IDAMAP 2008 Intelligent Data Analysis in Biomedicine and Pharmacology

Artificial Intelligence in Medicine 2008, Washington DC

Program Co-Chairs
John Holmes and Allan Tucker

1 Introduction

Welcome to IDAMAP 2008, the thirteenth workshop on Intelligent Data Analysis in Biomedicine and Pharmacology, held in conjunction with the American Medical Informatics Association Annual Meeting 2008 in Washington DC.

The IDAMAP workshop series is devoted to computational methods for data analysis in medicine, biology and pharmacology that present results of analysis in the form communicable to domain experts and that somehow exploit knowledge of the problem domain. Such knowledge may be available at different stages of the data-analysis and model-building process. Typical methods include data visualization, data exploration, machine learning, and data mining.

Gathering in an informal setting, workshop participants will have the opportunity to meet and discuss selected technical topics in an atmosphere which fosters the active exchange of ideas among researchers and practitioners. The workshop is intended to be a genuinely interactive event and not a mini-conference, thus ample time will be allotted for general discussion.

2 Program

The scientific program includes a selection of papers presented throughout the workshop with the following themes:

-*Bayesian Networks and Decision Support*

-*Prediction and Classification*

-*Natural Language Processing and Text Mining*

-*Data Mining and Analysis*

We are delighted to have an invited talk from Dr Paola Sebastiani who is an associate professor in Biostatistics at the Boston University School of Public health and adjunct associate professor in the Bioinformatics Program. She is an expert in Bayesian data analysis, Time-series and Bioinformatics.

3 Program Committee

- Ameen Abu-Hanna, Academic Medical Center, Amsterdam, The Netherlands
- Riccardo Bellazzi, University of Pavia, Italy
- Carlo Combi, University of Verona, Italy
- Janez Demsar, University of Ljubljana, Slovenia
- Michel Dojat, Universite Joseph Fourier, Grenoble, France
- Dragan Gamberger, Rudjer Boskovic Institute, Croatia
- Werner Horn, Medical University of Vienna, Austria
- John H. Holmes, University of Pennsylvania School of Medicine, USA (chair)
- Jim Hunter, University of Aberdeen, UK
- Elpida Keravnou-Papaeliou, University of Cyprus, Cyprus
- Matjaz Kukar, University of Ljubljana, Slovenia
- Pedro Larranaga, University of the Basque Country, San Sebastian, Spain
- Nada Lavrac, J. Stefan Institute, Slovenia
- Xiaohui Liu, Brunel University, UK
- Oscar Luaces, Universidad de Oviedo at Gijon, Spain
- Peter Lucas, Radboud University Nijmegen, The Netherlands
- Marco Masseroli, Politecnico of Milan, Italy
- Silvia Miksch, Danube University Krems, Austria
- Lucila Ohno-Machado, Harvard Medical School and M.I.T., Boston, USA
- Niels Peek, Academic Medical Center, Amsterdam, The Netherlands
- Francois Portet, University of Aberdeen, Scotland
- Marco Ramoni, Harvard Medical School, Boston, USA
- Steve Rees, Aalborg University, Denmark
- Paola Sebastiani, Boston University School of Public Health, USA
- Yuval Shahar, Ben-Gurion University of the Negev, Israel
- Stephen Swift, Brunel University, UK
- Allan Tucker, Brunel University, UK (chair)
- Blaz Zupan, University of Ljubljana, Slovenia

4 Acknowledgements

We would like to thank our invited speaker, Dr. Paola Sebastiani, all of the paper authors, and all of the members of the program committee. We would especially like to thank the Knowledge Discovery and Data Mining Working Group of the American Medical Informatics Association, the Intelligent Data Analysis and Data mining Working Group (WG3) of the International Medical Informatics Association, the AMIA Working Group Steering Committee, and Dasha Cohen of the AMIA staff for their support in the organization of IDAMAP 2008.

IDAMAP 2008 Program
Friday, November 7, 2008
Georgetown West, Washington Hilton

Morning Session

IDAMAP 2008 Program
Friday, November 7, 2008
Georgetown West, Washington Hilton

Afternoon Session

1:40 pm	**Paper session:** ***Natural Language Processing and Text Mining***	
	E.J. Goldsmith, S. Mendiratta, R. Akella, K. Dahlgren Natural Language Query in the Biochemistry and Molecular Biology Domains Based on Cognition Search	45
	P. Langlais, F. Yvon, P. Zweigenbaum Translating Medical Words by Analogy	51
	D. Li, K. Schuler, G. Savova A Comparison between CRFs and SVMs in Disorder Named Entity Recognition in Clinic Texts	57
	A.T. McCray, K. Lee Exploring semantic locality patterns in the biomedical literature	63
3:00 pm	**Break**	
3:20 pm	**Paper session:** ***Data Mining and Analysis***	
	S. Pavlidis, A. Payne, S. Swift An Improved Methodology for Pathway Based Microarray Analysis Based on Identification of Individual Pathways Responsible for Gene Regulation	69
	S. Concaro, L. Sacchi, C. Cerra, P. Fratino, R. Bellazzi Temporal Data Mining for the Analysis of Administrative Healthcare Data	75
	L.J. Moniz, A. Buczak, J. Copeland, H. Rolka, J. Lombardo, S. Babin, B. Feighner Data-Driven Hybrid Method for Synthetic Electronic Medical Records Generation	81
	J.L. Lustgarten, V. Gopalakrishnan, W.R. Hogan, S. Visweswaran Improving a Knowledge Base for Use in Proteomic Data Analysis	87
	G.H. Merrill, P.B. Ryan, J.L. Painter Applying a UMLS/SNOMED-based Drug Ontology for Observational Pharmacovigilance	90
5:00 pm	**Closing**	
6:00 pm	**Social Event**	

After the first wave of genome wide association studies: what's next?

Paola Sebastiani

Department of Biostatistics

Boston University

The technology of SNP arrays has moved the field of genetics from candidate gene studies to genome-wide association studies, and the last two years have produced hundreds of manuscripts reporting results of genome wide association studies for a variety of heritable traits. What do the results tell us? Unfortunately very little, and the genetic community is ready to go on board of the next high throughput technology - deep sequencing - in the hope of discovering those rear variants that may be causative of complex and common diseases. However, the massive amount of genetic data produced by these studies remains largely unexplored, because of the challenge of mining and modeling these massive data sets and the focus on limiting false positives rather than false negatives that has limited the analysis. Genetic data can be used for risk prediction modeling, and the intelligent data analysis community is well powered for this type of studies. Also, integration of genetic data with other gene products can speed up the discovery of disease mechanism and identify targets for further studies. This use of genetic data is however limited by the major issue of subject identifiability that prevents access to individual data. I will discuss these issues and show examples from research on rare disease in which subjects' identifiability is a major obstacle to data access.

An Application of Bayesian Network Models to Discover Productivity Patterns in Primary HealthCare

Masoumeh T. Izadi, Robyn Tamblyn, Dale Dauphinee, and David L. Buckeridge
McGill University, 1140 Pine Ave West
Montreal, Quebec Canada H3A 1A3

mtabae at cs.mcgill.ca, robyn.tamblyn at mcgill.ca, dale.dauphinee at mcgill.ca
david.buckeridge at mcgill.ca

Abstract

It is widely recognized that accurate measures of the productivity of health care providers are essential for the healthcare planning. Productivity estimation should be based on efficient and valid inferences from healthcare and health services data. However, inferential analysis of these data is complicated by factors such as confounding, random error, classification error, and selection bias, all of which increase uncertainty. To date, productivity analysis has neither taken into account the full range of dynamic variables that are involved, nor accounted for their uncertainty and complex interactions. In this paper, we suggest the use of Bayesian networks (BNs) as a compact and efficient tool to represent different aspects of uncertainty inherent in productivity assessment among primary care physicians. We demonstrate that BNs provide a powerful framework for data analysis in this domain using demographic and clinical practice data from Quebec's Medicare system. The predictive performance of BNs is compared with well-known methods frequently used in medicine and health sciences. We also discuss some of the challenges and opportunities for future work.

1 Introduction

Aging populations, rising patient expectations, and changing approaches to health care have increased the demand for primary care services in many countries. At the same time, the number of primary care practitioners is limited, and there is an increasing gap between the supply and demand for services by primary care physicians. Effective workforce planning to tackle this issue, must: (1) project the needs of the population, and (2) define the health care supply resources required to meet those needs. In this study we focus on the supply model. The supply model should be based on current patterns in and assumptions about the skills and productivity of health care professionals.

Currently in the US, the Physician Requirement Model (PRM) and the Physicians Supply Model (PSM) [Altarum, 2000; DHHS, 2006] are used to predict the necessary supply of physicians services. These models use traditional linear and moving average projections to estimate required services in the future. Projections using these models assume that the specialty choice, practice behavior, and level of productivity of all new graduates will continue the same way as in the past. The American Academy of Family Physicians (AAFP) has its own needs-based workforce policy to identify the workforce needs for family physicians. Recent studies show that productivity is the most influential parameter in projection of supplier services. For instance, a 0.5% increase in productivity per year would decrease the number of additional MDs required by 20% over 15 years [Birch *et al.*, 2007]. Commonly used methods of physician workforce projections are based on the assumption that all physicians are equally productive, productivity is increased for all physicians each year by a fixed amount, and that productivity is uniformly distributed proportional to the demand. Yet, the evidence supports none of these assumptions. Physician productivity is highly variable, and is influenced by physician age, gender, attitudes towards lifestyle, and other dynamic variables. In reality, physician distribution is far from uniform, with the majority of physicians choosing urban areas to practice, despite the existence of highly publicized problems in under-served areas.

Physician productivity, as a key determinant of medical workforce supply, is influenced by many variables. These variables include characteristics of physicians and places in which they practice. Most previous work on productivity analysis has focussed on finding the association between a few demographic features and a single measure of work performed (e.g. physician's patient visits, total hours worked, or total dollars generated), mainly using a simple regression model [Dreyer *et al.*, 2007]. Although regression models are useful tools for analysis of clinical and administrative data, some important features of true processes are difficult to model with standard regression methods. Most notably, it is difficult to model complex dependencies between explanatory variables and to model multiple outcomes [Hutcheson and Sofroniou, 1999]. In addition, most existing studies included only a few physicians and limited data, so the findings may reflect random variations rather than broad patterns. While discovering associations using these methods can be used to gain background knowledge about some parts of the problem, it is difficult to see the whole picture without an integrated model that learns profile-based patterns of practice behavior [Moore, 2002]. In fact in some cases, lacking appropriate models and making invalid assumptions have had undesirable consequences. For instance, projections of oversupply made in 1980 by Graduate Medical Education National Advisory Committee (GMENAC) and those made in the early 1990s using health maintenance organizations (HMO) staffing patterns were seriously in error [Chan, 2003; Curson, 2003; Weiner, 2002]. Therefore, where they exist, the prediction methods are rudimentary [O'Connor, 1993]and better tools for measuring and predicting medical manpower are needed for evaluating, a priori, how different policy decisions could impact the resources available. While the need for more advanced methodologies is recognized in this domain [Lewin, 2006; Chan, 2003; Curson, 2003; Joyce *et al.*, 2004], planning agencies and policymakers continue to struggle with how to put this into practice.

We propose a different approach to productivity analysis using machine learning techniques. Specifically, we propose to use Bayesian networks (BNs) [Jensen, 1996; Pearl, 1988] to model the workforce productivity and to relate productivity to demographic characteristics of physicians and their geographical distribution, considering a range of dynamic variables and their complex interactions. BNs are designed to complement traditional hypothesis-driven methods of data analysis. They can perform predictions under various scenarios, and show the results of possible interventions immediately. Our objective in developing BN models for productivity is to develop a multi-purpose tool that mines practice profiles and more realistically represents our knowledge about the primary care supply system. In particular, we want to represent attributes that are of interest to decision makers and health administrators in determining physician staffing requirements or physician distribution over different regions.

2 Development of the Bayesian Network for Productivity

In this study, we are interested in measuring different indicators of productivity among primary care physicians. Therefore, we use multiple outcome measures that are not necessarily independent of each other. To be able to address this problem in a regression framework, we would need to make several independent models one for each indicator. BNs facilitate use of multiple independent regression models by introducing a single model which includes all outcomes.

2.1 Practice Indicators Definition

Individuals and organizations have used a variety of productivity measures, but one traditional measure of productivity is the number of patient encounters. This measure is often accompanied by measures of work time, such as the total days the physician works in a year. Another traditional measure of physician productivity is dollars generated to the practice. Some of the traditional measures of productivity have limitations that inhibit cross-physician comparison. In this study we consider the total number of days worked, and the total number of patients visits in each practice year as outcome measures for productivity. Another measure related to productivity is how successful a physician is in maintaining his/her patients. This measure is called continuity of care and it reflecs patient satisfaction and physician accessibility. Ideally, high productivity should exist while maintaining high continuity of care. Continuity of care is defined in a number of ways. We computed this index as follows:

$$COC_{\text{per patient}} = \frac{\text{number of visits to study physicians}}{\text{number of visits to all physcisians}}$$

then the COC index for each study physician is computed as the mean $COC_{\text{per patient}}$ over all his/her patients.

2.2 Data

We used medical licensing and billing data for a cohort of 864 primary care physicians who passed the Quebec family medicine certification examination (QLEX) between 1990 and 1993 and entered practice in the "fee-for-service" scheme in Quebec, Canada before 1998. Four administrative databases, linked by encrypted beneficiary identifiers, were used to assess practice performance by analyzing billing by the physicians for their 3.4 million patients treated in the universal health care system in Quebec over the first four years of practice. The registrant database provides patient age, sex, postal code, and date of birth. The medical services database provides the type, location, diagnosis, treating and referring physician, and date of all services delivered on a fee-for-service basis by Quebec physicians. For each physician, the medical services claims files were used to identify all patients seen by the physician from the earliest licensure to four years later. To minimize the variation of productivity due to the effect of years of experience in practice, we have considered the year in practice as unit of time. This means that the first year of practice for study physicians varies in calender years. We also considered people in the same salary scheme to avoid variation in practice volume due to the salary effect [Kinder, 2001; Weiner, 2004].

2.3 Model Specification

The set of variables considered in the model, their description and their corresponding values for mean and standard deviation are displayed in Table 1. the outcome variables considered as indicators of productivity include continuity of care, days, and visits. The rest of the variables are chosen based on their potential to predict the outcome measures. All variables were categorized based on their empirical distribution. We categorized the certification exam score variable into quartiles. The continuity of care variable was categorized as: high index > 0.2 and low index < 0.2. Number of days worked were broken into three levels: high > 220, medium $144 - 220$, and low < 144; and number of patient visits as: high > 3000, medium $1000 - 3000$, and low < 1000. For the geographical place of practice, we tracked all of the cohorts by the use of Metropolitan Influence Zones (MIZs) in order to determine whether the majority of time were worked at an urban, rural, or suburban establishment. Four post-graduate training centers in Quebec were selected as categories for the GradSchool variable, along with two additional categories, one for other north American schools and one for the other international medical graduates (IMGs).

Current structure learning methods require variables with no missing values in the data set. Therefore, we initially found the skeleton structure from the data with no missing values using Max-Min Hill-Climbing (MMHC) [Tsamardinos *et al.*, 2006] algorithm. Then, we refined the structure with the help of domain experts to capture the domain knowledge about associations among the variables. More precisely, we reversed the direction of the edges based on the Marvok equivalence property where possible, in order to illustrate the causalities. The model obtained is shown in Figure 1. The network parameters were learned using the EM algorithm in NeticaTM version 2.0.

Table 1: Variables used in the model for 864 MDs.

Variable	Description	Range	Mean(std)
AGE	age at certification exam	[20,42]	26.2 (3.3)
Gender	gender	binary	N/A
GradSchool	institution of graduation	{1,...,6}	N/A
Score	z-score of certification exam	[-3.97 ,2.75]	-0.041(1.06)
Geographical	geographical place of practice	{urban,suburban,rural}	N/A
COCIndex	continuity of care index	[0,0.58]	0.25 (0.06)
Days	number of work days per year	[0,346]	183 (100)
Visits	number of patient visits	[0,8573]	2417 (1523)

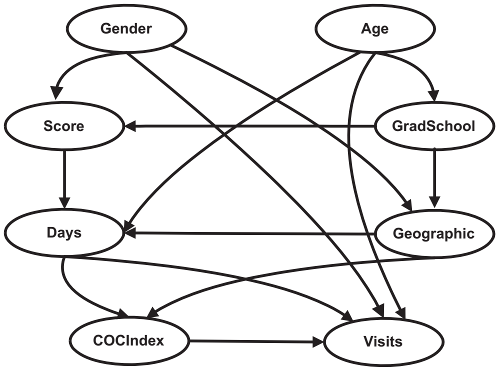


Figure 1: The Bayesian network model for productivity

3 Results

We illustrate here three class of tasks which are possible to perform by the network.

3.1 Predictions and Predictive Evaluation

In order to estimate future services produced by physicians with particular characteristics, we need to predict the outcome measures of productivity. To gain confidence in the network in terms of prediction and obtain generalizable results, we used the network to predict different outcome measures under three network conditions: (1) three outcome variables were included and we made predictions about one outcome at a time, $O3P1$; (2) three outcome variables were included and we made predictions about all three outcomes at the same time, $O3P3$; (3) only one outcome variable was included in the network and we made predictions about that outcome, $O1P1$. The prediction results obtained with the BNs were compared with the results obtained from a Naive Bayes model (NB) and a logistic regression model (LR)[Hosmer and Lemeshow, 2000], that included all covariates and one outcome measure. The assumptions in NB and LR will be violated if we have more than one outcome measure in a single model. Therefore, for each outcome variable we should have a separate model and compare the results with BNs. This is the reason why we considered $O3P3$ and $O1P1$ to have a fare comparison with LR and NB. We performed this comparative evaluation only for the continuity of care index outcome here. We used 10-fold cross validation for prediction of outcome measures using different models. For regression models, we imputed missing values using multivariate imputation by chained equations [Oudshoorn *et al.*, 1999] and randomly sampled from observed values within a column. A logistic regression model was fit to each training set with the COCIndex variable as the outcome and all variables other than visits and days worked as predictors. The inclusion of age, geographical information, and graduate school in the LR model significantly improved the fit of the model while score, and gender did not. The estimated regression coefficients were used to predict the COCIndex variable for each person in the test set. Figure 2 presents the mean of 10 independent runs of 10-fold cross-validation experiments for prediction accuracy with all the models using the receiver operating characteristic (ROC) curves. The area under an ROC curve (AUC) was also considered, as an alternative to error rates, to measure the quality of models. Different models may have the same error rate but different AUC values (the relationship between the AUC values and error rates can be found in [Cortes and Mohri, 2004]). The AUC results related to different models are shown in Figure 2.

These results illustrate the benefits that Bayesian networks offer over NB and LR. The $O3P1$ model outperforms other cases as the information on other outcomes are taken into account. The results also imply that BNs still have greater prediction accuracy than NB and LR in the other two conditions. The three simultaneous outcome predictions propagates more uncertainty through the network than prediction of one outcome. Therefore, the result of the AUC value for BNs on $O3P3$ is slightly lower than the one for $O1P1$, although the difference is larger in their error rates. The results for LR and NB are close in measuring the error rate as we expected (40% for NB versus 42% for LR). However, NB seems to perform more poorly than LR in terms of AUC statistics. This is because algorithms which are designed to minimize the error rate may not lead to the best possible AUC values in general [Cortes and Mohri, 2004]. Table 2 presents the error rate for multiple classifications by different network conditions and NB. Between the three outcome measures, days of work seems to be the most difficult to predict and this is consistent among all models. The $O3P1$ has the best results for predicting all outcome measures.

3.2 Identifying patterns

In addition to prediction of a particular variable, the BN model can be used to identify significant practice patterns

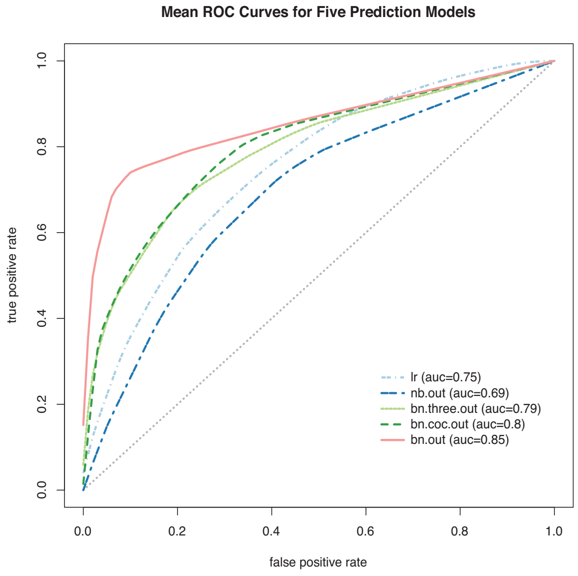


Figure 2: Performance evaluation of different models by their ROC curves and their AUC values. The AUC of the naive Bayes model=0.69, logistic regression=0.75, BN with simultaneous prediction of the three outcomes=0.79, BN with prediction of COC while other outcome variables are not included in the network=0.8, and BN with prediction of COC while other outcome variables are included in the network=0.85

Table 2: Error rate for predicting three indicators of productivity using different models

Indicator	Model			
	O3P1	O1P1	O3P3	NB
COCIndex	16%	24%	28%	40%
Days	25%	35 %	36 %	64%
Visits	22%	30%	34%	58%

and variations around key variables such as continuity of care, practice volume, and geographical distributions during years of practice. Particular patterns of interest include: population patterns e.g. aging and geographical distributions; provider patterns e.g. IMGs and gender; behavioral patterns e.g. adaptation to practice. As there is a considerable variance in the practice volume and continuity of care over the physicians population, experts are often interested in identifying the major sources of variance and loss of practice efficiency [Miller *et al.*, 2001; O'Neill and Kuder, 2005]. Another important pattern is the increasing percentage of female physicians. Quebec has the highest percentage of women entering medicine in North America.

We examined patterns by age and gender as these two demographic features generally influence workforce planning. In this study, we considered the distribution of outcome variables among age-gender groups. Our results show that women tend to work for fewer number of days on average (male: 199± 97 and female 171 ± 100). While this is consistent with the current understanding of the system according to the domain experts, the results show that not all women have a lower level of productivity than all men. Our analysis on age-gender groups show in particular, women in certain age group $27-33$ are more likely to work for lower number of days and to produce lower number of services, which perhaps is related to changes in their family status. Younger females and younger males are more likely to have similar productivity levels, however, our findings show that their likely geographical distribution varies significantly. For instance, males less than 24 years of age (at the graduation time) tended to be more active in rural areas than their female counterparts.

Table 3: Policies for recruiting 100 MDs

Action	Result
IMGs only	Days → +16% Visits → +19%
Uniform gender distribution	Days → +14% Visits → +18% Coc → +4%

3.3 Implication in policy making

The real power of using Bayesian networks for productivity analysis lies in the capability to perform what-if impact analysis on the productivity measures. That is, to determine the potential impact of different policies on productivity. For demonstration purposes we assume that it is possible to add 100 physicians to the current distribution. We need to analyze the impact of different policies for recruiting and deploying these physicians. Using our network of Fig.1. we performed simulations to determine productivity under policies shown in Table 3. The assumptions, which can be translated as actions taken by the policy, are provided in the first column of this table and in the second column we observe the resulting changes in the outcome measures. If we add 100 international medical graduates, IMGs, to the current distribution of physicians, then total worked days over all physicians increases by 16%, and total visits over all physicians increases by 19%. Whereas, if we add 50 male and 50 female physicians, then the total worked days over all physicians increases by 14% and total visits over all physicians increases by 18%. This is while the continuity of care increases by 4%.

It is also possible to conduct exploratory testing through "what-if" analysis by instantiating selected nodes and observing how the probabilities of other nodes change. To illustrate, we can inquire at the network, for instance: if physicians generate high number of visits (> 3000 visits per year), then what would be the likely gender distribution of these physicians? We then observe 64% male and 36% female. We might also be interested in the geographical distribution these physicians would likely follow, and we observe 21% rural , 31% suburban , and 48% urban. It should be noted that BNs can also be used as an automated decision making tool by introducing a set of actions and defining the utility of those actions using the framework of influence diagrams. In the current application, however, it is not straightforward to define estimates of utilities for possible workforce actions. If one developed a demand model to define services and resources required for a particular population, and a model for estimation of cost per MD hiring/training for each policy action, then it would be possible to estimate utilities and use the resulting influence diagram as a decision support tool.

4 Discussion

The BN methodology presented is capable of assessing productivity-influencing profiles and makes it possible to evaluate the impact of different health care policies efficiently. Using a BN in this context allows both quantitative and qualitative data to be incorporated in the process in order to project the supply of services based on real data, not based on assumptions and ad hoc strategies. A Bayesian network can also provide feedback to an analyst on forecasts and help reduce bias in those forecasts. Our experiments illustrate how different results in services produced can be achieved through different policies with a constant increase in the number of primary care providers.

The empirical results obtained for predicting productivity reveal that Bayesian networks can give better predictions than logistic regression models. While there is currently no general characterization of the class of problems in which one of these techniques work better than the other, this is considered a domain-dependent issue (empirical comparison of these methods on several UCI databases and the intrinsic differences between learning the logistic regression model (maximizing the conditional likelihood) and the optimal Bayesian network (maximizing the full likelihood) are discussed in [Grossman and Domingos, 2004; Mitchell, 2005; Roos *et al.*, 2005]). In this study, we only selected available variables which where expected to be interesting from the view point of health services. We intend to explore broader range of data and extra variables such as: choice of practice setting (i.e. CLSC clinics, outpatients, and etc), length of time spend per patient, service line (i.e. proportion of patients in different service groups including pediatrics, mother and child, elderlies, and etc). It should be noted that the inclusion of other extra variables, which contain useful information about productivity indicators, such as family status or financial status may result in a different mode and perhaps more accurate predictions. However, this type of data is not available.

In analysis of observational data in general, confounding is an important issue. The framework of graphical models has been used with increasing frequency for model interpretability, for multi-way analysis, and to address concerns of confounding in a variety of research settings. Bayesian networks, as causal directed acyclic graphs, can control for confounding variables through the detection of independencies and algorithms to determine conditional dependencies on a set of variables. In particular, moralization for a node, which is connecting all its parents by an edge, is easily performed in BNs to control for confounding. This is an advantage over statistical models in which confounding has to be adjusted by complicated statistical techniques such as matching, stratification, multi-variable adjustment, and propensity scores.

In a deployed implementation of this model, one faces problems such as: computational complexity, limited and incomplete data, feature selection, and good resolution for the variables. In this study we relied upon expert opinion to discretize some of the variables. Discretization causes loss of information, in general, and results will be sensitive to the discretization thresholds. It is particularly important in our case due to the mixture of discrete and continuous variables. Learning a linear regression model for discrete variable given continuous parents is proposed as a remedy in this case. Although, the advantage is that it works with no information loss caused by discretization, it can only model linear dependencies.

Physicians' productivity depends in part on the patient population. Future physician requirements, must take into account likely changes in the volume, mix, and case acuity of patient workload resulting from the aging of the population and other demographic and administrative factors. New measures of physician services related to the quality of care outcomes are needed as well. Future productivity measures also need to account for all that a physician does in practice and for participation in activities that are not easily valued in an objective way. This work is currently being extended to include additional variables that will help to incorporate the additional determinants of productivity. Bayesian networks can scale well for the purpose of this application. We are also interested in combining multiple element models at the patient level, physician level and service level into an overall model by relational Bayesian modeling. While computational costs of the compact representations were low for our problem, we believe these may be a bottleneck for larger models, so this should be an area of further research and more sophisticated learning methods will be required to address this problem.

5 Conclusions

This paper discussed some of the key determinants of productivity and introduced the use of a Bayesian network methodology for productivity analysis. The increase in the intake of medical schools, which take seven years to produce family physicians, is not sufficient for filling the gap between supply and demand for primary care services. Efficient planning together with increasing physicians productivity will help achieve this goal in a more efficient manner. The proposed framework provides a systematic method to represent uncertain information that analysts must manage for efficient planning. Future research will include linking results from extended productivity network analysis to quality assessment and decision making.

References

[Altarum, 2000] Altarum. Report of work performed on the physician supply model. Report by Altarum Institute to the National Center for Health Workforce Analysis, Bureau of Health Professions, Health Resources and Services Administration, 2000.

[Birch *et al.*, 2007] S. Birch, G. Kephart, G. Tomblin-Murphy, L. O'Brien-Pallas, R. Alder, and A. MacKenzie. Human resource planning and the production of health: A need-based analytical framework. *Canadian Public Policy*, (Special number), 2007.

[Chan, 2003] B. Chan. Physician workforce planning: what have we learned? lessons for planning medical school capacity and img policies: the canadian perspective. In *International Medical Workforce Conference*, 2003.

[Cortes and Mohri, 2004] C. Cortes and M. Mohri. Auc optimization vs. error rate minimization. In *NIPS*, 2004.

[Curson, 2003] J. Curson. Physician workforce planning: what have we learned? lessons for planning medical school capacity and img policies. In *International Medical Workforce Conference*, 2003.

[DHHS, 2006] U.S. DHHS. Department of health and human services. physician supply and demand: Projections to 2020, October 2006.

[Dreyer *et al.*, 2007] JF. Dreyer, GS. Zaric, SL. McLeod, C. Anderson, and ME. Carter. Predictors of emergency physician workload. *Acad. Emerg. Med.*, May(14), 2007.

[Grossman and Domingos, 2004] Daniel Grossman and Pedro Domingos. Learning bayesian network classifiers by maximizing conditional likelihood. In *ICML*, 2004.

[Hosmer and Lemeshow, 2000] D. Hosmer and S. Lemeshow. *Applied logistic regression*. Wiley, 2000.

[Hutcheson and Sofroniou, 1999] G. Hutcheson and N. Sofroniou. *The multivariate social scientist: Introductory statistics using generalized linear models*. Sage Publications, 1999.

[Jensen, 1996] FA. Jensen. *An introduction to bayesian networks*. Springer, 1996.

[Joyce *et al.*, 2004] CM. Joyce, JJ. McNeil, and JU. Stoelwinder. Time for a new approach to medical workforce planning. *MJA*, 180(7):343–346, 2004.

[Kinder, 2001] K. Kinder. Assessing the impact of payment method and practice setting on german physicians' practice patterns. *Ambul. Care Manage*, 24(2):11–18, 2001.

[Lewin, 2006] Lewin. The physician workforce: Projections and research into current issues affecting supply and demand. Lewin Group report prepared for the Bureau of Health Professions, Health Resources and Services Administratio, 2006.

[Miller *et al.*, 2001] WL. Miller, RR. McDaniel, BF. Crabtree, and KC. Stange. Practice jazz: understanding variation in family practices using complexity science. *J. Fam. Prac.*, 50(10):872877, 2001.

[Mitchell, 2005] Tom M. Mitchell. Generative and discriminative classifiers: Naive bayes and logistic regression. *Machine Learning*, 1:1–17, 2005.

[Moore, 2002] KJ. Moore. A productivity primer. *Family Practice Management*, 9(5):72–73, 2002.

[O'Connor, 1993] K. O'Connor. Management implication of physician practice patterns: Strategies for managers. *Hospital and Health Serv. Admin.*, 38(4):572–577, 1993.

[O'Neill and Kuder, 2005] L. O'Neill and J. Kuder. Explaining variation in physician practice patterns and their propensities to recommend services. *Med. Care Res. Rev.*, 62(3):339–357, 2005.

[Oudshoorn *et al.*, 1999] C. Oudshoorn, V. Buuren, , and V. Rijckevorsel. Flexible multiple imputation by chained equations of the avo-95 survey, 1999.

[Pearl, 1988] J. Pearl. *Probabilistic reasoning in intelligent systems : Networks of plausible inference*. Morgan Kaufmann, 1988.

