Utilization of Administrative Registers using Statistical Knowledge Discovery

Reijo Sund

National Research and Development Centre for Welfare and Health (STAKES),
P.O.Box 220, 00531 Helsinki, Finland
Reijo.Sund@stakes.fi

Abstract. The increase in the volume of data collected for administrative purposes has introduced new difficulties in extracting useful information to support decision making. The aim of this study is to present some new methodological approaches for the utilization of administrative registers in research and in production of statistical information by combining fruitful ideas from the fields of statistics and data mining. Since the most important contribution of these ideas lies in understanding of phenomena, problems and data in knowledge discovery processes, the ideas are first described as steps in that process and then also illustrated with a practical real-world example.

1. Introduction

The resources of providers of health and social services are limited. The effective use of these resources needs administrative planning and political willingness. Moreover, political issues and the increasing role of cost-effectiveness considerations have recently, among other things, introduced new pressures to improve health and social service systems (see e.g. [1]).

To support decision making and quality control there is a demand for scientifically valid and comparative information. On the other hand, the information revolution has made it possible to effectively collect and store huge data sets (see e.g. [2], [3], [4]).

For example, the administrative health registers in Finland contain massive amounts of detailed data of good quality. However, raw data as such are of little value since in the context of problem solving the results are important and not the actual data. To fulfil the need of information, an important aim in the development of administrative information systems development has recently been the improved utilization of register based data (see e.g. [5]).

In principle, data can be converted to useful information, if answers to relevant problems can be extracted from the data by appropriate means (see e.g. [6]). However, the growing size of data sets has brought out challenges for the traditional statistical approaches (see e.g. [7]). In addition, from the statistical point of view it is important to notice that administrative registers are so called second hand data sets, i.e. the data have already been collected and there is no more possibilities to design

optimal data collection strategies to solve some particular problem. It is also unrealistic to assume that the understanding of complex methodological details is the only way to get results in the practical data analysis. The need of information together with the lack of statisticians has lead to the development of alternative, more concrete, ways to analyze data (see e.g. [8], [9]). In other words, statistics is no more alone in the field of data analysis as Huber [10] points out: "The most data analysis is done by nonstatisticians, and there is much commonality hidden behind diversity of languages. Rather than to try to squeeze the analysis into a too-narrow view what statistics is all about, statisticians ought to take advantage of the situation, get involved interdisciplinary, learn from the experience, expand their own mind, and thereby their field, and act as catalysts for the dissemination of insights and methodologies. Moreover, the larger the data sets are, the more important the general science aspects of the analysis seem to become relative to the 'statistical' aspects."

1.1 Aim of this study

The aim of this study is to present some new methodological perspectives for the utilization of administrative registers in research and in production of statistical information by combining ideas mainly from the fields of statistics and data mining. The most important contribution of these ideas lies in understanding of phenomena, problems and data in knowledge discovery processes.

First the concept of statistical knowledge discovery process is defined in order to allow the embedding of ideas to a suitable framework. After that certain parts of statistical knowledge discovery process are examined from the point of view of administrative register data. A systems approach and an event history framework are described in the 'understanding the problem' part. In the 'data understanding' part generalized event sequence as well as data structures, censoring and practical and statistical interpretations related to it are presented. The 'data preprocessing' part reviews common preprocessing techniques: data abstraction, cleaning, integration and reduction. Also some preprocessing tasks and methods suitable for event history framework, including an 'interesting pattern remapping' technique, are suggested. After that the ideas are illustrated with a practical real-world example and finally the key points of this paper are discussed in the conclusions.

1.2 Previous research

There is a huge amount of research related to knowledge discovery in databases (see e.g. [9]). Essentially the starting point here is the mining of administrative data with scientific interpretation [11] related to certain phenomena [12]. Practical example used here is from the field of health services research, which is closely connected to health care and medicine (see e.g. [13], [14], [15], [16], [17], [18]). Other relevant references are given in the text while the specific issues are presented.

2. Statistical Knowledge Discovery Process

The increase in the volume of collected data has presented new difficulties in extracting useful information to support decision making. The traditional manual data analysis has become insufficient, and methods for efficient analysis strategies are needed. In order to find sensible solutions to real-world problems, the whole problem solving must be done in a systematic way. So called knowledge discovery process is a formalization of a complex problem solving process which fundamentally requires human participation (see e.g. [19], [20]). There is even a proposed global standard for knowledge discovery process available [21]. Fundamental ideas in knowledge discovery process seem to be closely related to traditional statistical thinking and statistical practice (see e.g. [22]). As a consequence, statistical knowledge discovery process can be thought to consist of the following interactive steps: understanding the problem, understanding data, data preprocessing, modeling, evaluation and reporting. In this paper the first three steps are examined more thoroughly.

