Rate stationärer Patientinnen/Patienten mit zerebrovaskulärer Erkrankung in 2 Jahren	741	587	756	633	Pat/100.000 EW	191	599	Personen
Rate stationärer Patientinnen/Patienten mit Krebserkrankung in 2 Jahren	1.559	1.891	1.835	1.890	Pat/100.000 EW	402	1.929	Personen
Einwohner/-innen pro Kassenarzt/-ärztin für Allgemeinmedizin	2.343	2.267	2.143	2.228	EW/A	11	45	Ngl. Ä
Anteil der ab 55-Jährigen bei Ärztinnen/Ärzten für Allgemeinmedizin mit Kassenvertrag	72,7%	40,0%	60,3%	53,4%	Anteil (in %)	8	18	Ngl. Ä
Einwohner/-innen pro Wahlarzt/-ärztin für Allgemeinmedizin	6.445	4.080	2.563	2.797	EW/A	4	25	Ngl. Ä
Kinder pro Kassenfacharzt/-ärztin für Kinderheilkunde**	/	7.700	5.207	4.534	EW(<15a)/A	/	2	Ngl. Ä
Einwohner/-innen pro Kassenfacharzt/-ärztin für Innere Medizin**	/	25.503	22.601	22.037	EW/A	/	4	Ngl. Ä
Distanz zum nächsten Akutkrankenhaus (inkl. Außenstellen, in Pkw-Minuten)	0	/	/	19	Pkw-Minuten	0	/	Pkw-Minuten
Anzahl der Apotheken im Einzugsbereich (exkl. Hausapotheken)*	3	17	/	10	Anzahl	3	17	Apotheken
Distanz zum nächsten Pflegeheim mit Pflegeplätzen (in Pkw-Minuten)***	0	/	/	10	Pkw-Minuten	0	/	Pkw-Minuten
Einwohner/-innen ab 65 Jahren pro Alten- bzw. Pflegeheim im Einzugsbereich	2.634	3.169	1.177	1.894	EW(>=65a)/Heim	2	6	Heime
Anteil Einwohner/-innen, die Kassenärzte/-ärztinnen für Allgemeinmedizin aufsuchen	73,7%	67,0%	65,1%	62,8%	Pat/EW (in %)	18.987	68.296	Personen
Kassenarztäquivalente für Allgemeinmedizin pro 100.000 EW (inkl. Ambulanzen)	50	51	50	47	ÄAVE/100.000 EW	12,9	51,7	ÄAVE
Kassenarztäquivalente für Kinderheilkunde pro 100.000 Kinder (<15a; inkl. Ambulanzen)	0	18	49	48	ÄAVE/100.000 EW(<15a)	0,0	2,7	ÄAVE
Kassenarztäquivalente für Innere Medizin pro 100.000 EW (inkl. Ambulanzen)	3	7	19	16	ÄAVE/100.000 EW	0,8	7,4	ÄAVE

Fußnoten

* Wert für Österreich als Median der Bezirkswerte definiert

** Bezirke ohne FÄ für KJju bzw. für IM mit §2-Kassenvertrag werden bei der Darstellung der Indikatoren nicht berücksichtigt

*** Wert für Österreich insgesamt als Median der Gemeindewerte definiert

Hinweis für die Betrachtung am Bildschirm

Ziehen Sie mit der Maus über den Indikator-Text für weitere Details.

Legende

ä = Jahre; A = Arzt/Ärztin; ÄAVE = ärztliche ambulante Versorgungseinheiten; Ngl. Ä = Niedergelassene Ärztinnen/Ärzte; Pat = Patientinnen/Patienten; EW = Einwohner/-innen;

[1] Erwartete Absolutwerte für das Jahr 2019

Erläuterung: Die "erwarteten Absolutwerte" werden aus den Indikatorenwerten (i.d.R. Verhältnisse zu Einwohnerzahlen) berechnet. Basis für die Berechnung der Absolutwerte sind die Einwohnerzahlen 2019.

12. The national public health reporting system: Web-based public health reporting in Sweden

Summary

Context

Sweden is a high-income country with about 10 million people in 2018. The Swedish public health policy (revised and adopted by the Swedish Parliament in 2018), has a clear focus on equitable health for all and an overall goal to reduce health inequalities within a generation. There are 21 regions in Sweden and 290 municipalities with a joint responsibility is to achieve this objective as part of the national policy for sustainable societal development and equitable health care. Sweden has 290 municipalities of sparsely populated in 21 regions. Among 290 municipalities, 72 (25 percent) have less than 10 000 inhabitants. These municipalities are particularly important for welfare services that have an impact on citizens' lifelong health, for example; childcare, schools, social service, elderly care and support to people with disabilities, some emergency services, environmental issues, urban planning and sanitation (waste and sewage). These municipalities have extensive responsibilities in the areas of public health; often find it hard to prioritize data collection and analysis.

Across the Swedish population, the levels of health are good and have improved in most areas over recent decades. However, some significant health disparities remain and vary depending on sex, age, educational level proxy for socio-economic status. Health disparities related to level of education are often larger among women and men and within a region than between regions. Over time, health disparities have increased in some cases, such as for life expectancy.

The overall mission of the Public Health Agency of Sweden is to promote good public health, to evaluate the impact of methods and strategies of public health, monitor the health status and factors influencing health. This is achieved through knowledge building and dissemination of knowledge, promoting health, preventing diseases and injuries and promoting effective protection against infection. The Public Health Agency of Sweden provides regularly a report of public health policy to the government as an annual report on the development of public health and its determinants. The annual report is published in March every year, mainly as a web-based report but also as a written summary report with the government as a primary target group (https://www.folkhalsomyndigheten.se/publicerat-material/publikationsarkiv/f/folkhalsans-utveckling-arsrapport-2020/). It provides an overall summary of the current situation of public health, new developments over time and its determinants in Sweden, with a special focus on reducing health inequalities. The report also includes an international outlook.
The web-based reporting makes the health information easily accessible to the municipalities, as well as gives them deeper knowledge about each municipality. The Public Health Agency of Sweden is currently in the process of reviewing, revising and updating the current National Public Health Reporting System. A new system, reflecting the new structure of the Parliament-endorsed Public Health policy, will be in place by the end of 2020.

Methods

The Public Health Agency of Sweden updates to the indicators-based public health reporting annually. For the analysis, the Agency uses data from various sources, such as registers and population-based surveys, undertaken both by the Agency itself and by other national

authorities. As data is collected with a Unique Persona Identifier (UPI) it is possible to link data from different sources to the very same person. Data is disaggregated and analysed by age, sex, educational level, county (municipality where possible), country of birth (Sweden, Nordic countries, EU/ESS, outside EU/ESS). This kind of reporting provides an overall description of the situation in public health and its determinants in Sweden as well as highlight the important aspects of developments in the area. This reporting highlights the main differences in health, living conditions and lifestyle habits to reduce the health inequalities among different groups in the population. The web-based reporting is easily accessible on the website. It enables deeper knowledge on each indicator and it complements the annual report. It is based on a selection of indicators and contains descriptive data, statistical analyses and interactive figures and maps. The national perspective is used as a starting point, but whenever possible, data can be stratified at regional and local levels. To highlight the differences between different groups, the results are reported by sex, age, education, country of birth, and county (and municipality level when data is available).

Results
Public health in Sweden remains strong overall, but there are significant health disparities in the population. Health varies between sex, age, education level and country of birth as well as between different geographical areas. The health differences are greater in intra-regions than those of inter-regions are. The list of indicators is described below as annexe 1.

Study limitations
This is not a study with limitations whereas this is an annual report based on indicator-based public health reporting. Sweden has the resources and knowledge to undertake this work. However, this is time-consuming, as one needs to undertake a fair bit of data analysis and quality control of data.
Disaggregation of data into groups based on the prohibited discrimination grounds, as well as the five national minorities, would be more than welcome. However, such data is due to legal and practical reasons not available for the whole population. There is on-going developmental work regarding many of the discrimination grounds such as disability. Such developmental work may lead to better monitoring of the health of national minorities in the future.

Conclusions
Web-based reporting is resource-effective way of reporting. Governmental policies are important to tackle health inequalities. However, the most important measures are the responsibility of the regions and the municipalities. Therefore, it is important to find efficient public health solutions at the local community level.

Added value of web-based reporting
Web-based reporting is a new and effective way of reporting for the Public Health Agency, with the possibilities to update information whenever new data is available.

Implications of available evidence
Easily accessible, continuously updated data and associated analysis are of great importance for public health planning at the local level. This type of reporting stimulates discussions and is useful to many of our target groups where most of the public health work is carried out, such as small municipalities and civil society organisations.

Annex 1: List of indicators

Note: data for all indicators can be disaggregated by age, sex, educational level, country of birth, geographic location (county, municipality if possible).

Total indicators (outcomes + determinants) = 30

Health Outcome (n= 19)

1. Deaths according to alcohol index
2. Pre-mature death
3. Death, various types of cancer (breast, lung, prostate, colon and rectal)
4. Deaths, cardiovascular diseases
5. Deaths, suicide
6. Deaths, drug poisoning
7. Deaths, intentional and unintentional injuries
8. Fall injuries among elderly
9. Acute Myocardial Infarction (AMI) incidence
10. Stroke incidence
11. Life expectancy
12. Imparied mental health, self-reported
13. Stress, self-reported
14. Anxiety, self-reported
15. Psychosomatic problems amongst pupils, self-reported
16. General health status, self-reported
17. Infant mortality
18. Violence-inflicted injuries
19. Overweight and obesity, self-reported

Living and working Conditions (n=8)

1. Low income standard
2. Low economic standard, children
3. Low economic standard, adults
4. Primary school diploma (eligible to apply for secondary school)
5. Upper secondary school diploma (eligible to apply for university)
6. Long-term unemployment
7. Occupational level in the population
8. Young adults neither in employment nor in education

Lifestyle Habits (n=3)

1. Alcohol, risk consumption (heavy drinking), self-reported
2. Cannabis use in the population, self-reported
3. Tobacco smoking, daily, self-reported

14. Improving Burden of Injury Estimates: Case study based on injuries in Wales, 2012 - 2017

Summary

Background

Injuries are a leading cause of death and disability across Europe. To ensure injury prevention resources are appropriately allocated, and evidence-based preventative interventions and policies are targeted at those injuries/demographic groups creating the greatest health burden; valid and reliable methods for quantifying disease and injury burden are required.