[Roos *et al.*, 2005] Teemu Roos, Hannes Wettig, Peter Grunwald, Petri Myllymaki, and Henry Tirri. On discriminative bayesian network classifiers and logistic regression. *Machine Learning*, (59):267–296, 2005.

[Tsamardinos *et al.*, 2006] I. Tsamardinos, L. Brown, and C. Aliferis. The max-min hill-climbing bayesian network structure learning algorithm. *Machine Learning*, (65):3178, 2006.

[Weiner, 2002] JP. Weiner. A shortage of physicians or a surplus of assumptions? *Health Affairs*, 21(1):160162, 2002.

[Weiner, 2004] JP. Weiner. Prepaid group practice staffing and us physician supply: lessons for workforce policy. *Health Affairs*, 2004.

Intelligent Systems for Interferon Treatment Decision Support in Chronic Hepatitis C Based on i-Biopsy

Alexandru George Floares
SAIA - Solution of Artificial Intelligence Applications
Str. Vlahuta, Bloc Lama C, Ap. 45, 400310
Cluj-Napoca, Romania
alexandru.floares@gmail.com

Abstract

In chronic hepatitis C, which can progress to cirrhosis and liver cancer, Interferon is the only effective treatment, for carefully selected patients, but it is very expensive. Some of the selection criteria are based on liver biopsy, an invasive, costly and painful medical procedure. Developing an efficient selection system, based on non-invasive medical procedures, could be in the patients benefit and also save money. We investigated the capability of a knowledge discovery in data approach, to integrate information from various non-invasive data sources - imaging, clinical, and laboratory data - to assist the interferon therapeutical decision, mainly by predicting with acceptable accuracy the results of the biopsy. The resulted intelligent systems, tested on 700 patients with chronic hepatitis C, based on C5.0 decision trees and boosting, predict with 100% accuracy the results of the liver biopsy. Also, by integrating other patients selection criteria, they offer a non-invasive support for the correct Interferon therapeutic decision. To our best knowledge, these decision systems outperformed all similar systems published in the literature and offer a realistic opportunity to replace liver biopsy in this medical context.

1 Introduction and Medical Background

The hepatitis C virus is one of the most important causes of chronic liver disease. It accounts for about 15% of acute viral hepatitis, 60% to 70% of chronic hepatitis, and up to 50% of cirrhosis, end-stage liver disease, and liver cancer. An estimated 150-200 million people worldwide are infected with hepatitis C.

Activity (necroinflammation) and fibrosis are two major histologic features of chronic hepatitis C included in the most used scoring systems, METAVIR and Ishak. These systems assess histologic lesions in chronic hepatitis C using two separate scores, one for necroinflammatory grade - METAVIR A (A for activity) or Ishak NI (NI for necroinflammatory) and another for the stage of fibrosis (F) - METAVIR F or Ishak F.

Liver biopsy is the gold standard for grading the severity of disease and staging the degree of fibrosis and the grade of necroinflammation. permanent architectural damage. Liver biopsy is invasive and usually painful; complications severe enough to require hospitalization can occur in approximately 4% of patients [Lindor, 1996]. In a review of over 68,000 patients recovering from liver biopsy, 96% experienced adverse symptoms during the first 24 hours of recovery. Hemorrhage was the most common symptom, but infections also occurred. Side effects of the biopsies included pain, tenderness, internal bleeding, pneumothorax, and rarely, death [Tobkes and Nord, 1995].

Transient elastography (FibroScan®) is an ultrasound imaging technique used to quantify hepatic fibrosis in a totally non-invasive and painless manner. It performs well in identifying severe fibrosis or cirrhosis, but is less accurate in identifying lower degrees of fibrosis.

In chronic hepatitis C, which can progress to cirrhosis, and liver cancer, Interferon is the only effective treatment, in carefully selected patients. Unfortunately, Interferon treatment is very expensive and the patients' selection criteria include fibrosis and necroinflammation assessed by liver biopsy, an invasive medical procedure.

As an example, the Romanian Ministry of Health's criteria, for selecting the patients with chronic hepatitis C who will benefit from Interferon treatment, are:

1. Chronic infection with HCV: antibodies against HCV (anti-HCV) are present for at least 3 months.
2. The cytolytic syndrome: the transaminases level is increased or normal.
3. Pathology (*biopsy*): the Ishak NI ≥ 4 and Ishak F ≤ 3.
4. The virus is replicating: the transaminases level is increased or normal, and anti-HCV are present, ant RNA-HCV $\geq 10^5$ copies/mililitter.

Developing an efficient selection system, based on non-invasive medical procedures, is important for the patients' benefit and could also save money.

The main questions are:

1. Can we extract and integrate information from various (non-invasive) sources, e.g. imaging, clinical, and laboratory data, to build systems capable to predict the biopsy results - fibrosis stage and necroinflammation grade - with an acceptable 90%-100% accuracy?
2. Can we integrate these predictions with other selection criteria, in a system capable to support the correct interferon treatment decision?

3. Can we quantify the end results of the Interferon treatment and use them in a system capable to identify the important selection criteria and their cutoff values?

For any traditional medical approach, the answer to this important question is *NO*. A knowledge discovery in data or data mining approach, based on computational intelligence tools, could be the foundation for a positive answer to the above and similar important medical questions. As it will be shown, the extraction and integration of information from various data sources is indeed possible, and the prediction accuracy of the resulted intelligent systems could even reach 100%. Also, the intelligent system for Interferon treatment decision support can be built and is effective.

In this way, an important medical protocol or workflow for patients management - Interferon treatment decision in chronic hepatitis C - is integrated with intelligent agents or modules. By letting this agents to learn the prediction of the end results of the Interferon treatment, they could reveal the biomedical variables correlated to various degree of treatment response, and also their cutoff values, delimiting the response patients'subgroups (work in progress). We developed a similar intelligent system for Interferon treatment decision support in chronic hepatic B (Floares, 2008 - submitted to "International Conference on Medical Informatics and Biomedical Engineering", Venice, Italy, October 29-31, 2008).

By far the most difficult problem of these investigations consists in predicting the results of liver biopsy [Floares *et al.*, 2008a], [Floares *et al.*, 2008b], (Floares, 2008 - accepted at "Intelligent Systems for Medical Decisions Support", CIBB 2008, 3-4 October, 2008, Vietri sul Mare, Salerno, Italy). We used several non-invasive approaches - routine laboratory tests and basic ultrasonographic features - with and without liver stiffness measurement by transient elastography (FibroScan®), to build intelligent systems for staging liver fibrosis and the grade of necroinflammation in chronic hepatitis C.

To the best of our knowledge, this is the first intelligent system, to support Interferon treatment decision in chronic hepatitis C, developed by integrating intelligent agents (modules) in the medical workflow, capable to predict the fibrosis stage and necroinflammatory degree with the highest published accuracy (100%). The fact that we reached similar results for hepatitis B and also in a different but similar problem - predicting prostate biopsy results in prostate cancer to support surgical treatment decisions (Floares et al., 2008 - accepted at Workshop on Computers in Medical Diagnoses, IEEE International Conference on Intelligent Computer Communication and Processing, August 28 - 30, 2008, Cluj-Napoca, Romania), corroborate our believe that this approach can become a standard one.

2 Building Intelligent Systems for Interferon Treatment Decision Support

2.1 Data Integration and Data Preprocessing

One of the key aspect of intelligent data analysis is in our opinion the integrating various medical data sources: clinical, imaging and lab data. Our experiments showed that isolated data sources do not usually contain enough information for building accurate intelligent systems. The main problems we found, in mining the medical data bases, were the small number of patients relative to the number of features, and the large extent of missing data. However, comparing to similar medical studies our dataset was quite large, with hundreds of patients.

The order of the pre-processing steps is important. Due to the above mentioned problems, one should avoid as much as possible the elimination of patients form the analysis during data pre-processing, and try to eliminate uninformative features first. If feature selection is performed first, even without using sophisticated methods for missing data imputation, the number of eliminated cases is smaller. For a recent exhaustive collection of feature selection methods see [Guyon *et al.*, 2006].

Feature selections was performed in three steps:

1. Cleaning. Unimportant and problematic features and patients were removed.
2. Ranking. The remaining features were sorted and ranks were assigned based on importance.
3. Selecting. The subset of features to use in subsequent models was identified.

In data cleaning, we always removed or excluded from the analysis the following variables:

- variables that have all missing values,
- variables that have all constant values,
- variables that represent case ID.

The following cases were always removed:

- cases that have missing target values,
- cases that have missing values in all its features.

The following variables were also removed:

1. Variables that have more than 70% missing values.
2. Categorical variables that have a single category counting for more than 90% cases.
3. Continuous variables that have very small standard deviation (almost constants).
4. Continuous variables that have a coefficient of variation CV < 0.1 (CV = standard deviation/mean).
5. Categorical variables that have a number of categories greater than 95% of the cases.

For ranking the features, "predictor" an important step of feature selection, also important for understanding the biomedical problem, we used a simple but effective method which considers one feature at a time, to see how well each feature alone predicts the target variable. For each feature, the value of its importance is calculated as (1 - p), where p is the p value of the corresponding statistical test of association between the candidate feature and the target variable. The target variable was categorical with more than two categories for all investigated problems, and the features were mixed, continuous and categorical.

For categorical variables, the p value was based on Pearson's Chi-square altfel, fara Pearson test of independence between X, the feature under consideration with I categories, and Y target variable with J categories. The Chi-square test involves the difference between the observed and expected frequencies. Under the null hypothesis of independence, the expected frequencies are estimated by $\widehat{N} = N_i \cdot N_j/N$. Under the null hypothesis, Pearson's chi-square converges asymptotically to a chi-squared distribution χ_d^2 with degree of freedom $d = (I-1)(J-1)$, and the p value is equal with the probability that $\chi_d^2 > X^2$, where $X^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}(N_{ij} - \widehat{N}_{ij})^2/\widehat{N}_{ij}$. The categorical variables were sorted first by p value in the ascending order, and if ties occurred they were sorted by chi-squared in descending order. If ties still occurred, they were sorted by degree of freedom d in ascending order.

For the continuous variables, p values based on the F statistic are used. For each continuous variable a one-way ANOVA F test is performed to see if all the different classes of Y have the same mean as X. The p value based on F statistic is calculated as the probability that $F(J-1, N-J) > F$, where $F(J-1, N-J)$ is a random variable that follows and F distribution with degrees of freedom $J-1$ and $N-J$, and

$$F = \frac{\sum_{j=1}^{J} N_j(\overline{x} - \overline{\overline{x})^2/(J-1)}}{\sum_{j=1}^{J}(N_j - 1s_j^2/(N-J)} \tag{1}$$

If the denominator for a feature was zero, the p value of that feature was set to zero. The features were ranked first by sorting them by p value in ascending order, and if ties occurred, they were sorted by F in descending order. If ties still occurred, they were sorted by N in descending order.

Based on the features' importance $(1-p)$, with p calculated as explained above, we ranked and grouped features in three categories:

1. important features, with $(1-p)$ between 0.95 and 1,
2. moderately important features, with $(1-p)$ between 0.90 and 0.95, and
3. unimportant features, with $(1-p)$ less than 0.90.

Some of the categorical features and also the target categorical variable have imbalanced distributions, and this can cause some modeling algorithms to perform poorly. We tested the influence on the prediction accuracy of several methods for dealing with imbalanced data (see [Hulse *et al.*, 2007] for a recent comprehensive review). Because the number of patients is small relative to the number of features, a very common situation in biomedical data bases, we only used oversampling methods and not undersampling methods. We also found that simple techniques such as random oversampling perform better than the "intelligent" sampling techniques. An exhaustive comparison of these methods ca be found in [Hulse *et al.*, 2007].

2.2 Intelligent Systems as Ensemble of Classifiers

For modeling, we first tested the fibrosis and necroinflammation prediction accuracy of various methods:

1. Neural Networks
2. C5.0 decision trees
3. Classification and Regression Trees
4. Support Vector Machines, expresia consacrata
5. Bayesian Networks.

Because physicians prefer white-box algorithms, we have chosen C5.0 decision trees, the last and improved version of the C4.5 algorithm [Quinlan, 1993], with 10-fold cross-validation.

Breiman's bagging [Breiman, 1996] and Freund and Schapire's boosting [Freund and Schapire, 1997] are examples of methods for improving the predictive power of classifier learning systems. Both form a set of classifiers that are combined by voting, bagging by generating replicated bootstrap samples of the data, and boosting by adjusting the weights of training cases. While both approaches improve predictive accuracy, boosting showed sometimes greater benefit. Unfortunately, boosting doesn't always help, and when the training cases are noisy, boosting can actually reduce classification accuracy. Naturally, it took longer to produce boosted classifiers, but the results often justified the additional computation. Boosting should always be tried when peak predictive accuracy is required, especially when unboosted classifiers are already quite accurate.

Boosting combines many low-accuracy classifiers (weak learners) to create a high-accuracy classifier (strong learner). We used a boosting version called *AdaBoost*, with reweighting; AdaBoost comes from ADAptive BOOSTing [Freund and Schapire, 1997].

Suppose we are given the training set data (X_1, F_1),..., (X_n, F_n), where n is the number of patients, the input $X_i \in \Re^p$ represents the p selected features in the preprocessing steps (image, laboratory data, etc.), and the categorical output F_i is the fibrosis stage (things are similar if necroinflammation is the output) according to one of the two scoring systems Metavir F and Ishak F, and assumes values in a finite set {F0, F1,..., Fk}, were k = 5 for Metavir F (from Metavir F0 to Metavir F4) and k = 7 for Ishak F (from Ishak F0 to Ishak F6). The goal is to find a classification rule $F(\mathbf{X})$ from the training data, so that given a new patient's input vector **X**, we can assign it a fibrosis degree F from {F0, F1,..., Fk} according to the corresponding scoring systems.

Moreover, we want to find the best possible classification rule achieving the lowest misclassification error rate. We assumed that the patients' training data are independently and identically distributed samples from an unknown distribution. Starting with the unweighted training sample, the AdaBoost builds a classifier which can be a neural network, decision tree, etc., that produces class labels - fibrosis degree. If a training data point (patient) is misclassified, the weight of that training patient is increased (boosted). A second classifier is built using the new weights, which are now different. Again, misclassified training patients have their weights boosted and the procedure is repeated. Usually, one may build hundred of classifiers this way. A score is assigned to each classifier, and the final classifier is defined as the linear combination of the classifiers from each stage.

With the above notations, and noting with I an indicator function, a compact description of the AdaBoost algorithm used is the following:

1. Initialize the patient weights $\omega_i = 1/n, i = 1, 2, \ldots, n$.
2. For $m = 1$ to M:
   (a) Fit a classifier $F^{(m)}(\mathbf{x})$ to training patients using weights ω_i.
   (b) Compute
   $$err^m = \sum_{i=1}^{n} \omega_i I(F_i \neq F^{(m)}(\mathbf{X}_i)) / \sum_{i=1}^{n} \omega_i. \quad (2)$$
   (c) Compute
   $$\alpha^{(m)} = \log \frac{1 - err^{(m)}}{err^{(m)}}. \quad (3)$$
   (d) Set
   $$\omega_i \leftarrow \omega_i \cdot \exp(\alpha^{(m)} \cdot I(F_i \neq F^{(m)}(\mathbf{X}_i)) \quad i = 1, 2, \ldots, n. \quad (4)$$
   (e) Re-normalize ω_i.
3. Output
   $$F(\mathbf{X}) = \arg\max \sum_{m=1}^{M} \alpha^{(m)} \cdot I(F^{(m)}(\mathbf{X}) = k). \quad (5)$$

For two-class classification problems AdaBoost could be very successful in producing accurate classifiers. The multi-class classification is more involved, and some technical criteria must be satisfied and experiments need to be done. While fibrosis stage and necroinflammation degree prediction are multi-class classification problems, the Interferon treatment decision is a binary one.

Therefore, it will be advantageous to take into account the largely accepted cutoff values for:

- fibrosis stage, e.g., Ishak F ≤ 3, and to build an intelligent system capable to predict if the fibrosis stage is either Ishak F ≤ 3 or Ishak F > 3,
- necroinflammatory degree, e.g., Ishak NI ≥ 4, and to build an intelligent system capable to predict if the necroinflammatory degree is either Ishak NI ≥ 4 or Ishak $F < 4$.

The intelligent system for the Interferon treatment decision support takes as inputs the outputs of the above systems. The decision is again binary, recommending or not the Interferon treatment. For the positive decision a series of other criteria, presented in section 1, must be satisfied. The proposed methodology is by no means restricted to the Romanian Ministry of Health's criteria, or even to this problem. On the contrary, we believe that this is a rather general methodology for building intelligent systems for medical decisions support.

The final intelligent systems are the result of a more detailed data mining predictive modeling strategy which is patented now, consisting mainly in:

- Extracting and integration information from various medical data sources, after a laborious preprocessing:
  - cleaning features and patients,
  - various treating of missing data,
  - ranking features,
  - selecting features,
  - balancing data.
- Testing various classifiers or predictive modeling algorithms.
- Testing various methods of combining classifiers.

2.3 Intelligent Virtual Biopsy and Intelligent Scoring Systems

Replacing painful, invasive, and/or costly procedures with intelligent systems, taking as inputs integrated data from non-invasive, usual, or cheap medical procedures, techniques and tests, and producing as output 90-100% similar results with the replaced techniques, is an important medical goal. We outline some general ideas, terms and concepts to characterize this new exciting enterprize.

The central new concept is *Intelligent Virtual Biopsy* (IVB), which designates an intelligent system capable to predict, with an acceptable accuracy (e.g., 90-100%), the results given by a pathologist, examining the tissue samples from real biopsies, expressed as scores of a largely accepted scoring system. As an alternative term we suggest *intelligent biopsy* or *i-biopsy*, were the term intelligent indicates that the system is based on artificial intelligence. To predict the pathologist's scores, the intelligent systems take as inputs and integrate various non-invasive biomedical data.

Also, to distinguish between the scores of the scoring systems of the real biopsy, and their counterparts predicted by the i-biopsy, we proposed the general term of *i-scores* belonging to *i-scoring systems*. In the gastroenterological context of these investigations, we have the following correspondences:

1. Liver intelligent virtual biopsy (IVB), or *liver i-biopsy* is the intelligent system corresponding to the real liver biopsy.
2. The *i-Metavir F* or *A* and *i-Ishak F* or *NI* correspond to the two liver fibrosis or necroinflammation scoring systems Metavir F or A, and Ishak F or NI respectively.
3. The *i-scores* are the values predicted by the intelligent systems for the fibrosis scores.

From a biomedical point of view, the most important general characteristics of the i-scores are exemplified for the Metavir F and Ishak F scores:

1. I-Metavir F or A and i-Ishak F or NI scores have exactly the same biomedical meaning as Metavir-F or A and Ishak-F or NI, scoring the same pathological features.
2. I-Metavir F or A and i-Ishak F or NI scores are obtained in a non-invasive and painless manner, as opposed to Metavir-F and Ishak-F.
3. The estimation of i-Metavir F or A and i-Ishak F or NI does not have the risks related to Metavir-F or A and Ishak-F or NI estimation via biopsy.

3 Results

We have built the following modules, components of the intelligent system for Interferon treatment decision support:

1. Module for liver fibrosis prediction,
   (a) according to Metavir F scoring system
      i. with liver stiffness (FibroScan®),
      ii. without liver stiffness (FibroScan®)
   (b) according to Ishak F scoring system
      i. with liver stiffness (FibroScan®),
      ii. without liver stiffness (FibroScan®).
2. Module for the grade of necroinflammation (activity) prediction, according to Ishak NI scoring systems

The fibrosis prediction module was first built using a dataset of 381 chronic hepatitis C patients and the METAVIR scoring system [Floares *et al.*, 2008a]. Now, it was tested on 700 chronic hepatitis C patients and the fibrosis is predicted according to METAVIR F or Ishak F scoring system. As we previously mentioned, in the interferon treatment decision system we used the binary version of the fibrosis and necroinflammation classifiers. For the version with liver stiffness, at the end of the preprocessing stage, besides liver stiffness, the relevant features for predicting liver fibrosis, according to Metavir scoring system, were: age, aspartate aminotransferase, gamma-glutamyltranspeptidase, cholesterol, triglycerides, thickness of the gallbladder wall, spleen area and perimeter, left lobe and caudate lobe diameter, liver homogeneity, posterior attenuation of the ultrasound, liver capsule regularity, spleen longitudinal diameter, the maximum subcutaneous fat, perirenal fat. Combining all these features, the intelligent system was able to predict each fibrosis stage with 100% accuracy.

In the mean time we have tried to reduce the number of features to at most ten, without sacrificing the accuracy, because some of our investigations showed that this is possible [Floares *et al.*, 2008b]; the results are very encouraging (manuscript in preparation).

We also wanted to investigate if it is possible to build intelligent systems, capable to predict fibrosis scores according to Metavir F and Ishak F scoring system, without using apparently a key source of information - the liver stiffness measured with FibroScan®. Such intelligent systems could be useful to those gastroenterology clinics having ultrasound equipment but not the expensive FibroScan®.

After feature selection, the relevant features for Metavir F prediction without Fibroscan®were: cholesterol, caudate lobe diameter, thickness of the abdominal aortic wall, aspartate aminotransferase, preperitoneal fat thickness, splenic vein diameter, time averaged maximum velocity in hepatic artery, time averaged mean velocity in hepatic artery, flow acceleration in hepatic artery, hepatic artery peak systolic velocity. Combining these 10 attributes, the boosted C5.0 decision trees were able to predict each fibrosis stage, according to Metavir F scoring system, with 100% accuracy, even without liver elastography.

The relevant features for predicting liver fibrosis according to Ishak F scoring system were: caudate lobe diameter, left lobe diameter, liver capsule regularity, liver homogeneity, thickness of the abdominal aortic wall, steatosis (ultrasonographic), cholesterol, sideremia, and liver stiffness. The boosted C5.0 decision trees were able to predict each Ishak fibrosis stage with 100% accuracy.

We also built a module for predicting the grade of necroinflammation according only to Ishak NI scoring systems, because Metavir A scoring system is less used. The selected features were: aspartate aminotransferase (ASAT), alanine aminotransferase (ALAT), left liver lobe diameter, hepatic artery acceleration time, hepatic vein Doppler waveform, liver capsule regularity, posterior attenuation of the ultrasound, liver parenchymal echogenity, and hepatic arterial pulsatility index. Combining these 9 attributes, the boosted C5.0 decision trees were able to predict each fibrosis stage, according to Metavir F scoring system, with 100% accuracy, even without liver elastography.

All these models have 100% accuracy, and at the moment of writing this paper, the intelligent systems were tested on 528 patients with chronic hepatitis C.

4 Discussions

The reasons for the relative disproportion between the number of patients and the number of features is that, at the beginning of these investigations, our multi-disciplinary team tried to define a large number of potentially important features. We intended to use a *data-driven* approach avoiding as much as possible restrictive a priori assumptions. Usually, this opens the door for potential surprises, e.g., previously unknown and unexpected relationships between fibrosis and various other biomedical features. There were some unexpected findings (results not shown) but they need further investigations.

The accuracy of the first experiments was about 60%, but a careful pre-processing increased the accuracy of the predictions with 20% to 25%. We tested five algorithms - Neural Networks of various types and architectures, C5.0 decision trees, Classification and Regression Trees, Support Vector Machines, with various kernels, and Bayesian Networks - with the default parameter settings. C5.0 accuracy was the highest, about 80%, and parameter tuning and boosting increase the accuracy to 100%, and AUROC to 1. Preprocessing and boosting drastically increased the accuracy from 60% to 100%.

A short comment about the meaning of 100% diagnostic accuracy seems to be necessary, because it confused many physicians who say that 100% accuracy is not possible in medicine. The meanings will be made clear more easy by means of examples. We have proposed intelligent systems predicting the fibrosis scores resulted from liver biopsy with 100% accuracy. Usually, an invasive liver biopsy is performed and a pathologist analyzes the tissue samples and formulates the diagnostic, expressed as a fibrosis score. The pathologist may have access to other patient's data, but usually these are not necessary for the pathological diagnostic. Moreover, in some studies it is required that the pathologist knows nothing about the patient. His or her diagnostic can be correct or wrong for many reasons, which we do not intend to analyze here. On the contrary, for the intelligent system some of the clinical, imaging and lab data of the patient are essential, because they were somehow incorporated in the system. They were used like features to train the system, and they are required for a new, unseen patient, because the i-biopsy is in fact a relationship between these inputs and the fibrosis scores.

Intelligent systems do not deal directly with diagnostic *correctness*, but with diagnostic prediction accuracy. In other words, the intelligent system will predict, in a non-invasive and painless way, and without the risks of the biopsy, a diagnostic which is 100% identic with the pathologist diagnostic, if the biopsy is performed. While the accuracy and the correctness of the diagnostic are related in a subtle way, they are different concepts. An intelligent system will use the information content of the non-invasive investigations to predict the pathologist diagnostic, without the biopsy. The correctness of the diagnostic is a different matter, despite the fact that a good accuracy is almost sure related with a correct diagnoses, but we will not discuss this subject.

The accuracy of the diagnosis, as well as other performance measures like the area under the receiver operating characteristic (AUROC), for a binary classifier system [Fawcett, 2004], are useful for intelligent systems comparison. From the point of view of accuracy, one of the most important medical criterions, to our best knowledge the proposed liver intelligent virtual biopsy or i-biopsy system outperformed the most popular and accurate system, FibroTest [Poynard *et al.*, 2007] commercialized by Biopredictive company. The liver i-biopsy presented in this paper is based on a five classes classifier, more difficult to build than binary classifiers; we also build binary classifiers as decision trees with 100% accuracy and mathematical models (work in progress, results not shown). Despite the fact that AUROC is only for binary classifiers, loosely speaking a 100% accuracy n classes classifier is equivalent with n binary classifiers with AUROC = 1 (maximal). In [Poynard *et al.*, 2007], a total of 30 studies were included which pooled 6,378 subjects with both FibroTest and biopsy (3,501 chronic hepatitis C). The mean standardized AUROC was 0.85 (0.82-0.87).

Moreover, in some circumstances the result of the liver IVB could be superior to that of real biopsy. When building the intelligent system, the results of the potentially erroneous biopsies, which are not fulfilling some technical requirements, were eliminated from the data set. Thus, the IVB predicted results correspond only to the results of the correctly performed biopsies, while some of the real biopsy results are wrong, because they were not correctly performed. Due to the invasive and unpleasant nature of the biopsy, is very improbable that a patient will accept a technically incorrect biopsy to be repeated. Unlike real biopsy, IVB can be used to evaluate fibrosis evolution, which is of interest in various biomedical and pharmaceutical studies, because, being non-invasive, painless and without any risk, can be repeated as many time as needed. Also, in the early stages of liver diseases, often the symptoms are not really harmful for the patient, but the treatment is more effective then in more advanced fibrosis stages. The physician will hesitate to indicate an invasive, painful and risky liver biopsy, and the patients is not so worried about his or her disease to accept the biopsy. However, IVB can be performed and an early start of the treatment could be much more effective. Moreover, we have obtained high accuracy results for other liver diseases, like chronic hepatitis B and steatohepatitis, for other biopsy findings, like necroinflammatory activity and steatosis (results not shown), and also for prostate biopsy in prostate cancer. These corroborate our believe that this approach can become a standard one.

References

[Breiman, 1996] Leo Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

[Fawcett, 2004] T. Fawcett. Roc graphs: Notes and practical considerations for researchers. technical report. Technical report, Palo Alto, USA: HP Laboratories, 2004.

[Floares *et al.*, 2008a] A. G. Floares, M. Lupsor, H. Stefanescu, Z. Sparchez, A. Serban, T. Suteu, and R. Badea. Toward intelligent virtual biopsy: Using artificial intelligence to predict fibrosis stage in chronic hepatitis c patients without biopsy. *Journal of Hepatology*, 48(2), 2008.

[Floares *et al.*, 2008b] Alexandru Floares, Monica Lupsor, Horia Stefanescu, Zeno Sparchez, Radu Badea, and Romania. Intelligent virtual biopsy can predict fibrosis stage in chronic hepatitis c, combining ultrasonographic and laboratory parameters, with 100% accuracy. *Proceedings of The XXth Congress of European Federation of Societies for Ultrasound in Medicine and Biology*, 2008.

[Freund and Schapire, 1997] Y. Freund and R. E. Schapire. A decisiontheoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

[Guyon *et al.*, 2006] Isabelle Guyon, Steve Gunn, Masoud Nikravesh, and Lofti Zadeh. *Feature Extraction: Foundations and Applications.* Studies in Fuzziness and Soft Computing. Springer, August 2006.

[Hulse *et al.*, 2007] Jason Van Hulse, Taghi M. Khoshgoftaar, and Amri Napolitano. Experimental perspectives on learning from imbalanced data. In *Proceedings of the 24 th International Conference on Machine Learning*, Corvallis, OR, 2007.

[Lindor, 1996] A. Lindor. The role of ultrasonography and automatic-needle biopsy in outpatient percutaneous liver biopsy. *Hepatology*, 23:1079–1083, 1996.

[Poynard *et al.*, 2007] Thierry Poynard, Rachel Morra, Philippe Halfon, Laurent Castera, Vlad Ratziu, Franoise Imbert-Bismut, Sylvie Naveau, Dominique Thabut, Didier Lebrec, Fabien Zoulim, Marc Bourliere, Patrice Cacoub, Djamila Messous, Mona Muntenau, and Victor de Ledinghen. Meta-analyses of fibrotest diagnostic value in chronic liver disease. *BMC Gastroenterology*, 7(40), 2007.

[Quinlan, 1993] J.R. Quinlan. *C4.5 : Programs for Machine Learning.* Morgan Kaufmann, 1993.

[Tobkes and Nord, 1995] A. Tobkes and H. J. Nord. Liver biopsy: Review of methodology and complications. *Digestive Disorders*, 13:267–274, 1995.

An artificial intelligence tool to predict fluid requirement in the ICU: A proof-of-concept study

Leo Anthony Celi MD
Laboratory of Computer Science
Massachusetts General Hospital
50 Staniford Street, Boston, MA, USA 02114
lceli@mit.edu

L. Christian Hinske MD
Decision Systems Group
Brigham and Women's Hospital
900 Commonwealth Ave, Boston, MA, USA 02215
chinske@mit.edu

Gil Alterovitz PhD
Children's Hospital Informatics Program
Children's Hospital
320 Longwood Avenue, Boston, MA, USA 02115
gil@mit.edu

Peter Szolovits PhD
Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
32 Vassar Street, Cambridge, MA, USA 02139
psz@mit.edu

Abstract

We propose a way of personalizing medicine in the ICU by using information derived from the application of artificial intelligence on a high-resolution database. Calculation of maintenance fluid requirement at the height of systemic inflammatory response was selected to investigate the feasibility of this approach. MIMIC II is a database of patients admitted to the Beth Israel Deaconess Medical Center ICU. Patients who were on vasopressors greater than 6 hours on day 1 were identified. Demographic and physiologic variables on day 1 were extracted; the outcome to be predicted is the total amount of fluid given on day 2. We represented the variables by learning a Bayesian network from the underlying data. Using 10-fold cross-validation repeated 100 times, the accuracy of the model was 77.8%. The network generated has a threshold Bayes factor of 7, which translates into $p < .05$ assuming Gaussian distribution of the variables. Based on the model, the probability that a patient will require a certain range of fluid on day 2 can be predicted. By better predicting maintenance fluid requirements based on the previous day's physiologic variables, one might be able to prevent hypotensive episodes requiring fluid boluses during the course of the following day.

