Lucerne University of
Applied Sciences and Arts
HOCHSCHULE
LUZERN

Data Science in Health Care

UK Drug Prescriptions (Final Project)

Philip Hochuli

Lucerne University of Applied Sciences and Arts
Master of Science (MsC) Applied Information and Data Science

Spring 2019

Final Project - What You Need to Know I

1. Your Final Project: What You Need to Know

Just Like in Our In-Class Project, You Will Work With Real Data from Health Care

- You will apply everything you have learned so far: Cleansing, restructuring and analyzing real-world data
- Goal: Ask questions to the data and gain insights
- Again, you may choose yourself what you want to investigate (I give you some ideas at the end)
- Choose a tool/ programming language you prefer (I discourage the use of Excel)

Hand-in Requirement: You Are Expected to Create a Presentation of at Most 15 Slides I

- 5 - 10 Min. time for presentation + 5 minutes to answer questions
- Also, you must hand in your code used for the project
- Optional: You may use an *appendix* in your presentation to hand-in additional material
- You are expected to hand in your presentation and your code by 23.59 o'clock **two days before the exam** at latest
  - For example, if your exam is scheduled for 4 July 2019 you have to hand in by 2 July 2019 at 23.59 o'clock (at latest)
  - If you don't hand in on time, you will receive grade 1
- Make sure your work is reproducible (I should be able to run/ test your code) and interpretable (use comments if necessary)

Your Slides Should at Least Cover the Following Content

- An explanation of how you proceeded and what you wanted to analyze (probably more than one question, depending on complexity)
- Main findings
- Major problems encountered
- How you evaluated whether your analysis is satisfying or not (this holds particular true for all model-based approach and visualizations)
- A realistic self-assessment of what you did well and where you see drawbacks in your work
- How you could improve your analysis
- Key takeaways and personal learnings

Your Grading Will Be Based on the Grading Matrix Published on Ilias

- You will be graded as a group
- There will be a second examiner present that day
- I won't answer any questions after the end of our regular lectures, unless of administrative nature

Any questions ?

2. Introduction to the Project: GP Drug Prescriptions in England

The National Health Service (NHS) Publishes Monthly Data on All Drugs, Dressings and Appliances Prescribed in England and Dispensed in the UK I

- Data from **all** General Practitioners (GPs)[1]
- Includes also prescription from specialist clinics, hospices, prisons, out of hours services and training units
- Only around 0.2% of England's prescriptions not included
- No information on patients
- No data on dispensing authorities (such as pharmacies)
- $>$ 10 Mio. rows of data (monthly)
- Published monthly, 7 - 8 weeks after the end of the month (e.g. mid-October for August data)
- Raw data at the unit of observation {GP, BNF-code}

[1] In a few words, GPs are quite similar to Swiss family doctors. In particular, they take a holistic approach to medicine. However, they have an even higher importance than Swiss family doctors and more responsibilities. UK citizens typically register themselves at a GP.

The Date Includes Information on Drugs, Dispensed Quantity and Costs

- Total number of items prescribed
- Net ingredient cost (NIC)
- Actual cost (to NHS)
- Total quantity (in units of the drug)
- SHA and PCT (geographic identifiers)
- BNF-code

Extract from the Data (July 2018)

Table 1: Sample from the NHS prescription data, 07/2018

PRACTICE	SHA	PCT	BNF_CODE	BNF_NAME	ITEMS	NIC
H83024	Q63	07V	0407010H0AABUBU	Paracet_Suppos 500mg	3	255.50
L84037	Q64	11M	0407020K0AAABAB	Diamorph HCl_Inj 10mg Amp	2	33.12
B86066	Q52	15F	0407010F0AAADAD	Co-Codamol_Cap 30mg/500mg	15	77.51
L81078	Q65	15C	0407020Q0AAEFEF	Morph Sulf_Cap 10mg M/R	2	6.94
B83661	Q52	02W	1001010P0AAADAD	Naproxen_Tab 250mg	5	8.80
B82628	Q50	03M	0106040M0BCACAA	Laxido_Oral Pdr Sach (Orange) S/F	39	209.27

Before We Move On: Understanding Pharmaceutical Products is Not Trivial. You Should Know Some Basics.