Traditionally, the impact of injuries on overall population health was quantified using simple, singular outcomes e.g. incidence of injury-related fatalities, hospital admissions or emergency department attendances. However, in recent years new composite measures combining mortality and morbidity data have become increasingly important in providing comparable, comprehensive summaries of the overall burden of disease and injury.

A key composite measure introduced in 1993 by the World Bank was Disability Adjusted Life Years (DALYs)[1]. DALYs aim to summarize population-level health burden for any disease/injury, by combining years of life lost (YLL) through premature mortality, with years lived with disability (YLD). One DALY equates to one year of "healthy" life lost. The calculation of injury-related DALYs requires several data elements, including information on non-hospitalized injury cases, hospitalized injury cases, injury-related fatalities, injury-specific disability weights (indicating the severity of an injury), and remaining years of life. Disability weights (DWs) are key to the calculation of DALYs, and several studies [2–4] have produced injury-related DWs using varying methods, which have resulted in substantially different weights. Injury-VIBES (Validating and Improving Injury Burden Estimates Study [5]) is one of the most comprehensive injury DW studies to date. DWs were derived from case-reported outcomes collected across several cohort studies, and DWs were generated at a higher granularity and for a greater number of injury groupings than in previous studies.

This case study aims to utilize these new Injury-VIBES DWs to generate new injury burden estimates for Wales and provide a methodology other countries can follow to produce comparable estimates.

Methodology

Establish health problem and population

The first step is to establish the health problem and population for which the burden will be estimated. For this case study, we explored the injury burden for the population of Wales, in the years 2012 - 2017. A key component of DALY calculations is the number of fatalities, hospitalized cases, and non-hospitalized emergency department (ED) attendances in the given population. We were able to access routinely collected records on all injury-related fatalities, admissions and Emergency Department (ED) attendances across Wales through a world-leading infrastructure, the Secure Anonymized Information Linkage (SAIL) databank [6,7]. The SAIL databank provides approved health analysts with access to millions of anonymized, individual level, linkable health records to support health-related research. This piece of work was conducted by The All Wales Injury Surveillance System [8], a group of researchers with a specific remit to support the reduction of injuries in Wales. The inclusion criteria for injury-related deaths, hospitalized and non-hospitalized injury cases is as follows: For death (Only individuals, living in Wales, with a valid Anonymized Information Linkage Field (ALF_E) in the SAIL

Databank were included. ALF_E's are a double encrypted version of an individual's NHS number, Valid 'date of death registration' between *01/01/2012 – 31/12/2017, o*nly Welsh residents (inclusion in the Welsh Demographic Service (WDS) dataset), valid gender code (1=Male, 2=Female), Age <=110, ICD10 codes recorded in the 'underlying cause of death' field in the ONS mortality dataset used to establish injury deaths. Injury ICD10 codes: S%%%, T00-T65, T704, T708, T709, T71, T750, T751, T754, T794, T795, T796, T797, T798, T799, V, W00-W41, W44-W77, W79-W99, X00-X29, X32-X49, X59-X99, Y00-Y05, Y08-Y32, Y35-Y36, F100, F110, F120, F130, F140, F150, F160, F170, F180, F190).

For injury admissions (Only individuals, living in Wales, with a valid Anonymized Information Linkage Field (ALF_E) in the SAIL Databank, were included. ALF_E's are a double encrypted version of an individual's NHS number, valid admission date between *01/01/2012 – 31/12/2017, o*nly Welsh residents (inclusion in the Welsh Demographic Service (WDS) dataset), valid gender code (1=Male, 2=Female), Age <=110). Only patients with the following admission method codes: 21-A&E or dental casualty department of the health care provider; 22– GP, after a request for immediate admission has been made directly to a hospital provider by a General Practitioner or deputy; 23-Bed Bureau; 24-Consultant clinic of this or another health care provider; 25-Domiciliary visit by Consultant; 27-Via NHS Direct Services; 28-Other means, including admitted from the ED department of another provider where they had not been admitted; 29-Emergency transfer. Only cases where the 1st episode, in the 1st admission within a person super spell, contains the following ICD10 codes: S00-S99, T00-T65, T704, T708, T709, T71, T750, T751, T754, T794, T795, T796, T797, T798, T799, F100, F110, F120, F130, F140, F150, F160, F170, F180, F190. Specifically, we followed the R/Z rule. A physical injury ICD10 code was required to either be in primary position in the first episode or, if not in primary position, then only an R/Z ICD10 code or NULL values could precede the injury code.

For non-hospitalized admissions (Only individuals, living in Wales, with a valid Anonymized Information Linkage Field (ALF_E) in the SAIL Databank, were included. ALF_E's are a double encrypted version of an individual's NHS number, valid attendance date between 01/01/2010 – 31/12/2017, only Welsh residents (inclusion in the Welsh Demographic Service (WDS) dataset), valid gender code (1=Male, 2=Female), Age <=110, only new attendances included (e.g. follow-up attendances excluded), attendances admitted to hospital (1/2), transferred to a different trust (3), the patient died in the department (10) or dead on arrival (11) excluded). Injury diagnosis code present in diagnosis positions 1-6 (EDDS codes or ICD10 codes as defined below) or treatment codes in positions 1-6).

Mapping disability weights to local injury diagnostic codes

Once all injury-related fatalities, admissions and ED attendances have been identified in the given population, the next step is to map injury-specific Disability Weights (DWs) from the chosen DW project to local injury diagnostic codes in these datasets. DWs indicate the severity of an injury, ranging from 0 (perfect health) to 1 (equivalent to death). In this study, we used DWs generated by the Injury-VIBES project. Injury-VIBES generated DWs for both hospitalized and non-hospitalized injuries, by several injury classifications (e.g. ICD-10, GBD (Global Burden of Disease), IDB (Injury Data Base) and Eurocost)[5]. Some of the local diagnostic codes in our inpatient and ED dataset mapped directly to Injury-VIBES DWs, whereas others required additional calculations to generate weighted averages, or expert opinion to select nearest matches. It is advised any mapping of Injury-VIBES DWs to local codes are reviewed by several relevant experts and clinicians, to ensure DWs are mapped as appropriately as possible.

We have also provided an additional table mapping Injury-VIBES DWs to the standardized European Injury Data Base format (IDB).

Some key mapping decisions in this case study:

1. It was agreed all superficial injuries, contusions and open wounds would be mapped to the combined open wounds and superficial injuries DW. This was an additional unpublished DW generated by the Injury-VIBES project statistician.
2. For non-hospitalised nerve/vascular injuries it was agreed the same combined open wounds and superficial injuries DW would be used.
3. It was agreed that all visceral injuries would be admitted, and any occurrence in the non-admitted dataset would likely be an error. As such, all non-admitted visceral injuries were assigned a DW of zero.
4. Where no DW existed for a dislocation/sprain/strain of a particular body part, it was agreed the lowest dislocation/sprain/strain DW would be used as a conservative estimate.
5. In instances where several DWs overlapped one injury, the most specific DW to the particular injury was selected.
6. In cases where no appropriate DW existed, the lowest hospitalised/non-hospitalised DW was used.

DALY calculations

DALYs (Disability Adjusted Life Years) are calculated as the sum of Years of Life Lost (YLLs) due to premature mortality and Years Lost due to Disability (YLDs).

DALY = YLL + YLD

YLLs are calculated as follows:

YLL = N x RYL

N = number of deaths

RYL = remaining years of life at age of death

For this case study, the number of injury fatalities was obtained from the Office for National Statistics (ONS) national mortality dataset [9], available within the anonymized SAIL databank [6,7]. The remaining years of life estimates were obtained from the Global Burden of Diseases project [10].

YLDs for injury burden are calculated as follows:

YLD = (I x (STDW + (LTDW x RYL))

I = Number of injury cases (separate calculations for both hospitalized and non-hospitalized cases as DWs differ depending on admittance status)

STDW = Short-Term Disability Weight (a weight factor reflecting the severity of the injury on a scale from 0 (perfect health) to 1 (dead) in the first 12 months post-injury)

LTDW = Long-Term Disability Weight (a weight factor reflecting the severity of the injury 12 months post injury)

RYL = remaining years of life 12 months post injury

For this study, the number of hospitalized and non-hospitalized injury cases (I) in Wales were obtained from data on inpatient admissions and emergency department attendances in the SAIL databank. Separate counts were made for individuals who were admitted, and those who only presented at EDs, as Injury-VIBES DWs differ dependent on admittance status. Both short-term (STDWs) and long-term DWs (LTDWs) from the Injury-VIBES study were mapped to both admitted

and non-admitted injury cases. Short-term DWs represent disability in the first year post-injury, and long-term DWs represent disability in the remaining years of life 12 months post-injury.
For hospitalized and non-hospitalized cases in which multiple injuries were diagnosed, the decision was made in this study to select the injury with the highest STDW and its corresponding LTDW. It should be noted that previous studies have found disability at 12 months post-injury increases with the number of injuries affecting the patient [11]. Therefore, we would recommend future studies consider exploring the impact of multiple injuries to avoid underestimating the injury burden.
However, an aspect of current DALY methodologies may artificially inflate burden estimates. Applying long-term disability (12 months post-injury) to the remaining years of expected life may not be a true reflection of long-term disability for all injuries. In the future, we are planning further analyses to explore curtailing long-term disability at remission of injury symptoms, rather than the end of life, to evaluate the impact on injury burden estimates.