1 Introduction

The gold standard in evidence-based medicine is a well-designed, well-executed multi-center prospective randomized controlled trial. However, in the ICU, it would be impossible to perform such a study to determine whether every diagnostic test, monitoring device or treatment intervention leads to improved patient outcomes. Even when such trials are performed and subsequently published, they very rarely, if ever, provide clear evidence upon which to base the management of an individual patient. Patients enrolled in these studies are heterogeneous, and conclusions are valid for the "average" patient. Unfortunately, each patient is unique in terms of how he responds to an intervention. He may not benefit, or worse, may be harmed by a medication, device or procedure that has been shown to correlate with a good patient outcome "on average". In addition, these studies investigate one treatment at a time. In reality, treatments are given simultaneously to a patient in the ICU and interact with each other. The nature of these interactions likely varies from patient to patient, and perhaps even within the same patient at different points in time.

Over the years, we have adopted a multitude of diagnostic tests, monitoring devices and treatments in the ICU based on underpowered studies, in most cases non-randomized, which demonstrate modest benefits on soft clinical endpoints or intermediate outcomes. It is unclear which of these interventions contribute to survival benefit. Despite all the medical advances available in the ICU, fewer than half of patients who experience severe sepsis are alive at 1 year [Yende and Angus, 2007]. Another study found that mortality of pneumococcal bacteremia has not changed over the past 50 years [Trampuz *et al.*, 2004]. Finally, acute renal failure treated in the ICU with renal support therapy still carries a mortality of 64-79%, not significantly changed over the decades [Landoni *et al.*, 2006]. Over the last decade, we have witnessed the electronification of health care delivery, and with it, the creation of large ICU databases of tremendous granularity and resolution. At the same time, the concept of personalized medicine emerged. The goal of personalized medicine is to provide the right treatment to the right patient at the right time. It involves integration of genomics, proteomics, metabolomics, systems biology, bioimaging and other disciplines in order to characterize the uniqueness of a patient and predict his risk of developing a disease or his response to treatment. It is a tool that can potentially optimize care customization in the ICU where it is needed the most, given how sick the patients are and how some treatments can lead to worse clinical outcomes. Unfortunately, the dynamic cytokine and neurohormonal milieu of the critically ill patient alters such processes as gene transcription and drug metabolism, rendering information derived during static non-diseased conditions of limited use. In this paper, we propose an alternative way of personalizing medicine in the ICU using empiric data. We chose prediction of fluid requirement of the critically-ill patient at the height of inflammatory response to explore the feasibility of this approach.

The first 72 hours is critical for ICU patients. Whether the patient is being admitted for sepsis, acute coronary syndrome, multiple traumatic injuries, intracranial hemorrhage, burns, or for post-operative care after open heart surgery or organ transplantation, this period is characterized by systemic inflammatory response fueled by a cytokine storm and is most vulnerable to episodes of hypotension and consequent reduced organ perfusion. Suboptimal fluid management during this critical period leads to release of more inflammatory cytokines and catecholamines that further worsen the hemodynamic status of the patient. As shown in a number of clinical studies, reduced tissue perfusion resulting from fluid under-resuscitation translates into increased illness severity and a longer ICU stay [Rivers *et al.*, 2001; Donati *et al.*, 2007]. In practice, clinicians "guess" the rate of maintenance fluids (usually in the range of 1-3 ml/kg/hr) by estimating fluid loss, a task that is very difficult in a critically ill patient because of the absence of a defined set of rules and guidelines for specific patient subsets in various clinical scenarios. In this study, we set out to see whether we can predict the total amount of fluid administered to a patient on day 2 in the ICU, given the physiologic data from the previous 24 hours.

2 Materials and Methods

The Laboratory of Computational Physiology at Massachusetts Institute of Technology (MIT) developed and maintains the Multi-parameter Intelligent Monitoring for Intensive Care (MIMIC II) database, a high resolution database of ICU patients admitted to the Beth Israel Deaconess Medical Center since 2003 that has been de-identified by removal of all Protected Health Information. An Institutional Review Board (IRB) approval was obtained from MIT to develop and maintain a de-identified ICU database that is made available to the public for research purposes. As this study only involved data modeling outside of a clinical context, no additional IRB approval was sought.

The MIMIC II database currently consists of data from more than 18,000 patients that has been de-identified and formatted to facilitate data-mining. The 3 sources of data are waveform data collected from the bedside monitors, hospital information systems and other third party clinical information systems.

Using the MIMIC II database, patients with significant systemic inflammatory response on admission were identified. We defined these patients as those who were on vasopressors agents for more than 6 hours during the first 24 hours of their admission.

For each patient, we obtained demographic data and physiologic variables during the first 24 hour period in the ICU. These variables included vital signs, those that affect and/or represent total body water, and those that reflect severity of illness. Rather than representing one state of each variable which is typically the worst value in severity scoring systems, we decided to include the following for each variable that we evaluated: all measurements obtained, mean, variance, maximum value, minimum value, number of measurements obtained, and the last measurement during the first 24 hours, as this reflects whether the patient is improving, stable or worsening compared to the worst value. Filtering was performed by deleting values that are outside physiologically feasible range.

Using R software (R version 2.7.2, The R Foundation for Statistical Computing, Auckland, New Zealand), linear regression was performed using stepwise forward selection procedure and with a 2:1 split sample approach (2/3 training data, 1/3 validation data). The total fluid administered during the second 24 hour period administered during the second 24 hour period in the ICU was selected as the outcome variable. Using Bayesware Discoverer (Bayesware Discoverer Version 1.0, Massachusetts Institute of Technology, Cambridge, MA, USA), we constructed a Bayesian network using the variables we have identified and extracted from the MIMIC II database. A maximum number of allowable parents was set at 10, the Bayes factor was set at 7, and the maximum number of states per variable was set at 20. The variables were divided into 4 quartiles according to frequency. Finally, the search order was arranged so that the outcome variable we are interested in, i.e. the total fluid intake for the second 24 hours, would have the largest number of nodes considered as potential parents. The accuracy of the model is calculated using ten-fold cross validation repeated 100 times.

3 Results

There were a total of 3014 patients who were on at least one vasopressor agent for a minimum period of 6 hours during their first 24 hours in the ICU and whose total fluid intake and output were recorded.

The distribution of the total fluid intake during the second day in the ICU, the outcome variable, was skewed towards the lower values (Figure 1).

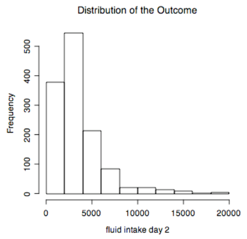


Figure 1. Distribution of Total Fluid Intake on Day 2

The values were therefore log transformed to approximate a more normal distribution for the linear regression model.

Using a stepwise forward variable selection on a 2:1 split sample approach, 14 variables were found to be predictive of the total amount of fluid given to the patient on day 2 in the ICU. The coefficients of the variables in the fitted model, their standard errors and the corresponding p values are shown in table 1.

Variables	Estimate	Standard Error	p value
Number of vasopressors	5.35e+02	2.38e+03	3.25e-06
Maximum heart rate	1.12e+01	3.65	6.43e-09
Maximum hemoglobin	5.71e+01	8.76e+01	0.0021
Minimum hemoglobin	-5.17e+02	9.36e+01	1.03e-10
Variance of hemoglobin	-1.92e+02	4.72e+01	4.01e-08
Total fluid intake on day 1	1.03e-01	2.14e-02	5.09e-05
Total fluid output on day 1	-1.99e-01	3.68e-02	1.50e-06
Most recent platelet count	1.19e+01	2.99	7.90e-08
Number of sedatives	2.74e+02	1.07e+02	7.73e-05
Age	-1.15e+01	4.68	0.0106
Mean platelet count	1.37e+01	2.97	4.17e-06
Minimum serum sodium	1.68e+02	4.21e+01	7.27e-05
Most recent serum sodium	8.33e+01	3.61+01	0.0212
Mean serum sodium	1.67e+02	6.29e+01	0.0080

Table 1. Coefficients of the Fitted Linear Regression Model

R-squared was calculated as a measure of the explained variation accounted for by the linear regression model. It is given by the formula $R^2 = 1 - (SS_{err}/SS_{tot})$ where SS_{err} is the residual sum of squares and SS_{tot} is the sum of squares differences from the mean proportional to the variance. The adjusted R-squared value calculated was 0.25, suggesting that very little of the observed variation can actually be explained by the model. The linear regression model suffered from large variation with a high standard residual error, making it suboptimal for clinical application. For this reason, we shifted to a Bayesian network model to represent our variables and our outcome.

Figure 2 illustrates the Bayesian network model generated from the MIMIC II database. For this particular data set, five variables were found to be correlated with the total fluid intake for the second 24 hours in the ICU: total fluid intake for the first 24 hours, number of vasopressor agents, mean systolic pressure, mean heart rate and mean serum sodium. Based on the model, the probability that a patient will require a certain range of fluid on day 2 (one of four quartiles) can be predicted given the values for the variables that are direct parents of our outcome variable. The accuracy of the model was found to 77.8% on ten-fold cross validation repeated 100 times.

The threshold or minimum Bayes factor is the smallest amount of evidence that can be claimed for the null hypothesis (no correlation between variables), or the strongest evidence against it on the basis of the observed data. This is the benchmark to compare it against a p value. The simplest relation between p values and Bayes factors are based on a Gaussian approximation. In that situation, the minimum Bayes factor is calculated with the same numbers used to calculate a p value [Berger, 1985]. The

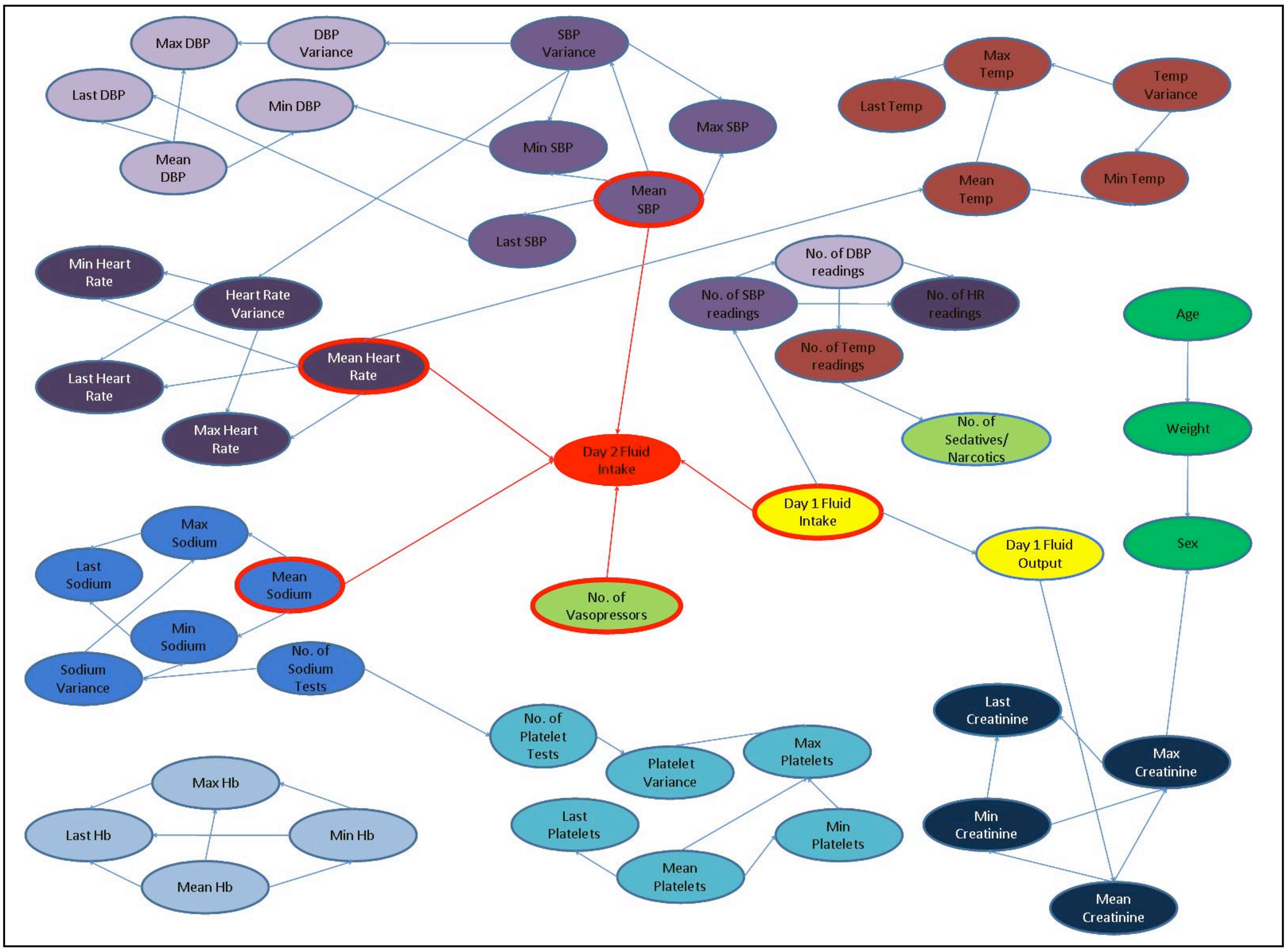


Figure 2. Bayesian network model predicting maintenance fluid requirement on day 2 in the ICU

formula is as follows: [Goodman, 1999]

$$\text{Minimum Bayes factor} = e^{-Z^{\wedge}2/2}$$

where Z is the number of standard errors from the null effect. This formula allows us to establish an exchange rate between minimum Bayes factor and p values in a Gaussian case. For a threshold Bayes factor of 7 which was used to generate our model, the corresponding p value is less than 0.05, assuming Gaussian distribution of the variables.

4 Discussion

Figuring out the fluid requirement to maintain an adequate intravascular volume (and optimal preload) is difficult at the time of critical illness. In practice, clinicians fear over-estimating this fluid requirement. This may contribute to the occurrence of hypotensive episodes especially during the period of maximal systemic inflammatory response. These hypotensive episodes may be averted by being able to predict more accurately the fluid requirement of the patient as the disease process evolves in response to treatment or as a result of healing.

The goal of this proof-of-concept study is to explore the feasibility of supplementing traditional evidence-based medicine, expert opinion and clinical intuition with empiric data. A Bayesian network was generated between physiologic variables obtained during the first 24 hours in the ICU and the total amount of fluid given on the second day in the ICU (maintenance fluid plus all the boluses the patients received) from a large database. A greedy search algorithm was used with the outcome variable of interest being evaluated first for potential parent nodes, and the demographic variables (age, sex and weight) being evaluated last. Given the values of the physiologic variables from day 1, the range of the total fluid given to the patient on day 2 can be predicted. Cross-validation was used to determine how well a Bayesian network model represents our data. In cross-validation, the model generated from the training set is evaluated against a previously unseen data. The accuracy of the model was 77.8%.

We suspect the reason why the accuracy of the Bayesian network generated from the data is not higher relates to the limitations of our methodology. The subset of patients included in the analysis likely still represents a heterogeneous group given that the reason why the patient was on vasoactive drugs was not considered. The patients likewise probably represent a wide spectrum as regards the degree of inflammatory response, with possible inclusion of patients who have minimal amount but got put on vasopressors nonetheless. What we would like to trial in the future once we have a larger database is to specify a more homogeneous group of patients in terms of demographic variables, co-morbidities and clinical scenario.

Another potential source of model inaccuracy in the ICU is data noise. This includes device-related artifacts (e.g. arterial blood pressure dampening), laboratory errors, missing data, and erroneous transcription, to name a few. Filtering during data pre-processing was performed to reduce, but not, obliterate the impact of noise. The choice of the Bayes factor (the likelihood of the model with links between the parent nodes and their children as compared to a model where the variables are independent) is thus crucial in preventing over-fitting when the data is unavoidably noisy.

The biggest limitation of this proof-of-concept study, however, is that clinical outcomes (e.g. resolution of acidosis, weaning from vasoactive agents, ICU length-of-stay, mortality) were not included in the generation of the model. The reason why we elected to exclude these variables from the analysis is that there are other variables that affect these clinical outcomes apart from fluid management (e.g. choice of antibiotics, timeliness of surgery if required, end-of-life decision made by the ICU team and the family). For this proof-of-concept study, we took a simple approach and focused on predicting how much fluid is given a patient given physiologic data obtained during the previous 24 hours, regardless of clinical outcome. It will be interesting to generate separate models using data from patients with good and poor clinical outcomes and compare these models to determine how much influence fluid management has on these outcomes.

A number of studies have looked at the application of artificial intelligence tools in the ICU. Barbini, Cevenini and colleagues [2007] compared different models in predicting ICU morbidity after cardiac surgery and found the Bayesian and logistic regression models to be superior to artificial neural network, scoring systems and k-nearest neighbor in terms of discrimination, generalization and calibration for this particular task. Bayesian network has also been used to predict prognosis of head injured patients in the ICU [Nikifordis and Sakellaropoulos, 1998], mortality of patients readmitted to the ICU [Ho and Knuiman, 2008], and likelihood of ventilator-associated pneumonia [Schurink *et al*., 2007] and other nosocomial infections [Schurink *et al*., 2005].

It is unlikely that we can replace clinician expertise with an intelligent software. We envision three important uses of artificial intelligence tools applied to empiric data. The first is to supplement clinical knowledge to support decisions in specialized, complicated problems where there may not be adequate evidence in the way of prospective randomized controlled trials. The second is to potentially accelerate acquisition of clinical intuition by junior doctors in the ICU by "learning" from their local database how senior intensivists managed identical patients in a specific clinical scenario. Finally, these tools might be of use for ongoing surveillance of medical device, medications and interventions for clinical outcomes (rather than surrogate endpoints) especially in the ICU where these are sometimes adopted without clear evidence of long-term benefit. We have taken a deductive approach in generating a model from empiric data. Combining such deductive approach with an inductive knowledge base from domain expertise in pathophysiologic processes and available ICU literature may provide a better tool in assisting clinicians in making decision for individual patients in specific clinical scenarios.

6 Conclusion

There are very few interventions performed in the ICU, whether for diagnostic, monitoring or treatment purposes, that are based on robust evidence. Even when prospective randomized controlled trials are available, they very rarely, if ever, provide clear evidence upon which to base the management of an individual patient. The project introduces the concept of using empiric data to obtain patient- and clinical scenario-specific recommendations in the ICU. Prediction of maintenance fluid is chosen as the problem domain to test the feasibility of the concept. A software application was developed that builds a model consisting of patients that are similar to an index patient in terms of age, gender, ethnicity, admitting diagnosis, severity score on admission and co-morbidities. Based on the model, physiologic variables that are directly correlated with the outcome variable of interest are identified. The idea is to provide the values of these predictor variables from the index patient to the model, and a predicted range of fluid requirement is obtained from the joint conditional probabilities. We plan to evaluate the effect of the availability of this information in the ICU in an intention-to-treat prospective observational study. An adherence-to-protocol and on-treatment analyses will be incorporated into the design of the study. Finally, allow us to modify one of the most often quoted aphorisms in modern medicine. The plural of anecdote is neither data nor evidence, unless one has a humongous database of anecdotes. In the absence of a prospective randomized controlled trial, and with the use of artificial intelligence, it may provide us with as close to an evidence as we can get.

Acknowledgments

This work has been supported in part by National Library of Medicine Training Grant # 2T15LM007092-16.

References

[Yende and Angus, 2007] Yende S, Angus D. Long-term Outcomes from Sepsis. *Current Infectious Disease Reports,* 9(5):382-6, September 2007.

[Trampuz et al., 2001] Trampuz A, Widmer AF, Fluckiger U, Haenggi M, Frei R, Zimmerli W. Changes in the epidemiology of pneumococcal bacteremia in a Swiss university hospital during a 15-year period, 1986-2000. *Mayo Clinic Proceedings*, 79(5):604-12, May 2004.

[Landoni et al., 2006] Landoni G, Zangrillo A, Franco A, Aletti G, Roberti A, Calabrò MG, Slaviero G, Bignami

E, Marino G. Long-term outcome of patients who require renal replacement therapy after cardiac surgery. *European Journal of Anaesthesiology*, 23(1):17-22, January 2006.

[Rivers et al., 2001] Rivers E, Nguyen B, Havstad S, Ressler J, Muzzin A, Knoblich B, Peterson E, Tomlanovich M; Early Goal-Directed Therapy Collaborative Group. Early goal-directed therapy in the treatment of severe sepsis and septic shock. *New England Journal of Medicine*, 345(19):1368-77, Novermber 2001.

[Donati et al., 2007] Donati A, Loggi S, Preiser JC, Orsetti G, Münch C, Gabbanelli V, Pelaia P, Pietropaoli P. Goal-directed intraoperative therapy reduces morbidity and length of hospital stay in high risk surgical patients. *Chest*, 132(6):1817-24, December 2007.

[Berger, 1985] James Berger. *Statistical Decision Theory and Bayesian Analysis.* Springer-Verlag, New York, New York, 1985.

[Goodman, 1999] Goodman SN. Toward evidence-based medical statistics. 2: The Bayes factor. *Annals of Internal Medicine*, 130(12):1005-13, June 1999.

[Barbini et al., 2007] Barbini E, Cevenini G, Scolletta S, biagioli B, Giomarelli P, Barbini P. A comparative analysis of predictive models of morbidity in intensive care unit after cardiac surgery – Part I: model planning. *BMC medical informatics and decision making*, 7:35, November 2007.

[Cevenini et al., 2007] Cevenini G, Barbini E, Scolletta S, biagioli B, Giomarelli P, Barbini P. A comparative analysis of predictive models of morbidity in intensive care unit after cardiac surgery – Part II: an illustrative example. *BMC medical informatics and decision making*, 7:36, November 2007.

[Nikifordis and Sakellaropoulos, 1998] Nikifordis G, Sakellaropoulos G. Expert system support using Bayesian belief networks in the prognosis of head-injured patients of the ICU. *Medical Informatics (London)*, 23(1):1-18, January-March 1998.

[Ho and Knuiman, 2008] Ho K, Knuiman M. Bayesian approach to predict hospital mortality of intensive care readmissions during the same hospitalisation. *Anaesthesia and Intensive Care*, 36(1):38-45, January 2008.

[Schurink et al., 2007] Schurink CA, Visscher S, Lucas PJ, van Leeuwen HJ, Buskens E, Hoff RG, Hoepelman AI, Bonten MJ. A Bayesian decision-support system for diagnosing ventilator-associated pneumonia. *Intensive Care Medicine*, 33(8):1379-86, August 2007.

[Schurink et al., 2005] Schurink CA, Lucas PJ, Hoepelman IM, Bonten MJ. Computer-assisted decision support for the diagnosis and treatment of infectious diseases in intensive care units. *Lancet Infectious Disease*, 5(5):305-12, May 2005.

Minimal Set of Attributes Required to Report Hospital-Acquired Infection Cases

Jimison Iavindrasana, Gilles Cohen, Adrien Depeursinge, Rodolphe Meyer, Antoine Geissbuhler
Department of Medical Informatics
University and Hospitals of Geneva
Rue Micheli-du-Crest 24, CH-1211 Geneva, Switzerland
{jimison.iavindrasana, gilles.cohen,adrien.depeursinge,rodolphe.meyer,antoine.geissbuhler}@sim.hcuge.ch

Abstract

Data collection for hospital-acquired infection prevalence study is resource consuming. Data mining techniques can be applied to data extracted from the hospital data warehouse in order to report potential cases to be reviewed by infection control practitioners (ICP). The objective of this paper is to investigate the minimal set of attributes required for an automated cases reporting. Information gain and SVM recursive feature elimination combined with a chi-square filtering were used to select the most important features in the prevalence database of the 2006 survey. The temperature and workload were included within the 20 most important features. These attributes are not well documented and removed from the list of important features. The results obtained with the resulting dataset were acceptable because the ICP will have to analyze the electronic health record of only 22.73% of hospitalized patients.

1 Introduction

Hospital-acquired infections or nosocomial infections (NI) are those infections acquired in a hospital, independently of the reason of the patient admission. NI appears after 48 hours after the patient admission. These infections may be related to medical procedures such as the implantation of infected urinary tracts or simply occur during the hospitalization where the micro-organisms are transmitted from other patients, medical staff or are a consequence of the hospital environment contamination.

In Switzerland, 70 000 hospitalized patients per year are infected and 2 000 deaths per year are caused by NI. A hospital aware of the quality of the patient care should have an infection prevention, control and surveillance program. The surveillance is the process of detecting these infections. Prevalence surveys are recognized as valid and realistic approaches of NI surveillance strategies [French et al, 1983]. Prevalence of NI is presented as prevalence of infected patients, defined as the number of infected patients divided by the total number of patients hospitalized at the time of study, and prevalence of infections, defined as the number of NIs divided by the total number of patients hospitalized at the time of study [Sax et al, 2002]. From these formulas, the prevalence survey is resource and labor consuming: the electronic health record (EHR) of all patients admitted for more than 48 hours the day of survey should be analyzed by infection control practitioners (ICP). If necessary, additional information is obtained by interviews with nurses or physicians in charge of the patient.

The University Hospitals of Geneva has been performing yearly prevalence survey since 1994. The prevalence database contains 83 attributes ranging from administrative information, demographic characteristics, admission diagnoses, comorbidities and severity of illness scores, type of admission, exposure to various risks of infection, clinical and paraclinical information, and data related to infection when present. One of the main characteristic of the prevalence data is the nature of the attribute values: most of them are nominal. An attribute value summarizes the presence or absence of a particular risk factor or a sign and symptom of an infection like central venous catheter or an antibiotic treatment that can be found in the EHR of the patient. Only the year of birth and the workload values are numerical. Another important characteristic of the prevalence data is the imbalance between the positive and negatives cases: there are around 10% of positive cases.

IT can bring a valuable support for NI surveillance. The hospital data warehouse contains all the data in the hospital operational system except the data of the day. A complex data processing may be implemented to extract all necessary information from the data warehouse, analyze and summarize the information and populate the prevalence database. This approach is too ambitious because of the nature of data to be analyzed: some of them may be found in free text or not well documented in the EHR and require an intervention of a specialist. The most realistic approach is to query the hospital data warehouse in order to populate the N most important item of the prevalence database (N<<83 where 83 is the number of attributes in the prevalence database) and apply data mining techniques to report "potential cases" to be reviewed by the ICP. The potential cases are those predicted as positive cases by a classification algorithm. The main advantage of this approach is the reduction of their workload, and will allow them to evaluate the presence of NI on subset of patients; they can have much more time to analyze the content of the patient record and

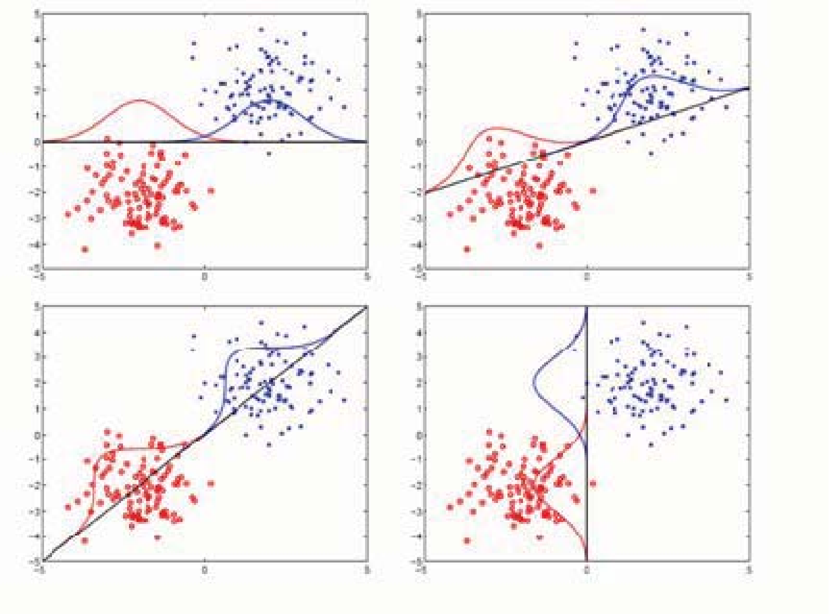


Figure 1: Illustration of Fisher's linear discriminant. The algorithm searches for the direction providing the best separation of the classes when projected upon. In this figure, the third image (bottom left) provides the best separation of the datasets.

may found new risk factors. Indeed, the most consuming time during a NI prevalence survey is the data collection, which represents approximately 800 hours per year [Cohen et al, 2004].

Various data mining techniques were applied at the University Hospitals of Geneva to support NI prevalence survey since 2002. We can quote among others the use of different form of the SVM algorithm optimization including asymmetrical margin approach [Cohen et al, 2003], one-class SVM [Cohen et al, 2004], a comparison of SVM with other classification algorithms [Cohen et al, 2006]. In this paper, we are analyzing the minimal set of features necessary to report potential cases to be reviewed by ICP. We will remove important features if they are not well documented in the EHR. The differences with the previous works lie in the objective of the experimentation, the methodology and the dataset used.

2 Background

2.1 Nosocomial infection prevalence data:

In this work, we performed a retrospective analysis of prevalence data collected at the University Hospitals of Geneva (6 hospitals and 2'200 beds) during the 2006 NI prevalence survey. The dataset contains five data categories: 1) demographic information, 2) admission diagnosis (classified according to McCabe [McCabe and Jackson, 1962] and the Charlson index classifications [Charlson et al, 1987]); 3) patient information at the study date (ward type and name, status of Methicillin-Resistant Staphylococcus Aureus portage, etc); 4) information at the study date and the six days before (clinical data, central venous catheter carriage, workload, infection status, etc) and 5) those related to the infections i.e. for infected patients (infection type, clinical data, etc.). In this study, we are interested in the four first categories of data as they are related to patient infection, which comprises 45 attributes. The dataset contains 1573 cases.

To homogenize the data values, we transformed all numerical data into nominal ones. The year of birth was converted into age and discretized into 3 categories (0-60; 60-75; >75) as in [Sax et al, 2002], and a new variable "hospitalization duration" was created. A Mann-Whitney-Wilcoxon statistical test on the workload value provides a significant difference between infected and non-infected patients. As it is the unique attribute having missing values (91 cases including 2 positive cases), all cases having no workload value were removed. The latter and the hospitalization duration were discretized using the minimum description length principle [Kononenko, 1995]. Patients admitted for less than 48 hours at the time of the study and not transferred from another hospital were also removed. The final dataset contains 1384 cases containing 166 positive cases (11.99%). And finally, all attributes were binarized. Let us call this dataset S.

2.3 Class imbalance problem

The class imbalance problem is an important problem in machine learning since the class of interest is represented with a small number of examples [Japkowicz and Stephen, 2002]. In the presence of imbalanced datasets, classification algorithms tend to classify the larger class accurately while generating more errors in the minority class. If a positive class has a ratio of 10%, a classification accuracy of 90% may be meaningless if the classification is not sensitive at all.

The class imbalance problem induces specific approaches to train classifiers and evaluate their performance. Two approaches were proposed to deal with the class imbalance problem in [Cohen et al, 2004; Estabrooks, 2004]. The first one is to modify the classification algorithm or at least use an algorithm able to deal with imbalanced data. The second resamples the data to reduce the imbalance effect. The latter has the advantage of being independent of any classification algorithm.

2.4 Fisher's Linear Discriminant

The basic idea behind linear discriminant algorithms is to find a linear function providing the best separation of instances from 2 classes. Fisher's linear discriminant (FLD) is looking for a hyperplane directed by *w*, which (i) maximizes the distance between the mean of the classes when projected on the line directed by *w* and (ii) minimizes the variance around these means [Fisher, 1936]. An illustration of this property is highlighted on the figure below (Figure 1).

Formally, FLD aims at maximizing the function:

$$J(w) = \frac{w^T S_B w}{w^T S_W w}$$

where S_B is the scatter matrix between classes and S_W the scatter matrix within classes. This equation permits to formulate FLD as an algorithm aiming at minimizing the variance within the classes and maximizing the variance between classes. An unknown case will be classified into the nearest class centroid when projected onto a hyperplane directed by *w*.

