Outbreak detection based on a tree-structured anatomic model for infection

Klaske van Vuurden[a], Carl-Fredrik Bassøe[b], Gunnar Hartvigsen[a]

[a] *Department of Computer Science, University of Tromsø, Tromsø, Norway*
[b] *Røde Kors Sykehjem, Bergen, Norway*

Abstract

When designing an outbreak detection system, it may be preferable to use existing medical data as input instead of requesting additional date from medical professionals. In this paper we propose using existing symptom data and reported immunological reactions in EPRs in combination with a model based on the anatomy of disease. We argue that these data for all patients in a geographical area are sufficient to indicate the increase in incidence of infectious diseases. We transform lexical patient data in a seven-step algorithm to a two-dimensional space representing the medical anomalies in a geographical area.

Keywords*: Epidemiological research, population surveillance, algorithms, syndromic surveillance, symptom model, patient clusters, immunological reactions*

Introduction

An infectious disease can potentially cause an increase in incidence over a period of time in a geographical area. To determine whether such an increase is indeed the result of an infectious disease, we need to identify which patients are afflicted by any anomalous infection, and what these patients have in common. To search for a common cause we need to examine the essence of the infection; i.e., the type of immunological reactions (IRs) caused by it, the body parts affected and the microorganism causing it (determined later by laboratory research). These basic elements of an infectious disease are represented in Figure 1 as introduced by Bassøe [1].

The elements that can be detected by the patient in an early stage of disease are affected body parts and immunological reactions. In recent years electronic patient records (EPRs) have be mined for symptoms, signs and laboratory results [2, 3]. In an EPR, we will find symptoms and immunological reactions reported by the patient or found by their medical contact person. From the symptoms we can deduce which body parts are likely involved. By collecting this information for all patients living in an area, we can calculate which patients display related symptoms and are potentially afflicted with the same infection.

In this paper we will demonstrate seven steps to transform symptom data into evidence for possible disease outbreaks, with the model that closely mimics the effect of disease on human anatomy [4].

Materials and Methods

From lexical data representing symptoms and IRs, extracted from EPR we build a map containing clusters in seven steps. Each step has an algorithm associated with it. Note that the correctness of these algorithms is not taken into account here, as it exceeds the scope of this paper. The steps are presented below.

From symptoms to EPR (1)

Symptoms will be added into a patient's EPR based on that patient's statement of symptoms to a medical professional. Alternatively, a social networking site such as Facebook could be used as a source for symptoms.

For each symptom a timestamp will be added to the symptom to distinguish time of onset. This will give us the possibility of discarding symptoms that are likely not topical anymore for any current infectious disease outbreak.

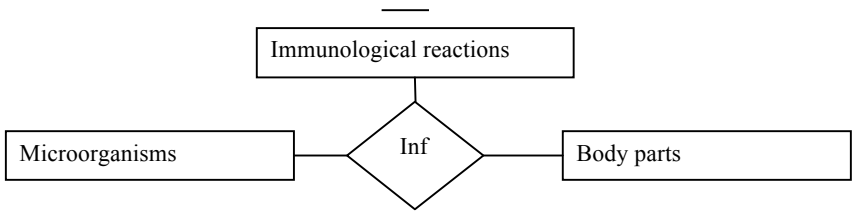

Figure 1 – General model of infections.

The EPR in use will be a problem-oriented like PROMED, which is described elsewhere [5]. This facilitates data extractions by eliminating all notes on other problems than infections. Direct access to signs of infection and immune reactions are obtained automatically [6, 7].

From EPR to tree (2)

Lexicographical analysis will be used to extract a patient's symptoms, reported over a certain time period from the EPR. We locate the body part(s) and the disorders associated with each of the symptoms. We have a predefined tree organizing all organs based on their hierarchical place in different systems within the human body. Based on this tree, we create a new tree to represent the organs likely involved in our patient's symptoms. Each node represents a body part and will get a label $\in\{0,1,2,\ldots\}$ based on the number of symptoms reported associated with that body part by that patient.