2.1 Understanding the problem

For problem solving, the problem has to be "matched" against the corresponding domain knowledge and data. Usually a statistician or data analyst is not an expert on the field where the problem arises. It means that the effective co-operation with domain experts is a fundamental part of knowledge discovery process. Here the characterization of the problem is formalized in a way which has been found to be easily understandable also for non-statisticians and, as a consequence, gives a common language for the experts of different fields.

A Systems Approach. Very often the starting point for a statistical modeling of any phenomenon is to characterize the important properties of the phenomenon, i.e. operationalize it as a system. A system is defined as a 'group of things or parts working together or connected in some way as to form the whole'. This definition of a system needs that the objects and restrictions related to the system are adequately described.

Time is essential factor in many problem domains. For example, disease processes evolve in time and patient records give the history of patients. For such dynamic phenomena, time should be explicitly considered in the system formulation.

Event History Framework. In fact, the importance of the longitudinal event history approach has long been recognized in many areas, especially in social sciences, econometrics and medical research (see e.g. [23], [24], [25], [26], [27]). In the analysis of dynamic phenomenon the focus is on the sequences of events which occur in time. Usually these event histories are described with the individuals' transitions across a set of discrete states in time. In the simplest case - known as survival analysis (see e.g. [28], [29]) - only one qualitative change is possible, for example the transition from a state 'alive' to the state 'dead'. However, in many cases one transition can not describe the dynamics of a phenomenon in a realistic way. For example the slow development of a disease, which finally reaches the state 'death', can not be modeled accurately using only the time between the first diagnosis and death.

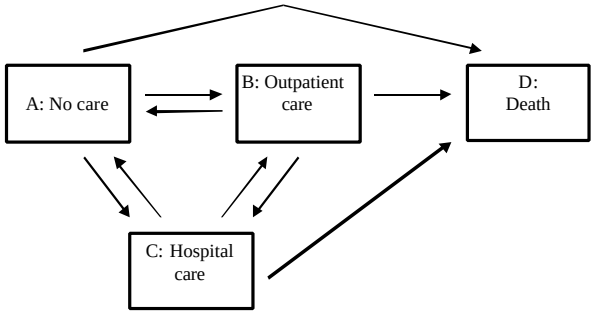


Fig. 1. A system as a directed graph

If the system is constructed in a reasonable way, the individuals' event histories can be considered as 'paths' through the system. Graphically the system can be presented as a directed graph whose nodes and edges have characteristics, which describing the properties of possible states and transitions in a system (Fig. 1).

2.2 Understanding data

Administrative data sets are massive second hand data sets. In general, data must have an appropriate form for storage and analysis purposes and for intuitive interpretation, as well. Here the essential parts of the traditional event history data structures including censoring (see e.g. [24], chapter 2) and event sequences originating from data mining (see e.g. [30], chapter 4) are combined to a generalized event sequence. Also some statistical models suitable for the analysis of generalized event sequence type data are presented, since with second hand data sets there are no more possibilities to design optimal data collection strategies to solve some particular problem, i.e. the data give the restrictions to analyzing possibilities.

Data structures. Event history data consist of observations of the form $(\tau_k,\mathbf{D}_k)$, where τ_k is an 'occurrence time' and $\mathbf{D}_k$ is an 'explanation' for the event (and n is the number of observations and $\tau_1 \le \tau_2 \le \ldots \le \tau_n$ and $\tau_i < \tau_j$ for at least one observation $i \ne j$ and i,j,k = 1,2,…,n).

Usually $\mathbf{D}_k$ consists of a set of attributes (variables). However, all the attributes do not necessarily contain important or interesting information. Some attributes can be unrelevant in relation to the solving of particular problem or can be easily derived from other attributes. It is often reasonable to divide up the relevant information into two subsets of attributes. The first subset defines an event type $\mathbf{E}_k$ and the other includes important 'covariate' information $\mathbf{i}_k$. For generality it is useful to allow a transformation f_k for the occurrence time τ_k. If not stated otherwise, $f_k(\tau_k) = \tau_k$ (k = 1,2,…,n).

Let m be the number of distinct occurrence times, t_i be the ith distinct occurrence time (i=1,2,…,m) and the event set $\mathbf{A}_i$ be the set of relevant information of observations occurred at the same time, i.e. $\mathbf{A}_i = \{(t_i,\mathbf{E}_k,\mathbf{i}_k)\}$, where i=1,2,…,m and for every i, k is over the observations for which $t_i = f_k(\tau_k)$.