In a classification task, an object is a member of exactly one class and an error occurs if the object is classified into the wrong one. The objective is then to minimize the misclassification rate. With FLD algorithm, the scatter matrix within classes S_W is evaluated on the training datasets. The matrix S_W may be singular, a regularization factor *r* ($0 \leq r \leq 1$) is introduced into the computation of the inverse of this matrix [Hastie et al, 2001]. The regularization factor *r* has to be optimized to minimize the misclassification error.

3 Material and methods

Two feature selection algorithms were used independently. The first based on the information gain (IG) of each attribute [Quinlan, 1986] and the second based on the combination of attributes using SVM Recursive Feature Elimination (RFE) [Guyon, 2002]. These 2 algorithms return all the features ranked by order of importance. To filter the most important ones, a chi-square statistic test was performed to filter the discriminative features to be retained for an evaluation with a classification algorithm. These feature selection algorithms were applied to 100 training sets build from the original dataset S. The significant attributes retained by both feature selection algorithms over the 100 training datasets were retained to build a second dataset S1. Afterwards, we removed the important features in S1 which are not well documented in the EHR to obtain a third dataset S2. We then evaluated the performance of the FLD algorithm on these two datasets. For classification purposes, we used the open-source toolbox MATLABArsenal[1]. This MATLAB package contains many classification algorithms and in particular the regularized FLD algorithm as described above. FLD was chosen as it has only one parameter so easier to optimize.

The methodology adopted to evaluate the performance of the FLD is inspired by the experimental setup described in [Rätsch et al, 2001]. One hundred (100) partitions of training and testing sets were generated with the data source S1 and S2 having respectively a ratio of 60% (approximately 830 cases) and 40% (approximately 554 cases). The original data distribution is kept in both partitions i.e. 11.99% of positive cases. Five balanced dataset (50% of positive cases and 50% of negative cases) were created from the first five training sets. A grid search algorithm is then applied to these down-sampled datasets using a 5-folds cross-validation to find the best parameters of the classification algorithm. The regularization factor *r* takes values from 2^{-20} to 1 during this process. The best parameter of each training set was the one providing the highest recall (i.e. the parameter permitting to predict highest rate of true positive cases) and the highest precision. The best value selected for the classification algorithm is the median of the 5 best parameters obtained with the five down-sampled data. The 100 training sets (having the original class distribution) are then used to train the FLD models with this best parameter. This process allowed us to build 100 models and to validate each of them on the corresponding testing set. The general performance of the classifier is computed as the mean of the 100 classification performances on the test sets. The performance of the classification algorithm with the 2 datasets (S1 and S2) is also compared with respect to the Mann-Whitney-Wilcoxon statistical test.

4 Results

4.1 Feature selection

Twenty (20) attributes were retained from the two feature selection algorithms; IG and SVM RFE returned the same features after the Chi-square filtering. The hospitalization duration up to 7.5 days, was retained as a discriminative attribute. Two admission diagnoses are discriminative: those classified as "non fatal" and "fatal in less than 6 months" according to the McCabe classification, transfer as admission, "congestive cardiomyopathy" and "diabet with organ affected" as comorbidities. In the third data category, the intensive care unit and obstetrical wards, absence or actual MRSA colonization are the most discriminative attributes. In the fourth data category, an antibiotic treatment, fever, a surgery, a stay at the intensive care unit during the hospitalization, a presence of artificial ventilation, urinary tract, central venous catheter and the 3 categories of workload value were significantly discriminative.

[1] http://www.informedia.cs.cmu.edu/yanrong/MATLABArsenal/MATLABArsenal.htm
(last accessed October 2008)

Table1: List and rank of features obtained after applying IG and SVM RFE followed by a Chi-square filtering. The first column provides the rank of each attributes. The two algorithms provided the same features but not with the same rank.

Rank	IG + Chi-square filtering	SVM RFE + Chi-square filtering
1	Antibiotic therapy	Antibiotic therapy
2	Fever	Hospitalization duration up to 7.5 days
3	Mechanical ventilation	Transfer from another hospital as admission
4	Urinary tract	Mechanical ventilation
5	Workload value > 91.5	McCabe score fatal < 6 months
6	Workload value <=45.5	Fever
7	Stay at the intensive care unit during hospitalization	Urinary tract
8	Central vein catheter	Diabetes with organ affected
9	Hospitalization duration up to 7.5 days	Congestive cardiomiopathy
10	Intensive care unit ward	Intensive care unit ward
11	Obstetrical ward	Workload value > 91.5
12	Surgery	Workload value <=45.5
13	McCabe score fatal < 6 months	Stay at the intensive care unit during hospitalization
14	No MRSA colonization	McCabe score non fatal
15	Actual MRSA colonization	Surgery
16	McCabe score non fatal	Actual MRSA colonization
17	Workload value between 45.5 and 91.5	Central vein catheter
18	Diabetes with organ affected	No MRSA colonization
19	Transfer from another hospital as admission	Workload value between 45.5 and 91.5
20	Congestive cardiomiopathy	Obstetrical ward

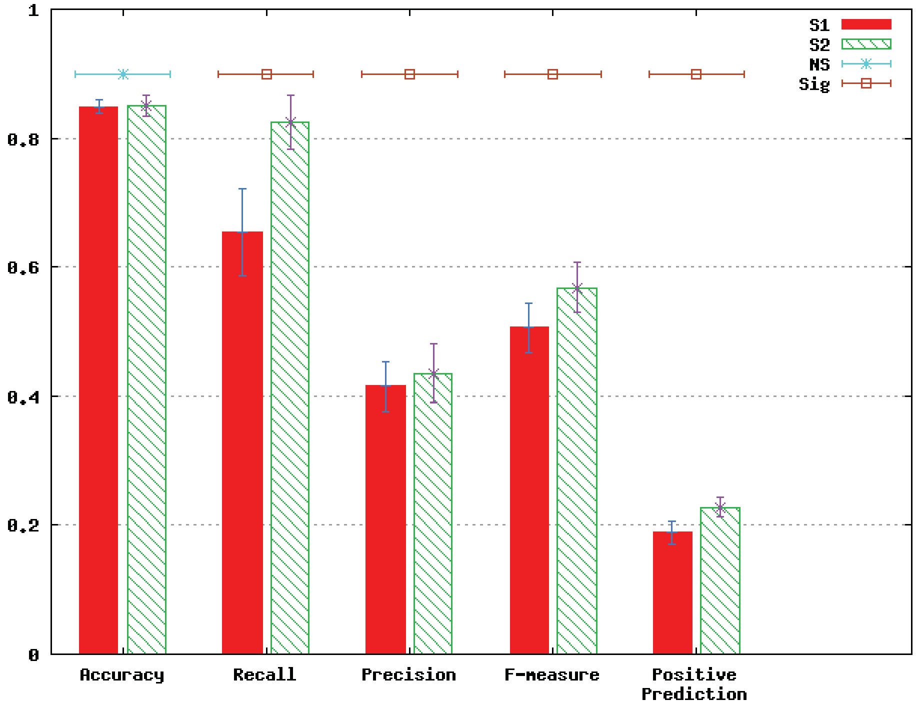


Figure 2. The mean and standard deviation of each performance metrics on the datasets S1 and S2. Sig (respectively NS) indicates (no) significant difference between the performance measure on the two datasets.

Table 1 summarizes the features returned by both IG and SVM RFE. The feature selections described above provided two clinical features, which are not always documented or at least not documented in a machine readable format in the clinical database: fever and workload value. These attributes were removed to create the second dataset S2.

4.2 FLD performances

The grid search algorithm applied on the two datasets S1 and S2 returned respectively *r* =0.5 and *r* =1 as the best parameter. The figure above (Figure 2) summarizes the performance metrics (recall, precision, f-measure, accuracy and the ratio of positive predictions) obtained with the two datasets in terms of their mean, standard deviation (SD) and the performance comparisons.

Dataset S1 and S2 permit to obtain respectively a mean recall (±SD) of 65.37% (±6.76) and 82.56 (±4.22), a precision (±SD) of 41.50% (±3.9) and 43.54% (±4.59), a f-measure (±SD) of 50.58(±3.83) and 56.87(±4.29) over the 100 data split realizations. The mean accuracy (±SD) for S1 and S2 are 84.83%(±1.04) and 85.04%(±1.65) and the positive prediction ratios are respectively 18.82% (±1.72) and 22.73% (±1.55).

According to the results above, if we query the hospital data warehouse with the features present in the dataset S1 and S2 and classify the results with the FLD algorithm, we can expect retrieving an average of (±SD) 65.37% (±6.76) and 82.56 (±4.22) of the infected patients. The mean numbers of potential cases (±SD) to be submitted to the ICP are respectively 18.82% (±1.72) and 22.73% (±1.55) of the hospitalized patients.

A Mann-Whitney-Wilcoxon statistic test provided a *p* value < 0.001 for accuracy, precision, f-measures and the positive prediction ratio. According to this test, there is a statistically significant difference between the accuracy, precision, f-measure and the ratio of positive prediction. The removal of the temperature and the workload features improved the performance of the FLD.

5 Discussion and conclusion

In this paper we investigated the minimal set of features necessary to report potential cases to be reviewed by infection practitioners. IG and SVM RFE were used to select these features and Fisher's linear discriminant was chosen as classification algorithm. The removal of the attributes characterizing fever and workload value significantly improved the performance of the classifier. This result may surprise as these attributes was retained by the IG and SVM RFE as important features to predict a NI. However, this phenomenon is not new in statistical and machine learning domain. This phenomenon is called redundancy or negative interaction [Kludas et al, 2008]. This redundancy cannot be evaluated with IG as it only evaluates quantity of information brought by each attribute to the value of the class. It was not also handled by the SVM RFE because it eliminates X features per iteration (1 in our case) and do not try all possible combinations of features.

The results we obtained with the S2 dataset are acceptable. The ICP will only review the EHR of 1/5 of hospitalized patients and we may reduce significantly their workload with the attributes in the S2 dataset. The precision value indicates that 43.54% of these patients are infected and they represent 82.56% of all infected patients. The precision rate is acceptable because the data of non-infected patients are necessary for statistical tests in order to identify the most important risk factors. The ICP will also have enough time to identify new risk factors from the infected patients' EHR and propose new preventive measures for the future.

5.1 Limits of this work

The evaluation of the discriminative power of the selected features was carried out using Fisher's linear discriminant algorithm because of its simplicity: one parameter (0<r<=1) to optimize during the grid search process. A comparison with other classification algorithms such as Support Vector Machines (SVM) and the Kernel FLD has to be carried out. The grid search algorithm for algorithm parameters optimization has high computational cost especially for classification algorithms with more than one parameter to optimize such as SVM or the Kernel FLD. A gradient descent method could converge more rapidly to the best parameter and can improve the generalization performance as described in [Chapelle et al, 2002].

5.2 Future work

The results obtained were promising and in the future, we plan to evaluate the discriminative power of the selected features with more than one classification algorithm. The result of these evaluations i.e. the minimal attributes required to predict most of the positive NI cases will be retained to build queries for the hospital databases in order to automatically report potential cases for the prevalence surveys. This automated nosocomial infection reporting will permit to conduct more prevalence surveys with less cost than the usual method for conducting prevalence studies.

Acknowledgments

The authors are grateful for the dataset provided by the infection control team at the Geneva University Hospital.

References

[Chapelle et al, 2002] Chapelle O, Vapnik V, Bousquet O, Mukherjee S. Choosing multiple parameters for support vector machines. Mach. Learning. 2002;46(1-3):131-59.

[Charlson et al, 1987] Charlson ME, Pompei P, Ales KL, MacKenzie CR. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. J Chronic Dis. 1987;40:373-83.

[Cohen et al, 2003] Cohen G, Hilario M, Hugonnet S, Sax H. Asymmetrical Margin Approach to Surveillance of

Nosocomial Infections Using Support Vector Classification. IDAMAP; 2003.

[Cohen et al, 2004] Cohen G, Hilario M, Sax H, Hugonnet S, Pellegrini C, Geissbuhler A. An application of one-class support vector machine to nosocomial infection detection. In Proceedings of MedInfo:2004;11:716-20.

[Cohen et al, 2006] Cohen G, Hilario M, Sax H, Hugonnet S, Geissbuhler A. Learning from imbalanced data in surveillance of nosocomial infection. Artificial Intelligence in Medicine. 2006;37(1):7-18.

[Estabrooks, 2004] Estabrooks A. A multiple resampling method for learning from imbalanced datasets. Comput Intell. 2004,20(1):18-36.

[Fisher, 1936] Fisher RA. The use of multiple measurements in taxonomic problems. Annals of Eugenics. 1936;7:179–88.

[French et al, 1983] French GG, Cheng AF, Wong SL, Donnan S. Repeated prevalence surveys for monitoring effectiveness of hospital infection control. Lancet. 1983;2:1021-23.

[Guyon et al, 2002] Guyon, J. Weston, S. Barnhill, V. Vapnik. Gene selection for cancer classification using support vector machines. Machine Learning. 2002;46:389-422.

[Hastie et al, 2001] Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference, and prediction. Springer. 2001.

[Japkowicz and Stephen, 2002] Japkowicz N, Stephen S. The class imbalance problem: a systematic study. Intell Data Anal J 2002;6(5):429-49.

[Kludas et al, 2008] Kludas J., Bruno E., and Marchand-Maillet S. Can Feature Information Interaction help for Information Fusion in Multimedia Problems? In Proceedings of MMIU:2008.

[Kononenko, 1995] Kononenko I. On biases in estimating multi-valued attributes. Eds.: Morgan Kaufmann. In Proceedings of the 14th International Joint Conference on Artificial Intelligence:1995.

[McCabe and Jackson, 1962] McCabe WR, Jackson GG. Gram-negative bacteremia, I: etiology and ecology. Arch Intern Med. 1962;110:847-55.

[Quinlan, 1986] Quinlan JR. Induction of decision trees. Machine Learning, 1, 81-106, 1986.

[Rätsch et al, 2001] Rätsch G, Onoda T, Müller KR. Soft margin for AdaBoost. Mach.Learning. 2001;42(3):287-320.

[Sax et al, 2002] Sax H, Pittet D, Swiss-NOSO Network. Interhospital Differences in nosocomial infection rates: importance of case-mix adjustment. Arch Intern Med. 2002;162(21):2437-42.

Prediction of the Pulse Pressure Trend in Elderly Using In-Home Monitoring Sensors: A Pilot Study

Mihail Popescu PhD[1,4], Elena Florea[1], Jean Krampe[3], Marjorie Skubic PhD[2], Marilyn Rantz PhD[3]
[1]Health Management and Informatics; [2]Electrical and Computer Engineering, [3]Sinclair School of Nursing; [4]MU Informatics Institute, University of Missouri, Columbia, MO

Abstract

We describe the possibility of employing the data generated by a continuous, unobtrusive nursing home monitoring system for predicting the trend of pulse pressure (PP) in elderly (PP=systolic BP-diastolic BP). We investigated several factors that influence PP trend prediction such as sensor type, number of initial training samples and resident medication. We conducted a retrospective pilot study on two residents of the TigerPlace aging in place facility, with age over 70, that had blood pressure measured between 100 and 300 times during a period of two years. The pilot study suggested that pulse pressure trend can be reasonably well estimated (average relative error of less than 10%) using apartment motion sensors.

1 Introduction

The proportion of elderly in the population is growing at a rapid rate in countries around the world. Many of these seniors prefer to live independently for as long as they are able, despite the onset of conditions such as frailty and dementia. Solutions are needed to enable independent living while enhancing seniors' safety and their families' peace of mind [Cuddihy et al., 2007], [Rowan et al., 2005].

Aging adults are often stereotyped as purposefully masking any decline in abilities to avoid outside intervention and this fact leads to the concern held by adult children about their aging parents [Rowan et al., 2005]. Elderly patients are particularly at-risk for late assessment of cognitive changes due to many factors: their impression that such changes are simply a normal part of aging, their reluctance to admit to a problem, their fear of being institutionalized and even the failure of physicians to fully assess their cognitive function due to the belief that no intervention is possible [Hayes et al., 2004]. Relying on self-report by the patient or their family is also unreliable. Current clinical monitoring approaches may miss important fluctuations in behavior and health state [Hayes et al., 2005]. In addition, presenting the long term patterns to the elders may encourage them to seek help [Morris et al., 2005].

In this paper we present one of the first attempts that we are aware of, of employing unobtrusive sensors deployed in the living environment to predict clinical status of the residents. When abnormal clinical values (of the pulse pressure in our case) are predicted the caregivers are alerted to a potential need for medical intervention.

2 Pulse pressure as a predictor of cognitive and functional decline.

Pulse pressure (PP) is defined as the difference between the systolic blood pressure (SBP) and the diastolic blood pressure (DBP). The majority of individuals older than 70 years have an increased pulse pressure resulting from age-related stiffening of the central elastic arteries and systolic hypertension. A high PP in the elderly predicts a higher risk of cardiovascular events [Peters et al., 2007], coronary heart disease [Franklin et al, 2001], renal disease, heart failure [Swaminathan and Alexander, 2006] and mortality [Safar et al., 2004]. A high PP has been related to cognitive decline in people over 50 years of age [Waldstein et al., 2008]. According to Safar et al. [Safar et al. 2004] , a PP of 60 mm Hg is a strong mechanical factor predicting cardiovascular mortality. As a conclusion, it seems that PP is a good clinical indicator of multiple conditions existent in older adults.

However, according to Peters et al. [Peters et al., 2007] and Swaminathan and Alexander [Swaminathan and Alexander, 2006], no practical cut-off value exists for differentiating normal pulse pressure from abnormal pulse pressure. PP seems to increase with age [Safar et al., 2004] and for any given age over 70, men have a 5%-10% higher PP than women [Swaminathan and Alexander, 2006].

Our sensor data capture external information (behavioral) about the resident that is subsequently reflected in the predicted PP. In previous work [Popescu et al. 2008] we considered elevated PP as a supervised classification problem. Among the three classifiers tried (SVM, neural nets and robust regression), the robust regression performed best, hence we decided to further used it for PP trend prediction. By continuously computing the pulse pressure and comparing it with the measured trend we may alert the nursing staff when some predefined variability limits are exceeded. This approach may provide additional blood pressure monitoring for the elderly persons susceptible to blood pressure variations during the time between two nursing visits. In addition, the comparison of the computed and measured pulse pressure trends over longer periods of time may provide additional warnings of abnormal unreported clinical events.

3 Methodology

TigerPlace [Rantz et al., 2005] is an independent living facility for seniors designed and developed as a result of collaboration between Sinclair School of Nursing, University of Missouri and Americare Systems Inc. of Sikeston, Missouri. A primary goal of TigerPlace is to help the residents not only manage their illnesses but also stay as healthy and independent as possible. Each resident included in the study has a Data Logger in his or her apartment that collects data from wireless sensors (Figure 1).

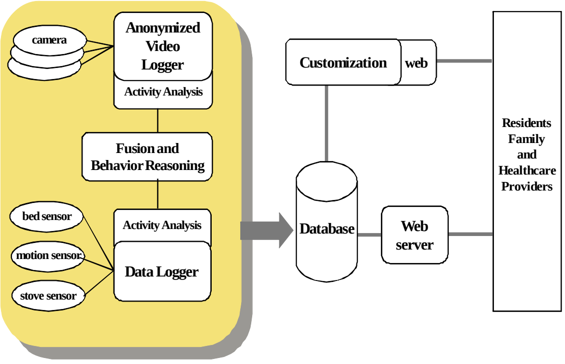


Figure 1. The sensor network. Motion and bed sensors were used in this study.

The Data Logger date-time stamps the data, and logs them into a file that is sent to a database on a secure server via a wired network connection. Fourteen networks (without video) have been installed in TigerPlace apartments; the video part of the network is currently under development.

The sensor network consists of several types of sensors mounted in different places throughout the residents' apartments, including motion sensors, bed sensors, and stove temperature sensor. The motion sensors are placed in various places, such as bathroom, bedroom, kitchen, living room, etc. and some of the residents have this type of sensor installed on the door of the refrigerator, kitchen cabinets and even drawers. They capture resident motion through his/her apartment by emitting a signal (firing) as often as there is movement around them. The bed sensors are in fact sets of sensors, composed of a pneumatic sensor strip across the bed and a motion sensor attached to the bed headboard (see[1]for a similar system). The sensor strip is able to keep track of the resident's movement in the bed, namely restlessness, pulse and breathing, as long as the resident is in the bed. The sensor strip and motion sensor attached to the bed are connected together and they function similarly to the motion sensors mentioned previously: they fire as long as they detect activity. Unlike the motion sensors, the bed sensor strip captures three types of activities, which are structured on three or four levels of severity.

The prediction of the PP trend is based on the intuition that if the resident does not feel well, his/her sleep and motion patterns are altered. In this study, we used the following values for PP prediction:

- the total number of motion sensor firings from the day (from 7 am to 9 pm) and night (from 9 pm to 7 am) previous to the PP measurement;
- the total number of bed restlessness (level 1- bed movement for 1 to 3 seconds) firings from the day (from 7 am to 9 pm) and night (from 9pm to 7am) previous to the PP measurement;
- the total number of low heart rate (HR<30 beats/min.) firings from the day (from 7 am to 9 pm) and night (from 9 pm to 7 am) before the PP measurement.

Although each considered resident lives alone in his apartment, some extra motion hits were possible due to housekeeping or occasional visits. In this research we did not consider factors that affect the total number of motion firing such as visitors, the duration of the sleep and the time out of the apartment. We are currently working on algorithms for detecting these factors and plan to account for them in the future. Instead, we considered the night sleep occurring from 9pm to 7am and we removed the samples where there were no sensor hits for at least three hours (the resident was probably out of the apartment).

The predicted values of the PP, $\hat{y}_i$, were calculated using a robust linear regression [Hogg 1979] as:

$$\hat{y}_i = \sum_{j=1}^{M} x_{ij}\beta_j + \beta_0, \qquad (1)$$

where $x_i \in R^M$ is the input vector of M sensor measurement, $M \in \{2,3,...,6\}$, and $\beta \in R^{M+1}$ is the vector of regression coefficients. The regression coefficients are computed using an iterative procedure that minimized the criterion:

$$S = \sum_{i=1}^{N} \Psi(y_i - \hat{y}_i) \qquad (2)$$

The robust regression is more resilient to outliers by replacing the default least square function $\Psi(t) = t^2/2$ with one that has decreasing weights for outliers. A popular choice for Ψ is Tukey's biweight (bisquare) function [Hogg 1979] $\Psi(t) = t(1-(t/k)^2)^2$ for $|t|\leq k$ and 0 else, where k=4.685 is a tuning constant. In this paper we used a robust regression implementation, *robustfit*, found in the statistical toolbox of MATLAB (http://www.mathworks.com). The algorithm used for calculating the PP trend has the following steps:

```
Step 1: Use $N_{start}$ PP measurements $\{y_i\}$ and the related sensor vectors $\{x_i\}$, $i=1,N_{start}$, to compute the first set of regression coefficients, $\beta$;
Step 2: For each new sensor input $i \in [N_{start+1}, N]$
        - Use $\beta$ to compute the PP estimate $\hat{y}_i$
        - if there is a measurement in that day $y_i$, update the regression coefficients, $\beta'$
        - set $\beta = \beta'$
        End For.
Step 3: Compute the trend of the predicted and measured values as the average of the previous week (7 previous values, if some days are missing).
```

We also computed, for the sake of the display, the predicted values from 1 to N_{start} based on the initial model.

The performance of the regression (goodness of fit) was evaluated by computing the average relative error (ARE) between the predicted PP values and the measured values as:

$$\mathrm{ARE} = \frac{\sum_{i=1}^{N_{test}} \frac{|y_i - \hat{y}_i|}{y_i}}{N_{test}} [\%], \qquad (3)$$

where N_{test} is the number of test (predicted) values.

4 Data set

The data available for the two residents considered in the study is shown in Table 1.

	Total records	*Out of the room*	*Total data set*
Male	93	52	41 (30 PP≥60)
Female	139	49	90 (35 PP≥60)

Table 1. The data for the two residents considered in the study

The "out of the room" data was due to the resident being out of his apartment for more than three hours. We are currently working on an algorithm that will reliably detect when the resident is out of his apartment which will increase both the PP prediction accuracy and the amount of available data.

We mention that in addition to the sensor data we have all the clinical records (medication, nursing visits, hospitalizations, etc) for both residents and some personal notes of the residents.

The results of the PP trend prediction for the above residents, Male and Female, are presented in next section.

5 Results

The main goal of this work was to determine the feasibility of the PP trend prediction based on the sensor data and to investigate if the predicted trend reflects the functional decline of the residents. We were interested in several computational aspects of the PP prediction:

- what choice of sensor data input provides the best prediction;
- what is the best choice of algorithm constants such as the number of training samples and weight threshold for our regression algorithm;
- what is the influence of other factors such as medication or other clinical variables.

The answers to the three above questions are shown in the next subsections.

5.1 Choice of sensors

We have tested various combinations of the three available sensors (motion, breathing and restlessness) during day and/or night time. The average relative errors for the tested combinations for the two available residents (Male and Female) are given in Table 2.

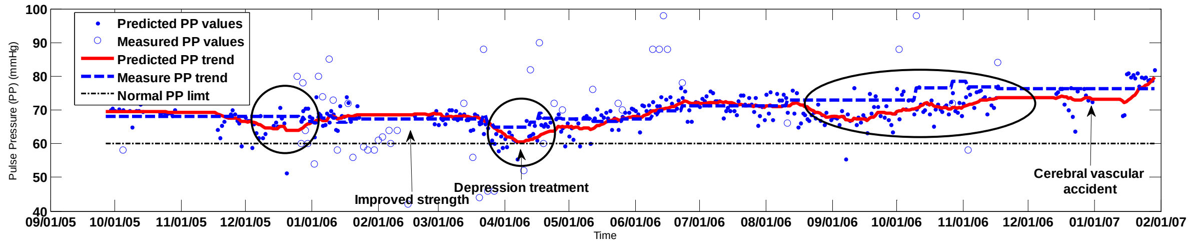


Figure 2. The predicted and measured pulse pressure for the Male resident.

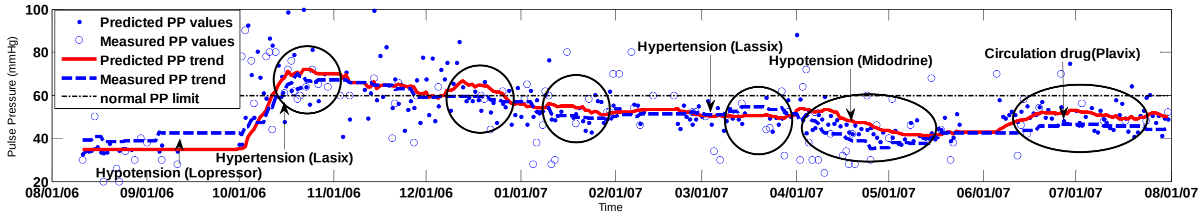


Figure 3. The predicted and measured pulse pressure for the female resident.

No.	*Sensor type*	*Male* [%]	*Female* [%]
1	motion	3.1	8.5
2	motion + restlessness	3.8	9.8
3	motion + restlessness + heart rate	10.5	11.5
4	motion + heart rate	7.4	10.7
5	Restlessness + heart rate	4.2	15.5
6	day motion + day restlessness	3.7	9.2

Table 2. The average relative error (ARE) for the PP trend for the two residents in the study, Male and Female.

The average relative errors from Table 2 are mostly below 10% which denotes that the PP trend can be predicted reasonably well. The day and night total motion (2 variables, row 1) seem to predict best the PP trend in both cases. However, due to the insufficient training data from this experiment, this choice might have won just because it had the least variables to train. The resulting PP plots for this case are shown in Figure 2 and 3. The Male resident (Figure 2) had a decreasing functional status during the period under observation: his PP was all the time over 60 with a continually increasing trend. In fact, he passed away one month after the observation period ended. The Female resident (Figure 3) had an episode of high blood pressure (October 2006) after which she recovered and currently, she is in stable condition. We see in these examples that PP can be used as an indicator of the long term functional decline.

5.2. Choice of algorithm constants

We chose the robust regression against the regular least square version due to the high variability of the measured data. For example, for the Female case, the PP on 10/11/06 is 80 while the next day is 44 and the second day is 80 again. The displayed values of both trends (predicted and measured values) were calculated by averaging seven consecutive measurements.

Next, we observed that an additional threshold on the function Ψ that excludes from regression calculation points with weights less than a given threshold WTHRESH, improves the ARE. By trial and error we found WTHRESH=0.8 which in most cases improved ARE by about 10% for both residents, for all choices of features.

Another parameter of our algorithm is the number of starting samples, N_{start}. This will give us an idea of the minimum number of samples that we have to collect in order to have a good prediction. The effect of the number of starting samples on our algorithm is shown in Table 3.

N_{start}	*7*	*12*	*26*	*33*	*40*	*46*
ARE[%] Female	9.1	7.7	7.4	6.7	6.9	7.1

Table 3. The effect of the starting number of samples on the trend prediction for the Female resident

From Table 3 we see that the number of starting samples that minimizes the error is about 30. This implies that for improving the prediction we need to have at least a month worth of blood pressure measurements for each resident (effort under way at TigerPlace).

5.3 Influence of medication or unknown clinical factors

Intuitively, medication that alter motion and sleep patterns might influence our prediction. However, we did not observe this effect in our case study. Instead, the predicted PP trend was able to capture the influence of hyper- and hypotension medication. In figure 3, we see that the predicted PP trend increases when Lopressor (drug that increases blood pressure) is administered (Aug 20, 2006) and decreases when Lasix (drug that lowers blood pressure) is taken.

Since medication did not seem to be a factor influencing the motion and sleep patterns, we could not explain certain differences between the predicted and the measured PP trend during a period of month or two (circled in Figure 2 and 3). Our hypothesis is that the differences between the two trends are produced by subtle changes in the sleep and motion pattern that were not reported by the resident. These differences might be able to provide a hint about the possibility of upcoming abnormal clinical events. In the Male example (Figure 2) the PP trend underestimated the measured PP previous and during the time when the resident had health issues such as depression (April 06) or cerebral vascular accident (CVA) (January 07). It is interesting that the resident noted in his journal during the period between September 06 and December 06 (previous the CVA) that he was having the "Same old night". Also, the blood pressure was measured only once during that time. Conversely, in the case of the Female resident (Figure 3) the predicted PP trend overestimated the measured PP trend in several instances when the resident felt better (10/11/06, 12/05/06, 05/02/06) and underestimated it when she felt worse (03/15/06- pneumonia) than expected.

6 Conclusion

We investigated a method for predicting the pulse pressure trend in elderly residents using unobtrusive monitoring sensors. The prediction of PP may help nursing staff provide interventions that might prevent grave clinical events such as heart attacks or strokes. The comparison between the measured and predicted PP trend may be also used for detecting abnormal clinical events in elders.

We tried several sensor inputs that could be used to predict PP trend such as the bed restlessness, room motion and low heart rate. The conclusion of our study was that room motion predicts reasonably well (ARE less than 10%) the PP trend.

However, our study has several limitations. First, the sample size and data sets were small. We plan to measure BP of three elders daily for three months in order to have a larger data set. Second, the sensors readings were influenced by factors such as the presence of visitors, time out of the apartment and sleep duration that we only partially accounted for. We are currently working on algorithms to detect and correctly integrate them in future work. Third, other conditions aside of PP (such as a bad meal) may influence the sensor readings. We are currently researching fuzzy rule based methods that would detect abnormal conditions regardless of their origin.

References

[Alawan et al., 2006] Alwan M., Dalal S., Mack D., Kell S., Turner B., Leachtenauer J, Felder T.. Impact of monitoring technology in assisted living: outcome pilot. IEEE Trans. on Inf. Tech. in BioMedicine, 10(1), pp. 192-198, 2006.

[Cudihy et al., 2007] Cuddihy P., Weisenberg J., C. Graichen and M.Ganesh. Algorithm to automatically detect abnormally long periods of inactivity in a home. Proc. of the 1st ACM SIGMOBILE Intl. Workshop, pp. 89-94, New York, 2007.

