Matthieu Da Col
15/10/2024

Data sourcing

The main goal of the study is to identify possible correlations or links between health resources and causes of deaths in the EU. By resources, we aim to focus on money spent by health providers, on the number of physicians in each region and on the number of available hospital beds. Some predictive models could be made on the future trends of death causes, or on needed health resources.

Contents

Matthieu Da Col
15/10/2024

1. Chosen data sets:

Summary:

All external data sets are owned by and from the Eurostat database collection, the statistical office of the European Union whose mission is to provide high-quality statistics and data on Europe. Eurostat produces European statistics in partnership with National Statistical Institutes and other national authorities in the EU Member States, pledge for a trustworthy source.

Description of the data sets:

Downloaded file	Title	Link
estat_hlth_cd_acdr2.tsv	Causes of death - crude death rate by NUTS 2 region of residence	Link
estat_hlth_sha11_hp.tsv	Health care expenditure by provider	Link
estat_hlth_rs_bdsrg2.tsv	Available beds in hospitals by NUTS 2 region	Link
estat_hlth_rs_physreg.tsv	Physicians by NUTS2 region	Link
estat_hlth_silc_08_r.tsv	Self-reported unmet needs for medical examination by main reason declared and NUTS 2 region	Link

Presentation of each data set:

1. **estat_hlth_cd_acdr2.tsv**
   a. Data collection method: The data are derived from the medical certificate of death, which is obligatory in the UE Member States. The COD data refer to the underlying cause which - according to the World Health Organization (WHO) - is "the disease or injury which initiated the train of morbid events leading directly to death, or the circumstances of the accident or violence which produced the fatal injury".
   b. Data contents: The crude death rate describes mortality in relation to the total population. Expressed in deaths per 100 000 inhabitants, it is calculated as the number of deaths recorded in the population for a given period divided by population in the same period and then multiplied by 100 000. Crude death rates are calculated for 5-year age groups. The variables are the yearly death counts from 2011 to 2021.
   c. Data limitations: the data between 1994-2010 and starting from 2011 are not always comparable (partly due to the different groupings of causes of deaths), but it has already been taken out of the study.
   d. Data relevance: This data set is highly relevant since the causes of deaths (COD) are to be compared to health resources.

2. **estat_hlth_sha11_hp.tsv**
   a. Data collection method: the System of Health Accounts (SHA) and its related set of International Classification for the Health Accounts (ICHA) is used for the collection of the data on health care expenditure. The SHA has the goals to constitute an integrated system of comprehensive, internally consistent, and internationally comparable accounts, which should as far as possible be compatible with other aggregated economic and social statistical systems.

   b. Data contents: The data set provides the information on multiple units of count of the expanses for the detailed heath providers, on a yearly basis as the variables.
   c. Data limitations: Its main limit is the geographical scope limited by countries and not on NUTS 2 as all the other data sets.
   d. Data relevance: The most relevant feature is the unit count based on the average euro per inhabitant and for the total of all providers.