Results

In 2017, injuries in Wales resulted in an estimated 11,631 Disability Adjusted Life Years (DALYs) per 100,000 population (Table 1), or 363,485 DALYs in total. DALY rates published in this study are between 4 to 6 times greater than DALY rates published elsewhere for the UK/Wales (Table 4 in discussion).
Injury burden appears to be reducing in Wales, decreasing by 12.7% between 2012 and 2017. A similar trend has not been observed for injury admission/attendance counts in Wales during this period, suggesting the reduced burden is not simply related to a decrease in the incidence of injuries. The reduction in burden could relate to several factors including less injuries in younger age groups, or a decrease in the severity of injuries. We will explore the reasons behind the decreasing burden in Wales in future analyses.

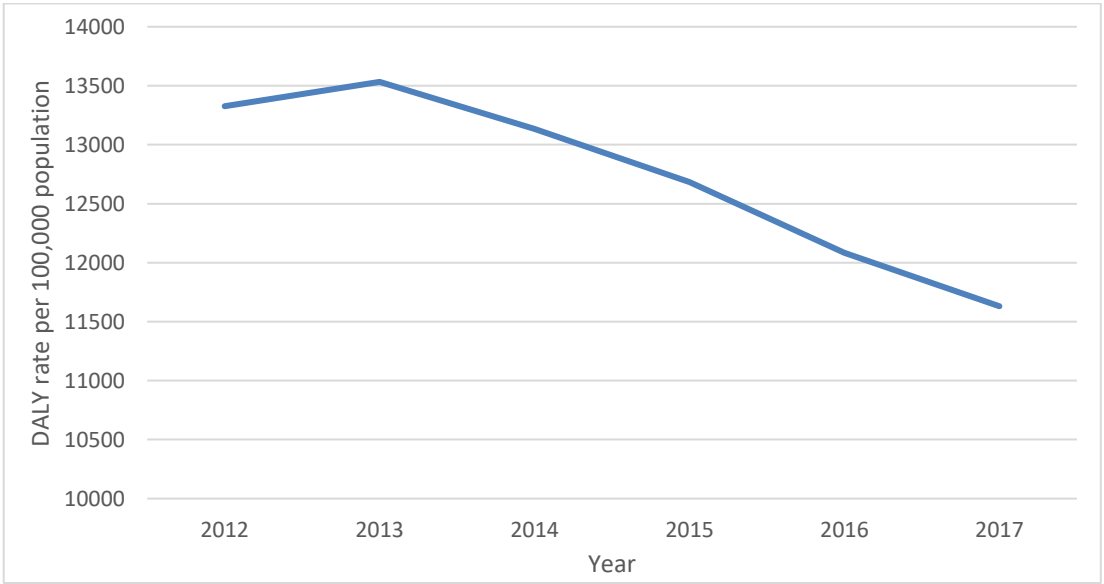


Figure1: Injury DALY rate per 100,000 population in Wales by year, from 2012 - 2017

Table 1: Injury DALY rate per 100,000 population in Wales by year, from 2012 - 2017

Year	DALY rate per 100,000 population
2012	13,327
2013	13,533
2014	13,134
2015	12,684
2016	12,082
2017	11,631

Children, and in particular young males, contribute the highest injury burden in Wales, with injury DALYs peaking in males aged 10-14 years (Figure 2, Table 2). Injury burden gradually decreases across the life course with females contributing a greater burden in the 65+ age groups compared to males.

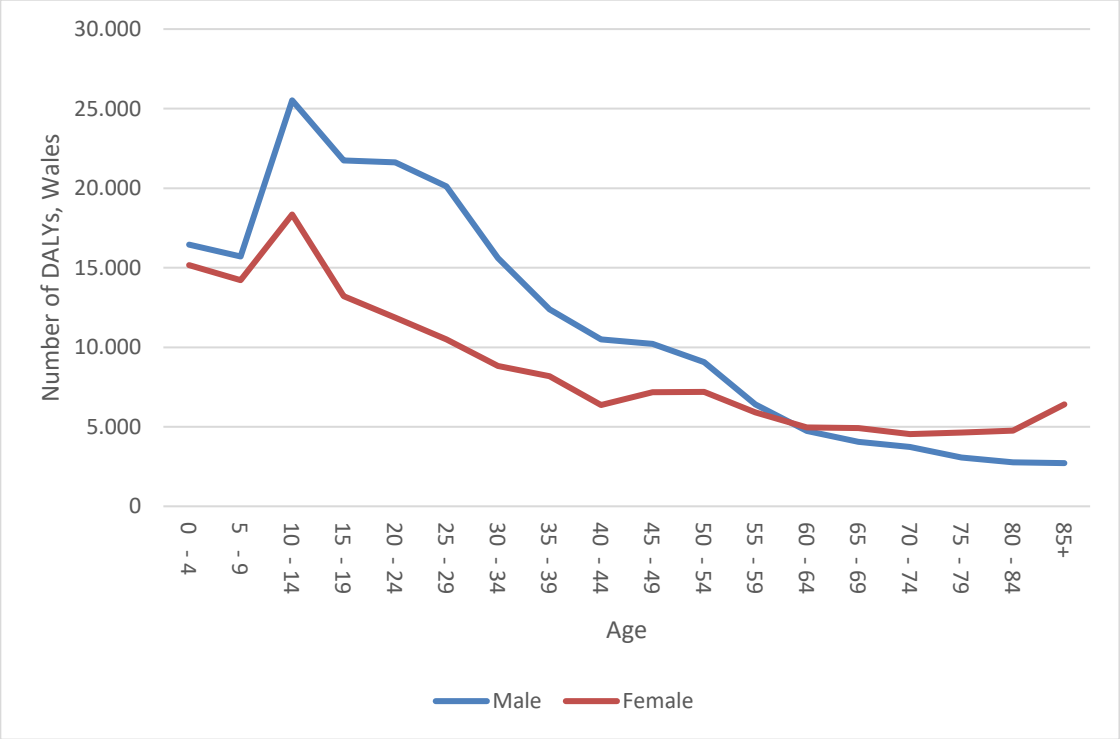


Figure 2: Injury DALY rate per 100,000 population in Wales by gender and age group in 2017.

Table 2: Injury DALY rate per 100,000 population in Wales by gender and age group in 2017.

Age	Male	Female
0 - 4	16,449	15,156
5 - 9	15,697	14,223
10 - 14	25,521	18,345
15 - 19	21,753	13,207
20 - 24	21,617	11,853
25 - 29	20,117	10,487
30 - 34	15,610	8,824
35 - 39	12,386	8,169
40 - 44	10,498	6,369
45 - 49	10,205	7,172
50 - 54	9,058	7,190
55 - 59	6,415	5,898
60 - 64	4,742	4,957
65 - 69	4,044	4,915
70 - 74	3,742	4,540
75 - 79	3,060	4,631
80 - 84	2,763	4,755
85+	2,716	6,402

Injury burden is greatest in the most deprived areas in Wales compared to least deprived, with individuals aged 25-29 living in the most deprived areas of Wales experiencing 2.5 times more injury DALYs compared to individuals living in the least deprived areas in this age group. Area level deprivation was assigned based on an individual's anonymized address (aggregated to area level in the SAIL system) and the 2011 Welsh Index of Multiple Deprivation [12]. The WIMD is an official Welsh Government measure of relative, area-level deprivation (1 = most deprived areas and 5 = least deprived).

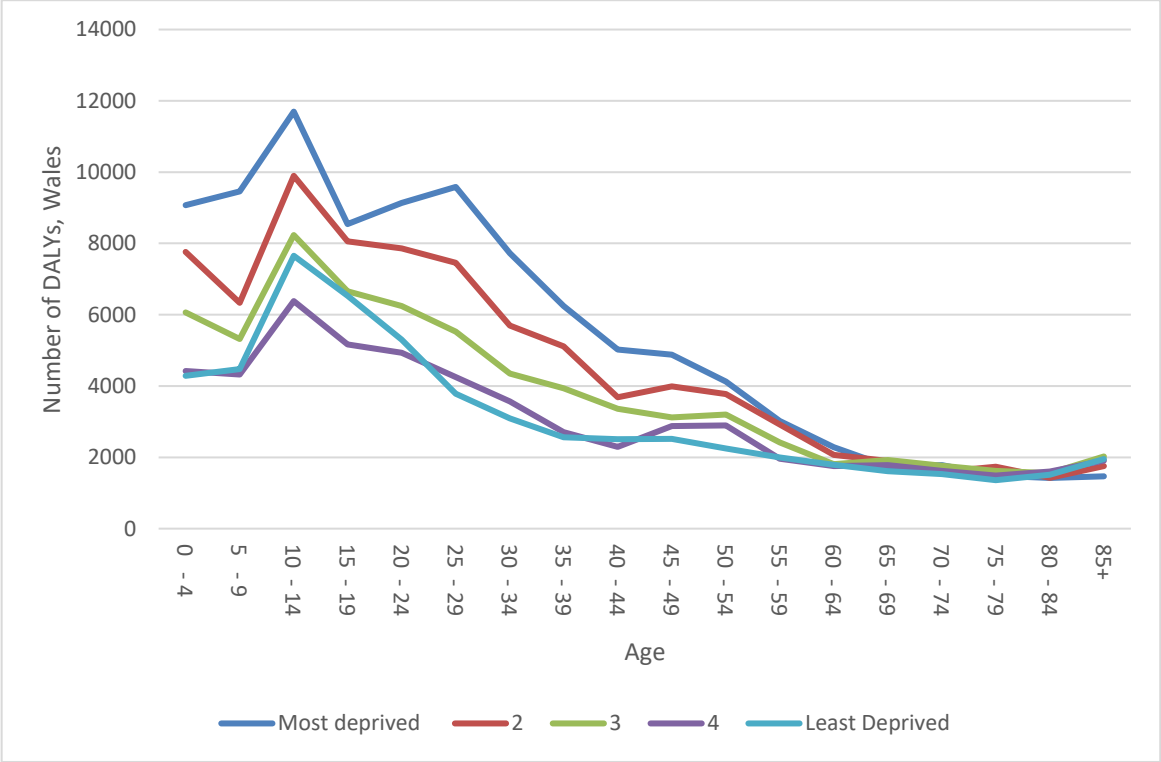


Figure 3: Injury DALY rate per 100,000 population in Wales by area-level deprivation and age group in 2017.