As a result we have a collection of trees representing the medical status of all patients in an area. Any trees containing only a root represent healthy patients

From tree to distance measurement (3)

For each new patient, or patient reporting new symptoms, we calculate the distance to other patients, based on the tree created in step 2. The distance between patient A and patient B is estimated by calculating how many steps (deletion, insertion, change of label) are necessary to change the tree representing patient 1 into the tree representing patient 2. This is called the Levenshtein distance between trees [8].

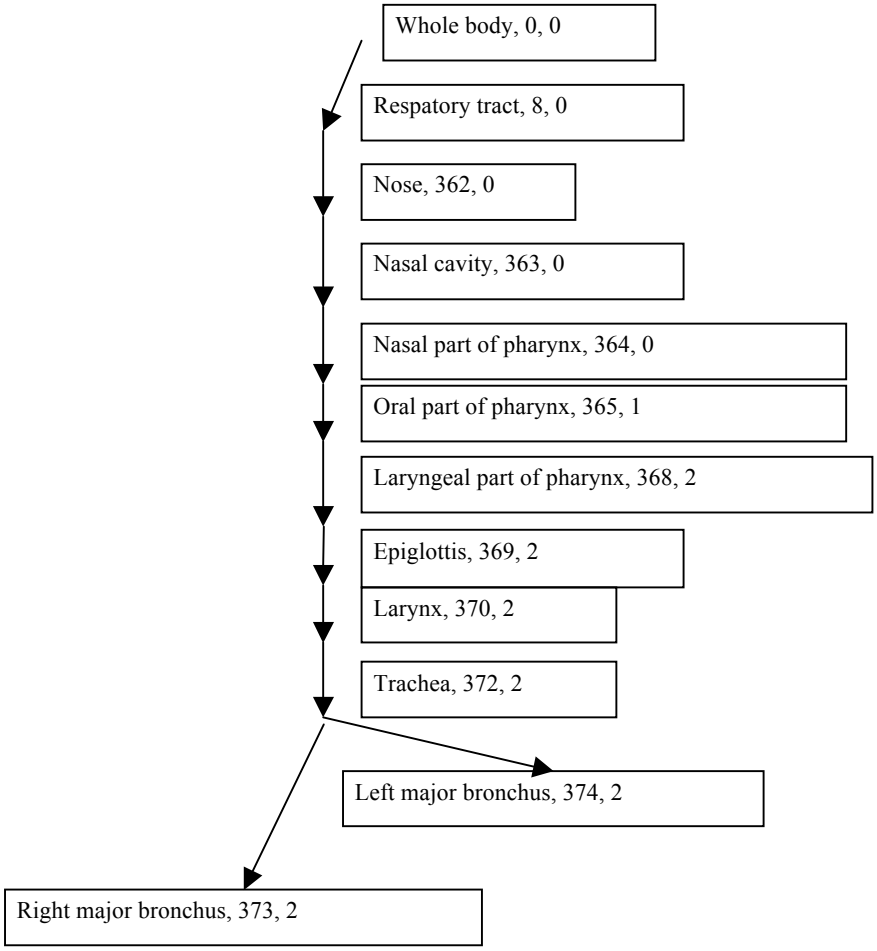

Figure 2 - Tree representation of patient with influenza pneumonitis

From distances to map (4)

After calculating all distances between all patients, we have attained a matrix of distances. We use an iterative push-and-pull algorithm to create a map of patients, approximately representing the relative similarities between their conditions. We create this map by randomly distributing points over a two-dimensional space, and adjusting for relative distances between points by push-and-pull operations [9]. This is an important step, since it facilitates a way to visualize the patient data of an entire geographical area.

Adjust map for additional data (5)

We then adjust the position of the patients on the map, based on data on additional symptoms reported in the EPR, that cannot be directly related to a specific body part (such as fever or high body temperature), their geographical proximity, family relation to one another and whether they belong in the same age group.