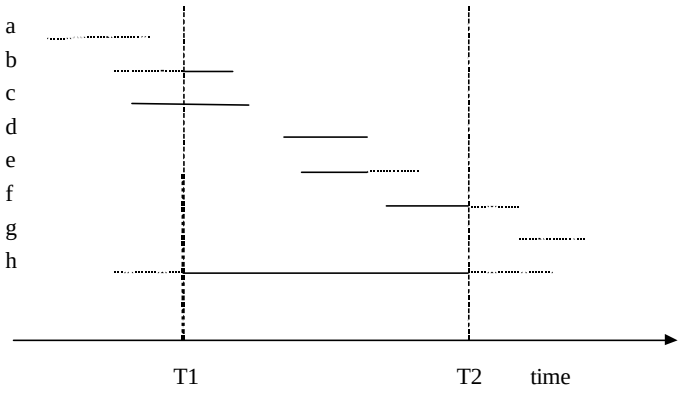


Fig. 2. Censoring

Censoring. In practice a data set is only a narrow window to the dynamics of a phenomenon. In other words the observations in the data set fulfill the condition $a < t_i < b$, where a and b are finite constants and i=1,2,…,m. The problems induced by a limited observation window are illustrated in Fig. 2, which shows examples of different types of 'censoring' for an individual's length of stay in some particular state (see also [27], 3-9). It is useful to include the type of censoring into data as an attribute, since the censoring must be taken care of in the analyses.

In the cases a and g there is no observed transitions to or from the state. This kind of censoring can be very problematic if the very first (or last) occurrence of some event type is considered more important than other occurrences (for example the first diagnosis of schizophrenia or the first back surgery operation). In the case b the transition to the state is not observed, but the transition from the state is known. In the case c both transitions are observed, but there is a potential problem, because the data outside the observation window are not complete for all individuals (typical situation for hospital discharge data). The case d corresponds to the uncensored observation. In the case e there has been an unknown or 'wrong' transition from the state (drop-outs or 'competing risk'). It is also possible that the follow up is ended before the occurrence of the event of interest (case f) or the transitions to and from the state are outside of the observation window (case h).

Generalized Event Sequences. The generalized event sequence **S** is defined to be a queue of event sets sorted by the (transformed) occurrence time, i.e. $\mathbf{S} = < \mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_m >$. In addition, let the conditional event sequence, to be called an event subsequence, be a sequence $\mathbf{S}_\theta = < \mathbf{A}_i \mid \mathbf{A}_i \in \mathbf{S}$ and condition θ is true $>$, where i=1,…,m.

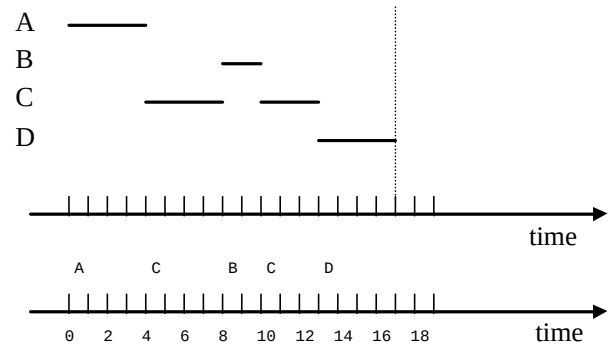
The definition of the generalized event sequence given above is very flexible and it can be used without any knowledge about the event history framework. However, using this framework generalized event sequences get a constructive and an intuitively clear interpretation. Moreover, transformations allowed in generalized event sequences make it possible to use the same data structure in the implementations of different statistical and data mining methods.

An Example of a Generalized Event Sequence. Fig. 3 shows an example of event history data and the corresponding event sequence, which could be a production of a system, let it be system P, presented in Fig. 1. There are also two graphical presentations for this particular event history. The first one is an event history description of transitions in a system where an individual actually stays in a current state until there is a transition to the another state. In the second one only the occurred events are marked to the figure. Since the time between two consecutive events is also the length of stay in a particular state, it is often very useful to include a 'length of stay' attribute into the covariate attributes $\mathbf{i}_k$, even though it can be easily calculated from the corresponding occurrence times.

In the definition of a generalized event sequence, parallel occurrences of events are allowed. In principle the systems approach can be made valid in this case by defining each combination of event sets to be a 'new' event type. In other words, it can be thought that the explanation for an event type is a combined description of events occurred at the same time. However, in many cases it is reasonable to classify the parallel events to different event types, since it is possible to find 'natural' interpretations for these parallel events.

Data in Fig. 3 is a description of movements of individual X in a system P. Let Q be a system with two states (state F: 'married' and state G: 'not married'). As a consequence, data corresponding to the movements of individual X in a system Q has the same form as the data in Fig. 3. In Fig. 4a is an example of that kind of situation. Now there are two event histories for individual X, one corresponding to the path through the system P and the other through the system Q. In this case it is known that the systems which 'generate' the data are parallel, but the both event histories can be combined into one event sequence (Fig. 4b).

