Identifying patient group characteristics using hospital diagnosis records

Niels Dalum Hansen[⋆†], **Kåre Mølbak**[‡], **Christina Lioma**[⋆], **Ingemar J. Cox**[⋆]

[⋆] University of Copenhagen, Department of Computer Science
[†] IBM Denmark
[‡] Statens Serum Institut

E-mail Niels Dalum Hansen: nhansen@di.ku.dk

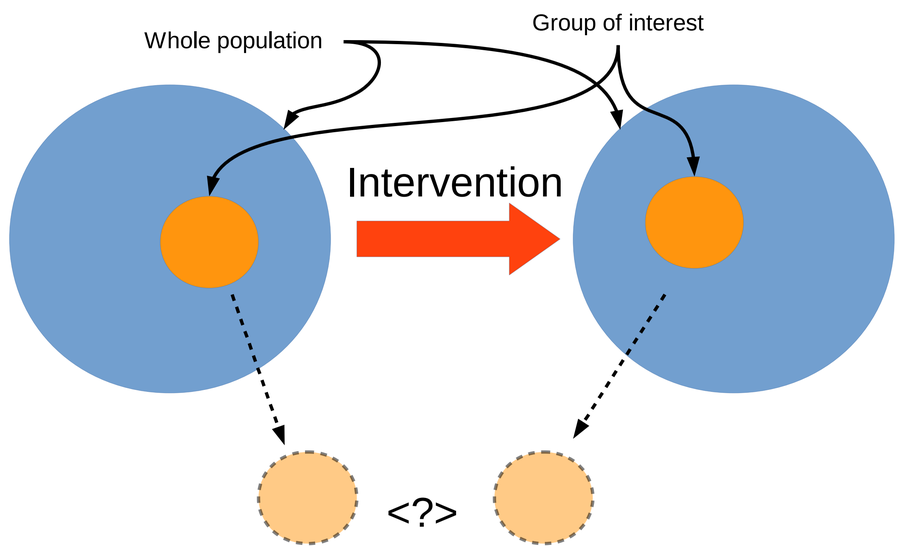
UNIVERSITY OF COPENHAGEN

Our problem

We would like to be able to identify ill-defined patient groups, specifically persons reporting to suffer from side-effects of the HPV vaccine. These are typically patients where there is disagreement on the correct diagnosis, but the patient group do share common characteristics.

Why is this interesting?

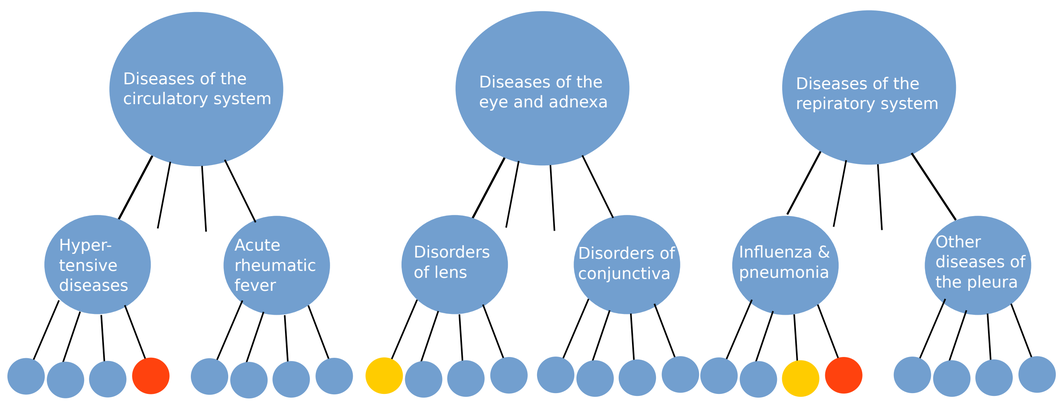
When studying the effects of health interventions on a large population a traditional epidemiological approach is to look at the changes in public health before and after the intervention. If the size of a patient group has changed, we know that the intervention might affect that patient group.



In the case of the persons reporting side-effects from the HPV vaccine, we would like to know if the persons represent an already existing group of patients. In addition we want to know whether or not this group has increased in size after the introduction of the vaccine.

