**University College London**

**Institute of Health Informatics**

**Data Methods for Health Research**

**GitHub link: https://github.com/UCLMScHDS2018/DMHR2018\_19\_assignment**

***Assignment A.***

***Question A.1.***

On a first instance, the information from practices located in the “City” column, corresponding to “London” was extracted (1). This provided data for only 56 practices, which gave the impression of the presence of a confounding factor in the address of the practices selected that was affecting the data filtered, since less than 60 practices seems to be too low of a number for supporting the primary care needs of over 8 million London inhabitants. Given these results, the individual data strings that are present in the "City" column were extracted to be analysed. A set of 3 strings that contain the word “London” ("London", "London colney" and "London road") were obtained. Then, the "Area" column was analysed with the same protocol to search for more practices that the previous filtering didn’t take into account that actually are based in London but their addresses were recorded in a different format. This new filtering gave the result of 876 practices, whose address format took "London" as the area in which they are registered, instead of the city. With both filters done, the data frames were concatenated to create a unique data frame that includes all the practices located in London, which allows to analyse all the 932 practices registered in the database.

The prescription data (1) gives 9,748,354 individual medicines, dressings and appliances that were prescribed by practices based in England on April, 2018. This data doesn't include private prescriptions, but only those whose cost is covered by the NHS. The "Items" column denotes the number of times an individual prescription was issued (2). The average medication was prescribed 9.12 times, and the most had been prescribed 5,147 times during April. The average cost of drugs prescribed in England was £65.97, with the most expensive costing £31,456. Although, this could be considered an outlier and more research needs to be done into it to possibly discard it. The "NIC" column indicates the net ingredient cost per 1 item, which is generally different than what the NHS actually covers since it takes different factors into its price. It does not take into account of any contract prices or discounts, dispensing costs, fees or prescription charge income, so the amount the NHS paid will be different (2).

A basic statistical description of the data from the patients registered at a practice throughout England (3) tells us that the average English practice has 8,154 registered patients, the one that has the least patients has only one, meanwhile the one that has the most has 72,227, although, this could be an outlier in the data.

The filtering of the England data with the London practice codes results in a new data frame with 748 practices. This is in contrast with the amount of codes obtained from the London "City" and "Area" data frames, which resulted in 932 practices. There is missing data for 184 practices, which could be explained by either practices that didn’t report on the month of interest, or that closed during April, or two or more that fused into only one practice. Also, dummy practices are included, created to identify prescribing in environments or circumstances outside of a general practice, such as specialist clinics, hospices, prisons, out of hours services and training units (4).

The describe function of tells us the following basic statistics about London practices:

-The average number of patients per practice in London is 7,816.

-The practice with the least patients has 8 registered.

-The maximum number of patients in a practice is 72,227. This could be an outlier from the data reported incorrectly.

-The median number tells us that half of the practices have more than 6,989 patients.

-The smallest quartile has less than 4,556 patients.

-The biggest quartile has at least 10,226 patients.

The total number of patients in London registered to a practice is 5,846,104. With the data from London practices extracted, it’s possible to filter the prescriptions data. The data reports the following basis statistics for April, 2018:

-A total of 806,139 individual item’s prescriptions were issued.

-The average item was prescribed 7.3 times.

-The item most prescribed was done 2,012 times.

-The average actual cost of the drugs prescribed in London was £53.60, with the most expensive costing £12,338. This could be an outlier.

-A sum of the items prescriptions results in 5,888,215 total prescriptions issued in London during April, 2018.

-The total cost of the prescriptions was £1,000,820,487.

In table 1 the 10 most prescribed items are reported. The most prescribed is the Influenza vaccination due to its seasonal trend. There are drugs usually prescribed for diabetes (Metformin), high blood pressure and CVD (Amlodipine, Atorvastatin), asthma (Salbutamol), and the widely used analgesic Paracetamol.