T_k	$\mathbf{E}_k$
0	A
4	C
8	B
10	C
13	D



$$\mathbf{S} = < \{(0,A)\}, \{(4,C)\}, \{(8,B)\}, \{(10,C)\}, \{(13,D)\} >$$

Fig. 3. An example of event history data and the corresponding event sequence

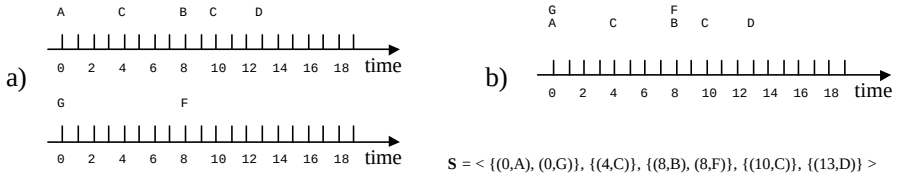


Fig. 4. An example of a situation with parallel occurrence times

In the case of two parallel systems the information concerning the 'source system' of an observation is a very valuable covariate. In general, there can, of course, be more than two data generating systems. If there is a need to restrict analyses to the observations from some particular system, that can be easily done using an event subsequence conditional to the corresponding system.

In practice there are always event histories for more than one individual. In that case it is trivial to include a covariate which identifies the individual, for example the social security number, while the whole data set still have the form of a generalized event sequence. Again it is possible to restrict the analyses to the arbitrary set of observations using event subsequences with suitable conditions.

Statistical Interpretations of a Generalized Event Sequence. From the statistical modeling point of view, a generalized event sequence can be interpreted as a sample path of a marked point process, if the occurrence times for all events are distinct. This leads to a very general family of statistical hazard-rate models suitable for censored data (see e.g. [31]).

In practice there is a lot of situations where so called calendar time of event occurrence is not important, since the 'real' information is the time between two consecutive events. In other words, the 'starting time' of a follow up can vary between individuals. A traditional solution, very common in survival analysis, is to transform the time axis from calendar time to the 'failure time'. This can be done using a transformation function $f_k(\tau_k) = \tau_k - b_k$, where b_k is the occurrence time of 'starting event' of a corresponding individual and k = 1,2,…,n (see e.g. [28], [29]).

Assuming that the probability of the next event type (state) depends only on time stayed in the current state, an appropriate choice for a model is a semi-Markov-model (see e.g. [32], [33]). Choosing the probabilities to change only on discrete time points, the semi-Markov-model can be formulated using the Markov chain, where the state space is expanded in a proper way (see e.g. [34]). In the traditional Markov chain the probability of the next event type (state) depends only on the current state (first order Markov property) and the probabilities are time-homogeneous (see e.g. [35], 372-427; [36], 61-84). As a matter of fact, the Markov chain interpretation corresponds to the situation, where the exact occurrence time is not important and only the order of observations matters. A generalized event sequence can be transformed to an event type sequence of this type using the transformation $f_k(\tau_k) = k$ (k = 1,2,…,n).

Moreover, choosing the transformation function f_k to be of the form $f_k(\tau_k) = c$, where c is an arbitrary constant and k = 1,2,…,n, the time dimension of the event sequence can be eliminated and the sequence reduces to data miners' classical market basket model (see e.g. [9], chapter 6) .

2.3 Data preprocessing

Usually massive second hand data sets contain so much information, domain specific features, inaccuracies and problems that raw data as such are not usable. To use data in problem solving, there must be understanding about the connections between the problem and data. In register based analyses the problem, domain knowledge and data also determine the most suitable model for the final problem solving. Moreover, most analyzing techniques are feasible to use only on moderately small data sets, since those typically need for access to the whole data set and the processing time will be directly proportional to the physical file size.

All in all, it can be stated that a sophisticated preprocessing incorporating non-technical domain knowledge in order to scale things down to a size fit for statistical analyses is the most important and time consuming part in register based data analysis.

Data abstraction. Often the connections between highly specific raw data and the highly abstract domain knowledge are so complicated that it is not possible to find any direct links between data and knowledge. In other words, an intelligent interpretation of raw data must be embedded into analyses, so that the resulting derived data set is at the level of abstraction corresponding to the current problem. Since noise is an unavoidable phenomenon also some kind of data validation and verification which makes use of knowledge should be performed. This kind of "intelligent" action is called data abstraction (see e.g. [17], chapter 2). For example, in the discovery of medical knowledge data is usually patient specific, while medical knowledge is patient independent and consists of generalizations that apply across patients.

Data cleaning. Real world data are very often more or less incomplete, noisy and inconsistent. Data cleaning deals with detecting and removing errors and inconsistencies from data in order to improve the quality of data (see e.g. [37]). Compared to data abstraction, cleaning is more data driven and technically oriented. In other words, corrections of erroneous and inconsistent codes as well as missing values can be usually made to the whole database, but data abstraction always results in problem specific derived data sets.