- **Active constituent/ ingredient** (in German: "Wirkstoff"): The chemical substance that makes a drug "work".
- **Drug**: A drug is a product. It consists of active constituents (usually one, not always) and "other substances" ("Hilfsstoffe""").
  - Example: Diclofenac is the active constituent used in Voltaren (PharmaWiki 2019)
- **Generics**: Use the same active constituent as original drugs in the same dosage and form (interpharma 2019). However, generics may differ from original products in terms of the "other substances" (Steiner 2006).
- **Presentation**: Combination of drug, dosage/strength and form (NHS-definition)

The NHS Uses BNF-codes for Classification of Prescriptions at the Presentation Level (Drug Name, Form and Strength or Size) I

- BNF (British National Formulary): Reference book with all drugs allowed for prescription in the UK
- 15 character code, e.g. *040702040AAAMAM* (Propranolol HCl_Tab 10mg)
- The first seven characters identify the chapter, section and paragraph of the BNF. This provides you **a lot** of information about what the drug is used for (i.e. anatomy, intention of drug and/or categorization).
  - Example: '0407020' = Central nervous system, antidepressant drugs, opioid analgesics
- The 8th and 9th character together identify the **chemical substance** (i.e. the active constituent(s))

The NHS Uses BNF-codes for Classification of Prescriptions at the Presentation Level (Drug Name, Form and Strength or Size) II

- The next four letters (up to the 13th) identify the product, strength and formulation of the prescription
- Finally, the last two letters are reserved to indicate whether the drug is *generic* or *branded*
  - For generics, the 14th and 15th letter repeat the 12th and 13th respectively

Note: There do exist some differences between the original BNF-codes and those employed by the NHS (check the readings). You will need to use the NHS-definition. This holds true since NHS added pseudo-chapters (20-23 plus 18 and 19), primarily to use dressing and appliances

Resources (Mandatory Readings in Bold, Recommended Readings in Italic)

- **Original source of data at NHS and essential information**: Link
- *More information on the data*: Link
- **Glossary of terms**: Link
- Database of original BNF-codes: Link
- Booklet on NHS's BNF-codes: Link
- *Explanation of BNF-codes as used by the NHS in general*: Link

3. Possible Questions to Investigate

Challenges and Recommendations

- Use only a single month of data (for comparability, please use July 2018)
  - Click here to access this data: Link
  - Download the ZIP-File which already contains the three data sets you will need
- Drop all observation which are not given a proper BNF-code (not 15 characters long)
- Don't evaluate actual cost, use NIC
- Don't evaluate quantity, look at number of prescribed items (ITEMS)
- You will process a lot of data. Make sure your computer can handle it.
- Some questions you want to investigate might be computationally too expensive. You might need to try different tools and libraries.

Basic Questions to Investigate I

- Find the number of presentations, chemical substances and, if possible, the number of drugs (you will need to combine data files and extract information from BNF-codes).
- Which five chemical substances have the largest number of drugs associated with them? Find these drugs! Why might there exist differences?
- Find the average number of drugs per substance. What about using a confidence interval on this average?
- Using a *word-cloud*, find the most common prescriptions for an individual GPs, a region or a medical chapter of the BNF. How would you compute this? Interpret what you see.
- Find the 100 most often prescribed substances (based on ITEMS). Looking at the top 10: Did you expect this result? What might this tell you about England's population? What might be possible problems in this measure?

Basic Questions to Investigate II

- Plot the distribution over the number of prescriptions (ITEMS) for GP practices. What might this tell you ?
- Compute and plot the number of prescribed drugs (not the sum of ITEMS) per GP. What might this indicator tell you ?
- Compute the number of prescribed substances or drugs per region. Illustrate your findings visually. Can you identify clusters? Would the use of an algorithm to identify clusters make sense in this case?