Table 1: The 10 most prescribed drugs in London on April, 2018.

|  |  |  |
| --- | --- | --- |
| **BNF code** | **BNF name** | **Items** |
| 1404000H0AAAKAK | Influenza\_Vac SplitViron Inact 0.5ml Pfs | 2,012 |
| 0103050P0AAAAAA | Omeprazole\_Cap E/C 20mg | 1,885 |
| 0601022B0AAABAB | Metformin HCl\_Tab 500mg | 1,654 |
| 0206020A0AAABAB | Amlodipine\_Tab 10mg | 1,578 |
| 0212000B0AAABAB | Atorvastatin\_Tab 20mg | 1548 |
| 0212000B0AAACAC | Atorvastatin\_Tab 40mg | 1,378 |
| 0206020A0AAAAAA | Amlodipine\_Tab 5mg | 1,313 |
| 0407010H0AAAMAM | Paracet\_Tab 500mg | 1,205 |
| 0301011R0AAAPAP | Salbutamol\_Inha 100mcg (200 D) CFF | 1,096 |
| 0601022B0AAABAB | Metformin HCl\_Tab 500mg | 1,090 |

The ten less frequent items prescribed comprise a list of 342,444 items prescribed only once in London practices. This data was filtered to exclude dressings and appliances and keep the information exclusively for drugs, which BNF code starts with two digits from 01 to 15 (5, 6). This resulted in a list containing 285,500 individual drugs (data frame “drugs\_1\_item\_lndn”). A sample of the ten most expensive drugs prescribed only once is presented in table 2. It is inferred that the high cost drugs of these drugs is one of the main reasons for being prescribed only once, being very specialized drugs, but since the data refers to only one month, these drugs could be prescribed again the following months.

Table 2: The 10 most expensive drugs prescribed only once in London during April, 2018.

|  |  |  |  |
| --- | --- | --- | --- |
| **BNF code** | **BNF name** | **Items** | **Actual Cost (£)** |
| 0208020N0AAAAAA | Phenindione\_Tab 10mg | 1 | 5,490.15 |
| 0908010C0AAAGAG | Levocarnitine\_Oral Soln Paed 1.5g/5ml | 1 | 4,105.39 |
| 0605010S0BBAJAT | Genotropin\_Inj 12mg Cart + Dil | 1 | 3,870.03 |
| 0602010M0AAAAAA | Liothyronine Sod\_Tab 20mcg | 1 | 3,660.46 |
| 0401010T0AAAEAE | Temazepam\_Oral Soln 10mg/5ml S/F | 1 | 3,142.76 |
| 0407020H0AAABAB | Dipipanone HCl/Cyclizine HCl\_Tab 10/30mg | 1 | 2,867.73 |
| 0605010S0BEAEBA | Norditropin SimpleXx\_Inj 30u 10mg/1.5ml | 1 | 2,761.6 |
| 0605010S0BEAEBA | Norditropin SimpleXx\_Inj 30u 10mg/1.5ml | 1 | 2,761.6 |
| 0403020Q0AAAAAA | Tranylcypromine Sulf\_Tab 10mg | 1 | 2,687.59 |
| 0802040AKAAABAB | Dimethyl Fumar\_Cap E/C 240mg | 1 | 2,546.63 |

***Question A.2***

For Cambridge the same data cleaning and filtering protocol was followed. For this city, a merge of "City" and "Area" data is not necessary as the one done for London, since there are big number of practices that are from Cambridgeshire but aren't in the city of Cambridge, resulting in data for 35 practices in total. From this data, it was possible to link the patients registered data to 30 practices. Reasons for this missing data are the same as noted in the previous question.

The describe function of tells us the following basic statistics about Cambridge practices:

-The average number of patients per practice in Cambridge is 10,029.

-The practice with the least patients has 568 registered.

-The maximum number of patients in a practice is 33,501. This could be an outlier from the data reported incorrectly.

-The median number tells us that half of the practices have more than 8,760 patients.