Table 3: Injury DALY rate per 100,000 population in Wales by area-level deprivation and age-group in 2017

Age	Most deprived	2	3	4	Least Deprived
0 - 4	9,074	7,762	6,063	4,422	4,284
5 - 9	9,464	6,335	5,319	4,325	4,477
10 - 14	11,697	9,897	8,235	6,383	7,654
15 - 19	8,541	8,061	6,655	5,168	6,535
20 - 24	9,135	7,858	6,241	4,936	5,299
25 - 29	9,583	7,452	5,528	4,255	3,785
30 - 34	7,728	5,701	4,348	3,566	3,092
35 - 39	6,244	5,109	3,935	2,705	2,563
40 - 44	5,023	3,684	3,358	2,293	2,508
45 - 49	4,880	3,987	3,119	2,876	2,515

50 - 54	4,126	3,778	3,197	2,898	2,248
55 - 59	3,019	2,919	2,416	1,961	1,998
60 - 64	2,284	2,070	1,805	1,752	1,788
65 - 69	1,742	1,902	1,927	1,779	1,609
70 - 74	1,779	1,589	1,760	1,624	1,530
75 - 79	1,487	1,738	1,617	1,489	1,359
80 - 84	1,421	1,425	1,557	1,599	1,516
85+	1,470	1,759	2,028	1,904	1,956

How do Injury-VIBES DALY estimates compare to other estimates?

The injury DALY estimates calculated for Wales in this case study are between 4 to 6 times greater than other Wales/UK injury DALY estimates (Table 4). Differences in methodological approaches and underlying disability weights account for these differences, with key factors including:

- Injury-VIBES used validated instruments to measure case-reported outcomes directly from injured patients. Panel designs, based on expert/public opinion can sometimes underestimate disability for injuries perceived to be 'less severe', and overestimate disability for injuries perceived to be more severe.
- Injury-VIBES provided DWs for a greater number of injury groups. In other studies, these missing injury groups would have been assigned a disability of zero.
- Previous studies have combined several conditions into single injury groupings. By providing individual level ICD10 DWs where possible, the Injury-VIBES study revealed injuries commonly grouped can result in substantially different outcomes and disability. For example, Injury-VIBES reported that clavicle fractures have a much lower disability than fractures of the humerus or scapula, and fractures of the distal radius are less disabling than fractures of the proximal radius.
- For admissions/attendances with several injury diagnoses, our methodology selected the injury with the highest short-term disability weight and corresponding long-term weight. Other studies have focused on the principal diagnosis, which may not have the highest disability.
- DALYs were calculated at an admission/attendance level. This means that individuals with multiple injury attendances/admissions in the time period will contribute multiple and larger YLDs.

Table 4: Injury DALY estimates per 100,000 population by study

Study	Population	Year	Injury DALYs per 100,000 population
The Burden of Injury Report in Wales[17]	Wales	2009	1,947
GBD 2017 estimate[18]	UK	2017	2,112
UKBOI study [19]	UK	2005	2,721
Injury-VIBES	Wales	2017	11,631

Study limitations

Several limitations should be considered when using DWs from the Injury-VIBES study, including a small number of cases for some DWs, differing follow-up rates and varying amounts of EQ-5D data available (the standardized instrument used to measure an individual's health status), study population limited to adults, comorbidities not taken into account, recovery within 3 months not considered, only high-income countries included, only primary diagnoses analyzed and certain injuries were underrepresented in the study population (e.g. penetrating injuries). Further research is required to establish the impact of these limitations on injury burden estimates.

Further, a current assumption of injury DALY calculations, is that disability at 12 months post-injury, is constant for the remainder of life. However, not all patients with continuing injury-related disability at 12 months, will have this disability for life. The AWISS are planning to conduct further analyses exploring the curtailment of long-term disability at remission of injury symptoms rather than the end of life. Using the SAIL databank, which facilitates individual-level health record linkage, we will link injury attendees/patients to prospective data on GP visits, outpatient appointments, etc. to make informed judgments on the likely curtailment of long-term disability for specific injuries. It will also be possible through SAIL to curtail disability at actual death, rather than expected death for individuals who have died in the study period.

Conclusions

This case study has produced injury burden estimates for Wales 4-6 times greater than previous estimates. Although limitations exist with current methods and the disability weights (DWs), the magnitude of difference suggests the burden of injuries in Wales, and potentially other countries is considerably higher than previously thought. Further, long-term disability is reported for all admitted injuries in the Injury-VIBES project, suggesting injury is often a chronic disorder and the burden of disease estimates should reflect this [5].

Added value of this study

Limitations aside, the methodology and disability weights presented in this case study are the most comprehensive to date. Future studies should consider exploring current limitations, the impact they may have on burden estimates, and seek to develop new approaches to minimize potential biases.

Implications of available evidence

The following document is designed to support countries, particularly those involved in the European Injury Data Base (IDB) project, to calculate improved injury burden estimates for their regions. To ensure the greatest reduction in the frequency and severity of injuries across Europe and globally, it is essential to have high quality, reliable estimates on the burden of injuries. Understanding the extent of the injury burden and demographic groups/injuries contributing to the greatest burden can ensure resources are allocated accordingly. This study has revealed that previous estimates most likely underestimate the true burden of injuries across Europe.

References

1 World Bank. World development report 1993: investing in halth. New York: Oxford University Press: 1993.

2 Meerding WJ, Looman CWN, Essink-Bot M-L, *et al*. Distribution and Determinants of Health and Work Status in a Comprehensive Population of Injury Patients. *J Trauma Inj Infect Crit Care* 2004;**56**:150–61. doi:10.1097/01.TA.0000062969.65847.8B

3 Van Beeck EF, Larsen CF, Lyons RA, *et al*. Guidelines for the Conduction of Follow-up Studies Measuring Injury-Related Disability. *J Trauma Inj Infect Crit Care* 2007;**62**:534–50. doi:10.1097/TA.0b013e31802e70c7

4 Wang H, Naghavi M, Allen C, *et al*. Global, regional, and national life expectancy, all-cause mortality, and cause-specific mortality for 249 causes of death, 1980–2015: a systematic analysis for the Global Burden of Disease Study 2015. *Lancet* 2016;**388**:1459–544. doi:10.1016/S0140-6736(16)31012-1

5 Gabbe BJ, Lyons RA, Simpson PM, *et al*. Disability weights based on patient-reported data from a multinational injury cohort. *Bull World Health Organ* 2016;**94**:806-816C. doi:10.2471/BLT.16.172155

6 Lyons R a, Jones KH, John G, *et al*. The SAIL databank: linking multiple health and social care datasets. *BMC Med Inform Decis Mak* 2009;**9**:3. doi:10.1186/1472-6947-9-3

7 SAIL Databank. https://saildatabank.com/ (accessed 16 Oct 2018).

8 All Wales Injury Surveillance System. www.awiss.org.uk (accessed 1 Oct 2019).

9 Deaths - Office for National Statistics. https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths (accessed 26 Mar 2020).

10 Bhalla K, Harrison J. Global Burden of Disease Burden Calculator. http://calculator.globalburdenofinjuries.org/ (accessed 23 Mar 2017).