3. **estat_hlth_rs_bdsrg2.tsv**
   a. Data collection method: the System of Health Accounts (SHA) and its related set of International Classification for the Health Accounts (ICHA) is used for the collection of the data on health care expenditure. The SHA has the goals to constitute an integrated system of comprehensive, internally consistent, and internationally comparable accounts, which should as far as possible be compatible with other aggregated economic and social statistical systems.
   b. Data contents: The data set contains three different units for bed counts, the raw number of beds, the number of inhabitants per bed and finally the number of beds per 100 000 inhabitants. The variables are the years from 1993 to 2022, covering the scope of the other data sets.
   c. Data limitations: Administrative data sources refer to registered health human resources and health care facilities. The underlying totality of institutions for which data collections are available may differ. In some countries, data may not be available for a subgroup of institutions (e.g. private hospitals) or professionals (e.g. practicing nurses).
   d. Data relevance: all three units allow for a normalized and comparable approach, making the dataset very valuable is assessing the health resources per region.
4. **estat_hlth_rs_physreg.tsv**
   a. Data collection method: the System of Health Accounts (SHA) and its related set of International Classification for the Health Accounts (ICHA) is used for the collection of the data on health care expenditure. The SHA has the goals to constitute an integrated system of comprehensive, internally consistent, and internationally comparable accounts, which should as far as possible be compatible with other aggregated economic and social statistical systems.
   b. Data contents: The data set contains three different units for physicians counts, the raw number of physicians, the number of inhabitants per physician and finally the number of physicians per 100 000 inhabitants. The variables are the years from 1993 to 2022, covering the scope of the other data sets.
   c. Data limitations: some values are flagged with a “d” for marking a definition difference, and some other with a “e” for estimated values.
   d. Data relevance: Physicians as part of the human resources of healthcare services will be a good indicator.
5. **estat_hlth_silc_08_r.tsv**
   a. Data collection method: The data is from a small portion of the European Statistics of Income and Living Condition (EU-SILC)
   b. Data contents: The variables refer to the respondent's own assessment of whether he or she needed the respective type of examination or treatment, but did not have it and if so what was the main reason of not having it. All indicators are expressed as percentage of the population.
   c. Data limitations: some countries do not have survey contestant on the NUTS 2 grain, making the percentage comparison interesting only on the country level.

d. Data relevance: since the data is from a annual survey, the sampling and the subjective elements might make it less relevant for the comparison purpose.

NOTA: *All code lists are available for download on this platform: https://ec.europa.eu/eurostat/databrowser/bulk?lang=en&selectedTab=codeList. Code lists of causes of death, of geographic regions and of reasons for unmet medical needs will be merged on a further step, once the 5 main datasets have been merged together.*

2. Data Profile

a. Variables kept and merged for the final data set

Category	Variable name	Data Type	count	mean	min	max
Independent variables	geo_code	object	491			
	country_region	object	455			
	year	object	11			
	sex	object	3			
	age	object	3			
Causes of death (in number of deaths per 100k inhabitants)	A-R_V-Y: All causes of death (A00-Y89) excluding S00-T98	Float64	45 540	1 903,77	44,08	10 509,70
	A15-A19_B90: Tuberculosis	Float64	41 193	1,42	0,00	54,76
	ACC: Accidents (V01-X59, Y85, Y86)	Float64	45 534	50,79	0,00	390,25
	ACC_OTH: Other accidents (W20-W64, W75-X39, X50-X59, Y86)	Float64	45 393	17,66	0,00	185,06
	A_B: Certain infectious and parasitic diseases (A00-B99)	Float64	45 492	29,28	0,00	330,37
	A_B_OTH: Other infectious and parasitic diseases (remainder of A00-B99)	Float64	45 435	25,98	0,00	330,37
	B15-B19_B942: Viral hepatitis and sequelae of viral hepatitis	Float64	41 859	1,65	0,00	44,98
	B180-B182: Chronic viral hepatitis B and C	Float64	23 865	1,05	0,00	37,71
	B20-B24: Human immunodeficiency virus [HIV] disease	Float64	35 868	0,62	0,00	56,40
	C00-C14: Malignant neoplasm of lip, oral cavity, pharynx	Float64	45 183	8,85	0,00	97,73
	C00-D48: Neoplasms	Float64	45 540	457,33	9,63	2 098,75
	C15: Malignant neoplasm of oesophagus	Float64	45 024	9,83	0,00	143,58
	C16: Malignant neoplasm of stomach	Float64	45 375	20,26	0,00	288,63
	C18-C21: Malignant neoplasm of colon, rectosigmoid junction, rectum, anus and anal canal	Float64	45 525	52,94	0,00	358,73
	C22: Malignant neoplasm of liver and intrahepatic bile ducts	Float64	45 435	17,05	0,00	166,88
	C25: Malignant neoplasm of pancreas	Float64	45 510	29,09	0,00	204,22
	C32: Malignant neoplasm of larynx	Float64	43 587	4,15	0,00	102,03
	C33_C34: Malignant neoplasm of trachea, bronchus and lung	Float64	45 540	91,24	0,00	610,71
	C43: Malignant melanoma of skin	Float64	45 027	5,52	0,00	65,06
	C50: Malignant neoplasm of breast	Float64	43 554	28,34	0,00	204,21