-The smallest quartile has less than 6,054 patients.

-The biggest quartile has at least 12,230 patients.

It is interesting to note that some of these numbers are higher for Cambridge than for London (i.e. average items prescribed and maximum), suggesting that there is a higher pressure by the Cambridge population on the primary care services, due to a lack of practices, and possibly healthcare professionals, relative to the city population, even though the item count for London is over ten times the one for Cambridge.

The number of all patients in Cambridge registered to a practice is 300,885.

-For Cambridge, the total number of prescriptions given by practices was 331,388.

-The average item was prescribed 8.5 times.

-The item most prescribed was done 2,087 times.

-The average actual cost of the drugs prescribed in Cambridge was £60.31, with the most expensive costing £8,099. This could be an outlier.

-For Cambridge, the total cost of the prescriptions given by practices was £65,134,534.

The ten most prescribed drugs in Cambridge are reported in table 3. There are drugs usually prescribed for high blood pressure and CVD (Amlodipine, Simvastatin), and the widely used analgesics Aspirin and Paracetamol, the Levothyrox for thyroid gland issues, and Omeprazole for digestive issues.

Table 3: The 10 most prescribed drugs in Cambridge during April, 2018.

|  |  |  |
| --- | --- | --- |
| **BNF code** | **BNF name** | **Items** |
| 0103050P0AAAAAA | Omeprazole\_Cap E/C 20mg | 2,087 |
| 0206020A0AAAAAA | Amlodipine\_Tab 5mg | 995 |
| 0212000B0AAABAB | Atorvastatin\_Tab 20mg | 955 |
| 0209000A0AAABAB | Aspirin Disper\_Tab 75mg | 817 |
| 0212000Y0AAADAD | Simvastatin\_Tab 40mg | 816 |
| 0602010V0AABZBZ | Levothyrox Sod\_Tab 100mcg | 740 |
| 0407010H0AAAMAM | Paracet\_Tab 500mg | 691 |
| 0212000B0AAABAB | Atorvastatin\_Tab 20mg | 685 |
| 0602010V0AABXBX | Levothyrox Sod\_Tab 50mcg | 678 |
| 0602010V0AABWBW | Levothyrox Sod\_Tab 25mcg | 644 |

As done for London, for Cambridge the prescriptions corresponding to drugs that were prescribed only once was extracted. A sample of the ten most expensive drugs prescribed only once is presented in table 4.

Table 4: The 10 most expensive drugs prescribed only once in Cambridge during April, 2018.

|  |  |  |  |
| --- | --- | --- | --- |
| **BNF code** | **BNF name** | **Items** | **Actual Cost (£)** |
| 0803043P0AAADAD | Lanreotide\_Inj 120mg/0.5ml Pfs | 1 | 1,737.94 |
| 0605010S0BCAGBS | Saizen\_Inj 20mg/2.5ml Cart | 1 | 1,289.83 |
| 090402000BBFHA0 | Elemental 028 Ex\_Pdr Sach 100g (3 Flav) | 1 | 1,111.4 |
| 0605010S0BBANAX | Genotropin MiniQuick\_Inj 800mcg Pfs Cart | 1 | 1,083.63 |
| 0401010ACBBAAAA | Xyrem\_Oral Soln 500mg/ml | 1 | 1,082.59 |
| 0605010S0BCAFBR | Saizen\_Inj 12mg/1.5ml Cart | 1 | 1,031.87 |
| 090402000BBIXA0 | Modulen IBD\_Pdr | 1 | 1,012.07 |
| 0605010S0BDAEAT | Humatrope\_Inj 36u Cart + Dil (12mg) | 1 | 1,001.61 |
| 090401000BBHKA0 | Phlexy-10\_Drink Mix Pdr Sach 20g(3 Flav) | 1 | 999.37 |
| 0604011D0AAALAL | Ethinylestr\_Tab 2mcg | 1 | 984.99 |

***Question A.3***

1. Across all practices, prescriptions for 1,330,453 different drugs belonging to Chapter 2 of the BNF were issued during April, 2018, amounting to a total of 26,449,832 prescriptions given by practices throughout England, and the total cost amounted to £5,448,656,348.
2. For antidepressant drugs, the total amount of the prescriptions given by practices was 5,715,873 and the total cost was of £925,174,735.