From map to layered map (6)

For each patient, we determine if they have experienced any immunological reactions. All groups of data points representing immunological reactions of the same type (e.g. type 3 for bacterial infections and type 4 for viral infections) can be represented in separate maps. The overlap of these maps will be the map from step 5.

From layers to clusters (7)

For each layer created in step 6, we detect clusters. If any of the clusters contains more than a predefined number of patients, there might be reason to believe that there is an outbreak of an infectious disease in the geographical area we are analyzing.

At this point the cluster of patients can be further analyzed, to determine overlap in symptoms and immunological reaction, to make a prediction about the nature of the outbreak, and take actions accordingly.

We suggest either k-means cluster [10] or visual detection. In the first case, visual detection can be done to crosscheck the result.

Summary

We have identified seven steps to transform symptom data into evidence for possible disease outbreaks. The model mimics the effect of disease on human anatomy. The steps are as follows: (1) From symptoms to EPR; (2) From EPR to tree; (3) From tree to distance measurement; (4) From distances to map; (5) Adjust map for additional data; (6) From map to layered map; (7) From layers to clusters.

In future work, we will add an eighth step in which we will study the source and direction of the outbreak after its existence has been established.

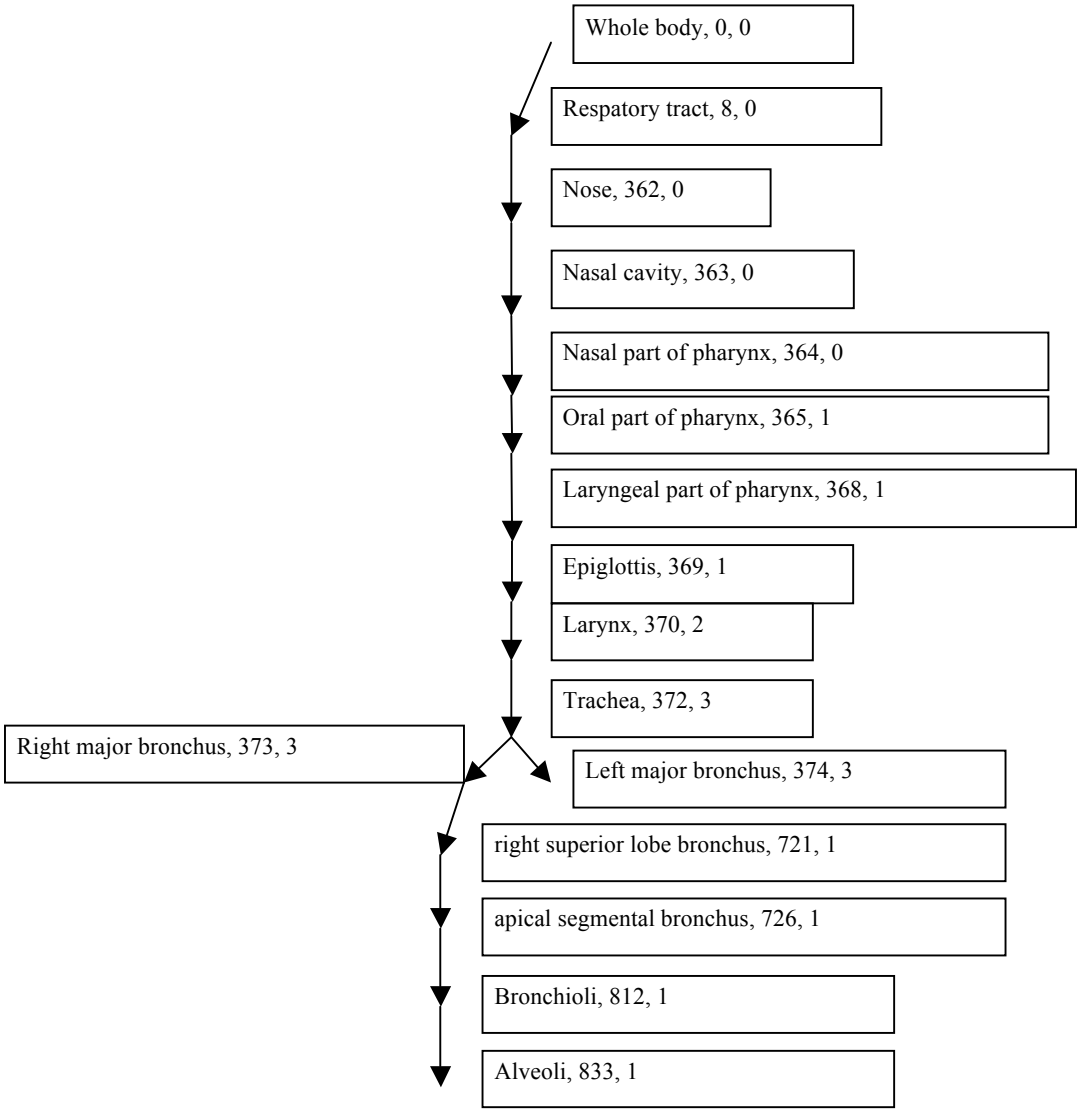

Figure 3 - Tree representation of patient with pneumococcal pneumonia.

Results

To demonstrate the seven steps of the algorithm, we implemented a basic version using Python. As examples for illustrating the system we have used the likely symptoms of patients with influenza pneumonitis, pneumococcal pneumonia, *E. coli* pyelonephritis and *E. coli* cystitis. The first two diseases primarily affect the respiratory tract. The latter two diseases affect the urinary tract. A virus causes the first disease, while bacteria cause the latter three. Patients infected with these diseases will therefore likely experience different IRs and symptoms affecting different body parts.