Data integration and reduction. Two other common types of preprocessing are called as data integration and data reduction. The idea in data integration is to include data from multiple sources in analyses and it is also known as record linkage (see e.g. [38], [39]).

Data reduction obtains a reduced representation of a data set which is much smaller in volume than the original data set, yet produces the same (or almost the same) analytical results (see e.g. [9], chapter 3). As a matter of fact, data reduction can consist of anything from simple database queries to very complicated modeling.

Preprocessing tasks and methods in event history framework. Since preprocessing is highly domain and problem specific, it is difficult to give any general suggestions concerning tasks and methods. Moreover, sometimes it can be difficult to distinguish what is preprocessing and what is modeling.

However, some non-trivial preprocessing tasks in the event history framework could be for example: "define" the state space for some system (what happens to patient after surgical operation?), find interesting and frequent combinations of patterns (which are frequent combinations of diagnosis and operation codes?), classify patterns to adequate hierarchies (which diagnoses relate to complications of a surgical operation?) and confirm the expert's "hypotheses" about the phenomenon from data.

Data mining tools from the “frequent pattern family” (see e.g. [40]) seem to be useful to these kinds of preprocessing purposes. The extensions of the basic ideas including templates (see e.g. [41]), multiple levels (see e.g. [42]) and measures of interestingness (see e.g. [43], [44]) and similarity (see e.g. [30]) are also useful in certain situations.

One very straightforward, but extremely useful, technique is so called interesting pattern remapping. The idea is to first to "forget" the time dimension of the generalized event sequence by using appropriate transformation. Since the whole data then reduces to market basket case, it is possible to use a levelwise search (see e.g. [45]) to extract the frequent patterns from data regardless of occurrence times. In the remapping phase interesting patterns can be classified to appropriate event types simply by "giving a new name" for each pattern. Finally the time dimension is restored and records without interesting event types are removed. This abstraction usually results in a notably reduced data set with an interpretation corresponding to the current problem.

3. Practical example

This example is a simplified extraction from a study, which aimed 1) to develop and implement register based performance indicators to measure the effectiveness of surgical treatment of hip fracture and 2) to evaluate and compare the effectiveness of health care providers. The complete results are reported elsewhere ([46], [47]). The study is a part of a larger project which aims to develop register based methods for measurement of effectiveness in specialized health care.

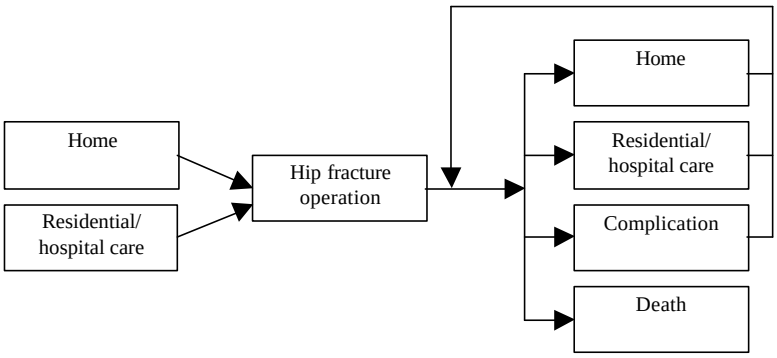


Fig. 5. System characteristics of remarkable life events related to hip fracture surgery

Understanding the problem. The first task in the project was to build up a research group consisting of experts of different fields. The group defined the actual problem more meticulously: the idea was to identify all hip fracture patients from the Finnish Health Care Register and follow the life events they encountered after hip fracture surgery according to register data.

A simplified system related to this particular problem is shown in Fig. 5. As it can be seen, the first hip fracture operation performed on an individual patient has a key role in the characterization of this phenomenon. Actually the state preceding the first hip fracture operation also matters from the clinical point of view, since the patients coming from home are usually in better condition than patients who are already in residential or hospital care. However, the major interest is in the events and pathways of care following the first hip fracture operation. In this case these events are classified into four categories. If everything goes well, the patient should get back home. Hip fracture is a serious condition for the elderly and it can be a starting point or catalyst also to other problems which may result in the need of residential or hospital care. Even fatal complications may follow hip fracture operation. This categorization was chosen, since the event types comprised the events of interest from the point of view of the original problem and even more importantly these life events are recorded in various registers.

For generality the system formulation also allows that there are multiple events after hip fracture surgery. As a matter of fact the pathways of care are more interesting than single events in situations, such as the cost-effectiveness evaluation. Only death in system formulation is an absorbing state, all other states can be followed by any other state, i.e. only the death state ends a path in the system. Moreover, even though these four states are distinct, they are not necessarily independent. For example death may be more probable after complication or it may not be very likely to get home if there is a decision for long term residential care. The absorbing death state causes even more problems, because it is not possible to say for any individual that there is an increased risk of complication after death. However, it is possible to speculate what would have happened, if death had not occurred. All in all, this kind of quite simple system definition seems to result in an extremely complicated competing risks model and more simplifications are needed in the actual modeling.