The amount of prescriptions as well as the spending on CVD medication by the NHS is over five times the amounts for antidepressant drugs. One of the reasons for this discrepancy is that the Chapter 2 of BNF covers a broader group of medications, easily surpassing the group for antidepressant drugs.

***Question A.4***

The total cost of the prescriptions given by all England practices during the month of April was £20,836,681,997.

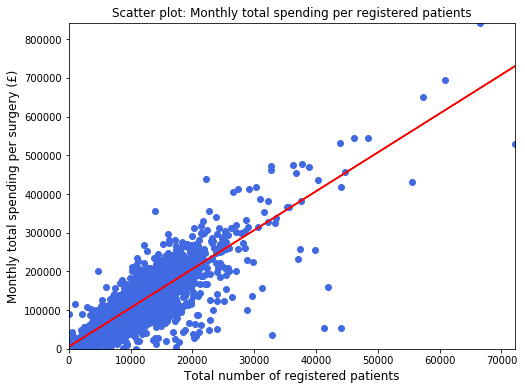


Figure 1: Scatter plot with monthly total spending per registered patients at an individual practice.

There appears to be points denoting practices that although register a number of patients over 30,000, report a monthly total spending that deviates a substantial amount from the trend line. These practices could be “dummy” practices or the result of a merger or splitting of practices. As seen in figure 2, the majority of practices issue prescriptions in the range of £10 to £12. Also, there are practices that are well above the average amount of patients registered. These practices could also be from mergers, or data reported incorrectly.

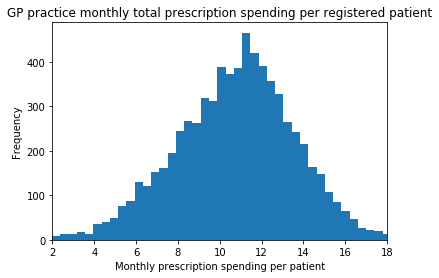


Figure 2: Distribution of monthly total prescription per registered patient at an individual practice. Note: a series of code snippets are included in the IPython Notebook in the attempt to fit the Gaussian curve.

***Assignment B***

***Question B.1***

Iceland's population in the year 2010 was 318,041, and its total deaths during the same year were 2,019.

New Zealand's population in the year 2010 was 4,367,360, and its total deaths during the same year were 28,649.

Italy's population in the year 2010 was 60,483,386, and its total deaths during the same year were 584,615.

***Question B.2***

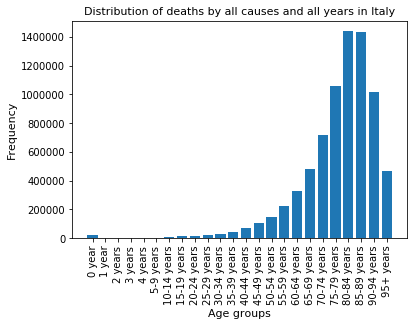
******

Figure 3: Distribution by age groups of all causes of death in Italy.

It is possible to observe a certain amount of child mortality, which is highest for new-borns at 0 years of age, probably because of complications during pregnancy, at birth or from congenital diseases. After that period, child mortality decreases and starts to grow again after 10 years of age. The group with the higher mortality is the cohort between 80 and 84 years old.

***Question B.3***

Lung cancer has been the major cause of mortality by neoplasms during the period covered by the data (table 5 and figure 4). The category by itself has more deaths than the next two groups combined. Being smoking the main cause of lung cancer (12) governmental policies against tobacco commercialisation and in favour of informing the general population remain as the most critical steps against the disease.