Available data

The data source is The National Danish Patient Register (LPR). All hospital visits are registered in this database with a personal identifier (CPR-number) and a code (ICD-10 code) identifying the specific diagnosis. The codes are ordered in a tree structure where at the top level there is a set of chapters and where each subsequent level contains more detailed diagnosis. A subset of the tree is illustrated below:



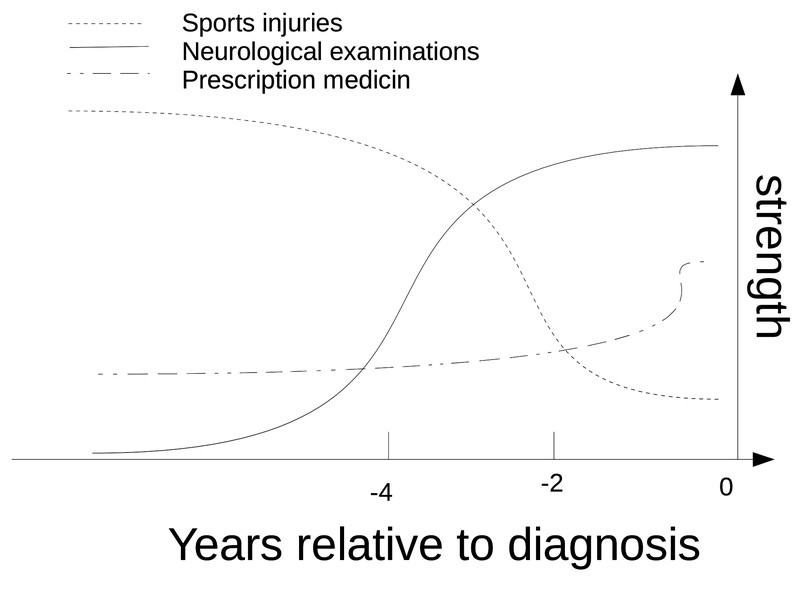
For well-defined diagnosis it is possible to select a set of ICD-10 codes and count the prevalence before and after the intervention. But for the ill-defined patient groups the doctors can assign an unknown set of diagnosis and likely also inconsistently. Two solutions might be possible.

Changing representation of the IDC-10 codes

For diseases that are not well-defined, we risk that the patient ends up with a diagnose in a "garbage" category. These categories contain diseases with inconsistent and vague symptoms. These categories exist within all of the medical specialties and are therefore far apart in the disease category tree. This does not necessarily mean that they are very different with respect to the symptoms. The red and yellow nodes illustrate two set of diagnosis that are similar though far apart. In such a case we would suspect that the textual description of the diagnosis similar. Moving from using the numerical and hierarchical structure of the ICD-10 codes to the textual description of the ICD-10 code might make it possible to get a better classification.

Disease progression

For girls with suspected side-effects of the HPV vaccine, there has been observed some common pre-diagnose characteristics. For example going from being very active in sports to suddenly having problems with dizziness and headaches. It might therefore be the case that the patient group share the same disease progression. The figure below illustrates a made up disease progression for girls with suspected HPV vaccine side-effects.



Temporal text mining

These two approaches can be combined within the context of temporal text mining [1]. First a set of topics, for example "sports injuries" or "neurological examinations" needs to be found. We model the disease descriptions as begin generated from a collection of word distributions, each distribution represents a topic, for example sports injuries. The generation of each diagnose description d follows the mixture model below:

$$p(w|d) = \lambda p(w|\Theta_B) + (1-\lambda)\sum_{j=1}^{k}(\pi_{d,j}p(w|\Theta_j)) \quad (1)$$

where w is a word, Θ_B is a background distribution of words, $\pi_{d,j}$ is the mixing coefficient for topic j for diagnosis d and Θ_j is the word distribution for topic j. The background distribution, Θ_B, is estimated from the collection of diagnosis representative of the general population. The final set of unknown variables $\{\Theta_j, \pi_{d,j} | d \in C, 1 \leq j \leq k\}$ needs to be inferred. λ is a hyper-parameter.

With a set words distributions and a collection of patients known to belong to a patient group, we can calculate the temporal strength of each topic. To do this all the time stamps of the diagnosis has to be normalized. In our case for example relative to when the patient reported a suspected HPV vaccine side-effect. Using the diagnosis descriptions with normalized time stamps the strength of each topic can be calculated as the likelihood that the diagnose descriptions were produced by the respective word distribution.

Classifying temporal text

Given the topics and the temporal topic strength we classifying new persons as belonging or not belonging to the patient group. Given a new collection of diagnosis, C_{new}, we can calculate a value, C_{score}, representative of how likely it is that the collection is generated from a specific set of topics. In our case C_{new} would correspond to the textual diagnose descriptions from a new patient. The function $words(C_{new}, t)$ extracts the words from diagnoses in C_{new} with time stamp t.

$$C_{score} = \sum_{t}^{T} \sum_{w \in words(C_{new},t)} \lambda p(w|\Theta_B) + (1-\lambda)\sum_{j=1}^{k}(\pi_j^t p(w|\Theta_j)) \quad (2)$$

Where T denotes the time steps and π_j^t the temporal strength of topic j.

Classification can be performed either directly using the C_{score} by applying a threshold, or the value can be used as input for other classifiers.

References

[1] Mei, Qiaozhu and Zhai, ChengXiang, *Discovering evolutionary theme patterns from text: an exploration of temporal text mining*, Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining, 2005