11 Gabbe BJ, Simpson PM, Lyons RA, *et al*. Association between the Number of Injuries Sustained and 12-Month Disability Outcomes: Evidence from the Injury-VIBES Study. *PLoS One* 2014;**9**:e113467.https://doi.org/10.1371/journal.pone.0113467

12 Welsh Government. Welsh Index of Multiple Deprivation (WIMD) 2011. 2011.

13 Swansea University. Injury Indicators for Wales Report 2019. 2019. https://www.awiss.org.uk/wp-content/uploads/2019/10/Injury-Indicators-for-Wales-Report-2019.pdf

14 Lyons RA, Turner SL, Lyons. The All Wales Injury Surveillance System (AWISS) Injury Indicator Data Quality Report. AWISS, Swansea University: 2019. https://www.awiss.org.uk/wp-content/uploads/2019/10/Injury-Indicator-Data-Quality-Report.pdf

15 Haagsma JA, Polinder S, van Beeck EF, *et al*. Alternative approaches to derive disability weights in injuries: do they make a difference? *Qual life Res an Int J Qual life Asp Treat care Rehabil* 2009;**18**:657–65. doi:10.1007/s11136-009-9484-0

16 GBD 2013 DALYs and HALE Collaborators, Murray CJL, Barber RM, *et al*. Global, regional, and national disability-adjusted life years (DALYs) for 306 diseases and injuries and healthy life expectancy (HALE) for 188 countries, 1990–2013: quantifying the epidemiological transition. *Lancet* 2015;**386**:2145–91. doi:10.1016/S0140-6736(15)61340-X

17 Jones S, Macey S, Lyons R, *et al*. The Burden of Injury In Wales. 2012.

18 GBD Results Tool. http://ghdx.healthdata.org/gbd-results-tool (accessed 26 Mar 2020).

19 Lyons RA, Kendrick D, Towner EM, *et al*. Measuring the population burden of injuries--implications for global and national estimates: a multi-centre prospective UK longitudinal study. *PLoS Med* 2011;**8**:e1001140–e1001140. doi:10.1371/journal.pmed.1001140

B. Examples related to ML method

14. Insight into antidepressant prescribing using open health data

Link to published article: bit.ly/2UqQXM3

Summary[15]

Background
The growth of big data is transforming many economic sectors, including the medical and healthcare sector. Governments worldwide have begun to include the impact of big data in their policy statements. The UK government recently stated that big data had "huge unrealized potential, both as a driver of productivity and as a way of offering better products and services to citizen-s". Despite this, research on the implications of big data to medical policymaking and service delivery remains relatively limited. With the exception of some public health surveillance and pharmacovigilance systems, the opportunities for better policymaking using big data are still largely unexplored. In this study, we provide an analysis of the potential application of big data and machine learning methods for the development of public policy and service delivery. In particular, we focus on the use of heterogeneous open data from a variety of online sources, including disease prevalence data, GP prescribing data and economic deprivation data. We examine how such datasets can be brought together and analyzed in such a way as to generate usable, actionable insights for clinicians, policymakers and the public. This study discusses the context of mental health and antidepressant prescribing in Northern Ireland and highlights its importance as a public policy issue.

Methods
A hypothesis is proposed, suggesting that the link between antidepressant usage and economic deprivation is mediated by depression prevalence. We tested this hypothesis by using open data drawn from multiple publicly available sources. Specifically, we examined the links between three major variables: economic deprivation, depression prevalence and antidepressant prescribing, to explore the correlations between them. We also examined correlations between these variables and other disease prevalence data, and with GP prescribing data for other drug groups. Finally, we applied a k-means clustering algorithm to determine if meaningful GP practice sub-groups could be identified from the overall dataset. An analysis of three heterogeneous open datasets is used to test this hypothesis: 1. Open prescribing data: Antidepressant prescribing data was downloaded from the Open Data NI portal, operated by the Northern Ireland Department of Finance. For this study, data for the twelve months of 2016 was used. 2. Economic deprivation data: The metric for economic deprivation used in this study was the Multiple Deprivation Measure (MDM), which is published by the Northern Ireland Statistics and Research Agency (NISRA). Unlike in other parts of the UK, individual GP practice data does not include deprivation measures. To link GP practices to the MDM data, the postcode for each practice was obtained from the Detail Data portal, an open data resource provided by the Northern Ireland Council for Voluntary Action. The online MySociety Mapit service was used to convert the postcode data into super output areas (SOA). 3. Disease prevalence data that is the number of patients per 1000 diagnosed as suffering from different diseases are available across most of the UK under the Quality Outcomes Framework (QOF), a collection of data, which is designed to measure GP performance to support GP payments. In Northern Ireland, the QOF data no longer includes disease prevalence figures, but fortunately, these are still published separately by the Department of Health. Since the prevalence data is linked directly to GP

practice identifiers, it allows for a direct comparison between the number of patients being diagnosed with depression and the amount of antidepressants being prescribed.
Jupyter Notebooks were used to algorithmically restructure, transform and merge datasets were required. Pandas and NumPy were used for the statistical analysis. Visualization of the data was also done through Jupyter Notebooks using MatPlotLib and Seaborn for charts and graphs and iPyLeaflet for maps. Correlations between the key variables were explored using the Pearson correlations and p-values. The scikit-learn library was employed to perform K-means clustering on the data. Correlations between key variables and several different clustering analyses were performed.

Results
Our results showed variation in prescribing patterns across GP practices such as the most heavily prescribed drugs are those for central nervous system disorders (including antidepressants), infections, and endocrine systems disorders (including insulin). Large variations in prescribing are visible across almost all drug categories. For the variation in disease prevalence across GP practices, the most visible variations appear in the diagnosis of more prevalent disease categories, such as depression, hypertension, asthma and diabetes. Economic deprivation was strongly correlated with central nervous system prescribing (r = 0.34) and moderately correlated endocrine system pre-scribing (r = 0.24). Antidepressant prescribing was strongly correlated with central nervous system prescribing (r = 0.54) and endocrine prescribing (r = 0.31). It was also moderately correlated with prescribing for cardiovascular (r = 0.25), infections (r = 0.29), obstetrics (r = 0.29) and musculoskeletal (r = 0.29). Clusters of GP practices based on prescribing behaviour and disease prevalence were also described and key characteristics are identified and discussed.

Discussion
This study explored correlations between three main variables – economic deprivation, depression prevalence and antidepressant prescribing – based on open GP prescribing data. The results showed that while there was a strong correlation between economic deprivation and antidepressant prescribing, the correlations between deprivation and prevalence and between prevalence and prescribing were weak. We, therefore, propose that the hypothesis that depression is a mediating factor between deprivation and prescribing is not supported by the available data. Moreover, the lack of a clear link between depression prevalence and prescribing rates suggest that the clinical basis for increased antidepressant prescribing requires further investigation.

Study limitations
The major challenges for the study were in identifying, integrating and analyzing open datasets from diverse sources. A range of variables cannot be fully explored in open data due to privacy or confidentiality considerations. Linking two classification systems of prescribing and GP practice data (i.e., ATC identifiers and BNF), lacks a formally standard method. This makes it difficult to make a valid comparison. Due to different data collection methods in various parts of the country, make it challenging to link and compare the data.

Conclusions
Mental health has been identified as a priority area for public investment, both within Northern Ireland and across the UK as a whole. Data analytics might be applied to open prescribing data, disease prevalence data and economic deprivation data to highlight issues related to public

policy and service delivery. The seeming weakness of the correlation between antidepressant prescribing and depression prevalence calls into question the medical basis for increasing antidepressant usage. Within Northern Ireland, where levels of antidepressant prescribing greatly exceed other countries, this is a particularly urgent concern. In order to address this problem, further data analysis might be used to identify anomalous prescribing patterns and thereby enable targeted interventions by the Department of Health. Such analysis could also help to identify hidden environmental or clinical factors, and to provide GPs with information to support clinical decision-making.

Added value of this study
This study highlights the impact of big data on other sectors suggests that data science can help to address some issues related to public policy and service delivery.

Implications of available evidence
Policymakers have also recognized the need to address effectiveness and efficiency in how these services are delivered. This analysis suggests that interesting correlations between disease prevalences and GP prescribing do exist, and may have useful implications for future policymaking. Moreover, the clustering of GP practices based on depression prevalence data implies that different populations respond to economic deprivation factors in different ways.

15. Using natural language processing to extract structured epilepsy data from unstructured clinic letter: development and validation of the ExECT (extraction of epilepsy clinical text) system

Link to published article: bit.ly/3byZtOK

Summary[16]

Background
Epilepsy is a common neurological disease with significant co-morbidity. Epilepsy research using routinely collected data currently tends to use sources such as primary care health records or hospital discharge summaries. The main disadvantage of these sources is that they do not contain detailed epilepsy information, for example, epilepsy subtype/syndrome, epilepsy cause, seizure type or investigation results. This limits the quality and type of epilepsy research questions that can be answered successfully. Clinic letters have been written electronically for decades and offer a wealth of disease-specific information to enhance routinely collected data for research. Although detailed disease (epilepsy) information is found in clinic letters, they are usually written in an unstructured or semi-structured format, making it difficult to automatically extract useful information.
Natural language processing (NLP) technology can be used to analyze human language and offers a potential solution for automated information extraction from unstructured letters. NLP is increasingly being used for healthcare information extraction applications; for example, to extract symptoms of severe mental illness and adverse drug events from psychiatric health records, to identify patients with non-epileptic seizures and for the early identification of patients with multiple sclerosis. For this study, we aim to use natural language processing techniques to extract detailed structured clinical information from unstructured epilepsy clinic letters to enrich routinely collected data.

Methods

Setting: We used manually de-identified and pseudonymised hospital epilepsy clinic letters to build and test the algorithm. The paediatric and neurology departments of a local general hospital serving half a million residents in Wales, UK, provided the clinic letters.

Design: We used the general architecture for text engineering (GATE) framework to build an information extraction system, ExECT (extraction of epilepsy clinical text), which used Bio-YODIE (i.e., GATE's biomedical named entity linking pipeline) and our customisations to map clinical terms to Unified Medical Language System (UMLS) concepts. The UMLS is a set of files and software, developed by the US National Library of Medicine that combines information from over 200 health vocabularies with over 3.6 million concepts and 13.9 million unique concept names. UMLS uses concept unique identifiers (CUIs) to identify senses (or concepts) associated with words and terms. Bio-YODIE applies several strategies to assign the correct UMLS sense to terms in the text, and, where necessary, disambiguates against several possible meanings for the same term. These strategies include term frequency, patterns of co-occurrence with other terms and measures of context similarity. We combined rule-based and statistical techniques to build ExECT system. We extracted nine categories of epilepsy information (i.e., epilepsy diagnosis, epilepsy type, seizure type, seizure frequency, medication, investigation and levels of certainty) in addition to clinic date and date of birth across 200 clinic letters.