[Morris et al., 2005] Morris M., Intille S. S., and Beaudin J. S., "Embedded assessment: overcoming barriers to early detection with pervasive computing. Proc. of PERVASIVE 2005, H. W. Gellersen, R. Want, and A. Schmidt, Eds., pp. 333-346, Springer-Verlag, 2005,

[Franklin et al. 2001] Franklin S.S., Larson M.G., Khan S.A., Wong N.D., Leip E.P., Kannel W.B. Does the relation of blood pressure to coronary heart disease risk change with aging? The Framingham heart study. Circulation, 103(9), 1245-1249, 2001.

[Hayes et al., 2004] Hayes T.L., M. Pavel, J.A. Kaye. An unobtrusive in-home monitoring system for detection of key motor changes preceding cognitive decline. Proc. of the 26th Annual Intl. Conf. of the IEEE EMBS, pp. 2480-2483, San Francisco, CA, 2004.

[Hogg 1979] Hogg RV. Statistical robustness: one view of its use and applications today. The American Statistician, Vol. 33, no 3, pp. 108-115, 1979.

[Peters et al., 2007] Peters R., Marero C., Pinto E., Beckett N., Hypertension in the very elderly. Aging Health, 3(4), 517-525, 2007.

[Popescu et al., 2008] Popescu M., Florea E., Skubic M, and Rantz M., "Prediction of Elevated Pulse Pressure in Elderly Using In-Home Monitoring Sensors: A Pilot Study," 4th IET International Conference on Intelligent Environments, University of Washington, Seattle, July 21-22, 2008.

[Rantz et al., 2005] Rantz M.J., Marek K.D., , Aud M.A., Johnson R.A., Otto D., Porter R. TigerPlace: A New Future for Older Adults. J. of Nursing Care Quality, vol. 20, no. 1, pp. 1-4, 2005.

[Rowan and Mynatt, 2005] Rowan J., Mynatt E. D.. Digital Family Portrait field trial: support for Aging in Place. Proc. of the SIGCHI Conf. on Human Factors in Computing Systems, pp. 521-530), New York: ACM Press, 2005.

[Safar et al., 2004] Safar M.E., Lajemi M., Rudnichi A., Asmar R., Benetos A. Angiotensin-converting enzyme D/I gene polymorphism and age-related changes in pulse pressure in subjects with hypertension. Arteriosclerosis, Thrombosis, and Vascular Biology, 24(4), 782-786, 2004.

[Swaminathan and Alexander, 2006] Swaminathan R.V., Alexander K.P.. Pulse pressure and vascular risk in the elderly: associations and clinical implications", Am. J. of Geriatric Cardiology, 15(4), pp. 226-232, 2006.

[Waldstein et al., 2008] Waldstein S.R., Carrington Rice S, Thayer J.F., et al. Pulse pressure and pulse wave velocity are related to cognitive decline in Baltimore Longitudinal study of aging. Hypertension,51, pp. 99-104, 2008.

Feature subsetting for biomedical document classification[*]

Manabu Torii
ISIS Center
Georgetown University Medical Center
Washington, DC 20057 USA

mt352@georgetown.edu

Hongfang Liu
Department of Biostatistics, Bioinformatics,
and Biomathematics
Georgetown University Medical Center
Washington, DC 20057 USA

hl224@georgetown.edu

Abstract

Document classification is a first step toward practical applications of text mining such as literature-based biological database curation. Powerful machine learning classifiers like support vector machines (SVMs) have enjoyed their success in document classification. For biological document classification, there is often a small vocabulary directly relevant to the target concepts and it derives a set of strong classification features for machine learning algorithms. Meanwhile, words outside of such a vocabulary can also contribute to classification performance subtly. Provided many word features with diverse utilities in biological document classification, we hypothesize that not all features can be exploited in machine learning algorithms when they are all present at once (model under-fitting). To address this issue, we used a random feature subsetting method, where small subsets of sampled features were used to train machine learning classifiers. Multiple classifiers trained on different subsets were combined into an ensemble classifier. To effectively incorporate features of diverse utilities, we introduced weighted sampling that favors features with high *information gain*. The experiments suggest that the proposed approach is effective in boosting classification performance and supporting biological database curation.

1 Introduction

Automated document classification has been used for literature-based database curation in the biomedical domain, e.g., immune epitope information (Wang, et al., 2007) or protein binding information (Donaldson, et al., 2003). Specifically, document classification systems have been developed to identify articles containing or not containing designated biological information as positive or negative instances, respectively. Unlike topic-classification where documents need to be classified into high-level topics (e.g., classification of news articles into topic categories such as business, entertainment, politics, science, and sports), the end goal of document classification for biological database curation is not to classify documents, but to support (manual) extraction of information from documents. Particular information targeted in the biological domain is often expressed with a relatively small vocabulary. For example, the vocabulary directly relevant to protein phosphorylation would include ‘kinase’, ‘site’, and ‘phosphorylate’ as well as their inflected/derived words, e.g., ‘phosphorylated’ or ‘phosphorylation’. Meanwhile, words outside of the directly relevant vocabulary can subtly contribute to document classification as well, e.g., experimental methods. Powerful machine learning classifiers such as support vector machines (SVMs) can accommodate a large number of word features without *over-fitting* to the training data (Joachims, 1998). However, machine learning classifiers may not be able to exploit all the available features when too many features are present (*under-fitting* (Sutton, et al., 2005)).

In this study, we experimented with a *feature subsetting* method to build an ensemble classifier (Bay, 1998; Sutton, et al., 2005). Feature subsetting, also known as feature/attribute bagging, is a technique to build an ensemble classifier consisting of machine learning classifiers trained on subsets of available features. This method improves performance of classifiers by reducing their *bias* and *variance* (Bay, 1998) and overcoming model under-fitting (Sutton, et al., 2005). For document classification with thousands of word features, manual selection of feature subsets is challenging, while completely random sampling would not be effective since there is often a small directly relevant vocabulary in biomedical document classification, without which performance of classifiers may be degraded. To overcome this issue, we considered *weighted* random feature subsetting, where features with higher *information gain* were sampled preferably.

In the following, we first explain data collections used in our experiments. Next we introduce our feature sampling method to build an ensemble classifier, followed by experimental results. We then summarize our work.

2 Data Collections

We used PTM (post-translational modification) data set developed at Protein Information Resource (PIR)[1]. The

[*] This work was partly supported by NSF IIS-0639062.

[1] http://pir.georgetown.edu/cgi-bin/ipkLitFt.pl?stat=12

PPI type	Annotated abstracts (positive)
Acetylation	937 (91)
Glycosylation	1,095 (162)
Hydroxylation	199 (48)
Methylation	259 (54)
Phosphorylation	669 (128)

Table 1. PTM data collections. MEDLINE abstracts were downloaded from PubMed as referenced in the PTM data collections. Some of the PMIDs in the collections were no longer found in PubMed, and thus the data collections were slightly smaller than the original collections.

data set consists of collections of MEDLINE abstracts and full-length articles gathered for curating protein PTM information. Each document was labeled by domain experts at PIR as either positive or negative according to the presence (positive) or absence (negative) of the designated PTM information. Early (smaller) versions of these collections were used in document classification studies by (Han, et al., 2006) and (Torii and Liu, 2007). In this study, from the latest version of the data sets, we used collections of MEDLINE abstracts for five PTM types (acetylation, glycosylation, hydroxylation, methylation, and phosphorylation). The number of MEDLINE abstracts in each data collection is found in Table 1.

3 Methods

In this section, we first describe the representation of documents as classification instances in SVMs. Then we introduce our feature subsetting method to derive an ensemble classifier.

3.1 Document Representation

We used stemmed words as classification features after removing the NCBI stop words[2] and rare words[3]. The stemming algorithm used was a simple *S-stemmer* implemented in PERL language that converts plural nouns and third-person singular verbs into normalized forms, e.g., “viruses” → “virus” or “binds” → “bind”. We then followed (Joachims, 1998) to represent documents as feature vectors. Namely, each document is represented as a vector of values, where a value at each position in the vector is a weighted frequency of a particular stemmed word. We used standard TF-IDF weighting (Jones, 1972)[4]. Each vector was then normalized to be a unit vector.

3.2 Weighted Random Feature Subsetting

Provided a large number of word features in document classification, individual features may not be exploited fully in machine learning systems. To address this issue, we built an ensemble classifier consisting of SVM classifiers trained using different subsets of available features.

Let M be the number of stemmed words available as classification features. We randomly sample R% of M features K times, and build K classifiers. Given that there is a small vocabulary essential in each biological document classification task, we should include words belonging to this vocabulary. Meanwhile, we also want to explore diverse features outside of this vocabulary. To reflect this intuition, we first calculated *information gain* (IG) of word features. IG has been commonly used for feature selection in document classification, e.g., (Joachims, 1998). Next, we sampled features M×R/100 times, where those with higher IG values were sampled more. Specifically, the following procedure was used to sample a set of features F ($\cup$ is a set union operator):

```
1.  F = {};
2.  i=0;
3.  Repeat M×R/100 times
4.    Randomly select one feature, w, among
        the top $M/2^i$ features sorted by IG;
5.    F = F $\cup$ {w};
6.    if $M/2^{i+1}$ becomes too small[5],
        then i=0, or i=i+1 otherwise;
```

Here, we sampled features with replacement, and the actual number of unique features was less than R% (see the result section). To build classifiers, LibSVM package (Chang and Lin, 2001) was used. We used the linear kernel, since it seemed to yield better performance in our preliminary experiments. We also used the option ‘-b’ provided in the software so that trained models output probability-like scores between 0 and 1 for predicted classes[6]. For each document, these scores were averaged as combined scores of an ensemble system.

4 Experiments and Results

To evaluate the proposed ensemble method, we conducted 2×5-fold cross-validation tests using the five PTM data collections. Namely, a data collection was split into five partitions, and each time used one partition as a test set and the remaining partitions as a training set (5-fold cross-validation test). This test was repeated twice for different partitioning, and an averaged Receiver Operator Characte-

[2] http://www.ncbi.nlm.nih.gov/books/bv.fcgi?highlight=stopwords&rid=helppubmed.table.pubmedhelp.T43

[3] Words that appear in less than three documents in the training corpus.

[4] A frequency of a stemmed word in the document is weighted with the log of the inverse of its document frequency in the training corpus.

[5] We used a heuristic condition: if $M/2^{i+1} < (M×R/100)/\log_2(M)$. The right hand side is roughly how many times we need to reset i=0 at line 6 if the condition is $M/2^{i+1} < 1$, or in other words how many times we repeat i=0 → i=1 → i=2 → … → i=$\log_2(M)$ during the loop. Noticing that we would sample more than $(M×R/100)/\log_2(M)$ times from top $(M×R/100)/\log_2(M)$ features in this manner, we naively chose it for our condition to reset i=0 so that we do not attempt to select features from a list of mostly pre-selected features in this sampling procedure.

[6] We used probability-like scores of LibSVM because they are normalized to [0, 1] and seemed amenable to the proposed ensemble method, compared to un-normalized SVM outputs.

ristic (ROC) curve over ten (2×5) runs was derived (Fawcett, 2006) [7].

We evaluated the weighted random feature subsetting described in the previous section and also completely random feature subsetting. In completely random feature subsetting, a list of features was shuffled, and M×R/100 features were selected for their use in machine learning systems. We also evaluated single SVM classifiers trained with top $M \times 1/2^i$ features ($i \geq 0$) sorted on IG.

For the acetylation, glycosylation, and phosphorylation data collections, two to three thousand word features were available during the cross-validation tests. For the hydroxylation and methylation collections, about one thousand word features were available. For the random subsetting method, we used fixed settings of R=10% and K=20 because usually little or no improvement was observed for R>10% or K>20, except for the case of hydroxylation (see the next section). In one run of the cross-validation test on the acetylation collection, for example, there were 2,836 features, and on average 197 unique features (6.9% of the available features)[8] were selected using the proposed sampling method. In fact, for all the data collections, ~6.9% of available features were sampled with the proposed sampling method. Figure 1 (Figure 1.a and 1.b) shows averaged ROC curves for ensemble classifiers and selected single classifiers for different PTM types..

5 Discussion and Conclusion

As in Figure 1, the weighted random subsetting method can derive ensemble classifiers superior to single SVM classifiers and completely random ensemble classifiers. The proposed method, however, yielded poor performance for the hydroxylation collection. Though the performance could be improved for different settings of R (e.g., R=50%), it did not yield better performance than the single classifier trained using all features. Similarly, the proposed method was not effective for the methylation collection either. These two data collections (hydroxylation and methylation) are much smaller than the other three collections. For each of the three collections with more abstracts, we repeated the experiments using 300 abstracts randomly picked from them. The results are shown in Figure 2. These figures show that the proposed method did not necessarily yield good performance for small data sets, and we may ascribe the poor performance for the hydroxylation and the methylation collections in the first experiments to their small data sizes.

To test the applicability of the proposed method to other data, we conducted an exploratory study using another data set publicly available in the domain. The Immune Epitope Database (IEDB) collection[9] created in the annotation of IEDB consists of over 20,000 MEDLINE abstracts (27.3% positive) (Wang, et al., 2007). We conducted a 10-fold cross-validation test over this collection. For the ensemble classifier, we used the settings of R=10% and K=20 without adjustment. There were over 20,000 word features during the cross-validation test. For example, 6.3% of available features were selected with the proposed sampling method in one of the ten runs. For single classifiers, as before, we used top $1/2^i$ features sorted on IG for $i \geq 0$. ROC curves for single classifiers were very similar for i=0 to 5 (100% to 3.1% of available features), but the performance degraded for $i \geq 6$ (1.6% or less features). As in Figure 3, the proposed approach still yielded improved performance on this data set.

In our experiments, we observed promising results for the weighted random feature subsetting method in biological document classification. We plan to investigate the selection of the variables K and R for different sizes of data collections.

References

Bay, S.D. (1998) Combining nearest neighbor classifiers through multiple feature subsets. *In Proc. of the Fifteenth International Conference on Machine Learning (ICML)*. Morgan Kaufmann Publishers Inc., pp.37–45.

Chang, C.-C. and Lin, C.-J. (2001) LIBSVM : a library for support vector machines (Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm).

Donaldson, I., Martin, J., de Bruijn, B., Wolting, C., Lay, V., Tuekam, B., Zhang, S., Baskin, B., Bader, G.D., Michalickova, K., Pawson, T. and Hogue, C.W. (2003) PreBIND and Textomy--mining the biomedical literature for protein-protein interactions using a support vector machine, *BMC Bioinformatics*, **4**, 11.

Fawcett, T. (2006) ROC graphs with instance-varying costs, *Pattern Recognition Letters*, **27**, 882-891.

Han, B., Obradovic, Z., Hu, Z.Z., Wu, C.H. and Vucetic, S. (2006) Substring selection for biomedical document classification, *Bioinformatics*, **22**, 2136-2142.

Joachims, T. (1998) Text Categorization with Support Vector Machines: Learning with Many Relevant Features. *In Proc of the European Conference on Machine Learning (ECML)*. Springer.

Jones, K.S. (1972) A statistical interpretation of term specificity and its application in retrieval, *Journal of Documentation*, **28**, 11-21.

Sutton, C., Sindelar, M. and McCallum, A. (2005) Feature Bagging: Preventing Weight Undertraining in Structured Discriminative Learning. *Center for Intelligent Information Retrieval Technical Report*. University of Massachusetts.

Torii, M. and Liu, H. (2007) Classifier ensemble for biomedical document retrieval. *Languages in Biology and Medicine (LBM)*.

Wang, P., Morgan, A.A., Zhang, Q., Sette, A. and Peters, B. (2007) Automating document classification for the Immune Epitope Database, *BMC Bioinformatics*, **8**, 269.

[7] We used vertical-averaging of ROC curves, i.e., we derived average true positive rates at fixed false positive rates, 0.05, 0.10, 0.15, 0.20,, ….

[8] We sampled with replacement using the procedure in Section 3.2. and thus the percentage of actually selected features (6.9%) was less than R (10%).

[9] http://www.biomedcentral.com/1471-2105/8/269/additional/

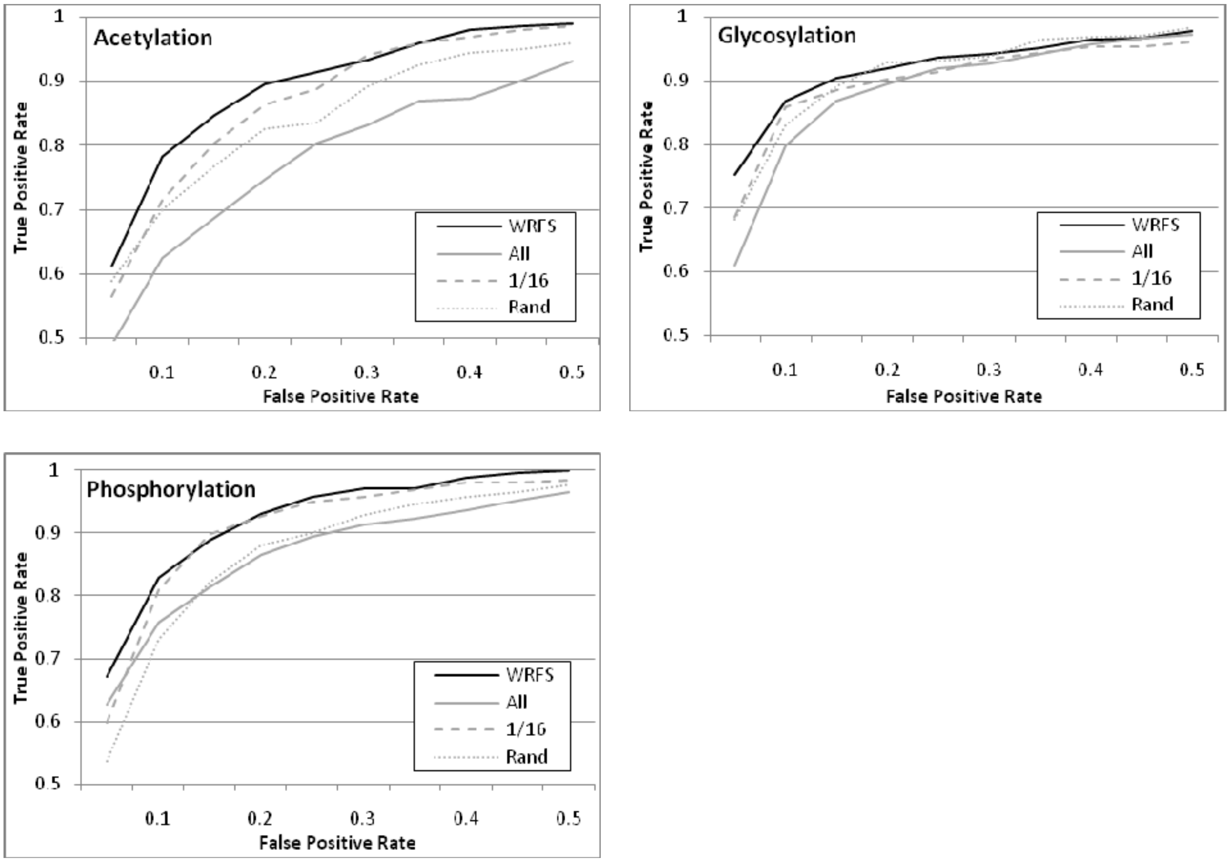


Figure 1.a ROC curves of classifiers for the three PTM collections (Acetylation, Glycosylation and Phosphorylation). The figures show the performance of single SVM classifiers using all features ("All"), the "best" single classifiers (ones with preferable ROC curves) using 1/16 of features ("1/16"), ensemble classifiers using weighted random feature subsetting ("WRFS") and ensemble classifiers using completely random feature subsetting ("Rand"). The curves are vertical averages of results obtained in 2×5-fold cross-validation tests. False positive rates greater than 0.5 are not shown because the results at high false positive rates are not of curators' interest in practice

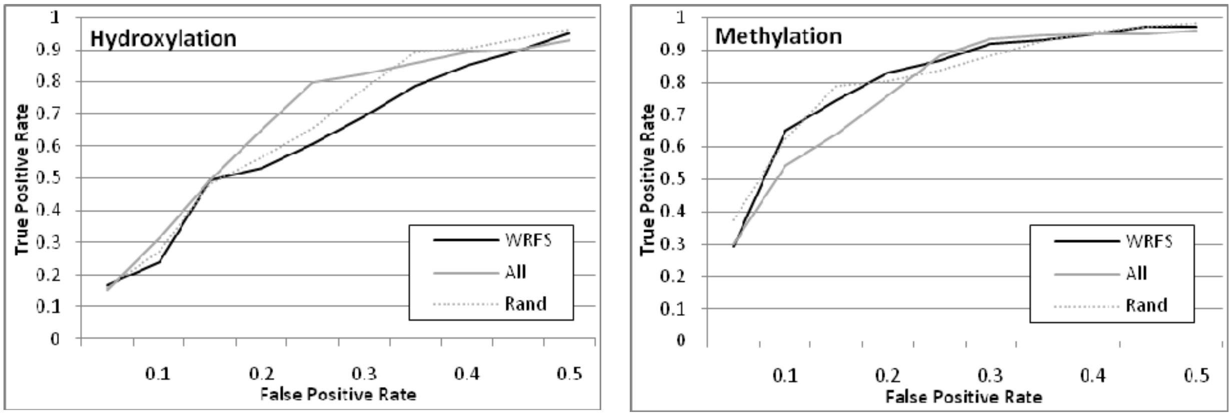


Figure 1.b ROC curves of classifiers for the two PTM collections with small data (Hydroxylation and Methylation). The figures show the performance of the best single SVM classifiers using all features ("All"), ensemble classifiers using weighted random feature subsetting ("WRFS") and ensemble classifiers using completely random feature subsetting ("Rand"). The curves are vertical average of results obtained in 2×5-fold cross-validation tests.

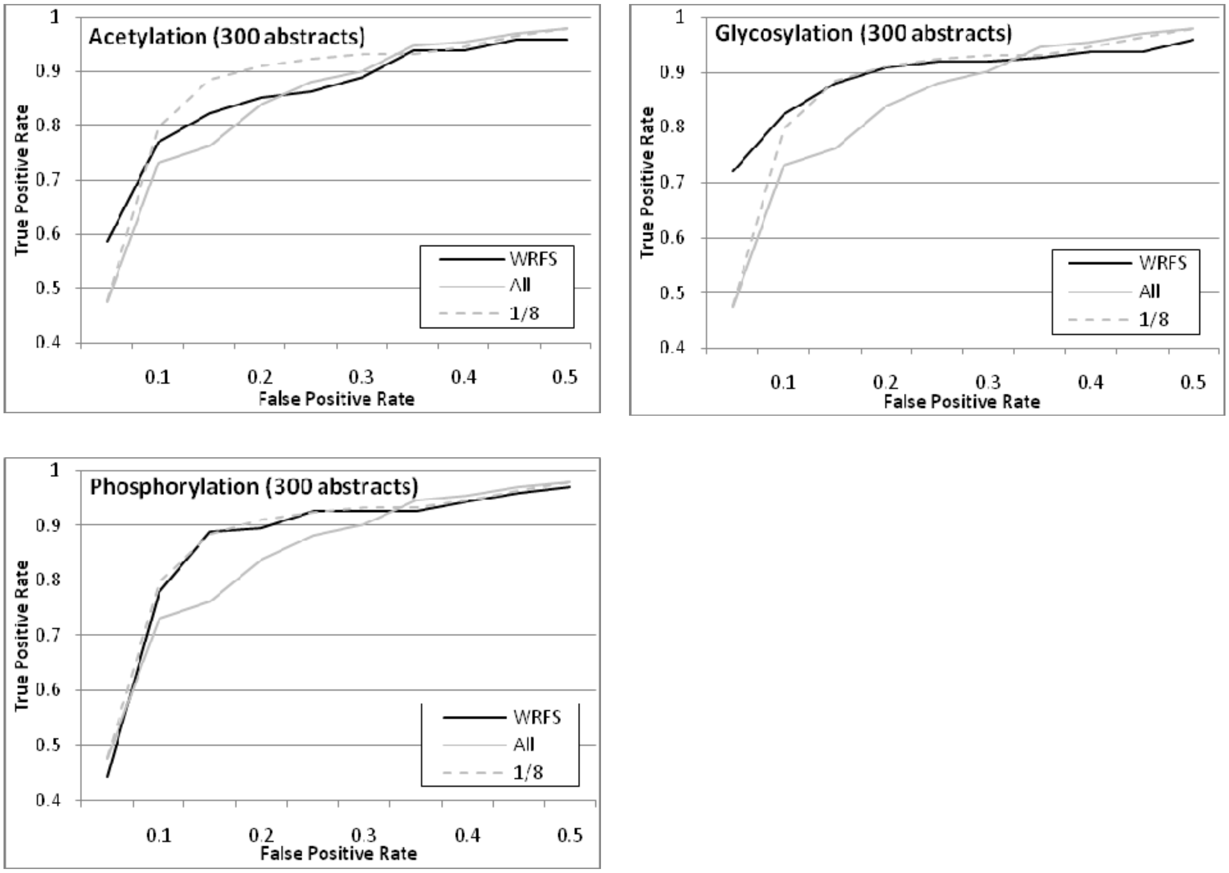


Figure 2. ROC curves of classifiers for small parts of PTM data collections. The figures show the performance of an ensemble classifier using weighted random feature subsetting ("WRFS"), a single classifier using all features ("All"), and the "best" single classifiers using 1/8 of all features ("1/8") on small document sets (300 abstracts) compiled from the acetylation, glycosylation, and phosphorylation collections. The curves are vertical average of results obtained in 2×5-fold cross-validation tests.

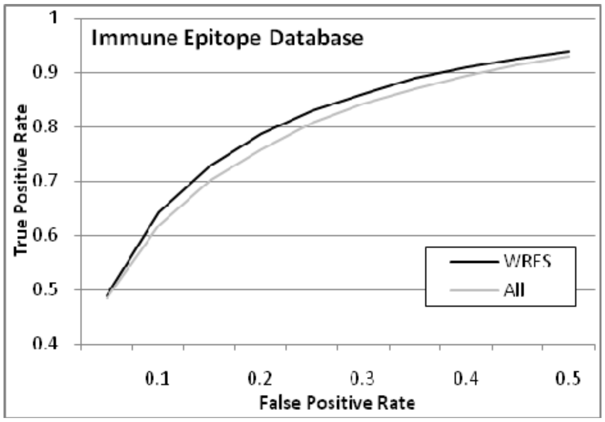


Figure 3. ROC curves for classifiers derived with the IEDB data collection. A 10-fold corss-validation test was conducted on the IEDB data collection to evaluate an ensemble classifier using weighted random feature subsetting ("WRFS") and the best single classifier using all features ("All").

Neonatal Sepsis Prediction with Machine Learning

Subramani Mani[1], Jörn-Hendrik Weitkamp, Constantin F Aliferis, Asli Ozdas, Huseyin Atakan Varol, Qingxia Chen, Randy Carnevale, Steven Steele
Vanderbilt University, Nashville, TN, USA

Abstract

Neonatal sepsis presents a challenging diagnostic and therapeutic problem to physicians. This study explores the application of machine learning (ML) algorithms for early detection of sepsis before blood culture results become available. Using data on 700 infants admitted to the neonatal intensive care unit (NICU) of a tertiary care hospital (Monroe Carell Jr. Children's Hospital at Vanderbilt) we show that ML algorithms can aid in early detection of sepsis thereby reducing unwanted antibiotic usage and lowering the risk of antibiotic resistance in the community. Following the best ML recommendation would result in prompt antibiotic treatment to an additional 18 babies with sepsis and prevention of probably unnecessary antibiotic treatment for 281 babies who did not have sepsis, when compared with physician providers.

1 Introduction and Background

Neonatal sepsis is a generalized infection occurring in newborn infants causing significant morbidity and mortality. It has an insidious onset with subtle clinical signs and symptoms but can quickly run a downhill course within 24-48 hours unless prompt antibiotic treatment is administered. More than half of the infants admitted to the neonatal intensive care unit (NICU) undergo sepsis evaluation (Gerdes, 2004). When sepsis is suspected blood is withdrawn for blood culture and the infant is started on antibiotics. Blood culture reports typically become available only after a period of 24-72 hours after blood is withdrawn. When the blood culture report becomes available the decision to continue or stop antibiotics is made based on culture positivity and the clinical profile of the baby. On an average for every culture positive sepsis result an additional ten or more babies receive antibiotic treatment (Gerdes, 2004) contributing to antibiotic resistance in the community and increased health care costs.

1.1 Sepsis Prediction

Many clinical and laboratory parameters have been studied for the task of sepsis prediction in newborns (Lam and Ng, 2008). Since sepsis in newborns has a low incidence with reference to the total number of babies delivered and the disease has high severity with rapid progression, we need a test that has high *sensitivity* (probability of test being positive given that sepsis is positive) and high *negative predictive value* (probability of sepsis being negative given that the test is negative) (Gerdes, 2004). However, no single clinical feature or laboratory test with high sensitivity and negative predictive value has been identified for accomplishing this task. Different non-specific tests such as C-reactive protein (a marker of inflammation), total white blood cell (WBC) count, absolute neutrophil count, the ratio of immature to total neutrophils (I/T ratio) have been evaluated in this context with limited success. The development of an effective sepsis prediction test in neonates has been described as an unmet medical need (Benjamin and Stoll, 2006).

1.2 Early and Late Onset Neonatal Sepsis

Based on the time of occurrence of signs and symptoms sepsis in the neonate can be categorized as early onset (within the first 72 hours after birth) and late onset (after 72 hours of birth). The etiology and evaluation varies for the early and late onset sepsis. Early onset sepsis is vertically transmitted from the mother to the baby during delivery and late onset sepsis is mostly of nosocomial origin (acquired in the hospital) (Benjamin and Stoll, 2006). However, in our study we include both early and late onset sepsis. The study investigates the feasibility of predicting sepsis with a high degree of certainty twelve to sixty hours before blood culture results become available (within twelve hours of blood withdrawal for blood culture).

The rest of the paper is organized as follows. In Section 2 we describe the dataset used in the study and in Section 3 we discuss the algorithmic and experimental methods employed for generating the results. In Section 4 we present the results and in Section 5 we discuss the implications of our findings. Limitations and future directions for our work are provided in Section 6 and Section 7 respectively.

2 Dataset

The data used in this study is from all babies admitted to the NICU in the Monroe Carell Jr. Children's Hospital at Vanderbilt University (VCH) over a period of 18 months starting from 01/01/2006. Out of the 1826 total admissions blood culture testing was performed on 952 infants. The final study sample consisted of 700 instances as explained below. The datasets used in this work include data

[1] Corresponding author (subramani.mani@vanderbilt.edu)

from NICU research database and data acquired from the electronic medical record system of the Vanderbilt University Medical Center (VUMC). The four datasets are listed below:

1. Antibiotics dataset has 1,952 records for 556 unique patients. Information such as start and end time, frequency of the antibiotics treatment and type of drug is stored in this dataset.
2. Microbiology test dataset contains 3,489 records for 952 unique patients. The time of the blood sample withdrawal and microbiology test results including the bacteria type and bacteria genus are included in this dataset.
3. Laboratory dataset contains 2,791,887 time-stamped measurements belonging to 71 different measured fields for 907 unique patients.
4. NICU dataset is a 210 variables by 1826 samples dataset containing variables such as birth weight, birth length, type of the delivery, Apgar score, etc.