Clustering Prescription Patterns (Advanced)

- Aggregate the data on SHA and compute the total number of prescribed items as well as the total NIC. Then, compute NIC/ITEM. Visualize your findings using an appropriate chart
  - Do you recognize clusters?
  - Use a cluster algorithm such as k-means to cluster your data. How many clusters did you find? Can you give them an interpretation? What do you learn from the shape of your clusters?
- In case you need a **real challenge**: Find those GPs with the most unusual patterns of drug prescriptions (in the cross-section, for a given month only)

Find Monthly Pairwise Correlations between Prescriptions (Chemical Substances, Drugs) Across Physicians (Advanced)

- What data structure did you use to answer this questions? I.e. how did you compute these correlations? Why?
- Use a correlation plot to visualize correlations between the 40 most often prescribed chemicals. What do you learn from this graph?
- Should you split your data into a train and test set before you do this? Why, why not?
- What if you repeat your computations for drugs, do you learn more ?

Watch out: These computation are very memory-intensive, think carefully how you approach this question.[2] I restricted my analysis to chemicals with at least 100 prescriptions (ITEMS) in a given month.

[2] There exist $>$ 11'000 drugs and $>$ 1'500 chemicals. This yields $>$ 60 Mio. unique drug combinations ($\frac{11000!}{10998!2!}$) and $>$ 1 Mio. chemical combinations

4. Showcase: Some Examples of Outcomes

Output Example 1: Word-cloud of Most Prescribed Chemicals

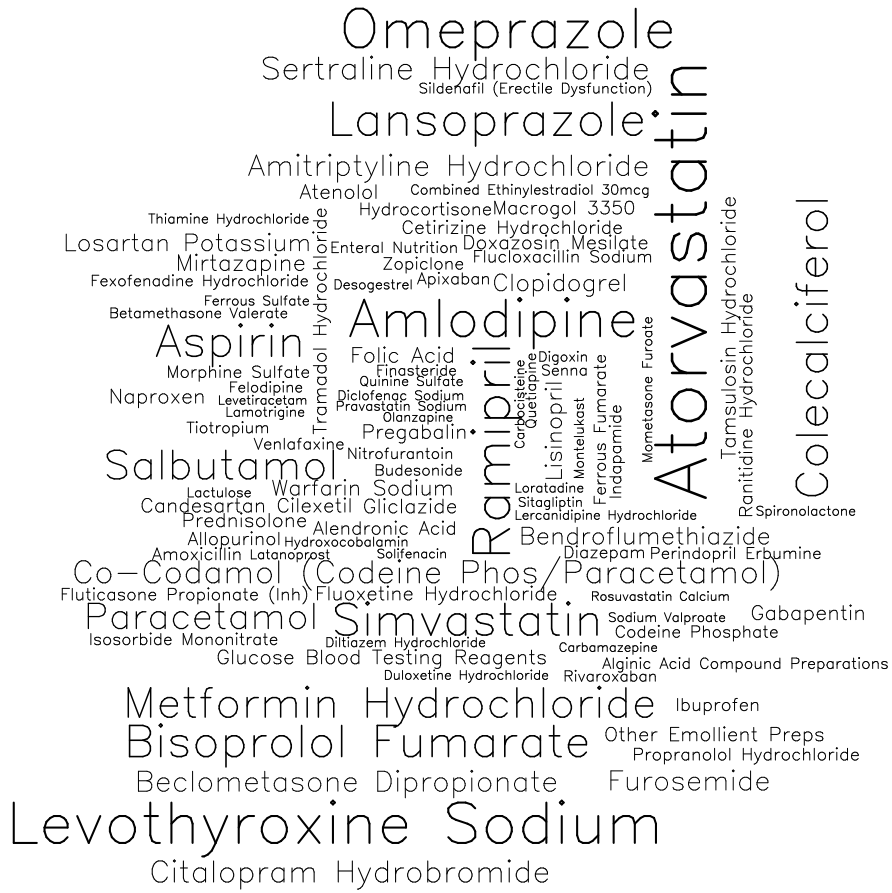

Figure 1: Most prescribed chemicals in July 2018. Source: Own computations based on NHS prescription data.