Table 5: Top five causes of death in Italy across all years. Codes obtained from ICD-10 Version 2016 (7).

|  |  |  |  |
| --- | --- | --- | --- |
| **ICD-10 Code** | **Cause** | **Deaths** | **Percentage of Overall Deaths** |
| C34 | Malignant neoplasm of bronchus and lung | 430,069 | 5.63 |
| C18 | Malignant neoplasm of colon | 182,802 | 2.39 |
| C50 | Malignant neoplasm of breast | 156,002 | 2.04 |
| C16 | Malignant neoplasm of stomach | 132,676 | 1.74 |
| C25 | Malignant neoplasm of pancreas | 132,125 | 1.73 |

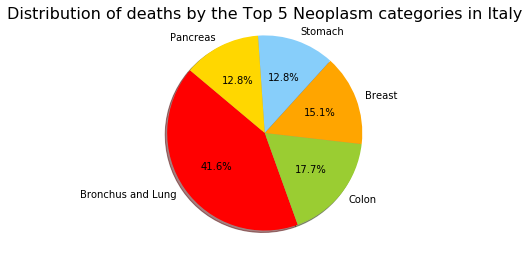
****

Figure 4: Top Five causes of death across all years and age groups. Percentages given as from between these five catgories.

***Question B.4***

There is a difference between age groups for mortality by neoplasms in Australia during 2010. The top five age groups that had the highest mortality during said year are presented in table 6.

Table 6: Top 5 age groups whose mortality was a result of neoplasms in Australia, 2010.

|  |  |
| --- | --- |
| **Age group** | **Deaths** |
| 80-84 years | 7,167 |
| 75-79 years | 6,291 |
| 70-74 years | 5,713 |
| 85-89 years | 5,520 |
| 65-69 years | 4,768 |

Given the age of the groups with the highest mortality in the population, this suggests that age is one of the biggest factors for appearance of neoplasms.

***Question B.5***

In the year 2010, 289 people out of every 100,000 people died from neoplasms in Italy. During the same period, 194 people out of every 100,000 people died of neoplasms in Australia. Italy has a frequency of neoplasm deaths, in the whole population, of almost 50% more than those from Australia.

The top five causes of death in Australia and Italy in the Neoplasm category during the year 2010 are detailed in table 7 and figure 5 (7). In contrast with the previous comparison against the total population, the main causes of death have very similar frequencies when are compared against the overall deaths during the same period. This suggests that the factors, and the susceptibility to these factors, are similar between both countries.

Table 7: Top 5 causes of death from the neoplasm category in Australia and Italy, 2010.

|  |  |  |
| --- | --- | --- |
| **Country** | **Cause** | **Frequency every 100,000 deaths** |
| **Australia** | C34 - Lung | 5,644.3 |
| C61 - Prostate | 2,255.5 |
| C50 - Breast | 1,997.6 |
| C80 - No specification of site | 1,939.7 |
| C25 - Pancreas | 1,696.5 |
| **Italy** | C34 - Lung | 5,763.8 |
| C18 - Colon | 2,488.3 |
| C50 - Breast | 2,093.3 |
| C25 - Pancreas | 1,798.1 |
| C16 - Stomach | 1,723.4 |