Antibiotics, microbiology test, and laboratory datasets are used to assign sepsis labels. A sepsis label is assigned for each unique patient that had at least one microbiology test and laboratory dataset record. Using a unique identifier for each sample, the information in the laboratory and the NICU datasets are combined to generate a dataset for performing machine learning to predict the sepsis labels. From the time-stamped measurements in the laboratory dataset, temporal variables are generated for a period of 60 hours (starting 48 hours before and finishing 12 hours after the first blood test) with 6 hour increments. The time of withdrawal of blood for blood test is denoted by t(0) and is not a 6 hour interval. For example, t(-48) denotes the 6 hour duration starting from 48 hours before t(0) and t(12) denotes the 6 hour slot ending 12 hours after t(0). A subset of the NICU variables is selected for the initial analysis. Note that all the selected NICU variables are available within 12 hours after the first blood test t(0). The objective of the sepsis classification is to predict whether a baby has sepsis 12 hours after blood for the microbiology test has been drawn. 781 (71x11) temporal variables from the laboratory dataset and 30 non-temporal variables from the NICU dataset were selected that met this specification. These variables and their descriptions are listed in Appendix A.

The dataset contains many missing values particularly in the temporal variables. Therefore temporal variables are converted to non-temporal scalar variables by taking the last non-missing entry for a temporal variable (See Fig. 1). This is usually called last observation carry forward (LOCF). The final dataset consists of 906 instances and 101 variables. The histogram showing the number of variables for different missing value ratios indicate that only 36 variables in the dataset have less than 10% missing values (See Fig. 2.). A subset of this dataset comprising of 700 random instances is used in this work for predicting the sepsis labels using machine learning. The remaining 206 samples are reserved to be used as an independent testing set in the later phases of the study.

t(-48)	t(-42)	t(-36)	t(-30)	…	t(6)	t(12)	LOCF
?	12	?	15	…	?	10	10
20	?	14	16	…	12	?	12

Fig. 1. Temporal variable to non-temporal variable conversion to reduce the missing value ratio in the dataset.

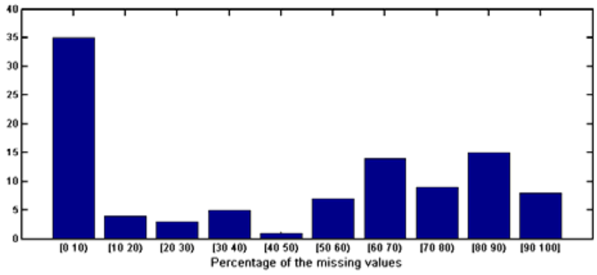


Fig. 2. Histogram showing the number of variables for different missing value ratios.

2.1 Sepsis Labeling Algorithm

We now describe the algorithm used to label the instances as Sepsis-true and Sepsis-false using all the available data. These labels were considered as gold-standard labels for our study purposes. The algorithm is based on the review article on neonatal sepsis (Gladstone et al., 1990) and the current best practice in the VCH NICU.

```
Sepsis Labeling Algorithm
-----------------------------------------------------------------------------
If all microbiology test results[2] are negative, assign sepsis label as false.
Else for each positive microbiology test result,
  If bacteria type is negative, start processing the next microbiology test result.
    If bacteria type is definite, assign sepsis label as true.
    If bacteria type is symptomatic, assign sepsis label as true if the baby has a laboratory condition and an observational condition.
      • A baby has a laboratory condition if one of the following conditions is true:
        o A second microbiology test result is positive for the same organism taken within 30 hours of the first test.
        o White blood cell count is less than 5,000 or more than 30,000 per microL.
        o Platelet count is less than 150,000 per microL.
        o Neutrophil count is less than 1,750 per microL.
      • A baby has an observational condition if one of the following conditions is true:
        o Antibiotic treatment more than 96 hours after the blood withdrawal for the microbiology test.
        o Pulse is less than 100 beats per minute (bradycardia).
        o Temperature is greater than 38°C (fever) or less than 36.5 °C (hypothermia).
Assign sepsis label as false, if sepsis label is not assigned.
-----------------------------------------------------------------------------
```

[2] Denotes blood culture test.

2.2 NICU demographics table

Some statistics on the demographics of the infants in the dataset is given in Table 1. Statistics for the whole study population is provided in column 2 and for the subgroup of very low birth weight (VLBW) babies in column 3.

Table 1 – Demographics of the study population.

	Whole Sample	VLBW (<1500 grams)
Number of infants	700	182
Birth weight (grams), median (25,75) percentiles	2180 (1445 , 3073)	1035 (770 , 1280)
Gestational age (weeks), median (25,75) percentiles	34 (31 , 38)	28 (26 , 30)
Male n (%)	403 (58%)	94 (52%)
Race (white) n (%)	551 (79%)	130 (71%)
Sepsis (positive) n (%)	81 (11.5%)	40 (22%)

3 Methods

Three classification algorithms were selected for the sepsis prediction (classification) task using the dataset. These algorithms are the support vector machine (SVM) (Vapnik, 1998), the Naïve Bayes (NB) classifier (Duda and Hart, 1973) and the decision tree (DT) classifier CART (Breiman et al., 1984). SVMs are considered as state-of the-art machine learning algorithms for classification by the machine learning community. The Naïve Bayes classifier has been used in many applications for clinical decision making. SVM based algorithms cannot handle missing values and therefore the dataset needs to be imputed. A simple mean value imputation method is employed to impute the dataset. For each variable in the training set, the mean and the mode of the non-missing values is assigned for the missing values of the continuous and nominal variables, respectively. For the missing values in the test sets, the mean and mode of the non-missing values in the training set are used.

In order to increase classification performance, it is a good practice to use feature selection algorithms which select a subset of the features that are highly predictive of the class. In this work, three feature selection algorithms are employed for feature selection. These are non-parametric one-way ANOVA: Kruskal-Wallis (KW) (Hollander and Wolfe, 1973), HITON-MB and HITON-PC algorithms (Aliferis et al., 2003a). For finding the optimal classifier and feature selection algorithm combination for the problem, a nested cross-validation (CV) procedure (Statnikov et al., 2005) is employed. In the nested cross-validation procedure, the parameters of the classifiers are optimized in the inner CV loop and the outer CV loop is used to evaluate the performance of the models (See Fig. 2). The number of features to be selected by KW algorithm is set from 5 to 25 with 5 increments. The nested CV design fully penalizes for feature selection. The maximum cardinality of the conditioning set for the HITON algorithms is set to 3 and the threshold values are selected to be 0.01 and 0.05. The SVM classifier is run with a polynomial kernel with the cost parameter C ranging from 0.01 to 100 with a multiplicative step of 10 and kernel degrees from 1 to 5. The number of the CV folds K is set to 10. The decision tree and Naïve Bayes classifiers do not have any parameters for optimization but their performances are also optimized using different feature selection algorithms with different set of parameters in the nested CV procedure. Area under the Receiver Operator Characteristics Curve (AUC) is selected as the performance metric since it provides an efficient metric independent of the class sizes and the classification threshold (Fawcett, 2006).

The nested CV procedure is implemented in Matlab. For the SVM classifiers, LibSVM (Chang and Lin, 2001) is called from Matlab. The WEKA implementation of the Naïve Bayes (Witten and Frank, 2005) algorithm is executed using Matlab scripts and the number of variables to be selected is set to 5, 10, 15, 20 and 25. The Matlab implementation of the Classification and Regression Trees is used for the decision tree classifier. HITON-MB and HITON-PC feature selection algorithms are called from Causal Explorer (Aliferis et al., 2003b).

```
Nested Cross-Validation Procedure:
1.Repeat for $K$ folds:
    •  Training set ($K-1$ partitions)
    •  Testing set (remaining partition)
  1.1 Repeat for $i = 1, \dots , K-1$:
      o Parameter optimization training set ($K-2$) partitions
      o Parameter optimization testing set (remaining partition)
      o Train the classifier $X$ on the parameter optimization training set using parameter set $g_i$ .
      o Test it on the parameter optimization test set.
      o Record $P(i)$, the average performance of $X$
      o for all the $K-1$ inner cross–validation loops.
  1.2 Determine the parameter set $g_j$ which maximizes $P(i)$ for $i = 1, \dots , K-1$.
  1.3Train the classifier X on the training set with parameter set $g_j$ .
      o Test the classifier obtained in 1.3 on the testing set.
2. Return $p$, the average performance of X over all $K$ testing sets.
```

Fig. 2. Nested CV procedure modified from GEMS nikov et al., 2005) for performance estimation in the outer loop and parameter optimization in the inner loop.

4 Results

The maximum, minimum and mean number of features selected for the SVM, NB and DT classifiers over the ten outer CV folds are summarized in Table 2. In all the cases, the features are selected using HITON-PC or HITON-MB algorithms.

Table 2 – Features selected for SVM, NB and DT Algorithms over 10 folds

	Minimum # of features	Maximum # of features	Mean # of features (std. dev.)
SVM	7	18	11.5 (3.6)
NB	7	11	9.1 (1.1)
DT	7	15	12.5 (3.3)

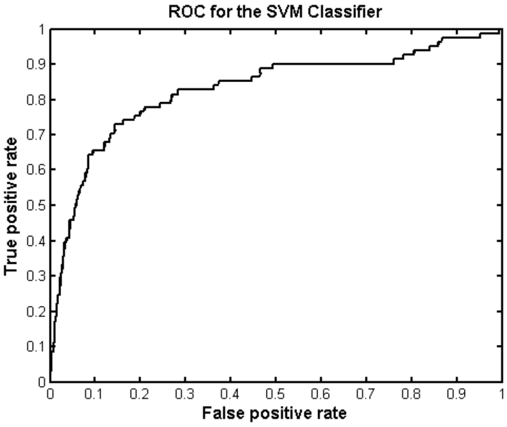


Fig. 3. ROC Curve using the whole sample for the SVM for the CV-fold with the nearest AUC score to the mean AUC over 10 folds (AUC = 0.82).

The mean AUC scores for the SVM, DT and NB classifiers for the 10-fold outer CV are 0.77 (s.d. 0.1), 0.72 (s.d. 0.13) and 0.75 (s.d. 0.11), respectively. The ROC curve for the SVM outer CV fold model (which is nearest to the mean AUC of the ten folds) when applied to the whole 700 sample training dataset is shown in Fig. 3. To assess the clinical impact of the ML approach we compared the sensitivity and specificity of the physicians and the ML algorithms as follows. We first define treatment sensitivity and treatment specificity for purposes of this comparison based on antibiotic treatment.

Treatment Sensitivity = (Number of sepsis babies treated) / (Number of sepsis babies in the whole dataset)
Treatment Specificity = (Number of non-sepsis babies not treated) / (Number of non-sepsis babies in the whole dataset)

For the calculation of the treatment sensitivity, a baby is considered treated if the physician started an antibiotic treatment between t(-48) and t(12)[3]. For the calculation of treatment specificity, a baby is considered not treated if no antibiotics are given to the baby between t(12) and t(120)[4]. The required data for the antibiotics treatment is acquired by processing the start and end time of the antibiotics treatments in the antibiotics dataset. Using the sepsis labels and antibiotics treatment information, the physician confusion matrix was created (Table 3). The confusion matrices for NB, DT and SVM algorithms were generated as explained below. To compare NB, DT and SVM confusion matrices to the physician confusion matrix, first the NB, DT and SVM confusion matrices are generated such that their treatment specificities are the same as the physician confusion matrix (Tables 4-6). This allows determining whether NB, DT and SVM perform better than the physicians with respect to the treatment sensitivity if they have the same treatment specificity of the physicians. Secondly, the NB, DT and SVM confusion matrices are generated such that their treatment sensitivities are the same as that of the physician (Tables 7-9). Then their treatment specificities are compared to the treatment specificity of the physician.

Table 3: Physician (Treatment Sensitivity 0.67; Treatment Specificity 0.27)

	Not treated	Treated
No sepsis	167	452
Sepsis	27	54

Table 4: Naïve Bayes with fixed treatment specificity (Treatment Sensitivity 0.78; Treatment Specificity 0.27)

	Not treated	Treated
No sepsis	167	452
Sepsis	18	63

Table 5: DT with fixed treatment specificity (Treatment Sensitivity 0.80; Treatment Specificity 0.27)

	Not treated	Treated
No sepsis	167	452
Sepsis	16	65

Table 6: SVM with fixed treatment specificity (Treatment Sensitivity 0.89; Treatment Specificity 0.27)

	Not treated	Treated
No sepsis	167	452
Sepsis	9	72

Table 7: Naïve Bayes with fixed treatment sensitivity (Treatment Sensitivity 0.67; Treatment Specificity 0.61)

	Not treated	Treated
No sepsis	378	241
Sepsis	27	54

Table 8: DT with fixed treatment sensitivity (Treatment Sensitivity 0.67; Treatment Specificity 0.49)

	Not treated	Treated
No sepsis	302	317
Sepsis	27	54

Table 9: SVM with fixed treatment sensitivity (Treatment Sensitivity 0.67; Treatment Specificity 0.72)

	Not treated	Treated
No sepsis	448	171
Sepsis	27	54

[3] Note that the goal is to predict sepsis before blood culture results become available.

[4] Antibiotic treatment may be started after blood culture results become available.

5 Discussion

The results indicate that machine learning (ML) algorithms can be used to predict sepsis in neonates within twelve hours of blood withdrawal for the blood culture test. The potential clinical impact of the ML models appears promising.

With the treatment specificity set at the level of the physician, the treatment sensitivity level of all the three algorithms exceeded that of the physician. When compared to the physician, NB, DT and SVM promptly treated 9, 11 and 18 additional babies respectively who had sepsis. When the treatment sensitivity was at the level of the physician, the treatment specificity of the algorithms also showed a marked increase. Compared to the physician, NB, DT and SVM prevented unnecessary treatment of the 211, 135 and 281 infants not having sepsis, respectively. This shows that ML algorithms can reduce unnecessary antibiotic therapy in the NICU while treating more infants with sepsis at an earlier stage.

Pearson's Chi-square test (one sided) was done for ascertaining whether the results of the machine learning algorithms are statistically significantly different from the physician's treatment decisions. The following *p* values were obtained: *p*=0.057 for NB with fixed specificity (Table 3 versus Table 4), *p*=0.025 for DT with fixed specificity (Table 3 versus Table 5) and p=0.0003 for SVM with fixed specificity (Table 3 versus Table 6). The tests for the Table 3 versus Table 7-9 (with fixed sensitivity) resulted in *p* values very close to zero showing great statistical significance. In general, all of the comparisons except one are statistically significant based on the cutoff *p* value of 0.05 when compared with the decisions made by physician (see Table 10).

Table 10: One sided Pearson’s Chi-square significance test results (*p* values) comparing machine learning algorithms with physician (Phy: Physician; Treat: Treatment).

	Fixed Treat Specificity			Fixed Treat Sensitivity		
	NB	DT	SVM	NB	DT	SVM
Phy	0.057	0.025	0.0003	<0.0001	< 0.0001	< 0.0001

The study that is closest to ours reported in the literature is a prospective study for sepsis prediction using heart rate characteristics, other clinical and laboratory measurements in a NICU setting (Griffin et al., 2005). Griffin et al. report an AUC of 0.82 for sepsis prediction using multivariable logistic regression based on a comparison between cases and controls. On the other hand our study population consisted of infants for whom a blood culture test had been ordered. Predicting sepsis from a study group consisting of only infants with some clinical suspicion of sepsis is a harder task. Additional evaluation of the methodology is needed before the results can be put to clinical use.

6 Limitations

The LOCF method that we used to convert temporal variables to non-temporal scalar variables can introduce bias when the reason for the missing observation depends on the unobserved value itself. Basically LOCF assumes that a value that was missing would be found to be identical to the previous value, with no error (Mallinckrodt et al., 2008).

7 Future Work

More sophisticated methodologies including maximum likelihood method with expectation-maximization algorithm, multiple imputation, or fully Bayesian approaches will be employed to extract information from the laboratory measurements with temporal data and studied with machine learning for better performance. Other feature selection and classification methods could be employed to increase the classification performance. The results of the SVM classification can be post-processed to create human-understandable models such as decision trees or rules.

Acknowledgments

We thank Yerbolat Dosbayev for creating an integrated dataset from multiple data sources for purposes of this study. We thank Alexander Statnikov, Shanthi Krishnaswami and Professor Frank Harrell for helpful discussions.

A List of Variables

No.	Variable Name	Original Dataset
1	Atyp Lymphs %	Laboratory
2	Ax Temp	Laboratory
3	Baso (ABS)	Laboratory
4	Base Excess Arterial	Laboratory
5	Base Excess Venous	Laboratory
6	Basophils %	Laboratory
7	Base Excess Capillary	Laboratory
8	Bicarbonate (Calc)	Laboratory
9	PCO2 Capillary	Laboratory
10	pH Capillary	Laboratory
11	Bilirubin Conjugated	Laboratory
12	CMB Temperature	Laboratory
13	C-Reactive Protein	Laboratory
14	CUM48 Transfusion	Laboratory
15	CUM Transfusion	Laboratory
16	Calcium Ionized	Laboratory
17	Creatinine Blood	Laboratory
18	EO Automated Abs	Laboratory
19	Eosinophil %	Laboratory
20	FlO2	Laboratory
21	Fibrinogen	Laboratory
22	Glucose Whole Blood	Laboratory
23	Glucose Blood	Laboratory
24	Bicarbonate (Calc)	Laboratory
25	PCV Blood	Laboratory
26	Heart Plus Pulse	Laboratory
27	Heart Then Pulse	Laboratory
28	Heart Rate	Laboratory
29	Potassium Whole Blood	Laboratory
30	Lactate Whole Blood	Laboratory

31	Lymphs %	Laboratory
32	Lymphs (Abs)	Laboratory
33	Atyp Lymphs (Abs)	Laboratory
34	Metamyelocytes (Abs)	Laboratory
35	Myelocytes (Abs)	Laboratory
36	Promyelocytes (Abs)	Laboratory
37	Metamyelo %	Laboratory
38	Mono (Abs)	Laboratory
39	Monocytes %	Laboratory
40	Myelocytes %	Laboratory
41	Nt Automated Abs	Laboratory
42	Sodium Whole Blood	Laboratory
43	Neutrophils %	Laboratory
44	Neutrophils (Abs)	Laboratory
45	O2 Saturation	Laboratory
46	O2 Saturation (Calc)	Laboratory
47	O2 Saturation (Venous)	Laboratory
48	Parental Transfusion	Laboratory
49	Inr	Laboratory
50	Patient PTT	Laboratory
51	Patient (PT)	Laboratory
52	Platelet Count	Laboratory
53	Promyelo %	Laboratory
54	Pulse Rate	Laboratory
55	Respiratory Rate	Laboratory
56	Ast Blood	Laboratory
57	Alt Blood	Laboratory
58	Bilirubin Total Blood	Laboratory
59	Total Hemoglobin	Laboratory
60	Temperature	Laboratory
61	Bacteria	Laboratory
62	Glucose Urine	Laboratory
63	Leukocyte Esterase	Laboratory
64	Nitrite	Laboratory
65	White Blood Cell Count	Laboratory
66	pCO2 Venous	Laboratory
67	pCO2 Arterial	Laboratory
68	Ph Arterial	Laboratory
69	Venous Ph	Laboratory
70	pO2 Arterial	Laboratory
71	PO2 Venous	Laboratory
72	Apgar Score (one minute)	Laboratory
73	Apgar Score (five minutes)	NICU
74	Apgar Score (ten minutes)	NICU
75	Maternal Anesthesia	NICU
76	Birthweight	NICU
77	CRIBscore	NICU
78	Chorioamnionitis	NICU
79	Diabetes	NICU
80	Resuscitation with Bag/Mask	NICU
81	Resuscitation with Cardiac Comp.	NICU
82	Resuscitation with Epinephrine	NICU
83	Resuscitation with Intubation	NICU
84	Resuscitation with Oxygen	NICU
85	Substance Usage	NICU
86	Fetal Monitoring	NICU
87	Gestational Age (weeks)	NICU
88	Gravida	NICU
89	Ethnicity of the mother	NICU
90	Length of the baby	NICU
91	Meconium in Amniotic Fluid	NICU
92	Mother's Age	NICU
93	Total babies in this pregnancy	NICU
94	Mother's Race	NICU
95	Preterm Labor	NICU
96	Number of prev. deliveries	NICU
97	Sex	NICU
98	Vaginal Delivery	NICU
99	Vaginal Presentation	NICU
100	Vertex Presentation	NICU
101	Is Sepsis	Class Label

References

Gerdes, J. S. (2004). Diagnosis and management of bacterial infections in the neonate. *The Pediatric Clinics of North America* 51: 939-959.

Lam, H. S. & Ng, P. C. (2008). Biochemical markers of neonatal sepsis. *Pathology* 40: 141-148.

Benjamin, D. K. & Stoll, B. J. (2006). Infection in Late Preterm Infants. *Clinics in Perinatology* 33: 871-882.

Gladstone, I. M., Ehrenkranz, R. A., Edberg, S. C. & Baltimore, R. S. (1990). A ten-year review of neonatal sepsis and comparison with the previous fifty-year experience. *The Pediatric Infectious Disease Journal* 9: 819.

Vapnik, V. N. (1998). *Statistical learning theory*: Wiley New York.

Duda, R. O. & Hart, P. E. (1973). Pattern Recognition and Scene Analysis. *John Willey and Sons, New York.*

Breiman, L., Friedman, J. H., Olshen, R. A. & Stone, C. J. (1984). *Classification and Regression Trees* Belmont: Wadsworth.

Hollander, M. & Wolfe, D. A. (1973). *Nonparametric statistical methods*: Wiley New York.

Aliferis, C. F., Tsamardinos, I. & Statnikov, A. (2003a). HITON: A Novel Markov Blanket Algorithm for Optimal Variable Selection. *AMIA... Annual Symposium proceedings [electronic resource]* 2003: 21.

Statnikov, A., Tsamardinos, I., Dosbayev, Y. & Aliferis, C. F. (2005). GEMS: A system for automated cancer diagnosis and biomarker discovery from microarray gene expression data. *International Journal of Medical Informatics* 74: 491-503.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters* 27: 861-874.

Chang, C. C. & Lin, C. J. (2001). LIBSVM: a library for support vector machines, 2001. *Software available at* http://www.csie.ntu.edu.tw/cjlin/libsvm.

Witten, I. H. & Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques*: Morgan Kaufmann, San Francisco.

Aliferis, C. F., Tsamardinos, I., Statnikov, A. & Brown, L. E. (2003b). Causal Explorer: A Causal Probabilistic Network Learning Toolkit for Biomedical Discovery. *International Conference on Mathematics and Engineering Techniques in Medicine and Biological Sciences (METMBS’03)*: 371–376.

Griffin, M. P., Lake, D. E. & Moorman, J. R. (2005). Heart Rate Characteristics and Laboratory Tests in Neonatal Sepsis. *Pediatrics* 115: 937-941.

Mallinckrodt, C. H., Lane, P. W., Schnell, D., Peng, Y. & Mancuso, J. P. (2008). Recommendations for the Primary Analysis of Continuous Endpoints in Longitudinal Clinical Trials. *Drug Information Journal* 42: 303-319.

Natural Language Query in the Biochemistry and Molecular Biology Domains Based on Cognition Search™

Elizabeth J. Goldsmith[†]||, Saurabh Mendiratta[†], Radha Akella[†,] and Kathleen Dahlgren||[§]

[†]Department of Biochemistry, The University of Texas Southwestern Medical Center at Dallas, 5323 Harry Hines Boulevard, Dallas, Texas 75390-8816. [§] Cognition Technologies, Inc, 6133 Bristol Parkway, Culver City, CA 90230.

Abstract

Motivation: With the tremendous growth in scientific literature, it is necessary to improve upon the standard pattern matching style of the available search engines. Semantic NLP may be the solution to this problem. Cognition Search (CSIR) is a natural language technology. It is used by asking a simple question that might be answered in textual data being queried. CSIR has a large English dictionary and semantic database enables the search process to be based on meaning rather than statistical pattern matching. Several features such as encoded synonymy, ontological relationships, phrases and seeds for word sense disambiguation should be particularly useful in improving access to MEDLINE.

Result: Here we have carried out several projects to "teach" the CSIR lexicon medical, biochemical and molecular biological language and acronyms from curated web-based free sources. Vocabulary from the Alliance for Cell Signaling (AfCS), the Human Genome Nomenclature Consortium (HGNC), the United Medical Language System (UMLS) Meta-thesaurus, and The International Union of Pure and Applied Chemistry (IUPAC) was introduced into the CSIR dictionary and curated. The resulting system was used to interpret MEDLINE abstracts. Meaning-based search of MEDLINE abstracts yields high precision (estimated at >90%), and high recall (estimated at >90%), where synonym, ontology, phrases and sense seeds have been encoded. The present implementation can be found at http://MEDLINE.cognition.com.

Contact:
Elizabeth.goldsmith@UTsouthwestern.edu
Kathleen.dahlgren@cognition.com

Introduction

With the increasing complexity of biomedical literature, several labs and companies have attempted to develop better search engines for MEDLINE (1-5). A few free sources are visible on the web e.g. Google Scholar (http://scholar.google.com/), Highwire press (http://highwire.stanford.edu/lists/freeart.dtl) whereas other relatively commercial sources of this information is present at Scopus (http://www.scopus.com/scopus/home.url), Ovid (http://www.ovid.com/site/index.jsp), and Infotrieve (http://www4.infotrieve.com/newMEDLINE/search.asp). Semantic NLP, which helps in understanding the meaning of the word in context is require to properly accessing the biomedical literature (4, 6-13). The problem with NLP technology is that it requires extensive knowledge and ability to manipulate it. Cognition semantic NLP is the only technology that has unraveled the full complexity of ordinary English. The goal in search technology is to create software that finds all the desired information (full recall) without producing undesired information (high precision). The pattern matching style of standard search engines is limited. It just matches string of letters to string of letters ignoring the context and meaning, thus misses the relevant information. Cognition's comprehensive semantic map includes words, phrases, synonyms and sense seeds (together with synonyms is able to select senses of ambiguous words), enabling the software to find desired content hidden in masses of other information.

Architecture of CSIR™

CSIR™ is a natural language processing (NLP) technology that has been under development for several years. The technology contains a broad semantic map of English based on word senses, their synonyms (6), hypernyms (higher nodes in an ontology) (7) and sense seeds (words triggering a particular word meaning). The CSIR Indexer uses its NLP component to build a cognitive model of the text in which all of the concepts (word meanings) of a document are indexed in an offline job. The index contains word strings for pattern matching as a default. The indexer relies on the dictionary, semantic map, morphological and syntactic tags, and database of synonyms and ontological relationships (Fig.1). At search time, CSIR interprets the query for meaning, and searches for the meaning of the query in the concept index. The patented meaning-based architecture and methods have been described previously (14-16).

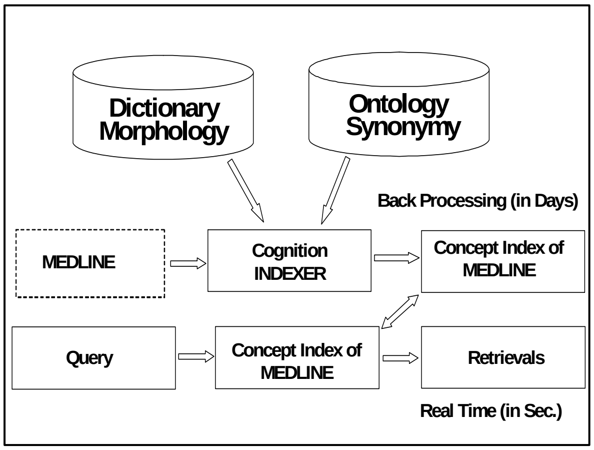


Figure 1: Architecture of CSIR

Since the original descriptions of this technology, significant improvements have been introduced, including sense disambiguation using sense seeds (8), phrase parsing (17), data compression and speed upgrades (18). The morphology and tokenization (word and phrase identification) components were built in-house (patent pending). The software also uses simple algorithms for phrasal parsing and concept clustering to improve document relevancy (precision). Demonstrations of CSIR are available at http://medline.cognition.com and http://wikipedia.cognition.com. The search engine should be used asking a straightforward question that might be answered in MEDLINE, such as "Oxidative stress in plants," “spectroscopy of amidohydrolases," or “Depression in aging.” Retrieval time on the 17 million MEDLINE abstracts is sub-second on Xeon Dual Core 3.0 GHz computers with 1 GB of RAM.

Methods

Ontology:

To augment the ontology for biochemistry and molecular biology, a top ontology was constructed by hand, based upon our own domain knowledge. Websites of curated biomedical terminology were crawled to their ontological attachments (as well as other data). Specialized programs were written to crawl each website. The ontological attachments were then mapped to our top ontology by hand.

Lexical and Concept Thesaurus Augmentation:

Biomedical terminological databases were crawled and the vocabulary (terms, phrases and acronyms) extracted, along with their synonyms. Vocabulary was checked for frequency in the MEDLINE abstracts and rank ordered for attention by curators. Curated terms and synonyms were automatically added to the CSIR semantic map. Acronym spell-outs were used as sense contexts for acronym meanings (9).

Precision and Relative Recall Test of CSIR vs PubMed.

Queries were formulated in formats consistent with either Cognition or PubMed (as a question for Cognition and as Boolians for PubMed). The total number of CSIR retrievals was recorded, and the relevance evaluated for the top 10 and top 20 retrievals, as assessed by the UT Southwestern team. The same queries were posed to PubMed for comparison (in a Boolean format: “genetic” AND “interaction” AND “BCL2”). To make the evaluation manageable, we used the “relative recall” technique, wherein full recall is estimated as the greatest number of retrievals achieved by either search engine. For example, one query was "genetic correlates of alcoholism". Of the first twenty CSIR retrievals, 16 were relevant. Thus CSIR's precision was 16/20 or 0.8. The number of retrievals for CSIR was 1,436. To extrapolate the good retrievals, we multiplied the precision ratio 0.8 times 1,436 to yield extrapolated recall of 1,149. The queries used here can be seen on the E.J. Goldsmith Lab webpage (http://hhmi.swmed.edu/Labs/bg/Cognition).

Results

Scale and Scope:

At the initiation of this project, a lexical evaluation of MEDLINE showed that CSIR was missing 66,000 tokens (words). Estimates of the total number of Biomedical terms is over a million, a much larger number, mostly phrases (10). Before this work, the CSIR Lexicon contained about 20,000 medical or biological terms (species, cells, anatomy, etc.). Here we added about 85,000 protein names, 35,000 chemical names, ontology for biochemistry and molecular biology possessing 2,400 nodes, and over 30,000 biomedical synonym classes. Together with other ongoing lexical augmentations, the detailed description of the entire Cognition semantic map is present in Table 1.

Cognition's Semantic Map
(Based on Computational Linguistic Science)

Word Stems	**506,000 Word stems**
Words and Phrases	**536,000 Word senses or concepts**
Meanings in context	**4,000,000 Semantic contexts**
Different Word Meanings	**17,000 Ambiguous word definitions**
Complex Word Series Meanings	**191,000 Phrases**
Ontology or Taxonomy	**7,000 Nodes**
Synonyms	**76,000 Thesaural concept groups**

Table 1: Cognition Dictionary by numbers

Ontology for Biochemistry and Molecular Biology

Ontologies need to be established at the desired granularity. We defined a top ontology for the biochemical and molecular biology domain that serves as a basis for capturing finer, more desired ontological nodes. Our top ontology, primarily for molecular entities, resembles SEMEDA (7), or TAMBIS (11). The very top of our ontology discriminates 'proteins," laboratory procedures," etc.; an intermediate level of protein and gene names was inspired by the ontology in the AfCS (eg. "binding protein," "g-protein", transcription-factors, etc), and by an ontology of terms in the HGNC that categorizes proteins and genes. (Table 2)

Table 2A: Ontology of Biochemical and Molecular Biology

A. **Piece of the Top Ontology for Biochemistry**
Macromolecule-node
Protein-stuff
antibody
binding protein
enzyme
Nucleic-acid
Laboratory-procedure
electrophoresis
Spectroscopy
B. **Ontology for protein kinases**
protein-kinases
protein-histidine-kinases
serine-threonine-kinases
AGC-kinases
STE-kinase
Tyrosine-kinase
ACK-kinase
EGFR-kinase
Tyrosine-Like-Kinase
MLK-kinase
RAF-kinase

Table 2B: Finer grained Protein Kinases ontology.