Measuring performance: We compared the results of our algorithm with a manual review of the letters by an epilepsy clinician. The manual review was performed by an epilepsy clinician (one of the authors: WOP) who was blinded to the algorithm results until the review was complete. We used predefined criteria for the manual review of epilepsy information items on nine elements (reported in table 1 of the published article). The core research team (authors: BF-S, ASL and WOP) reviewed every disagreement between the manual review and ExECT, and a consensus was obtained from the group on the correct annotation based on our pre-defined guidelines (reported in table 1 of the published article). We measured performance on both a *per item* and a *per letter* basis.

Analysis and statistical tests: To analyze the results, we used *precision, recall* and *F1 score* to measure the accuracy of ExECT. Precision is defined as the proportion of the instances extracted by the algorithm, which are true, recall is the proportion of true instances extracted by the algorithm and F1 score is the unweighted harmonic mean of precision and recall: (2×precision× recall)/(precision+recall).

Results

We identified 1925 items of information with overall precision, recall and F1 score of 91.4%, 81.4% and 86.1%, respectively, from 200 letter on *per item basis*. Precision and recall for epilepsy-specific categories on *per item* basis were: epilepsy diagnosis (88.1%, 89.0%), epilepsy type (89.8%, 79.8%), focal seizures (96.2%, 69.7%), generalized seizures (88.8%, 52.3%), seizure frequency (86.3%–53.6%), medication (96.1%, 94.0%), CT (55.6%, 58.8%), MRI (82.4%, 68.8%) and electroencephalogram (81.5%, 75.3%).

Study limitations

A small number of letters sourced from one health board was used. This limit the number of writing style and letters structures available to validate the algorithm. Therefore, the generalizability of the algorithm to population-level data and other diseases is limited at present but is possible with further work. Moreover, it is difficult to account for the variability of the language used to express patient information in clinic letters. Some items of information such as seizure frequency and investigations require many complex rules where patterns are hard to predict. Further work could be focused on employing machine-learning methods to complement

a rule-based approach; however, this would require a significant amount of time to annotate the large number of documents required for such a task. All disagreements between ExECT and manual annotation were reviewed by the research team as a whole but we only used one clinician to review the letters, which might have added bias to how the validation set was annotated.

Conclusions
We have built an automated clinical text extraction system (i.e., ExECT) that can accurately extract epilepsy information from free text in clinic letters. This can enhance routinely collected data for research in the UK. The information extracted with ExECT such as epilepsy type, seizure frequency and neurological investigations is often missing from routinely collected data. We propose that our algorithm can bridge this data gap enabling further epilepsy research opportunities.

Added value of this study
We used a gold standard data set of de-identified clinic letters to build and test ExECT, using a novel of the technique of NLP, from which we accurately extracted epilepsy information for research. We can now iteratively develop ExECT over larger sets of clinic letters and use it to extract detailed epilepsy information for research on a population-level basis. This technique is easier and quicker to extract useful information for research.

Implications of available evidence
This approach can also be adopted to develop algorithms for other diseases and potential clinical applications, for example, efficiently extracting relevant clinical information from historical letters to aid clinicians. This technique could facilitate clinical practice to record patient information in a structured manner.

C. Examples related to both data linkage and ML method

16. Proactive advising: a machine-learning driven approach to vaccine hesitancy in Croatia

Link to published study: https://bit.ly/3dmLe0e

Summary[17]

Background
Despite being nearly eradicated, measles has surged to 20- year high in 2018. Two Croatian counties (out of 21) were part of this research and in one of them, the vaccination rate of infant doses of MMR (I & II) had dropped from 95% in 2010 to below 60% in 2017. The current response to vaccine rejection in these counties was based on a notification system, whereupon receiving a notification on the missed mandatory vaccine, health administrators contact the family. Previous research has shown that the dissemination of health information related to vaccination targeting broad public implemented via public health communication channels, had no significant effect on vaccine hesitancy that emphasizes the need to carefully select who to message, message content, timing and delivery method. Therefore, the main objectives of the study were to assess whether or not effective machine learning models can be built to predict

if an individual child is at risk of not receiving the MMR vaccination, and whether these models can be deployed as an Early Warning and Monitoring System. This would allow healthcare policymakers to intervene pre-emptively and more accurately.

Methods

We used two data sources i.e., Electronic Health Records and census data, of two counties that have reported a low vaccination rate. We used the following variables: demographic and personal information, geographic information, visits to the doctor, infant vaccination record, sibling vaccination history and personal medical history. We performed several pre-processing procedures necessary to prepare the data set to develop this approach:

- continuous variables were coded/rescaled to have a mean of zero and standard deviation of one, missing variables were filled using variable's mean value
- binary variables were coded with 1 for true, and 0 for false, missing values were filled with 0,5
- categorical variables were encoded using a one-hot encoding scheme, missing variables were filled with 0
- separate binary variable was created to indicate systematic missing values
- outliers were removed

A set of three models was tested: ridge, LASSO (Least Absolute Shrinkage and Selection Operator) regression and gradient-boosted decision trees. We applied a grid search method to select the optimal model and hyperparameters among the three models. All models were trained using a dataset including students who began their first grade between 2011 and 2015. We used a cohort of students who began school in 2016 as a validation dataset. For each model, the following evaluation metrics were recorded: AUC, accuracy, average log-likelihood, precision and recall. Since some variables used were changed over time and models were trained to make predictions at multiple time points.

Results

All models outperformed the baseline on AUC (Area Under the Curve). When evaluating models based on precision, *k* was chosen to be 20%, meaning that the results show the precision for the top 20% most at-risk children. With respect to AUC, the overall best-performing model was the gradient-boosted decision trees model (0.83 for county 1, and 0.76 for county 2. County baseline values were 0.72 and 0.63, respectively). AUC for the ridge model was 0,81 for county 1, and 0,73 for county 2. AUC for the LASSO model was 0,81 for county 1, and 0,73 for county 2.
The results are shown in the figures below. Even though the gradient model was slightly better than the other 2 models, the LASSO model was chosen to be used in an Early Warning and Monitoring System because of the easier understanding and usage of the LASSO model by the healthcare providers. This was considered justified because the gradient model was only 3% better than the LASSO model.

FIGURE 1. AUC FOR COUNTY 1, SELECTED MODEL IN BLUE

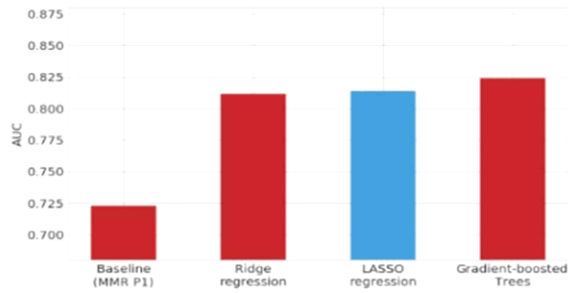
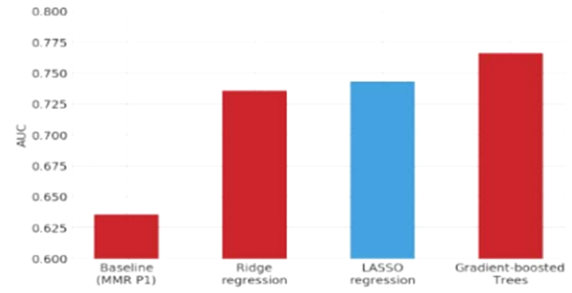


FIGURE 2. AUC FOR COUNTY 2, SELECTED MODEL IN BLUE



Discussion

The results of this showed that using Electronic Health Records, effective models for predicting vaccine hesitancy could be developed. Despite achieving somewhat lower precision, the LASSO regression model was implemented in the Early Warning and Monitoring System prototype dashboard to minimize the "black box" effect. This model is well equipped to answer a variety of policy and healthcare questions by offering predictions at four-time points e.g. before April 1st when applications for school enrolment begin and then in the fall right before the school begins.

Study limitations

There are a few challenges related to the implementation of the Early Warning and Monitoring System. Early Warning and Monitoring System must remain connected to the public health institute server to provide accurate and up-to-date risk score predictions. This linkage could encounter a potential security issue in the system. For future research, the authors propose Randomized Control Trial that would evaluate the effect of using the Early Warning and Monitoring System on vaccine uptake among hesitant groups, as compared to the "business-as-usual" system i.e., the current entirely reactionary system where healthcare providers get a notification when vaccine rejection already occurs. Unfortunately, further, development has been stopped due to legal interoperability issues.

Conclusions

This study explored the use of a machine learning approach to Electronic Health Record data to predict which families will be hesitant to vaccinate their children against Measles Mumps Rubella (MMR I& II). Machine learning predictions were better performing than status quo methods. The LASSO model was successfully implemented into a prototype of the Early Warning and Monitoring

System, which can guide healthcare policymakers to take action in dealing with vaccine-hesitant families.

Added value of this study
The study used a machine-learning technique for a better understanding of what affects vaccine hesitancy.

Implications of available evidence
The authors have developed a prototype Early Warning and Monitoring System dashboard based on the LASSO regression model, which policymakers can use to better target public messaging, start early interventions, keep under surveillance regions with a dense concentration of children at high risk of not receiving the MMR vaccine, etc.