Examples of trees representing the affected organs of the patients with the first two diseases are shown in Figure 2 and 3. In these figures, the affected organs are followed by their code in our system and the number of symptoms likely affecting them.

When we transform data for patients with these four diseases, using algorithm 1 through 4, we arrive at a two-dimensional map visualizing how closely the medical status of the patients is related.

Since this data is random and any additional medical data describing direct similarities between patients will be arbitrary, we have not included step 5 in this example. Furthermore, we have not included a layered map of immunological reactions in this paper. However, it can easily be derived from Figure 4; taking into account that only Influenza pneumonitis gives IR reactions indicating a virus infection. The others indicate a bacterial infection. We can visually divide the two types of infection.

In figure 4 we can also visually identify approximate clusters for the different diseases.

Discussion

The results presented here strongly suggest that with minimal medical knowledge, clusters of patients can be detected and a prediction of types of infectious diseases with high incidence can be made. Due to the two-dimensional structure of the output data, a quick visualization can be made, which opens for the possibility of rapid surveillance of infectious diseases in clinical practice.

Since the input data represents the anatomical structure of disease concisely and objectively, only a few parameters are needed as input, and all parameters are readily available in an EPR. In particular, it is unproblematic to automatically extract data on immunological reactions [6]. The simplicity of the combined algorithms seems promising.