Output Example 2: Word-cloud by GP

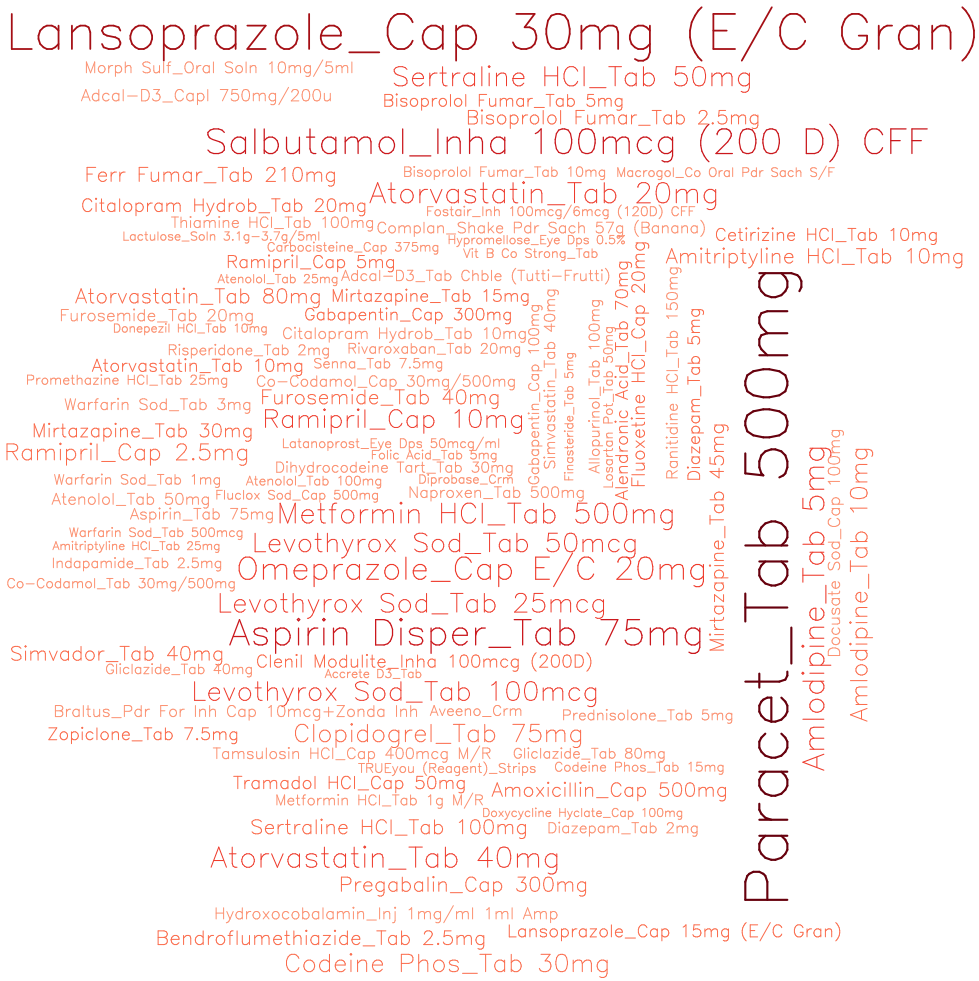

Figure 2: Most prescribed drugs for GP A81001. Source: Own computations based on NHS prescription data.