Understanding and preprocessing data. A cohort of patients with hip fracture in 1998 or 1999 were identified in the Finnish Hospital Care Register using a simple diagnosis group abstraction (all patients with at least one ICD-10: S72 diagnosis in 1998 or 1999). Using the unique person identity codes of the patient cohort, data on all inpatient and outpatient hospital care and deaths for this cohort were obtained from the Finnish Health Care Register, data warehouse of the Finnish Hospital Benchmarking Project and the National Causes of Death Register. The results of these straightforward database queries were integrated into a new data set containing 167 952 records for 17 099 patients.

Each record in this data set corresponds to one care episode in hospital (or death), not any actual event of the system, i.e. each observation includes information, such as patient and hospital ID-numbers, age, sex, area codes, diagnose and operation codes as well as dates of admission and discharge (or death). Data cleaning was performed

in order to correct impossible simultaneous hospital episodes, systematic errors in the use of symptom vs. cause diagnoses and some missing or erroneous attribute values in area codes.

Many types of censoring occurs in this data set: the first hip fracture operation (or other important event) can be outside the observation window, some hospital episodes may have begun before 1998, follow-up ends in the end of 1999 (there is census data available for the last day of every year in the Finnish Health Care Register) and the death may end the follow-up of a patient.

Operation codes corresponding to hip fracture surgery were abstracted into two different operation types. Using this and the diagnosis group abstraction of hip fracture, hip fracture operations were identified from the data. Since the state preceding hip fracture operation was also important, the histories of the patients were traced backwards and the preceding state was classified to be home or residential/hospital care using more complicated temporal data abstraction.

The forward direction abstractions were even more complicated and all event types needed special abstractions and techniques. For example, the acute complication events were identified using interesting pattern remapping, in which all clinically relevant complication diagnoses were remapped to one event type.

Modeling, evaluation and reporting. For the statistical modeling purposes it was assumed that any acute complication event after hip fracture operation is an outcome which reflects the effectiveness of the surgical treatment. In addition, deaths and the upper limit of the observation window were considered to cause censoring to the event of interest. With these assumptions the modeling reduced to standard survival analysis (see e.g. [29]) where the variables of interest are time between hip fracture operation and a complication or censoring event, and the censoring indicator. These variables were calculated for all patients over 60 years, who had been living at home before surgery. The final preprocessed data set had 8824 records, each containing relevant variables for one patient.

Since the outcome was an acute complication, the hazard function of acute complication occurrences was estimated. According to the hazard function, the probability of acute complications was higher during the first 30 day period after surgical operation. This finding based on the data corresponded to the domain knowledge and gave some evidence that the data abstraction was done in a proper way.

In spite of the fact that time dimension includes a lot of information, the actual effectiveness indicator related to acute complication was defined to be the rate of acute complications during the 30 day period after the surgical operation. The 30 day cumulative risk of acute complications obtained from the distribution function of censored event times was found to be suitable estimator for the rate (see e.g. [29], 48-52). The estimated 30 day acute complication rate was 13 % (95% CI: 12.3%-13.7%).

The most useful information was obtained, when rates were evaluated for health care providers, such as hospitals or hospital districts. This kind of profiling analyses allowed comparisons of effectiveness between the providers. In order to adopt more realistic, medically based criteria for judging the performance and improved case-mix adjustments, more sophisticated profiling analysis methods, such as control charts [48] and hierarchical multilevel Bayes-models [49], were used [46], [47].

4. Conclusions

In this paper, some new methodological approaches for the utilization of administrative registers in research and in production of statistical information were presented. This was done by combining ideas mainly from the fields of statistics and data mining. Since the most important contribution of these ideas lies in understanding of phenomena, problems and data in knowledge discovery processes, the ideas were described as steps in the statistical knowledge discovery process.

The first part in this process was the understanding of the problem. There were three main points in that part: the problem solving always needs domain knowledge, a systems approach generates a common language for the experts of different fields and event history framework is a well developed choice for outlining the characteristics of dynamic phenomena.

The second part was about the understanding of data. The key ideas were: Second hand register data has a form of a generalized event sequence, which is identical to the structure of event history data, and implicitly defines possible model types for problem solving. In practice, data are always censored, but they still give a glance to the data generating process, which have an intuitive interpretation in the event history framework. Moreover, a wide variety of suitable data structures for traditional models can be yielded from the generalized event sequence as special cases using the presented time transformation property.

In the third part, data preprocessing was discussed. Some suggestions of suitable tools were given and a pattern remapping technique was formulated. The ultimate message was that sophisticated preprocessing incorporating the non-technical domain knowledge in order to scale things down to a size fit for statistical analyses is the most important and time consuming part in the register based data analysis.