17. Artificial intelligence for diabetes research: Development of type I/II classification algorithm and its application to surveillance using a nationwide population-based medico-administrative database in France (manuscript under revisions)

Manuscript under revision at Diabetes Research and Clinical Practice

Summary

Background
Diabetes is one of the leading causes of morbidity and mortality worldwide. Public health surveillance of diabetes is fundamental to reduce its global burden. Over the last decades, Big Data has emerged offering new opportunities for surveillance. Big Data refers to a massive volume of information collected from different sources, characterized by the three Vs: volume, velocity and variety. One example of Big Data source for public health surveillance is the French national health data system, the SNDS (*Système National de Données Santé*). In the SNDS, individual, updated and exhaustive health information from the whole French population (66 million people) is electronically collected. It includes information on claims from out-of-hospital healthcare consumption and information on hospital stays from public and private hospitals.
Currently, a validated algorithm based on antidiabetic drug reimbursement identifies people with pharmacologically treated diabetes. This algorithm has very good performances (sensitivity 97.3%, specificity 99.9% and accuracy 99.9 %) but is not able to distinguish type 1 from type 2 diabetes. Differentiating type 1 and type 2 diabetes is crucial in diabetes surveillance since the two types of diabetes present relevant differences in terms of prevention, the population at risk, the natural history of the disease, pathophysiology, management and risk of complications. Artificial Intelligence, especially Supervised Machine Learning, might overcome this limitation by developing an innovative algorithm to classify pharmacologically treated type 1 and type 2 diabetes cases. Supervised machine learning includes different methods where classification or predictive algorithms are developed by linking known features to assess targets using a data set where these targets are characterized. The algorithm is then applied to further data sources where the targets are unknown.
The objectives of this study were to develop a type1/type 2 diabetes classification algorithm using Artificial Intelligence and to estimate type 1 and type 2 diabetes prevalence in France.

Methods

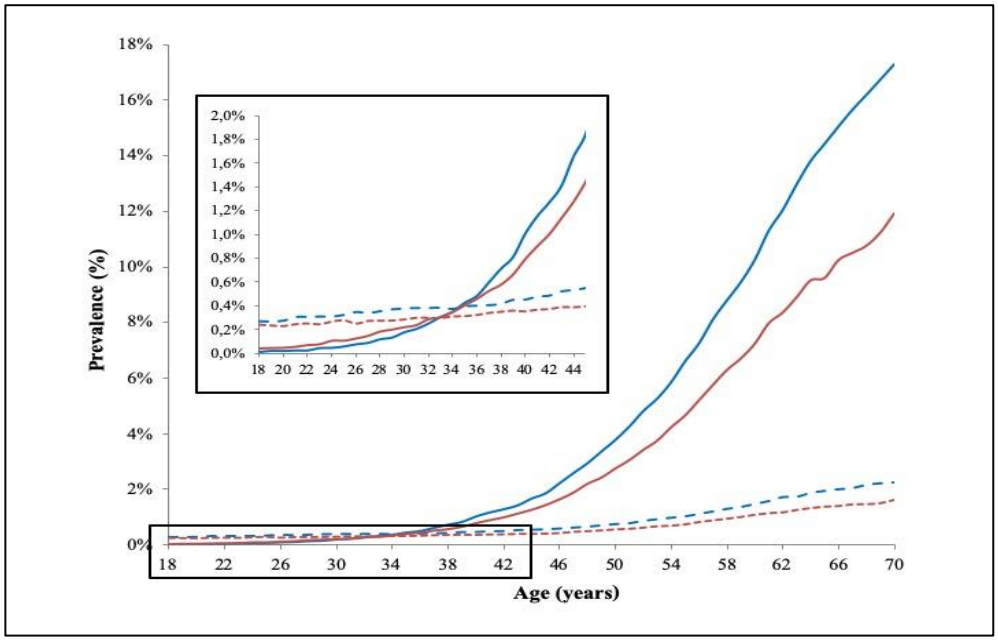
The population-based general-purpose cohort CONSTANCES was used to develop an algorithm for distinguishing type 1 from type 2 diabetes based on SNDS data. Since 2012, the CONSTANCES cohort has recruited 200,000 participants, compromising a representative sample of the French population aged between 18 to 69 years (at inclusion). The final data set compromised all diabetes cases from the Constances cohort (n=951).

A Supervised Machine Learning method based on eight steps was used: 1. selection of final data set; 2. target definition (Type 1 as positive target and type 2 as a negative target of diabetes cases); 3. coding features (A total of 3481 continuous features from SNDS data were coded); 4. splitting final data set in training/testing data set (i.e., 80% training and 20% testing data set); 5. features selection (After removing all features with a variance equal to zero, the ReliefExp score was estimated based on the relevance of each feature to differentiate between target positive and target negative group. The ReliefExp method is noise-tolerant and is not affected by feature interactions. The remained features were ranked using the ReliefExp score.); 6-8. Algorithm training, validation and selection of algorithms (The following types of models were applied to the training data set: Linear Discriminant Analysis (LDA), Logistic regression, Flexible Discriminant analysis (FDA) and C.5 decision tree (C5). For each model, the features were selected using three different thresholds of ReliefExp Score: 0.35, 0.1 and 0.05. After a first validation of the algorithms using the training data set (k-fold cross-validation), the algorithms' performances were assessed using the testing data set. The estimated performances for each algorithm were: sensibility, specificity, Kappa, F1 score and Areas Under the Receiver Operating Characteristics (AUROC curve. Finally, we retained a single model based on three criteria: its performance, its computational parsimony and its transferability to further databases.).

The selected algorithm was applied to SNDS data to estimate type 1 and type 2 diabetes prevalence among adults aged from 18 to 70 years.

Results

Among the 3481 SNDS features, only 14 were selected to train the different algorithms. The final algorithm was a Linear Discriminant Analysis model based on the number of reimbursements over the last year of fast-acting insulin, long-acting insulin and biguanides (specificity 97 % and sensitivity 100%). In 2016, the prevalence in France for type 1 diabetes was 0.3% (0.36% in men and 0.29% in women) and 4.4% (5.03% in men and 3.72% in women) for type 2. In the following figure, type 1 and type 2 diabetes prevalence are presented by a 1-year age group and by sex. Before 32-34 years, the prevalence of type 1 diabetes was higher compared to the prevalence of type 2 diabetes but after, type 2 diabetes prevalence increased exponentially with age, reaching rates of 18 % and 12 % among men and women respectively in the 70 years age group. Regarding gender, the prevalence rates of type 1 diabetes were higher among men across all age groups while for type 2, the prevalence among women was higher until the age of 32, at which there was a shift to men.



Distribution of type 1 and type 2 diabetes prevalence (%) in France among adults aged 18 to 70 years by sex and age

Dotted lines: type 1 diabetes; Solid lines: type 2 diabetes; blue lines: men; red lines: women

Study limitations

The CONSTANCE cohort includes only adults aged between 18 to 70 years, so the performance of this algorithm for other age groups might be different. The other type of diabetes such as Latent autoimmune diabetes of adulthood or Maturity Onset diabetes were not assessed in the phase of target definition. The algorithm shall be adopted over the years as the care may change over time.

Conclusions

Through Supervised Machine Learning methods, we developed a type1/type 2 classification algorithm based on the number of reimbursements of fast-acting insulin, long-acting insulin and biguanides over the last year. This algorithm has very good performances, as well as high transferability to prescription or medical claims databases from other countries.

Added value of this study

The study used a machine-learning technique to develop the algorithm classifying diabetes cases 1 and 2. This algorithm allowed to estimate for the first time the prevalence of type 1 and type 2 diabetes in France from 18 to 70 years with high precision.

Implications of available evidence

Artificial Intelligence opens new scopes in diabetes research and prevention.

VI. Discussions

We have described 17 studies as inspiring examples, which applied data linkage (13 studies), machine learning methods (2 studies) and both data linkage and machine learning approaches (2 studies). These examples cover the following domains of public health: obesity, injury, psychotic illness, disability and chronic health conditions, industrial pollution and cancer, suicidal prevention, health inequalities, cardiovascular diseases, occupational health of cancer patients, mental health, primary care, epilepsy, vaccine hesitancy and diabetes.

These studies used both health and non-health data sources according to the pre-defined research questions and objectives. Various **statistical methodologies** were applied to estimate health indicators. Some of these studies applied classical statistical approaches without artificial intelligence such as the Multilevel linear regression model, LASSO (Lease Absolute Shrinkage and Selection Operator), exploratory analysis, blinder Oaxaca decomposition and Poisson models at multilevel. Some studies applied more advanced statistical methods using artificial intelligence approaches such as Natural language processing, supervised machine learning, k-means clustering, exploratory spatial analysis and GIS (Geographical Information System).

Some **study limitations** were highlighted in general and we classified them as follows:
1. Study design (e.g., causality, misclassification of exposure-outcome, bias, limitations related to the age of study sample, use of isotropic model of exposure)
2. Data linkage (e.g., different data collection methods in different areas make it difficult to link and compare the data, lack of standard methods for data collection)
3. **Data sources** (e.g., lack of data on health inequalities at local levels, readily unavailability of data related to employment, education, occupation and socioeconomic status)
4. Data quality (e.g., lacking completeness of information for some routinely collected data sources, unavailability of certain information to improve the results of some analyses, lacking information on the secondary cause of death, exclusion of some groups for whom no linkage could be done due to lack of identifier number);
5. Data privacy (e.g., certain variables cannot be explored due to privacy or confidentiality issues, legal interoperability issues to link various data sources)

These examples highlight the added value of using innovative techniques (i.e., data linkage and/or AI), which improves completeness and comprehensiveness of information to guide the health policy process, effective patient care, health services management and effective health surveillance. Some studies applied artificial intelligence (i.e., machine learning and natural language processing techniques) to analyzing large datasets more efficiently and with high precision.
The authors of these studies highlighted ***added values/strengths of*** using data linkage and machine learning techniques, which can improve the method, data analysis and public health surveillance:
Methods: High statistical power with a large representative sample and prospective data collection allow studying long-term evolution of chronic diseases over time

Data analysis: More efficiently analysing big datasets (B. Cleland et al., 2018), High accuracy in estimation of health indicators (e.g. prevalence of diabetes I & II) (S. Fuentes et al., 2020), Highlight unobserved trends and patterns (K. E. Mason et al., 2017)

Public health surveillance: Potential risk groups to certain health outcomes (A. Zelviene et al., 2019), Variability in health inequalities from national to regional levels by the educational gradient (T. Lesnik et al., 2019), surveillance of cardiovascular diseases (L. Palmieri et al., 2019), Presence/absence of excess risk of a disease linked with residential proximity to industrial pollution (P. Fernandez-Navarro et al., 2019)

The results of these studies have important **implications in public health** improving health surveillance, prevention strategies, health care services and health policy process:

1. Public health surveillance: improved surveillance of CVD and planning prevention programmes (L. Palmieri et al., 2019); classifying diabetes cases type 1 and type 2 from a large health dataset and estimating relevant prevalence's (S. Fuentes et al., 2020); updating the list of chronic medical conditions that may impair driving skills (L. Orriols et al., 2014); evaluating the risk of cancer mortality in relation to residential proximity and industrial pollution (P. Fernandez-Navarro et al., 2019) and web-based reporting to improve public health activities at small municipalities levels (R. C. Wigand et al., 2019).