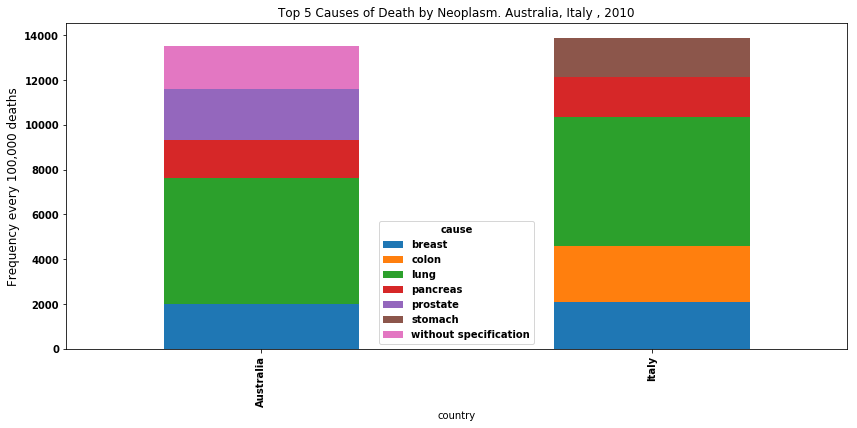


Figure 5: Stacked bar plot of the top five neoplasm related deaths during 2010 in Australia and Italy.

It is of interest to note that for both countries, lung and bronchus cancer is one the highest causes of death between neoplasms, corroborating what has been reported in previous studies on a worldwide level (8, 9). Also, three of the top causes for Italy, pancreas, colon and stomach, are commonly related to lifestyle habits, like diet and alcohol consumption. Breast cancer is predominant in both countries, in contrast with neoplasms of prostate, which is only one of the top causes in Australia.

Another point for interest for the data from Australia is the fact that carcinomas without specification site came between one of the top causes of neoplasms deaths. These are neoplasms where the site of origin which origin is unknown and the disease has already disseminated and nowadays remains one principal causes of neoplasms deaths in the country (7, 10, 11).

**References:**

1. Practice Level Prescribing - presentation level data - April 2018. <https://digital.nhs.uk/data-and-information/publications/statistical/practice-level-prescribing-data/april-2018>. Accessed 29/12/2018.
2. Practice level prescribing - glossary of terms. <https://digital.nhs.uk/data-and-information/areas-of-interest/prescribing/practice-level-prescribing-in-england-a-summary/practice-level-prescribing-glossary-of-terms>. Accessed 16/01/2019.
3. Patients Registered at a GP Practice, April 2018. <https://digital.nhs.uk/data-and-information/publications/statistical/patients-registered-at-a-gp-practice/patients-registered-at-a-gp-practice-april-2018-special-topic---registered-patients-compared-to-the-projected-resident-population-in-england>. Accessed 29/12/2018.
4. Practice level prescribing data: more information. <https://digital.nhs.uk/data-and-information/areas-of-interest/prescribing/practice-level-prescribing-in-england-a-summary/practice-level-prescribing-data-more-information>. Accessed 19/01/2019.
5. Prescribing Data: BNF Codes. EBM Datalab, University of Oxford. <https://ebmdatalab.net/prescribing-data-bnf-codes/>. Accessed 19/01/2019
6. Open Prescribing. <https://openprescribing.net/bnf/>. Accessed 19/01/2019
7. ICD-10 Version: 2016. <https://icd.who.int/browse10/2016/en>. Accessed 16/01/2019.
8. Didkowska J, Wojciechowska U, Mańczuk M, Łobaszewski J. Lung cancer epidemiology: contemporary and future challenges worldwide. Ann Transl Med 2016;4(8):150. doi: 10.21037/atm.2016.03.11.
9. Torre, L. A., Bray, F. , Siegel, R. L., Ferlay, J. , Lortet‐Tieulent, J. and Jemal, A. (2015), Global cancer statistics, 2012. CA: A Cancer Journal for Clinicians, 65: 87-108. doi:10.3322/caac.21262.
10. 2019 ICD-10-CM Diagnosis Code C80.0. <https://www.icd10data.com/ICD10CM/Codes/C00-D49/C76-C80/C80-/C80.0>. Accessed 20/01/19
11. Cancer of unknown primary statistics, Australian Government. [https://unknown-primary-cancer.canceraustralia.gov.au/statistics. Accessed 20/01/19](https://unknown-primary-cancer.canceraustralia.gov.au/statistics.%20Accessed%2020/01/19)
12. Causes, lung cancer, NHS. <https://www.nhs.uk/conditions/lung-cancer/causes/>. Accessed 20/01/19