Finally the ideas were illustrated with a practical example. It was seen that register based data analysis becomes very complicated and challenging even in simple situations. The perspectives presented in this paper propose many practically useful approaches to solve fundamental problems encountered in the analysis of massive second hand data sets.

All in all, every data analyst should remember that "statistics (data) are not collected, but produced; research results are not findings, but creations" [50].

References

1. Delasie, L., Kastelein, A., van Merode, F, Vissers, J.M.H. (eds.): Managing Health Care under Resource Constraints. European Journal of Operational Research 105 (1998)
2. Shortliffe, E.H., Perreault, L.E. (eds.): Medical Informatics : Computer Applications in Health Care and Biomedicine, Second Edition. Springer-Verlag, New York (2001)
3. Nykänen, P.: Decision Support Systems from a Health Informatics Perspective. PhD thesis, Department of Computer and Information Sciences, University of Tampere, Acta Electronica Universitatis Tamperensis 55, Tampere (2000)
4. Shani, M.: The impact of information on medical thinking and health care policy. International Journal of Medical Informatics 58-59 (2000) 3-10

5. Nenonen, M., Nylander, O.: A Theoretical Framework for Health Information Systems. Themes 3/2001. National Research and Development Centre for Welfare and Health (STAKES), Helsinki (2001)
6. Dudewicz, E.J., Mishra, S.N.: Modern Mathematical Statistics. Wiley Series in Probability and Mathematical Statistics, John Wileys & Sons, New York (1988)
7. Friedman, J.H.: The Role of Statistics in the Data Revolution. Bulletin of the International Statistical Institute, 52nd Session, Proceedings, Book 1 (1999) 121-124
8. Hand, D.J.: Data Mining : New Challenges for Statisticians. Social Science Computer Review 18 (2000) 442-449
9. Han, J. & Kamber, M.: Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers, San Fransisco (2001)
10. Huber, P.J.: Massive Datasets Workshop: Four Years After. Journal of Computational and Graphical Statistics 8 (1999) 635-652
11. Fayyad, U., Haussler, D., Stolorz, P.: Mining Scientific Data. Communications of the ACM 39 (1996) 51-57
12. McCarthy, J.: Phenomenal Data Mining: From Data to Phenomena. SIGKDD Explorations 1 (2000) 24-29
13. Cios, K.J. (ed.): Medical Data Mining and Knowledge Discovery. Studies in Fuzziness and Soft Computing, Physica-Verlag, Heidelberg (2001)
14. Goodall, C.R.: Data Mining of Massive Datasets in Healthcare. Journal of Computational and Graphical Statistics 8 (1999) 620-634
15. Katz, B.P. (ed.): Measuring Quality, Outcomes, and Cost of Care Using Large Databases. Annals of Internal Medicine 127 (1997) Number 8, Part 2
16. Lavrac, N.: Selected techniques for data mining in medicine. Artificial Intelligence in Medicine 16 (1999) 3-23
17. Lavrac, N., Keravnou E., Zupan B.: Intelligent data analysis in medicine. In: Encyclopedia of computer science and technology, Vol. 42 (Supp. 27), 113-157, Marcel Dekker, New York (2000)
18. Øhrn, A.: Discernibility and Rough Sets in Medicine: Tools and Applications. PhD thesis, Department of Computer and Information Science, Norwegian University of Science and Technology, NTNU report 1999:133, IDI report 1999:14, Trondheim (1999)
19. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P.: The KDD Process for Extracting Useful Knowledge from Volumes of Data. Communications of the ACM 39 (1996) 27-34
20. Brachman, R.J., Anand, T.: The Process of Knowledge Discovery in Databases. In: Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R. (eds.): Advances in Knowledge Discovery and Data Mining. AAAI Press, Menlo Park (1996)
21. Chapman ,P., Clinton, J., Kerber, R., Khabaza, T. Reinartz, T., Shearer, C. Wirth, R.: CRISP-DM 1.0 : Step-by-step data mining guide. The CRISP-DM consortium (2000) http://www.crisp-dm.org/
22. Phannkuch, M., Wild, C.J.: Statistical Thinking and Statistical Practice: Themes Gleaned from Professional Statisticians. Statistical Science 15 (2000) 132-152
23. Allison, P.D.: Event History Analysis - Regression for Longitudinal Event Data. Quantitative Applications in the Social Sciences series 46, Sage Publications, Beverly Hills (1984)
24. Blossfeld, H.-P., Rohwer, G.: Techniques of Event History Modeling: New Approaches to Causal Analysis. Lawrence Erlbaum Associates, Mahwah (1995)
25. Clayton, D.G.: The Analysis of Event History Data: A Review of Progress and Outstanding Problems. Statistics in Medicine 7 (1988) 819-841
26. Lancaster, T.: The Econometric Analysis of Transition Data. Econometric Society Monographs 17, Cambridge University Press, Cambridge (1990)
27. Yamaguchi, K.: Event History Analysis. Applied Social Research Methods Series 28, Sage Publications, Newbury Park (1991)
28. Kalbfleisch, J.D., Prentice, R.L.: The Statistical Analysis of Failure Time Data. John Wiley & Sons, New York (1980)