2. Prevention-based: developing **s**uicide prevention strategies (A. Zelviene et al., 2019); modifying residential environments to better facilitate healthy lifestyles to reduce overweight and obesity (K. E. Mason et al., 2017) and improving surveillance of under-coverage of MMR vaccination (A. Bell et al., 2019).

3. Health care services: evaluation of health outcomes and health services (K. Lloyd et al., 2015); planning the strategies for health care cost containment in context of chronic conditions and disability (J. Van der Heyden et al., 2015); planning of new interventions at primary care units using GIS technology (S. Mathis-Edenhofer et al., 2019); improving the delivery of health care services using open prescribing data sources (B. Cleland et al., 2018) and facilitate clinical practice to record patient information in a structured manner using natural language processing technology (B. F. Shadrach et al., 2019).

4. Health policy process: to reduce health inequalities (T. Lesnik et al., 2019); to provide evidence-based support to the professional reintegration policies and vocational rehabilitation programs for cancer patients (R. L. Kiasuwa-Mbengi et al., 2019); to develop GBS typology to translate the evidence into policy to improve mental health (A. Mizen et al., 2018)

We did not perform an exhaustive search to find more inspiring examples using data linkage and artificial intelligence. Nevertheless, this study provides an overview of various statistical methods used, study limitations, added value and implications of their results for public health. We recommend performing more research by developing new methodological approaches using data linkage and artificial intelligence. This would lead to producing sound evidence for improving public health surveillance, health care and policy development processes.

VII. Perspectives

We intend to develop the methodological guidelines using these inspiring examples and a generic case study already performed using linked data and MLT (ref). These guidelines would provide a

systematic and rational approach to estimating health indicators using linked data and machine learning techniques.

VIII. Conclusions

These inspiring examples would support countries to share different experiences and to learn from each other. Furthermore, these examples would help countries to develop, adopt and integrate innovative approaches using data linkage and artificial intelligence to estimating health indicators. These examples also allow comparing various innovative approaches used across MSs. This would improve the quality and comparability of health information across European countries. Eventually, the evidence produced by using innovative techniques would guide policymakers to make better decisions.

IX. References

1. BRIDGE-Health. https://www.bridge-health.eu/. 2014.
2. InfAct. Informaiton for Action: https://www.inf-act.eu/background. 2018.
3. Harron K, Dibben C, Boyd J, et al. Challenges in administrative data linkage for research. *Big Data Soc*. 2017;4(2):2053951717745678-2053951717745678.
4. Delnord M, Szamotulska K, Hindori-Mohangoo AD, et al. Linking databases on perinatal health: a review of the literature and current practices in Europe. *European Journal of Public Health*. 2016;26(3):422-430.
5. Machine Learning Technique: https://www.expertsystem.com/machine-learning-definition/. 2017.
6. Lloyd K, McGregor J, John A, et al. A national population-based e-cohort of people with psychosis (PsyCymru) linking prospectively ascertained phenotypically rich and genetic data to routinely collected records: Overview, recruitment and linkage. *Schizophrenia Research*. 2015;166(1):131-136.
7. Haneef R, Delnord M, Vernay M, al e. Innovative use of data sources: A Cross-sectional study of Data Linkage Practices across European Countries *Archives of Public Health (manuscript under revision)*. 2020.
8. Bradley CJ, Penberthy L, Devers KJ, Holden DJ. Health Services Research and Data Linkages: Issues, Methods, and Directions for the Future. *Health Services Research*. 2010;45(5p2):1468-1488.
9. Mason KE, Pearce N, Cummins S. Associations between fast food and physical activity environments and adiposity in mid-life: cross-sectional, observational evidence from UK Biobank. *The Lancet Public Health*. 2018;3(1):e24-e33.
10. Orriols L, Avalos-Fernandez M, Moore N, et al. Long-term chronic diseases and crash responsibility: A record linkage study. *Accident Analysis & Prevention*. 2014;71:137-143.
11. Van der Heyden J, Van Oyen H, Berger N, De Bacquer D, Van Herck K. Activity limitations predict health care expenditures in the general population in Belgium. *BMC Public Health*. 2015;15(1):267.
12. Palmieri L, Barchielli A, Cesana G, et al. The Italian register of cardiovascular diseases: attack rates and case fatality for cerebrovascular events. *Cerebrovasc Dis*. 2007;24(6):530-539.
13. Kiasuwa-Mbengi RL, Nyaga V, Otter R, de Brouwer C, Bouland C. The EMPCAN study: protocol of a population-based cohort study on the evolution of the socio-economic position of workers with cancer. *Archives of Public Health*. 2019;77(1):15.
14. Mizen A, Song J, Fry R, et al. Longitudinal access and exposure to green-blue spaces and individual-level mental health and well-being: protocol for a longitudinal, population-wide record-linked natural experiment. *BMJ Open*. 2019;9(4):e027289.

15. Cleland B, Wallace J, Bond R, et al. Insights into Antidepressant Prescribing Using Open Health Data. *Big Data Research*. 2018;12:41-48.
16. Fonferko-Shadrach B, Lacey AS, Roberts A, et al. Using natural language processing to extract structured epilepsy data from unstructured clinic letters: development and validation of the ExECT (extraction of epilepsy clinical text) system. *BMJ Open*. 2019;9(4):e023232.
17. Bell A, Rich A, Teng M, al e. Proactive advising: a machine learning driven approach to vaccine hesitancy: https://research.unl.pt/ws/portalfiles/portal/16465551/08904616.pdf. 2019.

X. Acknowledgements

We acknowledge the great support from InfAct partners who shared the case studies with us for this task of WP9:
Stefan Mathis-Edenhofer, Claudia Hable (Austrian National Public Health Institute GÖG); Herman Van Oyen (Sciensano, Belgium); Domina Vusio, Jelena Dimnjakovic, Jakov Vukovic (National Institute of public health, division of health informatics and biostatistics, Croatia); Romana Haneef, Anne Gallay, Sonsoles Fuentes, Sandrine Fosse-Edorh (Department of Non-Communicable Diseases and Injuries, Santé Publique France, Saint-Maurice); Sofian Kab (Population-Based Epidemiological Cohorts Unit, INSERM UMS 011, Villejuif, France); Emmanuel Cosson (Department of Endocrinology-Diabetology-Nutrition, AP-HP, Avicenne Hospital, Paris 13, University, Sorbonne Paris Cité, CRNH-IdF, CINFO, Bobigny, France, Sorbonne Paris Cité, UMR U1153 Inserm/U1125 Inra/Cnam/Université Paris 13, Bobigny, France); Rok Hrzic (Department of International Health, Care and Public Health Research Institute - CAPHRI, University[2] of Maastricht University, Maastricht, The Netherlands); Unim Brigid, Luigi Palmieri (Department of cardiovascular, Endocrine-metabolic Diseases and Aging,Italy); Ausra Zelviene, Rita Gaidelyte (Institute of Hygiene, Department of Health information Centre and Health Statistics, Lithuania); Tina Lesnik, Metka Zaletel, Tatjana Kofol Bric (NIJZ, Slovenia); Isabel Noguer (ISCIII, Spain); Hanna Lobosco (National Institute of public health, Sweden); Smantha Turner, Ronan A Lyons (Farr Institute of Health Informatics Research, CIPHER, College of Medicine, Swansa University, Wales).

XI. Appendices

A. Annex 1:

Name of the Member State: X-Member State
Status of studies: Previous or an ongoing study
Domains: Mortality or morbidity related health outcomes for non-communicable diseases, health interventions and their effectiveness, environment, nutrition, social behaviour/lifestyle, m-health, use of GIS (use of Geographical Information System), use of genomic data, health data governance or privacy, etc.

The following elements should be described in the selected studies:

- Context
- "Justification" to be an inspiring example
- Objective
- Methodology
    - **a.** Type of study design: Case study, report or methodological study
    - **b.** Applied settings: National or sub-national or metropolitan/territorial levels
    - **c.** Name and type of data sources used i.e., individual-level or aggregated data
    - **d.** Type of linkage (i.e., deterministic or probabilistic or both)
    - **e.** Name of advanced statistical techniques applied
    - **f.** Input variables
    - **g.** Validation and imputation of missing values (if applicable)
- Health indicators: Expected health outcome or intervention indicators or composite indicators estimated
- Results: Estimated health indicators and their interpretation
- Discussion: main results, limitations, implications in public health [Health status monitoring/health system performance] and recommendations
- Specific implications in health policy development/process (if relevant)
- Conclusion

Sciensano | Rue Juliette Wytsmanstraat 14 |
1050 Brussels | Belgium| e-mail: infact.coordination@sciensano.be|
Website: www.inf-act.eu |Twitter: @JA_InfAct