Introducing new language from existing databases:

Web-based sources of biomedical terminology were: acronyms from http://medstract.med.tufts.edu (6), the molecules and genes defined by the AfCS database (19), the Human Genome Nomenclature Consortium (20), the UMLS Metathesaurus and the International Union of Pure and Applied Chemistry (IUPAC) enzyme names. The acronym database and UMLS were selected for their wide coverage. We selected the AfCS and HGNC databases because the curators captured natural word usage, and have encoded a gross molecular ontology as well as some synonymy. The IUPAC database was chosen because the ontology has been constructed carefully. Some of the larger databases were avoided because we noted numerous errors and short and redundant acronyms, requiring too much curation. Since some acronyms were added to the semantic map in earlier projects, a challenge was to add only new senses (21). The database published at http://medstract.med.tufts.edu was used. We curated 16,256 acronyms, removing rarely used acronyms (usage cutoff of 20), and very redundant acronyms. This resulted in 15,657 acronyms with 16,858 total meanings.

We introduced vocabulary from the UMLS Metathesaurus. We built a map from the Metathesaurus ontology to our existing ontology, and then introduced the UMLS vocabulary into the lexicon automatically. Multi-sense words were inspected by a linguist to prevent duplication. Synonyms, with the appropriate senses, were introduced to the Concept Thesaurus automatically.
This database includes both nouns and verbs covering biological sciences and medicine, amounting to 88,423 word senses, and 76,816 synonyms.
We then obtained additional word senses, all nouns, from the Alliance for Cell Signaling (www.alliance.org) (19). This source is current, curated and offers ontological entries, giving 15,661

new or improved word senses. The adoption of this vocabulary was accomplished through a combination of automated tasks and expert curation. Duplicates were curated. Unknown vocabulary was then added to the semantic map automatically, including ontological attachments and synonyms. Data from the HGNC (www.genenames.org) (20) has also been partially introduced. About 30 ontologies of protein families in HGNC have been imported, including AKAPs, ADAM proteases, bcl, BRCA, channel proteins, P450s, tubulins, ubiquitin ligases, phosphatases, TNF-receptors, histones, SMADs, and so on. We also introduced the IUPAC enzyme names and EC numbers, over 6,000 names. These were chosen because of the well-thought-out ontology that may be accessed with the EC numbers. A difficulty with this augmentation is the lack of natural language usage and lack of synonymy.

Missing words by frequency

The numbers of words or tokens present in MEDLINE by missing in the Cognition dictionary were counted. Unknown works with frequency greater than 100 were curated; there were only 800 of these. The remainder gave the frequency distribution shown in Fig. 2. As can be seen in Fig. 2, capturing the words with frequency greater than 20 is desirable. At this writing, we have introduced most words with frequency greater than 50.

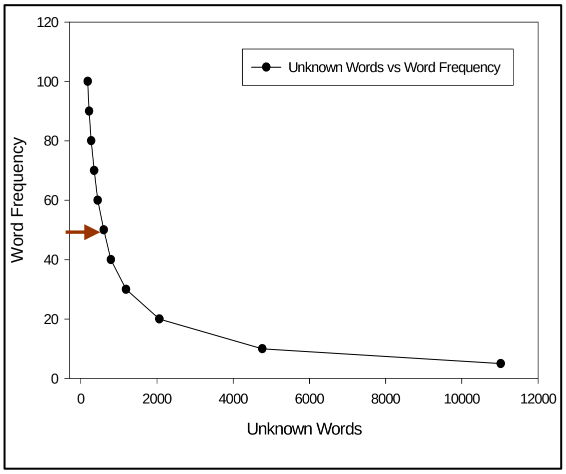


Figure 2 shows the Coverage of MEDLINE

Verbs

MEDLINE abstracts were also searched to find verbs, which were curated to find words (such as express, silence, translocate, spin, bait, prey) that have domain specific-meanings. This project has led to 225 new word senses. The added verb definitions contribute to improved precision through word sense disambiguation, and will be useful when full sentence parsing is included in CSIR (12)**.**

Precision and Recall Test

Fifty queries for MEDLINE were formulated as simple questions in the areas of biochemistry, molecular biology and medicine. The UT Southwestern team tabulated the relevance of the retrievals in http://MEDLINE.cognition.com and compared them with those of PubMed (http://pubmed.com) retrievals. As described in the methods, queries were formatted to conform to the two different search engines, and the relative recall method was used for evaluation. As can be seen Cognition did better by both precision and recall measures (Table 3). The reader, however, is perhaps the best judge of the relative performance of the search engines.

Table 3 Precision and Recall: Comparison between Cognition and Pubmed.

Cognition vs MEDLINE search	Cognition good/20	Cognition bad/20	Total	Pubmed good/20	Pubmed bad/20	Total
Genetic correlates of alcoholism	18	2	1436	6	14	44
DNA repair and aging	17	3	1220	11	9	1265
Drugs for fibromyalgia	17	3	1484	9	11	220
Genetic interactions of BCL2	18	2	876	8	11	19
Oxidative stress in plants	18	2	3122	9	11	3197
spectroscopy of amidohydrolases	17	3	861	7	13	1142
Benzene induced neuropathy	18	2	220	6	1	7
Birth defects from glycol ether	16	4	20	13	7	61
Depression in aging	19	1	13381	7	13	3658
Symptoms of type II diabetes mellitus	18	2	241	7	13	24704
Menopause and depression	18	2	696	11	9	1146
Treatment for bronchiectasis	18	2	2163	6	14	3207
OCD and anorexia	20	0	176	14	6	247
Proteolysis in SARS virus entry	4	0	4	2	0	2
Total	280	60	18433	125	127	34080
	Cognition			MEDLINE		
Precision	0.90			0.50		
Recall (*Assume total recall is the total of the cognition retrievals)	0.99			0.54		

Bootstrapping ontological attachments:

Most of the vocabulary derived from the acronym database and the UMLS had poor (very general) ontological attachments (eg, “amino-acid”). About 80,000 of 136,000 protein names were poorly attached. Attachments of well-classified words were spread to their synonyms resulting in 20,000 better attachments. A bootstrapping method took substrings as triggers; for example, “helix-loop-helix” as a substring of “transcription-factor-15-basic-helix-loop-helix” suggests an attachment to the node “helix-loop-helix.” This attachment was then assigned to the synonyms “bHLH-EC2-protein” and “paraxis”.

Discussion

We think that the natural language approach of CSIR has an important role in future access to textual information in the biomedical domain. This effort is our first pass at introducing biochemical and molecular biology terms into the CSIR lexicon. Other sources of new words will come from tracking user queries, evaluation of MEDLINE, and other curated databases. CSIR works equally well on full-text as on abstracts. It can be used to read full-text papers and other databases using text. This work contributes to precise interpretation of biomedical texts for purposes of search (1, 3, 25), research (4) and data mining (2, 26).

Uses and Applications of CSIR:

It is useful to review which linguistic processes produce these improved results. Morphology improves recall, so that the user can state a query term in one of its morphological variants, and CSIR automatically finds all other forms, as in phosphorylate and phosphorylation. Synonymy improves recall because one member of a synonym class retrieves documents with any of its members, as in “CD116,” "GMHCFS receptor alpha subunit," etc. Ontological reasoning improves recall as the software reasons down from higher-level concepts to lower-level concepts. For example, you can query "what MAP kinase phosphorylates ATF2" and get documents with “ERK” and “p38” which are kinds of MAP kinases. Sense disambiguation improves precision because only the documents that contain the query terms in the meanings intended by the user are retrieved. Phrase parsing improves both precision and recall. It improves precision by avoiding retrievals that happen to contain parts of a phrase in various positions, but not as the phrase. So "RNA", "binding" and "protein" might all appear in an

abstract that has nothing to do with RNA binding proteins. It improves recall because it enables the mapping of synonym relations between phrases, and between phrases and acronyms, as in "TUBB" and "beta-tubulin". Biomedical language also possesses ontological relationships for proteins, genes, the Tree-of-Life animals, diseases, etc. CSIR includes the function of downward reasoning in ontologies. Thus, CSIR NLP technology can help to solve problems in medicine by finding material about specific instances of general concepts such as “heart disease medicine”.

Areas for improvement

Precision is lowered when words are difficult to disambiguate, such as "Bad", which is an apoptosis protein, but at present is recognized as the ordinary English "bad". It will be relatively easy to address missing terms by frequency. We will use the methods of Tsuruoka (27) for future term recognition, synonymy expansion and evaluation of coverage. Automatic discovery of additional normalization rules, as in Wellner (2005) and Yoshimasa (2008)(22, 23) would be a further step. . Efforts directed toward database integration may provide useful definitions, synonymy and ontology in molecular biology (13). We also plan to introduce additional parsing functions (24), (12) which should improve the precision of Cognition Search.

Acknowledgements

We thank Ron Taussig for pointing out the Alliance for Cell Signaling website and other discussions. UMLS resources licensed (number 21817A334). The work in E. J. Goldsmith’s group was carried out under contract with Cognition, Technologies, Inc.

References:

1. Vanhecke TE, Barnes, M.A., Zimmerman, J., Shoichet, S. PubMed vs. HighWire Press: A head-to-head comparison of tow medical literature search engines. Computers in Biology and Medicine. 2007; 37:1252-8.

2. Divoli A, Attwood, T.K. “BiolE sentences - Extracting informative sentences from the biomedical literature.” Bioinformatics 2005; 21(9):2138-9.

3. Doms A, Schroeder, M. “GoPubMed: exploring PubMed with the Gene Ontology”. . Nucleic Acids Research 2005; 33.

4. Fontelo P, Liu, F., Ackerman, M. “askMEDLINE: a free-text, natural language query tool for MEDLINE/PubMed. BMC Medical Informatics and Decision-Making 5:5, 2005.

5. Matthew E. Falagas, 1, Eleni I. Pitsouni, George A. Malietzis and Georgios Pappas. Comparison of PubMed, Scopus, Web of Science, and Google Scholar: strengths and weaknesses. FASEB 2008; 22:338-42.

6. Wren JD, Chang JT, Pustejovsky J, Adar E, Garner HR, Altman RB. Biomedical term mapping databases. Nucleic Acids Res. 2005 Jan 1; 33(Database issue):D289-93.

7. Kohler J, Schulze-Kremer S. The semantic metadatabase (SEMEDA): ontology based integration of federated molecular biological data sources. In Silico Biol. 2002; 2(3):219-31.

8. Hatzivassiloglou V, Duboue PA, Rzhetsky A. Disambiguating proteins, genes, and RNA in text: a machine learning approach. Bioinformatics. 2001; 17 Suppl 1:S97-106.

9. Yu H, Kim, W., Hatzivassiloglou, V., and Wilbur, W. Disambiguating biomedical abbrevations. ACM Transactions on Information Systems (TOIS). 2006; 24(3):380-404.

10. Bodenreider O. Lexical, terminological and ontological resources for biological text mining. Ananiadou S, McNaught, J., editor: Artech House; 2006.

11. Baker PG, Goble CA, Bechhofer S, Paton NW, Stevens R, Brass A. An ontology for bioinformatics applications. Bioinformatics. 1999 Jun; 15(6):510-20.

12. Pustejovsky J, Castano J, Zhang J, Kotecki M, Cochran B. Robust relational parsing over biomedical literature: extracting inhibit relations. Pac Symp Biocomput. 2002b:362-73.

13. Philippi S, Kohler J. Using XML technology for the ontology-based semantic integration of life science databases. IEEE Trans Inf Technol Biomed. 2004 Jun; 8(2):154-60.

14. Dahlgren K, McDowell, J., and Stabler, E.P. Knowledge Representation for Commonsense Reasoning with Text. Computational Linguistics. 1989; 15:149-70.

15. Dahlgren K. Interpretation of Textual Queries Using a Cognitive Model. Ehrlbaum; 1992.

16. Dahlgren K, editor. Improving Precision and Recall with Linguistic Semantics. Proc Semantic Technology Conference; 2007; San Jose, CA.

17. Kornai A. Mathematical Linguistics. Springer; 2008.

18. Witten IH, Moffat, A.M., and Bell, T.C. Managing Gigabytes of Data. New York, NY. Morgan Kaufmann.; 1999.

19. Gilman AG. Cross talk: interview with Al Gilman. Mol Interv. 2001 Apr; 1(1):14-21.

20. Wain HM, Lush M, Ducluzeau F, Povey S. Genew: the human gene nomenclature database. Nucleic Acids Res. 2002 Jan 1; 30(1):169-71.

21. Wren JD, Garner HR. Heuristics for identification of acronym-definition patterns within text: towards an automated construction of comprehensive acronym-definition dictionaries. Methods Inf Med. 2002; 41(5):426-34.

22. Wellner BC, J and Pustejovsky, J. . “Adaptive string similarity metrics for biomedical reference resolution”. Proc ACL-ISMB Workshop on Linking Biological Literature, Ontologies and Databases: Mining Biological Semantics; 2005. p. 9-16

23. Yoshimasa T, McNaught, J and Ananiadou, S. . “Normalizing biomedical terms by minimizing ambiguity and variability. BMC Bioinformatics 9(Suppl 3). 2008(S2).

24. Friedman C, Kra P, Yu H, Krauthammer M, Rzhetsky A. GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles. Bioinformatics. 2001; 17 Suppl 1:S74-82.

25. Eaton AD. HubMed: a web-based biomedical literature search interface. Nucleic Acids Research 2006; 34.

26. Lee S, Yang, L., Jianrong, L., Friedman, C., Lussier, Y.A. . "Discovery of protein interaction networks shared by diseases”. . Pacific Symposium on Biocomputing; 2007. p. 76-87.

27. Tsuruoka Y, McNaught J, Tsujii J, Ananiadou S. Learning string similarity measures for gene/protein name dictionary look-up using logistic regression. Bioinformatics. 2007 Oct 15; 23(20):2768-74.

Translating Medical Words by Analogy

Philippe Langlais
Université de Montréal, Dept I.R.O.,
H3C 3J7 Montreal, QC, Canada
felipe@iro.umontreal.ca

François Yvon
Université Paris-Sud 11 &
CNRS, LIMSI, F-91403 Orsay
yvon@limsi.fr

Pierre Zweigenbaum
CNRS, LIMSI, F-91403 Orsay &
Inalco, ERTIM, F-75012 Paris
pz@limsi.fr

Abstract

Term translation has become a recurring need in medical informatics. This creates an interest for robust methods which can translate medical words in various languages. We propose a novel, analogy-based method to generate word translations. It relies on a partial bilingual lexicon and solves bilingual analogical equations to create candidate translations. To evaluate the potential of this method, we tested it on several datasets for five language pairs (10 translation directions). At best could the approach propose a correct translation for up to 67.9% of the input words, with a precision of up to 80.2% depending on the number of selected candidates. We compare it to previous methods including word alignment in parallel corpora and edit distance to a list of words, and show that these methods can complement each other.

1 Introduction

Term translation has become a recurring need in medical informatics. With the expansion of multilingual societies, the same information needs to be described in several languages (see, *e.g.*, the development of Spanish in the United States). Cross-language information retrieval enables users to search in a language and obtain documents in a different language, generally relying on query translation [Hersh and Donohoe, 1998]. To keep pace with the evolution of international medical terminologies, local terms must be provided for new or changed concepts (generally originally described in English). For instance, MeSH thesaurus updates are translated worldwide into multiple languages [Nelson *et al.*, 2004]. The SNOMED CT nomenclature is being translated into several languages, including French (`http://sl.infoway-inforoute.ca/content/dispPage.asp?cw_page=snomedct_e#8`). This requires a continuous effort to find the most suitable translations for new terms, often linked to new concepts.

A number of Natural Language Processing methods have been proposed to help find translations of words and terms. As suggested for another lexical task [McDonald, 1993], they can be divided into internal and external methods.

Internal methods rely on the words themselves, *i.e.*, their morphology. Observing that many European languages have similar medical words of Greek and Latin origins, a transducer may be designed to translate, *e.g.*, English words into French words. Such a transducer may be learnt by induction on a sample of *{word, translation}* pairs [Claveau and Zweigenbaum, 2005]. Morphological knowledge can be leveraged to decompose words into morphemes or *subwords* so that translation is only needed for this smaller set of items [Markó *et al.*, 2006; Namer and Baud, 2005]. A limitation however is that the set of subwords must be specified to begin with.

External methods take advantage of the context in which words occur. In a multilingual context, a suitable context is given by parallel text corpora: sets of bitexts, *i.e.*, *{text, translation}* pairs. Aligning sentences and words in these parallel texts performs a kind of reverse engineering which tracks back the lexical decisions and knowledge of human translators, down to word translations. This kind of method has been tested on multilingual medical terminologies such as ICD-10 [Baud *et al.*, 1998; Nyström *et al.*, 2006] and on documents extracted from a multilingual web site [Deléger *et al.*, 2006]. A well-known limitation is the relative scarcity of parallel corpora, which motivates the recourse to the more readily available "comparable" corpora: corpora of texts which are generally in two different languages, but cover the same theme (*e.g.*, smoking cessation). However, finding word translations in comparable corpora [Chiao and Zweigenbaum, 2002] is much more difficult than in parallel corpora.

The present work explores the use of a different internal method: analogical learning [Lepage, 1998; Stroppa and Yvon, 2005]. As the above-mentioned methods of this type [Claveau and Zweigenbaum, 2005; Markó *et al.*, 2006; Namer and Baud, 2005], it is trained on an initial bilingual lexicon and relies on the formal similarity of medical words in some languages to propose new translations; in contrast to external methods, it can generate translations for unseen words. In this paper, we examine how analogical learning performs on medical words. We evaluate it on a similar dataset as earlier, comparable work [Claveau and Zweigenbaum, 2005], and study its complementarity to an external method such as the above [Deléger *et al.*, 2006] and to a non-generative internal method based on edit distance. We also investigate the viability of the approach for translating from and into various languages, including morphologically rich languages such as Finnish.

This paper is organized as follows. We first introduce the principles of analogical learning on which our system relies, and describe the corpora used to test the method. We then present its evaluation, with a comparison to two other methods. We discuss the results and suggest an articulation of these different types of method, summarize our contribution and conclude with perspectives for further work.

2 Methods

2.1 Formal Analogy

A *proportional analogy*, or analogy for short, is a relation between four items noted $[x : y = z : t]$ which reads as "x is to y as z is to t". Among proportional analogies, we distinguish*formal analogies*, that is, those we can identify at a graphemic level, such as [*adrenergic beta-agonists*, *adrenergic beta-antagonists*, *adrenergic alpha-agonists*, *adrenergic alpha-antagonists*]. Formal analogies can be defined in terms of factorization [Stroppa and Yvon, 2005].

Définition 2.1 *Let x be a string over an alphabet Σ, a* factorization *of x, noted f_X, is a sequence of n factors $f_X = (f_X^1, \ldots, f_X^n)$, such that $x = f_X^1 \bullet f_X^2 \bullet f_X^n$, where $\bullet$ denotes the concatenation operator.*

We thus define a formal analogy as follows. Intuitively, this definition states that (x, y, z, t) are made up of a common set of alternating substrings.

Définition 2.2 $\forall (x, y, z, t) \in {\Sigma^\star}^4$, $[x : y = z : t]$ ***iff*** *there exists factorizations $(f_X, f_Y, f_Z, f_t) \in ({\Sigma^\star}^d)^4$ of (x, y, z, t) such that, $\forall i \in [1, d]$, $(f_Y^i, f_Z^i) \in \{(f_X^i, f_t^i), (f_t^i, f_X^i)\}$.*

It is routine to check that it captures the example analogy introduced above, based on the following factorizations:

$$\begin{array}{lcl}
f_x & \equiv & (\text{adrenergic bet, a-agonists}) \\
f_y & \equiv & (\text{adrenergic bet, a-antagonists}) \\
f_z & \equiv & (\text{adrenergic alph, a-agonists}) \\
f_t & \equiv & (\text{adrenergic alph, a-antagonists})
\end{array}$$

In the sequel, we call an *analogical equation* an analogy where one item (usually the fourth) is missing and we note it $[x : y = z : ?]$. Lepage [1998] proposes a method to solve analogical equations, that is, to generate the missing fourth item. Stroppa and Yvon [2005] describe a generalization of this algorithm, which accounts for the definition of formal analogy we gave above. More precisely, they show that the set of solutions to an analogical equation is a rational language; that is, we can build a finite-state machine to recognize them.

We implemented such a solver in this work, the details of which are beyond the scope of this paper. It is important to realize though that very often, there is not one single solution to an analogical equation, but many of them. For instance, *spondylite*, *ispondylte*, *ispndylote*, *spndyloite*, and *itespondyl* are 5 of the 110 solutions to the equation [*chondropathie* : *spondylopathie* = *chondrite* : ?]. Though all these forms verify this equation, only the first is a French word (and the *natural* solution to the equation).

2.2 Analogical learning

Let $\mathcal{L} = \{(i, o) \,|\, i \in \mathcal{I}, o \in \mathcal{O}\}$ be a learning set of observations, where $\mathcal{I}$ (resp. $\mathcal{O}$) is the set of possible forms of the input (resp. output) linguistic system of the application. We denote $I(u)$ (resp. $O(u)$) the projection of u into the input (resp. output) space; that is, if $u = (i, o)$, then $I(u) \equiv i$ and $O(u) \equiv o$. For an incomplete observation $u = (i, ?)$, the inference procedure consists in:

1. building $\mathcal{E}_\mathcal{I}(u) = \{(x, y, z) \in \mathcal{L}^3 \mid [I(x) : I(y) = I(z) : I(u)]\}$, the set of input triplets that define an analogy with $I(u)$.
2. building $\mathcal{E}_\mathcal{O}(u) = \{o \in \mathcal{O} \mid \exists (x, y, z) \in \mathcal{E}_\mathcal{I}(u) \text{ s.t. } [O(x) : O(y) = O(z) : o]\}$ the set of solutions to the equations obtained by projecting the triplets of $\mathcal{E}_\mathcal{I}(u)$ into the output space.
3. selecting candidates among $\mathcal{E}_\mathcal{O}(u)$.

Since the first two steps of this inference procedure are generating candidate solutions, we call them the *generator*. Step 3 is responsible for selecting candidates, and is therefore called the *selector*. Let us illustrate this with the word pair (*spondylitis*, ?) whose second term should be the French translation of *spondylitis*. The following analogical proportions are identified in (1): that written above, [*adenomalacia* : *adenitis* = *spondylomalacia* : *spondylitis*], [*arthropathy* : *arthritis* = *spondylopathy* : *spondylitis*], etc., where (*adenomalacia*, *adénomalacie*), (*adenitis*, *adénite*), (*spondylomalacia*, *spondylomalacie*), etc., are in our bilingual lexicon, but not (*spondylitis*, ?). Analogical equations such as [*adénomalacie* : *adénite* = *spondylomalacie* : ?] are thereby formed and solved in (2), producing solutions among which (3) correctly selects *spondylite*.

2.3 Practical issues

Although simple in principle, analogical learning involves important practical issues. There are basically two problems that must be solved for the approach to work. The first one stems for the identification of the triplets in the input space that form with the unknown term an analogy (step 1). This is an operation a priori cubic in the number of input objects, about 10,000 here). Langlais and Patry [2007] describe an approach where the problem is turned into solving a quadratic number of analogical equations, which is still too time consuming in applications such as ours. To alleviate time issues, the authors propose to sample forms in the input space. In this work, we applied a technique described in [Langlais and Yvon, 2008] which allows to solve the problem in a time roughly linear in the input space size. Here again, the description of this technique is beyond the scope of this paper. Suffice it to say, that thanks to this technique, we can solve step 1 of analogical learning exactly, that is, we can identify *all* the analogies (involving the form to translate) present in the input space.

The second problem stems for the potentially high number of forms produced by the generator. These forms arise in part because the solver generates many solutions, as we already discussed. The fact that several input analogies, and in turn, several target equations are being considered while translating a single form exacerbates the problem. To our knowledge, there is no known satisfactory solutions to this issue yet. In this work, we simply keep the count with which a given solution is generated. The top-ranked solutions are those proposed by the analogical system. This simple selector has the advantage that it allows to investigate how far down the list we must go to find the oracle solution (see next section).

3 Experimental protocol

3.1 Material

We ran our experiments with several goals in mind. First, we wanted to test how our approach is impacted by the size of the training material. Therefore, we collected two different sized corpora (MASSON and MESH) for the French-English language pair. Second, we wanted to check whether analogical learning is better suited for specific language pairs. We were also interested in observing whether it is more suited to translate into a morphologically rich language (such as Finnish) or the other way round. We therefore considered a bench of language-specific datasets (MESH). Last, we also compared the analogical method to a corpus-based alignment method, and compiled for this purpose the HEALTH dataset.

MASSON We used a list of French medical words and their English translations obtained from the Masson medical dictionary (`http://www.atmedica.com/`). The same initial list was used in the work of Claveau and Zweigenbaum [2005], but we did not keep words which were identical in French and English. We selected 13,392 word pairs with a normalized edit distance between 0.02 (differing by 1 character) and 0.67 (rather distant, such as (*toux, cough*). This list was randomly split into SEARCH (80%), DEV (10%) and TEST sets (10%, *i.e.*, 1,306 words) (DEV was not used in the part of the work presented here).

MESH The Medical Subject Headings (MESH) is the thesaurus used by the US National Library of Medicine to index the biomedical scientific literature in the MEDLINE database. Its preferred terms are called "Main Headings" (synonym terms are called "Entry Terms"). We collected pairs of source and target Main Headings (TTY[1] = 'MH') with the same MeSH identifiers (SDUI). [2] We considered five language pairs: two European close ones (English-French and English-Spanish), two distant ones (Finnish-English and Swedish-English) and one pair involving different scripts (Russian-English).[3] The resulting MESH datasets contain roughly half the pairs of terms we collected in MASSON. We randomly split each dataset in two parts: 90% into SEARCH and the remaining 10% into TEST.

HEALTH To investigate the performance of a corpus-based alignment method, we compiled two parallel corpora: one was obtained from the Health Canada English-French bilingual website (`http://www.hc-sc.gc.ca/`, 142,441 distinct target words), the other consists of 7,260 pairs of English and French abstracts of French journal articles published in about 350 French medical journals (about 3 million words in total, 107,441 distinct target words).

3.2 Evaluation

We computed the following measures to evaluate our analogical translation device:

Coverage is the proportion of input words for which the system can generate translations. If N_t words receive translations among N, coverage is defined as $\frac{N_t}{N}$.

Precision : among the N_t words for which the system proposes an answer, precision is the proportion of those for which a correct translation is output. The system proposes a ranked list of translations for each input word. Depending on the number of output translations k that one is willing to examine, a correct translation will be output for N_k input words. Precision at rank k is thus defined as $P_k = \frac{N_k}{N_t}$.

Recall is the proportion of the N input words for which a correct translation is output. Recall at rank k is defined as $R_k = \frac{N_k}{N}$.

We additionally compared, in ideal conditions, the recall of our (internal) method to that of an (external) word alignment method and to a non-generative internal method based on edit distance. *Word alignment* takes a parallel corpus of texts and their translations and aims to determine which pairs of (source, target) words are in a translation relation in the corpus. Ideally, if an input word is in the source corpus, its translation can be identified in the target corpus. Our ideal test therefore consists in checking whether a given input word occurs in the source part of our parallel corpora. *Edit distance* computes a distance between two words based on their common and distinct characters [Levenshtein, 1966]. Since in our setting, source and target words are often formally similar, given a list of potential target words, a candidate translation of an input word is the target word which is closest to it in terms of edit distance. An ideal situation for that method is one where all correct translations are included in the list of potential target words. We built such a list by adding the target part of our test MASSON set to the list of words in the English part of the HEALTH corpora (total of 229,695 unique words).

4 Results

The algorithm was applied to translate the terms of the TEST material, searching analogies (step 1) in the SEARCH set, solving the resulting analogical equations (step 2) then ranking solutions according to frequency (step 3).

Influence of the corpus size The contrast between the small (MESH) and large (MASSON) French-English datasets can be observed in Table 1. Out of the 1,306 terms of MASSON, 1,092 source words obtained translations, which yields a coverage of 83.6%. Precision ranges from $P_1 = 34.8\%$ to $P_{25} = 80.2\%$, while recall ranges from $R_1 = 29.1\%$ to $R_{25} = 67.9\%$. For the MESH test set, only 199 terms out of 509 ones received translations,

[1] In the UMLS Metathesaurus tables, the TTY field codes the type of the term. Its values depend on the source terminology.

[2] We did not collect pairs of entry terms because we do not know how to pair actual translations among the possibly numerous entry terms of a given main heading.

[3] Russian MeSH is normally written in Cyrillic, but some terms are simply English terms written in uppercase Latin script (e.g., *ACHROMOBACTER*). We removed those terms.

Table 1: Performance of the approach on MESH and MASSON (FR→EN).

	$nb.$	$Cov.$	P_1	R_1	P_{25}	R_{25}
MASSON	1306	83.6	34.8	29.1	80.2	67.9
MESH	509	39.1	46.2	18.1	61.3	24.0

yielding a coverage of 39.1%. The precision ranges from 46.2% to 61.3%, while recall ranges from 18.1% to 24%.

Clearly, the training size impacts the approach. Analogical learning can identify more input analogies in larger datasets, therefore proposing translations for more terms. This eventually comes at a price at rank 1: more noisy translations are produced for the largest dataset (see P_1). But allowing the system to propose more solutions clearly shows the advantage of searching through a larger datset.

Influence of the language pair We investigated on the MESH datasets the influence of the language pair and the translation direction. In total, we ran 10 translation sessions that are summarized in Table 2.

Table 2: Performance of analogical learning as a function of the translation direction (MESH).

	nb	$Cov.$	P_1	R_1	P_{25}	R_{25}
FI→EN	701	44.2	49.0	21.7	65.5	29.0
FR→EN	509	34.4	46.3	15.9	63.4	21.8
RU→EN	784	48.6	38.1	18.5	61.7	30.0
SP→EN	624	46.0	42.5	19.6	60.6	27.9
SW→EN	592	41.0	46.1	18.9	64.2	26.4
FI←EN	701	42.8	44.3	19.0	63.7	27.2
FR←EN	509	39.1	46.2	18.1	61.3	24.0
RU←EN	784	47.1	44.4	20.9	67.2	31.6
SP←EN	624	39.7	44.0	17.5	66.1	26.3
SW←EN	592	40.9	45.0	18.4	64.5	26.4

Overall, we observe that analogical learning offers comparable performances for all translation directions, although some fluctuations are observed. Somehow surprisingly, the largest coverage rates are observed when translating from and into Russian. This shows that analogical learning is not bounded to translate closely related languages only, not even is it designed to treat languages that share the same scripts. We also note that it is not affected by translating into a morphologically rich language, such as Finnish or Swedish.

Comparison to corpus-based alignment Among the 1,306 source words of MASSON, 262 can be found in the source part of the Health Canada corpus and 233 in the source part of the abstracts corpus (Table 3): in total 479 source words could ideally be translated by word alignment in these two corpora, *i.e.*, a recall of 36.7%, assuming the presence of suitable target words and a perfect aligner.

Table 3: Ideal recall by word alignment in two medical corpora.

	HEALTH	Abstracts	Total
Number of words	262	233	479
Recall (%)	20.1	17.8	36.7

Comparison to edit-distance Comparing the source (test) words of the MESH dataset to the target words in the TEST material by edit distance (Table 4), we measured that 93.7% of the correct translations were found in top position (MASSON). This shows that French and English terms are very close. When TEST set target words are merged into the list of target Health Canada words, this ideal recall decreased to 75.9%, and to 73.3% when adding the words of the abstracts corpus.

Table 4: Ideal recall by edit distance in TEST words and two medical corpora (MASSON). HealthC stands for the Health Canada corpus.