	C53: Malignant neoplasm of cervix uteri	Float64	30 633	4,11	0,00	99,12
	C54_C55: Malignant neoplasm of other parts of uterus	Float64	30 801	9,08	0,00	99,68
	C56: Malignant neoplasm of ovary	Float64	30 879	13,02	0,00	137,17
	C61: Malignant neoplasm of prostate	Float64	30 933	48,41	0,00	536,40
	C64: Malignant neoplasm of kidney, except renal pelvis	Float64	45 177	10,05	0,00	124,22
	C67: Malignant neoplasm of bladder	Float64	45 330	16,09	0,00	207,72
	C70-C72: Malignant neoplasm of brain and central nervous system	Float64	45 393	10,65	0,00	76,52
	C73: Malignant neoplasm of thyroid gland	Float64	43 593	1,42	0,00	34,35
	C81-C86: Hodgkin disease and lymphomas	Float64	45 378	12,13	0,00	93,04
	C88_C90_C96: Other malignant neoplasm of lymphoid, haematopoietic and related tissue	Float64	45 288	8,65	0,00	158,38
	C91-C95: Leukaemia	Float64	45 390	14,96	0,00	100,81
	C: Malignant neoplasms (C00-C97)	Float64	45 540	441,77	9,42	2 002,94
	C_OTH: Other malignant neoplasms (remainder of C00-C97)	Float64	45 531	52,15	0,00	329,95
	D00-D48: Non-malignant neoplasms (benign and uncertain)	Float64	45 246	15,66	0,00	203,15
	D50-D89: Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism	Float64	45 171	5,44	0,00	100,17
	E10-E14: Diabetes mellitus	Float64	45 423	47,54	0,00	751,06
	E: Endocrine, nutritional and metabolic diseases (E00-E90)	Float64	45 504	60,65	0,00	948,23
	E_OTH: Other endocrine, nutritional and metabolic diseases (remainder of E00-E90)	Float64	45 285	13,26	0,00	410,05
	F01_F03: Dementia	Float64	42 795	67,83	0,00	676,65
	F10: Mental and behavioural disorders due to use of alcohol	Float64	40 971	4,51	0,00	135,31
	F: Mental and behavioural disorders (F00-F99)	Float64	45 450	71,61	0,00	681,74
	F_OTH: Other mental and behavioural disorders (remainder of F00-F99)	Float64	43 323	3,57	0,00	76,67
	G20: Parkinson disease	Float64	45 336	16,82	0,00	136,44
	G30: Alzheimer disease	Float64	45 381	39,87	0,00	901,47
	G_H: Diseases of the nervous system and the sense organs (G00-H95)	Float64	45 534	77,82	0,00	1 048,17
	G_H_OTH: Other diseases of the nervous system and the sense organs (remainder of G00-H95)	Float64	45 522	21,35	0,00	177,13
	I20-I25: Ischaemic heart diseases	Float64	45 540	254,54	0,00	2 851,90
	I20_I23-I25: Other ischaemic heart diseases	Float64	45 534	159,73	0,00	2 645,85
	I21_I22: Acute myocardial infarction including subsequent myocardial infarction	Float64	45 477	94,96	0,00	1 130,26
	I30-I51: Other heart diseases	Float64	45 540	174,02	0,00	2 688,40
	I60-I69: Cerebrovascular diseases	Float64	45 531	166,55	0,00	2 051,52
	I: Diseases of the circulatory system (I00-I99)	Float64	45 540	732,86	0,00	5 861,08
	I_OTH: Other diseases of the circulatory system (remainder of I00-I99)	Float64	45 540	137,78	0,00	1 956,77
	J09-J11: Influenza (including swine flu)	Float64	39 312	2,36	0,00	144,38