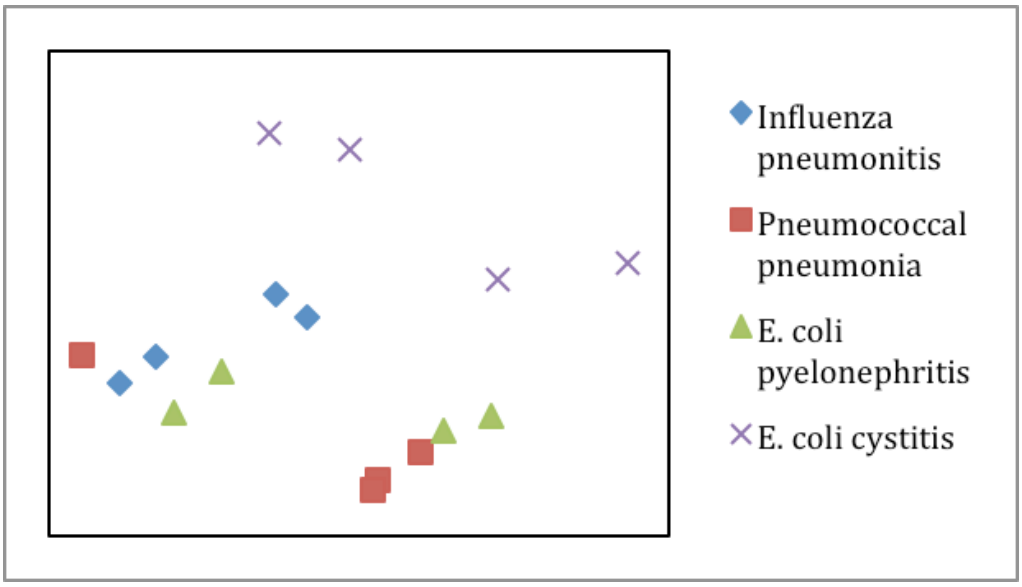

Figure 4 - Example of a two-dimensional map of patients.

Figure 4 shows correlation between data representing patients with similar health profiles. More correlation will likely be found if step 5 of our algorithm is taken into account, where other important overlapping health data will push the points together. However, the data used for this paper is simulated and adding further similarities between patients would be trivial.

In a follow up of this research, we will include real data and integrate an automatic mapping from symptoms to likely affected organs.

Conclusion

The development of strategies for the detection of infections before the onset of the symptoms is critical, especially given the limitations of the current disease surveillance systems that are based only on people's awareness of their health status. This is particularly important for vulnerable population groups, such as people with diabetes, that their health status is altered significantly when they are infected. Here, we highlighted the lessons learned in our project, the difficulties of this approach as well as our future plans. We hope that this work will provoke the thoughts for new directions within the disease surveillance field.

Acknowledgments

This work was supported in part by Norwegian Research Council Grant No. 174934 (Tromsø Telemedicine Laboratory).

References

[1] Bassøe, C-F. *A data structure for decision support systems, medical systems and clinical decision making*, MEDINFO (2007).

[2] Warrer, P, Hansen, EH, Juhl-Jensen, L, Aagaard, L. *Using text-mining techniques in electronic patient records to identify ADRs from medicine use.* Br J Clin Pharmacol. (2012);73:674-84.

[3] van den Branden M, Wiratunga N, Burton D, Craw S. *Integrating case-based reasoning with an electronic patient record system.* Artif Intell Med. (2011);51:117-23.

[4] Bassøe, C-F. *Combinatorial Clinical Decision-making,* PhD Thesis, University of Bergen, (2007).

[5] Botsis T, Bassøe CF, Hartvigsen G. *Sixteen years of ICPC use in Norwegian primary care: looking through the facts. BMC Med Inform Decis Mak* (2010);10:11.

[6] Rasmussen J-E, Bassøe C-F. *Semantic analysis of medical records. Meth Inform Meth* (1993);32:66-72.

[7] *Bassøe C-F. Automated diagnoses from clinical narratives: A medical system based on computerized medical records, natural language processing and neural network technology. Neural Networks* (1995);8:313-319.

[8] Levenshtein, V.I. *Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics Doklady* **10**(1966), 707–10.

[9] Cocx, TK., Kosters, WA. *A distance measure for determining similarity between criminal investigations,* 6th Industrial Conference on Data Mining (2006).

[10] MacQueen, JB. *Some Methods for classification and Analysis of Multivariate Observations. Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*. University of California Press (1967), 281–297.

Address for correspondence

Klaske van Vuurden, Department of Computer Science, University of Tromsø, 9037 Tromsø, Norway. E-mail address: Klaske.van.Vuurden@uit.no