Output Example 3: Clustering GPs

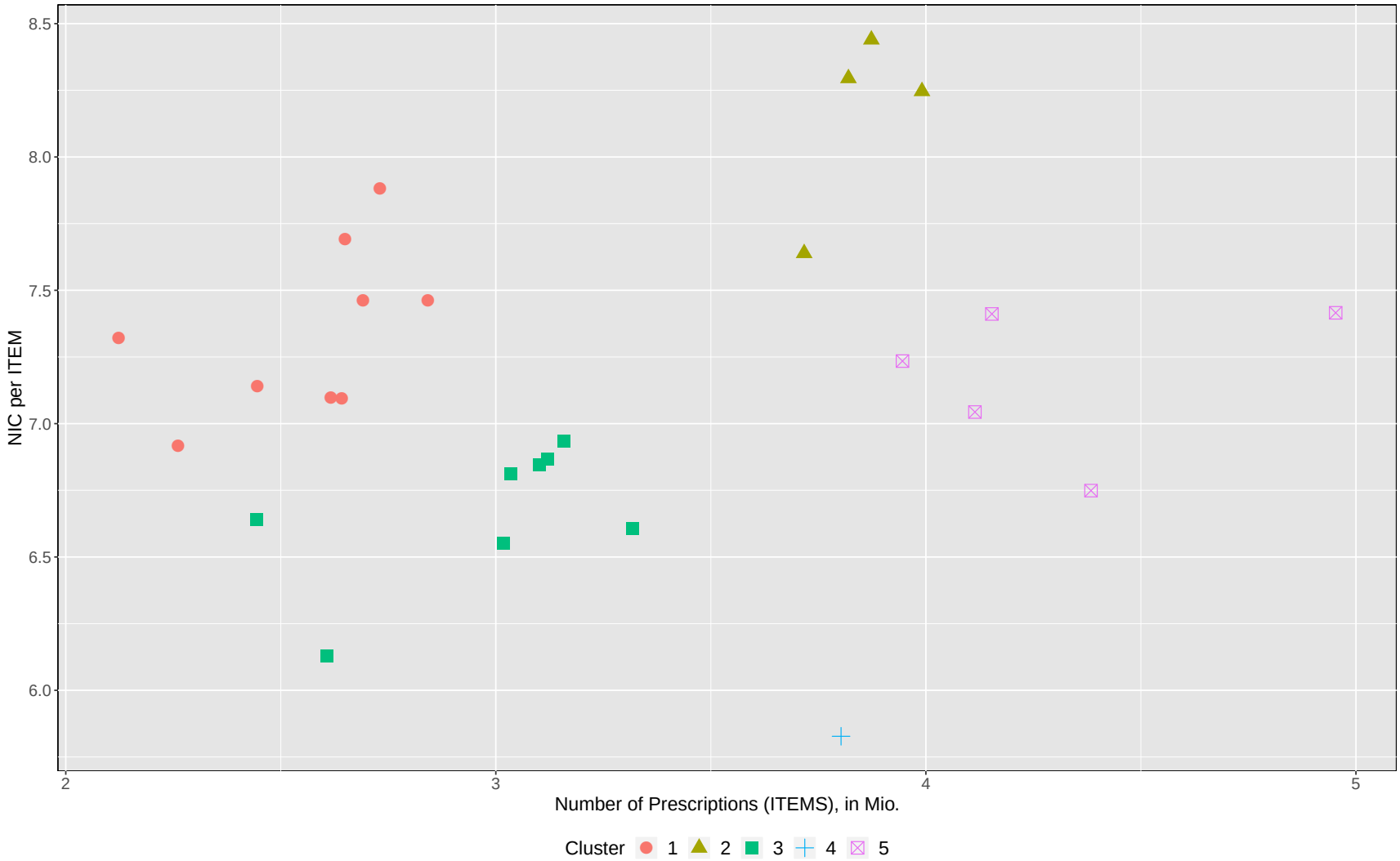

Figure 3: K-Means clustering applied on GP-prescription data. Source: Own computations based on NHS prescription data.

References

interpharma. 2019. “Generika | Interpharma.” Accessed March 23. https://www.interpharma.ch/medikamente/1597-generika.

PharmaWiki. 2019. “PharmaWiki - Diclofenac.” Accessed March 23. https://www.pharmawiki.ch/wiki/index.php?wiki=Diclofenac.

Steiner, Claudia. 2006. “Gesundheitswesen - Generika sind gleich und doch anders.” *Schweizer Radio und Fernsehen (SRF)*. March 6. https://www.srf.ch/sendungen/puls/gesundheitswesen/generika-sind-gleich-und-doch-anders.