	J12-J18: Pneumonia	Float64	45 468	53,79	0,00	1 096,31
	J40-J44_J47: Other lower respiratory diseases	Float64	45 540	69,56	0,00	1 155,54
	J40-J47: Chronic lower respiratory diseases	Float64	45 540	72,60	0,00	1 185,23
	J45_J46: Asthma and status asthmaticus	Float64	43 770	3,17	0,00	78,21
	J: Diseases of the respiratory system (J00-J99)	Float64	45 540	167,16	0,00	1 623,65
	J_OTH: Other diseases of the respiratory system (remainder of J00-J99)	Float64	45 468	38,87	0,00	554,57
	K25-K28: Ulcer of stomach, duodenum and jejunum	Float64	44 739	5,32	0,00	79,05
	K70_K73_K74: Chronic liver disease	Float64	45 420	18,94	0,00	237,10
	K72-K75: Chronic liver disease (excluding alcoholic and toxic liver disease)	Float64	26 616	10,98	0,00	233,28
	K: Diseases of the digestive system (K00-K93)	Float64	45 540	70,72	0,00	538,16
	K_OTH: Other diseases of the digestive system (remainder of K00-K93)	Float64	45 495	46,66	0,00	501,80
	L: Diseases of the skin and subcutaneous tissue (L00-L99)	Float64	43 797	3,65	0,00	85,87
	M: Diseases of the musculoskeletal system and connective tissue (M00-M99)	Float64	45 156	9,15	0,00	135,70
	M_OTH: Other diseases of the musculoskeletal system and connective tissue (remainder of M00-M99)	Float64	44 982	7,03	0,00	132,55
	N00-N29: Diseases of kidney and ureter	Float64	45 468	30,17	0,00	392,49
	N: Diseases of the genitourinary system (N00-N99)	Float64	45 501	42,09	0,00	436,10
	N_OTH: Other diseases of the genitourinary system (remainder of N00-N99)	Float64	44 958	12,08	0,00	182,72
	O: Pregnancy, childbirth and the puerperium (O00-O99)	Float64	18 297	0,12	0,00	4,94
	P: Certain conditions originating in the perinatal period (P00-P96)	Float64	44 856	1,80	0,00	28,73
	Q: Congenital malformations, deformations and chromosomal abnormalities (Q00-Q99)	Float64	45 108	2,54	0,00	97,59
	R95: Sudden infant death syndrome	Float64	31 137	0,17	0,00	6,43
	R96-R99: Ill-defined and unknown causes of mortality	Float64	44 532	31,62	0,00	1 844,23
	R: Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified (R00-R99)	Float64	45 489	69,70	0,00	2 146,23
	RHEUM_ARTHRO: Rheumatoid arthritis and arthrosis (M05-M06,M15-M19)	Float64	41 802	2,32	0,00	63,31
	R_OTH: Other symptoms, signs and abnormal clinical and laboratory findings (remainder of R00-R99)	Float64	43 752	40,16	0,00	1 105,80
	TOXICO: Drug dependence, toxicomania (F11-F16, F18-F19)	Float64	29 361	0,44	0,00	12,36
	U071: COVID-19, virus identified	Float64	7 722	187,91	0,00	2 192,41
	U072: COVID-19, virus not identified	Float64	6 204	10,28	0,00	440,73
	U_COV19_OTH: COVID-19, other	Float64	2 703	1,70	0,00	197,76
	V01-Y89: External causes of morbidity and mortality (V01-Y89)	Float64	45 537	69,22	0,00	425,23
	V01-Y89_OTH: Other external causes of morbidity and mortality (remainder of V01-Y89)	Float64	40 863	2,59	0,00	74,69
	V_Y85: Transport accidents (V01-V99, Y85)	Float64	45 297	7,10	0,00	134,98
	W00-W19: Falls	Float64	45 390	21,98	0,00	260,63
	W65-W74: Accidental drowning and submersion	Float64	42 735	1,50	0,00	50,57