29. Cox, D.R. & Oakes, D.: Analysis of Survival Data. Chapman and Hall, London (1984)
30. Moen, P.: Attribute, Event Sequence, and Event Type Similarity Notions for Data Mining. PhD Thesis, University of Helsinki, Department of Computer Science, Series of Publications A, Report A-2000-1, Helsinki (2000)
31. Arjas, E.: Survival Models and Martingale Dynamics. Scandinavian Journal of Statistics 16 (1989) 177-225.
32. Janssen, J. (ed.): Semi-Markov Models: Theory and Applications. Plenum Press, New York and London (1986)
33. Janssen, J., Limnios, N. (eds.): Semi-Markov Models and Applications. Kluwer Academic Publishers, Dordrecht (1999)
34. Howard, R.A.: Dynamic Probabilistic Systems, Volume I: Markov Models, Volume II: Semi-Markov and Decision Processes. John Wiley & Sons, New York (1971)
35. Feller, W.: An Introduction to Probability Theory and Its Applications, third edition. John Wiley & Sons, New York (1968)
36. Ross, S.M.: Applied Probability Models with Optimization Applications. Dover Publications, New York (1970/1992)
37. Rahm, E., Do, H.H.: Data Cleaning: Problems and Current Approaches. Bulletin of the IEEE Computer Society Technical Committee on Data Engineering 23 (2000) 3-13
38. Winkler, W.E.: The State of Record Linkage and Current Research Problems. U.S. Bureau of the Census, Technical Report 4/1999 (1999)
39. Alvey, W., Jamerson, B. (eds.): Record Linkage Techniques -- 1997. Proceedings of an International Workshop and Exposition, March 20-21, 1997, Arlington, VA. Federal Committee on Statistical Methodology, Office of Management and Budget, Washington DC (1997)
40. Mannila, H., Toivonen, H.: Knowledge Discovery in Databases: The Search for Frequent Patterns. Book Draft, Department of Computer Science, University of Helsinki (1998)
41. Klemettinen, M., Mannila, H., Ronkainen, P., Toivonen, H., Verkamo, A.I.: Finding Interesting Rules from Large Sets of Discovered Association Rules. In: Nabil, R.A., Bhargava, B.K., Yesja, Y. (eds.): Proceedings of 3th International Conference on Information and Knowledge Management (CIKM '94), ACM Press (1994) 401-407
42. Han, J., Fu, Y.: Discovery of Multiple-Level Association Rules from Large Databases. In: Proceedings of International Conference on Very Large Data Bases (VLDB'95), pages 420-431, Zurich (1995)
43. Sahar, S.: Interestingness via what is not interesting. In: Proceedings of KDD '99, San Diego, Ca, USA (1999)
44. Silberschatz, A., Tuzhilin, A.: What Makes Patterns Interesting in Knowledge Discovery Systems. IEEE Transactions on Knowledge and Data Engineering 8 (1996) 970-974
45. Mannila, H., Toivonen, H.: Levelwise Search and Border of Theories in Knowledge Discovery. Data Mining and Knowledge Discovery 1 (1997) 241-258
46. Rissanen, P., Sund, R, Nordback, I., Rousi, T., Idänpään-Heikkilä, U.: Lonkkamurtuman hoidon vaikuttavuuden rekisteriperusteinen mittaaminen ja vertailu (Register based measurement of effectiveness of surgical treatment of hip fracture, in Finnish). Aiheita sarja 2002. National Research and Development Centre for Welfare and Health (STAKES), Helsinki (2002)
47. Sund, R. et al.: Health Care Provider Profiling: Effectiveness of Surgical Treatment of Hip Fracture. Nordic Health Econometric Workshop, August 22, Helsinki, Finland (2002)
48. Adab, P., Rouse, A., Mohammed, M.A., Marshall, T.: Performance league tables: the NHS deserves better. British Medical Journal 324 (2002) 95-98
49. Marshall, E.C., Spiegelhalter, D.J.: Institutional Performance. In: Leyland, A.H., Goldstein, H. (eds.): Multilevel Modelling of Health Statistics. Wiley Series in Probability and Statistics, John Wiley & Sons, Chichester (2001)
50. Irvine, J., Miles, I., Evans, J.: Introduction: Demystifying Social Statistics. In: Irvine, J., Miles, I., Evans, J. (eds.): Demystifying Social Statistics. Pluto Press, London (1979)