	X40-X49: Accidental poisoning by and exposure to noxious substances	Float64	43 716	2,93	0,00	52,37
	X60-X84_Y870: Intentional self-harm	Float64	45 342	12,61	0,00	131,61
	X85-Y09_Y871: Assault	Float64	41 415	0,91	0,00	62,47
	Y10-Y34_Y872: Event of undetermined intent	Float64	36 270	3,41	0,00	220,24
Heath resources	inhabitants_per_physician	Float64	2 857	337,10	110,40	1 407,10
	total_physicians	Float64	2 964	12 461,49	27,00	376 852,00
	physicians_per_100K_inhabitants	Float64	2 857	343,20	71,07	905,79
	euros_per_inhabitant	Float64	347	3116,83	307,69	9 482,43
	inhabitants_per_bed	Float64	2 629	250,82	79,44	2 712,41
	total_beds	Float64	2 736	18724,83	11,00	672573,00
	beds_per_100K_inhabitants	Float64	2 629	478,67	36,87	1 258,83
Unmet medical needs (in percentage from the annual survey)	Didn't know any good doctor or specialist	Float64	1 489	0,13	0,00	2,60
	Fear of doctor, hospital, examination or treatment	Float64	1 489	0,22	0,00	2,20
	No time	Float64	1 489	0,55	0,00	4,20
	No unmet needs to declare	Float64	1 489	93,38	77,90	99,90
	Other reason	Float64	1 489	1,16	0,00	17,50
	Too expensive	Float64	1 489	2,13	0,00	16,00
	Too expensive or too far to travel or waiting list	Float64	1 489	3,42	0,00	17,40
	Too far to travel	Float64	1 489	0,23	0,00	2,30
	Waiting list	Float64	1 489	1,05	0,00	15,00
	Wanted to wait and see if problem got better on its own	Float64	1 489	1,13	0,00	7,40

b. Cleaning process

All data sets have been through the same cleaning process, namely:

- striped from white spaces in the values and the headers
- reduced to only the variables from 2011 to 2021 in order to match the scope of the limited range of data on the main data set of the causes of deaths
- striped from all letters in the numerical values, even though it means losing some nuances on the totals (some marks on estimations or different counting methods are all erased)
- Duplicates have been looked for, with no duplicate to deal with
- Null values are many: no values have been used to implement them (adding zeros would reduce the means), and the unknow values arise mostly on the geographical scope differences, having data for countries but not the detailed numbers for the regions. There is no repartition key available from the data at hand. Since EUROSTAT does not implement missing values, we chose to keep the methodology of the data owner.

c. Merging process

The data sets have been merged on the key holding the year and the geographics. Note that the variables of ages (more the 65y, less than 65y and total), and of gender (Man, Female and total) have been kept only for the sake of studies or models on the causes of deaths.

d. Limitations and ethics

The data from the EUROSTAT is already anonymized, so no PII data have been delt with. The EU allows individuals to perform studies on their datasets: creating some machine learning models and visual dashboards is legal on these resources.

Some lack of detailed data, on some regions and on some variables, will ask for prudence when comparing values from countries. Also, the removed flags for estimates and different counting methods might create some weird answers on some countries and some years.

3. Questions to explore

- Do the countries with the more healthcare resources per inhabitant endure less deaths for some particular deaths causes?
- Which healthcare resource is the most important facing deaths for people younger than 65y/o?
- Are all regions equal in term of death causes?
- Are physicians equally spread across the regions and does it impact the number of deaths?
- Where are people with the most unmet medical needs and does it affect their death rate?
- Do people die from the same causes across Europe?
- How the number of available hospital beds evolved from 2011 to 2021 in Europe?
- Are expenditures correlated to the number of physicians, beds and unmet medical needs?
- Which countries in Europe are the most/least efficient in providing health care from its resources?