	TEST	+HealthC	+Abstracts
Corpus size	1,306	143,433	229,695
Found	1224	991	957
Recall (%)	93.7	75.9	73.3

5 Discussion

On MASSON (FR→EN), analogical learning could identify a correct translation for up to 67.9% of the source test words, with a corresponding precision of 80.2%. Given the simplicity of frequency ordering used in place of step 3 in the present experiments, we expect the system to perform better in terms of precision if a better strategy is devised. Ongoing work on using a classifier to select candidate solutions shows that we can boost the precision of candidates to 90% with little or no loss in recall.

Even if we used the same dataset (MASSON), a precise comparison with Claveau and Zweigenbaum [2005] is difficult however, since their TEST set, although taken from the same superset, was quite different from ours as it contained pairs of identical words. Their best attainable precision was 75% when test words were randomly selected as in the present work, but included 10–12% of identical words. They do not report the corresponding recall.

Examples of successful analogies on MASSON are shown in the first part of Table 5. Example 1 shows how a translation where a word ending is involved (*-ie* / *-ia*) leverages an example with a prefix switch (*exo-* ↦ *ecto-*), itself licensed by another word pair (*exosquelette* ↦ *ectosquelette*). Example 2 ilustrates how an (*-ic* ↦ *-oid*) change in English is generated for *dermic* by analogy to (*lupic* ↦ *lupoid*), thereby producing a translation for French *dermoïde*. Examples 3–4 and 5–6 show multiple paths to support the same candidate translations, with analogies based on different words and suffixes. Note that "suffixes" and "prefixes" as mentioned in this paragraph are only an a posteriori description of the results of the algorithm: no morphemic knowledge was given to the system.

Word alignment in our two parallel medical corpora could at best identify 36.7% of the source words. Indeed, additional or larger parallel corpora might be found, but it is

Table 5: Example analogies supporting correct translations

	source	triplets for analogical equations	target
1	exocardie FR→EN	<ectosquelette,ectocardie,exosquelette> <ectoskeleton,ectocardia,exoskeleton>	exocardia
2	dermoïde FR→EN	<lupique,lupoïde,dermique> <lupic,lupoid,dermic>	dermoid
3	immunisation FR→EN	<volatil,immun,volatilisation> <volatile,immune,volatilization>	immunization
4	immunisation FR→EN	<neutralité,immunité,neutralisation> <neutrality,immunity,neutralization>	immunization
5	périvésiculite FR→EN	<ombilical,périombilical,vésiculite> <umbilical,periumbilical,vesiculitis>	perivesiculitis
6	périvésiculite FR→EN	<odontogramme,vésiculogramme,périodontite> <odontogram,vesiculogram,periodontitis>	perivesiculitis
7	alpha-cyclodextrins EN→SW	<beta-endorphin,alpha-endorphin,beta-cyclodextrins> <betaendorfin,alfaendorfin,betacyklodextriner>	alfacyklodextriner
8	iodoproteins EN→FI	<phosphates,iodates,phosphoproteins> <fosfaatit,jodaatit,fosfoproteiinit>	jodoproteiinit
9	pneumopericardium EN→SP	< hydrothorax,hydropneumothorax,pericardium> < hidrotórax,hidroneumotórax,pericardio>	neumopericardio
10	polysaccharides EN→FR	<liposarcoma,sarcoma,lipopolysaccharides> <liposarcome,sarcomes,lipopolyoside>	polyoside
11	bronchoscopy EN→RU	<arthrography,arthroscopy,bronchography> <артрография,артроскопия,бронхография>	бронхоскопия
12	buxus EN→RU	<cistaceae,buxaceae,cistus> <ладанниковые,самшитовые,ладанник>	самшит

known that their size is not indefinitely extendable. This illustrates that methods which rely on spotting known words are limited by the available material where such words can be found. An interesting point however is that the intersection between the words that can be translated by word alignment and those that can be translated by analogy only counts 161 words: assuming each method performed optimally, their union could translate 1,205 words and obtain 92.3% recall. For instance, the following words were translated only by analogy: *ablépharie, abrachiocéphalie, acroesthésie, actinocongestine*, while the following were found only in the parallel corpus: *acuité, acyclique, acétylation, acétyltransférase*.

Compared to alignment, edit distance had an easier task since all target words are included in the search list. Had we not added the list of target words, edit distance would have had a much lower potential recall. A more realistic test would consist in using for a candidate list a large corpus such as the target part of our parallel corpora. This would then come close to the above word alignment experiment, where the test of identity with a source word would be replaced with one of edit distance with a target word—although a monolingual corpus would be sufficient, and could therefore be much larger.

On the MeSH datasets, we observed that analogical learning is not impacted by the language pair being treated, nor by the translation direction. In particular, translating into a morphologically rich language does not seem to be a problem. This contrasts with statistical machine translation which is known to perform poorly when translating into a highly inflected language such as Finnish. Examples of successful analogies on MeSH for various language pairs are shown in the bottom part of Table 5. Russian correspondences often happen to rely on transliterations (11), but not systematically (12, buxus/самшит). The latter shows that the method does not rely on character correspondences between translated word pairs, but only on the synchronous existence of analogical equations in both languages.

As it is often the case with corpus-based approaches, analogical learning is impacted by the size of the training material. Based on our contrastive experiment on Masson and MeSH, we observe this trend for analogical learning as well. More examples allow the approach to identify a larger set of input analogies, yielding in turn a better coverage.

The analogical method is the only one of those tested here which can generate translations for unseen words. The resolution of an analogical equation combines the known words in the equation to create a new, hypothetical word which solves it. Identifying and solving a large number of such analogical equations builds cumulative support for the most promising hypotheses.

A way to improve the analogical method would be to provide it with knowledge on morphemes or “subwords,” as prepared in previous work [Namer and Baud, 2005; Deléger *et al.*, 2007]. This could be used to enforce morphemic boundaries when generating analogical equation solutions and therefore reduce the number of generated forms, or to perform a posteriori filtering of candidate translations in step 3.

Analogy-based word translations were used to help machine translation in [Langlais and Patry, 2007]. The top candidate translation was kept and directly fed to the MT system. The authors report small but consistent improvements. Candidate translations can also be used to help specialized terminologists prepare term translations. In that setting, a terminologist would examine the top n translation candidates and select the most relevant translation (or produce another one) to include in the target terminology.

6 Conclusion

We introduced an analogy-based method to generate word translations and evaluated its potential on medical words. Its precision can be quite good once a stronger selection component is integrated in its last step. Its recall is lower, with an upper bound at 68% (MASSON) in the current experiments. We saw that it can be increased by a combination with complementary, existing methods based on attested words, such as word alignment in parallel corpora or edit distance with a large word list. It has the distinctive ability to generate translations for unseen words.

We tested our method on different language pairs involving morphologically rich languages (such as Finnish and Swedish) as well as languages with different scripts. We observed that the approach does not seem to be impacted by the language pair considered. Lepage [1998] gave evidence that the approach works as well for Asian languages. We therefore plan to test the present method on more distant language pairs including Japanese or Chinese.

Another perspective is to tackle the direct translation of multiword terms: our analogical solver can work directly on such terms without having to first segment them into words. This should be particularly interesting in the context of medical terminologies.

References

[Baud *et al.*, 1998] Robert H. Baud, Christian Lovis, Anne-Marie Rassinoux, Pierre-André Michel, and Jean-Raoul Scherrer. Automatic extraction of linguistic knowledge from an international classification. In Branko Cesnik, Charles Safran, and Patrice Degoulet, editors, *Proceedings of the 9 th World Congress on Medical Informatics*, pages 581–585, Seoul, 1998.

[Chiao and Zweigenbaum, 2002] Yun-Chuang Chiao and Pierre Zweigenbaum. Looking for French-English translations in comparable medical corpora. *Journal of the American Medical Informatics Association*, 8(suppl):150–154, 2002.

[Claveau and Zweigenbaum, 2005] Vincent Claveau and Pierre Zweigenbaum. Translating biomedical terms by inferring transducers. In Elpida Keravnou Silvia Miksch, Jim Hunter, editor, *Proceedings 10th Conference on Artificial Intelligence in Medicine Europe*, volume 3581, Berlin / Heidelberg, 2005. Springer.

[Deléger *et al.*, 2006] Louise Deléger, Magnus Merkel, and Pierre Zweigenbaum. Contribution to terminology internationalization by word alignment in parallel corpora. In *Proceedings AMIA Annual Fall Symposium 2006*, pages 185–189, Washington, DC, November 2006. AMIA.

[Deléger *et al.*, 2007] Louise Deléger, Fiammetta Namer, and Pierre Zweigenbaum. Defining medical words: Transposing morphosemantic analysis from French to English. In *Proc MEDINFO 2007*, volume 129 of *Studies in Health Technology and Informatics*, pages 152–156, Amsterdam, 2007. IOS Press.

[Hersh and Donohoe, 1998] William R. Hersh and L. C. Donohoe. SAPHIRE International: a tool for cross-language information retrieval. *Journal of the American Medical Informatics Association*, 5(suppl):673–677, 1998.

[Langlais and Patry, 2007] Philippe Langlais and Alexandre Patry. Translating unknown words by analogical learning. In *EMNLP-CoNNL*, pages 877–886, Prague, Czech Republic, 2007.

[Langlais and Yvon, 2008] Philippe Langlais and François Yvon. Scaling up analogical learning. In *22nd COLING*, Manchester, England, 2008.

[Lepage, 1998] Yves Lepage. Solving analogies on words: an algorithm. In *COLING-ACL*, pages 728–734, Montréal, Canada, 1998.

[Levenshtein, 1966] V. I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, pages 707–710, 1966.

[Markó *et al.*, 2006] Kornél Markó, Robert Baud, Pierre Zweigenbaum, Lars Borin, Magnus Merkel, and Stefan Schulz. Towards a multilingual medical lexicon. In *Proceedings AMIA Annual Fall Symposium 2006*, pages 534–538, Washington, DC, November 2006. AMIA.

[McDonald, 1993] David D. McDonald. Internal and external evidence in the identification and semantic categorization of proper names. In Branimir Boguraev and James Pustejovsky, editors, *Corpus Processing for Lexical Acquisition*, pages 61–76. MIT Press, Cambridge, MA, 1993.

[Namer and Baud, 2005] Fiammetta Namer and Robert Baud. Predicting lexical relations between biomedical terms: towards a multilingual morphosemantics-based system. In *Studies in Health Technology Information*, volume 116, pages 793–798. IOS Press, Amsterdam, 2005.

[Nelson *et al.*, 2004] Stuart J. Nelson, Michael Schopen, Allan G. Savage, Jacque-Lynne Schulman, and Natalie Arluk. The mesh translation maintenance system: Structure, interface design, and implementation. In Marius Fieschi, Enrico Coiera, and Yu-Chuan Jack Li, editors, *Proceedings 10 th World Congress on Medical Informatics*, volume 107 of *Studies in Health Technology and Informatics*, pages 67–69, Amsterdam, 2004. IOS Press.

[Nyström *et al.*, 2006] Mikael Nyström, Magnus Merkel, Lars Ahrenberg, Pierre Zweigenbaum, Hokan Petersson, and Hans Ahlfeldt. Creating a medical English-Swedish dictionary using interactive word alignment. *BMC Medical Informatics and Decision Making*, 6(35), October 2006.

[Stroppa and Yvon, 2005] Nicolas Stroppa and François Yvon. An analogical learner for morphological analysis. In *9th Conf. on Computational Natural Language Learning (CoNLL)*, pages 120–127, Ann Arbor, MI, 2005.

A Comparison between CRFs and SVMs in Disorder Named Entity Recognition in Clinic Texts

Dingcheng Li[1], **Karin Schuler**[2], **Guergana Savova**[2]
Department of Linguistics, University of Minnesota, Minneapolis, Minnesota[1]
College of Medicine, Mayo Clinic, Rochester, Minnesota [2]
lixxx345@umn.edu, Karin.Schuler@mayo.edu, Savova.Guergana@mayo.edu

Abstract

Named Entity Recognition is a critical component in any information extraction system. A number of solutions have been proposed ranging from rule-based to purely statistical approaches. In this paper, we present a comparative study between two machine learning methods, Conditional Random Fields and Support Vector Machines, both of which have shown state-of-the-art results in the general and biomedical domains. We explore their applicability to clinical free text. A number of features including dictionary look-up, bag-of-words, part-of-speech tags, numerical features, capitalization and their combinations are tested. We evaluated their performance against a set of gold standard named entities and then against the named entity components (using Inside-Outside-Begin notation). Our baseline is a dictionary look-up method with SNOMED-CT as the knowledge base. Evaluations of both techniques show that CRFs outperforms SVMs for this task.

1 Introduction and Background

Named entity recognition (NER) is the discovery of textual mentions, or named entities (NEs), that belong to the same semantic class. In the general domain, example classes are people and organization names while in the biomedical domain NEs examples include diseases, signs/symptoms, anatomical signs, and drugs. NER has been investigated with a variety of techniques such as rule-based, e.g., dictionary look-up algorithms [Tsuruoka *et al.*, 2007] and machine learning such as Conditional Random Fields (CRFs) [McCallum and Li, 2003] and Support Vector Machines (SVMs) [Ji *et al.*, 2002]. Performance is high especially as applied to scholarly text and newswire narratives [Leaman and Gonzalez, 2008]. Clinical free-text, on the other hand, exhibits characteristics of both informal and formal linguistic styles. This, in turn, has proven to be a challenge for both rule-based and machine learning approaches when applied to the task of NER. Dictionary look-up algorithms have the potential to work very well when applied to the clinical domain as the list of medical NEs is likely to be finite. However, medical language follows the linguistic variations found in the general language coupled with the coining of new terms, especially new drug names, and the extensive use of abbreviations and acronyms. Thus, a comprehensive medical dictionary becomes a goal difficult to achieve. Meanwhile, the multitude of synonyms and polysemous terms create ambiguities making the unique mapping between a textual mention and a dictionary a hard problem. Therefore, machine learning methods hold the promise to generalize this variability through the contextual patterns the NEs follow which are difficult to capture through explicit rules only. CRFs and SVMs are two machine learning techniques which can handle multiple features, or characteristics, to be considered during learning. CRFs' main strength lies in their ability to include various unrelated features, while SVMs' advantage is in the inclusion of overlapping features. Leaman [Leaman and Gonzalez, 2008] create a NER system for biomedical literature with CRFs. Large number of features, such as lexical, dictionary, part-of-speech, lemmatization and numeric normalization features are used. Their system achieves good results in formal medical literature (F-measure reaches 81.96% for the bio-creative data). Ji [Ji *et al.*, 2002] applies SVM to biomedical NER. The data they uses is GENIA corpus[1]. Since in NER, the data is always unbalanced (non-NEs are always much more than NEs), they use class splitting techniques to split non-NEs into many smaller categories to balance the training input. The results they report are better than maximal entropy (about 50% for GENIA).

Due to the fact that both CRFs and SVMs have advantages in incorporating features and there exist both unrelated features (such as spelling and grammatical features) and overlapping features (such as features as neighboring words) in a text, we make comparisons on the performance of CRFs and SVMs as applied to the task of NER from clinical text. Specifically, we focus on the discovery of disease/disorder NEs.

2 Dataset and Its Representation

The dataset in our study is a corpus of 160 manually annotated clinical notes for the disorder NEs developed by the Mayo Clinic NLP group [Ogren *et al.*, 2008]. The total number of words in the corpus is 47,975 with a median word count of 249 words per note. The NE annotations are the result of a consensus between four human annotators and contain 1,556 annotations using 658

[1] Available via http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA

unique concept codes from a subset of SNOMED-CT[2] corresponding to human disorders. Inter-annotator agreement was calculated on annotations from 100 of the documents for span (90.9%), concept code (81.7%), context (84.8%), and status (86.0%) agreement. The annotation schema we used to train and test the models is the widely accepted IOB (inside-outside-begin) notation [Leaman and Gonzalez, 2008]. There are four IOB labels in our experiments: sentence, begin_NE, inside_NE and outside_NE. They represent the sentence delimiter, the first token of a NE, non-initial tokens of a NE and non-NE tokens respectively. Since many NEs involve more than one word and even a short sentence, it is hard for any token-based parser to identify all these entities simultaneously. The IOB format which defines a NE component based on its position presents a solution to that problem. In addition, this notation is able to support consecutive entities. For example, the NE ”congestive heart failure”, in the IOB format has three labels: begin_NE for ”congestive”, inside_NE for ”heart” and inside_NE for ”failure”. The gold standard NE format has one label: disorder for ”congestive heart failure”. In our project, we used 118 notes for training and the remaining 42 for testing. Therefore, there are 1264 NEs in the training data and 292 NEs in the testing data. Since in our project, we use the IOB format, all multiword NEs are divided into begin_NEs and inside_NEs. The 1264 NEs in the training data and the 292 NEs in the testing data correspond to 2087 and 468 IOB-formatted NE components respectively.

3 Conditional Random Fields and Support Vector Machines

CRFs are statistical machine learning models widely used in NLP. They are conditional distributions $p(y|x)$ with an associated undirected graphical structure. $p(y|x)$ is defined as a normalized product of potential functions, each of which has the following form [Lafferty *et al.*, 2001]:

$$exp(\sum_j \lambda_j t_j(y_{t-1}, y_t, \mathbf{x}, t) + \sum_k \mu_k s_k(y_t, \mathbf{x}, t)) \quad (1)$$

where $t_j(y_{t-1}, y_t, \mathbf{x}, t)$ is a transition feature function of the entire observation sequence and the labels at positions t and $t-1$ in the label sequence; $s_k(y_t, \mathbf{x}, t)$ are a state feature function of the label at position t and the observation sequence; and λ_j and μ_k are parameters to be estimated from training data. In a Hidden Markov Model (HMM)-like CRF, these two sets of parameters correspond to the log values of a transition model and sensor model respectively. Given the potential functions, a linear-chain conditional random field is a distribution that takes the form:

$$p(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} exp(\sum_j \lambda_j t_j(y_{t-1}, y_t, \mathbf{x}, t) + \sum_k \mu_k s_k(y, \mathbf{x}, t)) \quad (2)$$

where $Z(\mathbf{x})$ is a normalization term. Compared with HMM, CRFs allow dependencies among the observation sequence achieved by defining the feature functions. For example, in an HMM-like CRF, the transition feature function is $\mathbf{1}\{y_t = i\}\{y_{t-1} = j\}$ and the sensor feature function is $\mathbf{1}\{y_t = i\}\{x_t = o\}$. Other dependencies are allowed, such as the dependencies between x_t and x_{t-1} by adding a feature $\mathbf{1}\{x_t = o\}\{x_{t-1} = q\}$ to the model.

The dependency among the observation sequences is exactly the main reason why CRFs outperforms HMMs. A deeper reason is that CRF is a discriminative model, while HMM is a generative model. By modeling the conditional distribution directly, it is not required to model $p(x)$ which allows arbitrary dependencies on the observation sequence which relaxes the strong independence assumption made by HMMs. In our work, we used a CRF implementation called Mallet [McCallum, 2002] developed by the NLP group at University of Massachusetts. It is a Java-implemented first-order CRF with a Gaussian prior variance of 0.5.

The second method we experimented with is SVMs [Cortes and Vapnik, 1997]. SVMs perform classification by constructing an N-dimensional hyperplane that optimally separates the data into two categories. Given a training set of instance-label pairs $(x_i;\ y_i)$; $i = 1...l$ where $x_i \in R^n$ and $y \in \{1, -1\}^l$, SVMs require the solution of the following optimization problem:

$$min_{w,b,\xi} \frac{1}{2} w^T w + C \sum_{i=1}^{l} \xi_i \quad (3)$$

It is subject to

$$\begin{array}{l} y_i(w^T \phi(\mathbf{x}_i) + b) \geq 1 - \xi_i \\ \xi \geq 0 \end{array} \quad (4)$$

$\mathbf{x}_i$ is the training vector, the function ϕ map it to a higher dimensional space. The SVM finds a linear separating hyperplane with the maximal margin in this higher dimensional space. In addition, for the mapping, SVM needs a kernel function defined as:

$$K(x_i x_j) \equiv \phi(\mathbf{x}_i)^T \phi(\mathbf{x}_j) \quad (5)$$

There are four widely used kernel functions: linear, polynomial, radial basis function (RBF) and sigmoid. In our project, we use the LibSVM [Chang and Lin, 2001] implementation of SVMs. It implements the four functions and supports multi-label classification. The goal is to learn a **w** parameterized discriminant function F over input/output pairs and to maximize this function over the response variable to make a prediction. The function F assigns a numerical score to a pair of state sequence and observation sequence. For a specific observation sequence, the predicated state sequence is the one that has the highest numerical score. The state sequence is computed by

$$y = argmax_y F(x, y; w) \quad (6)$$

In our task of clinical NER, the extracted features from clinical free text are varied (see Section 4.) Hence, the relation between class labels and attributes is non-linear. This is the reason why we choose the RBF kernel function for our experiments.

[2] http://www.snomed.org/. Last accessed 3/12/2008

4 Features

This section describes the features along with the feature combinations we used to train the CRF and SVM models. For better comparison, we extract almost same features for both CRFs and SVMs.

Test No	Features
1	dictionary look-up
2	dictionary look-up + BOW + Orientation + distance (Window 5)
3	dictionary look-up + BOW + Orientation + distance (Window 3)
4	dictionary look-up + BOW + POS + Orientation distance (Window 5)
5	dictionary look-up + BOW +POS + Orientation distance (Window 3)
6	dictionary look-up + BOW +POS + Orientation +distance (Window 3) +bullet number
7	dictionary look-up + BOW + POS +Orientation +distance(Window 3) + measurement
8	dictionary look-up + BOW + POS + Orientation+ distance (Window 5) + neighboring number
9	dictionary look-up + BOW +POS + Orientation + distance (Window 3) + neighboring number
10	dictionary look-up + BOW +POS + Orientation + distance (Window 3) + neighboring number+ measurement
11	dictionary look-up + BOW +POS + Orientation (Window 3) + neighboring number + bullet number + measurement
12	dictionary look-up + BOW +POS + Orientation +distance (Window 3) + neighboring number + bullet number + measurement + capitalization

Table 1: Feature combination and test numbering

The experiment is done by browsing through each word in the clinic texts. For each word, a feature vector which involves all the features described as above table is set up. The meaning of these features is illustrated as follows. Dictionary look-up: a binary value feature (0 or 1) that represents if the NE is in the dictionary or not. SNOMED-CT was used as our dictionary. Bag of Words (BOW): a feature used to display which words surround the word present as well as the word itself. Usually, we select 3-5 neighboring words from its left and from its right (or 6-10 together). This is also called the context feature. Namely, from it, we can get information about word co-occurrences. Part-of-speech tags (POS) of BOW: the POS tags of the context words. This can provide information about the surrounding syntactic context of the target word and the POS assignment of the target word itself. Window size: number of tokens representing context surrounding the target word. For example, a window $size = 5$ will take 5 tokens to the left and 5 tokens to the right of the target word. Their combination is the context representation within that window size. Orientation: the location of the feature in regard to the target word. Possible values are left (L) or right (R). Distance: the proximity of the feature in regard to the target word. In our project, its maximum value is dependent on the window size.

Below is an example illustrating the features for the target word 'patient', in the sentence "The patient has a headache."

word_patient, pos_NN, word_L1_the, pos_L1_DET, word_R1_has, pos_R1_V, word_R2_a, pos_R2_DET, word_R3_headache, pos_R3_NN

In this example, word_L1_the means that the word 'the' is the first word on the left of the target word 'patient' and pos_L1_DET means that its part of speech is determiner.

Capitalization: the four possible token-based values are all_Upper _Case, all_Lower_Case, mixed_Case and initial_Upper_Case.

Bullet_Number: a feature referring to the presence or absence of bullet numbers used to format lists of entries in the clinical text. In clinical notes, physicians may often list a few possible cause of a symptom or list a few suggestions of treatments. Numbers used in such a list may provide some clues for potential named entities.

Measurement: a feature referring to the presence of a number found in conjunction with a measurement unit, e.g. mg or cm. Such a feature is often closely related to a medicine or a disorder. Thus, it may be a useful feature in the discovery of a named entity in clinic notes.

Neighboring Number: a feature referring to the presence of numbers around the target word (except bullet numbers.) For example, in "The patient weights 150 pounds.", "patient" has both measurement (pounds) and neighboring number features (150). We experimented with a variety of features and combinations of features shown in Table 1.

In addition, as model descriptions show CRFs, which are derived from HMMs have taken dependences between neighboring words into considerations. Its implementations are in fact a finite state machine. But SVM only aims at classifier data into two or more categories. Dependences between words are not involved. In order to catch the dependence, we add previous predicted labels as a feature to the current feature vectors.

5 Evaluation Metrics

We report recall, precision and F-score metrics as defined below. Table 2 shows the relation between the evaluation metrics.

	Gold	Standard	
	True	False	
Positive	True positive	False positive	Positive Predicative Value
Negative	False negative	True negative	Negative Predicative Value
	Sensitivity	Specificity	

T able 2: Evaluation Metrics (7)

$$recall = \frac{\#OfTruePositives}{\#OfTruePositives + \#OfFalseNegatives} \quad (8)$$

$$precision = \frac{\#OfTruePositives}{\#OfTruePositives + \#OfFalsePositives} \quad (9)$$

$$F\_score = \frac{2 * Precision * Recall}{Precision + Recall} \quad (10)$$

We apply the evaluation metrics to both the IOB format and the gold standard NE format (GSNE). The CRFs and SVM models are trained on the IOB format and then the output is converted to the GSNE format.

The features were annotated using the Mayo Clinic Information Extraction System [Savova *et al.*, 2008] built on the IBM's Unstructured Information Management Architecture[3] framework. Our baseline is the dictionary lookup method (Test 1 from Table 1).

6 Results

We trained and evaluated CRFs and SVMs with the 12 feature combinations from Table 1. Evaluation was done against the IOB and GSNE format.

6.1 Results with IOB notation: CRFs

Figure 1 shows the results from our runs of CRFs and the IOB notation. Our baseline (Test 1) is the dictionary lookup. The F-scores, recall and precision are 0.604, 0.468 and 0.852 respectively. When BOW is applied to a window of 5 in Test 2 results improve sharply. More than 0.15, 0.17 and 0.08 points are added to the F-score, recall and precision respectively. In number of NEs, the increase is from 257 to 323 in Predicted NEs and from 219 to 299 in Correct Predicted NEs. Test 3 with a window size of 3 yields better results than Test 2. Its F-score increases slightly, the main contribution coming from the improved recall. In Test 3 precision decreases slightly which demonstrated a precision-recall trade-off. Increased recall implies the possibility of increase in both true positives and false positives. If the addition in false positives surpasses that of true positives, it will lead to the reduction of precision. Results improve even further when capitalization is added as a feature in test 12. The F-score, recall and precision rise remarkably: 0.858, 0.774 and 0.963 respectively. The number of predicted NEs and correctly predicted NEs increase dramatically compared with Test 9 (from 336 to 376 in the number of output NEs and from 312 to 362 in the number of correct NEs). The reason capitalization features are strongly discriminative is supported by the observation that disorder NEs display capitalization patterns: 1) capitalization of the first NEs token, 2) capitalization of all NEs tokens, and 3) mixed capitalization.

6.2 Results with IOB notations: LibSVMs

Figure 2 shows the results from our runs of SVMs and the IOB notation. RBF was our choice kernel function; we also experimented with the four kernel functions and the different parameters provided by LibSVM. RBF kernel function is:

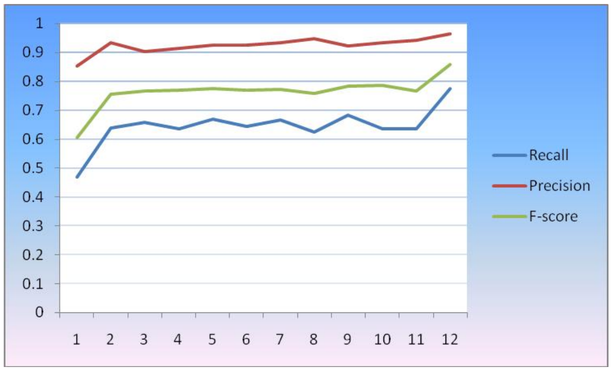


Figure 1: Evaluation Results of CRFs from 12 feature combinations (IOB format)

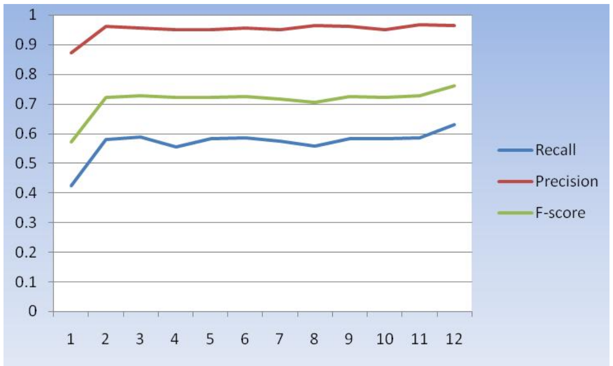


Figure 2: Evaluation Results of SVMs from 12 feature combinations (IOB format)

$$K(x, y) = exp(-\lambda \|x - y\|^2) \quad (11)$$

where λ is a parameter we need to pass in. In addition, a penalty parameter C for each (see formula (3)) is defined. LibSVM provides a v-fold grid search method for the training data on the two parameters using cross-validation. The grid search finds the best C and λ within some range. In our experiments, grid search gives us λ to 0.25 and C to 1. Some changes are made due to the data nature as discussed below. Our training data is heavily skewed towards non_NEs which are represented by about 95 % of the data with the remaining 5% for begin_NEs and inside_NEs. We increased the C margin by setting $C = 1$ for non_NE and $C = 5$ for begin_NE and inside_NE. That is, if non_NE is misjudged as begin_NE or inside_NE, the penalty is 1 while for a begin_NE or inside_NE misjudgement, the penalty is 5.

Compared with results from CRFs, we can see that the general trend is quite similar. The result of Test 1 (our dictionary look-up baseline) has F-score of 0.643. The improvement over the baseline outperforms CRF results. In number of NEs, the sum of true positives and false negatives is 294 and the true positives are 245. But in later experiments, although we observe a few jumpings in test 2, test 3 and test 12, the increasing is much lower than CRFs. Discussion will be given in next session on why this happens.

[3] http://incubator.apache.org/uima/. Last accessed 3/12/ 2008.

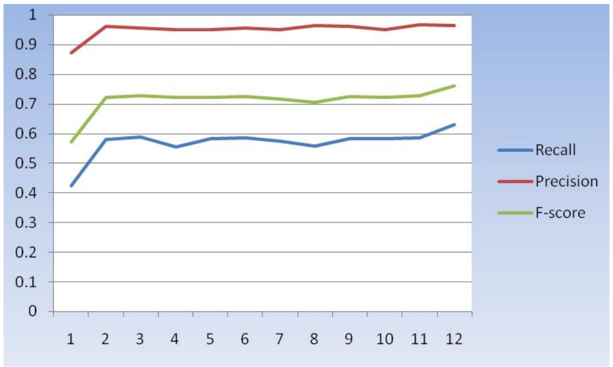


Figure 3: Evaluation Results of CRFs from 12 feature combinations (GSNE format and exact and partial match)

6.3 Results with Gold Standard Named Entities Notation (GSNE):CRFs

Due to the time limitation, we only transformed the IOB format into the GSNE format for CRFs. Figure 3 is the evaluation results of CRFs against our GSNE annotations when both exact and partial matches are considered. Partial matches are meaningful for clinical NER as they usually translate into discovering a broader but still valid term. For example, the NE ”congestive heart failure”, in the IOB format has three labels: begin_NE for ”congestive”, inside_NE for ”heart” and inside_NE for ”failure”. The gold standard NE format has one label: disorder for ”congestive heart failure”. Exact match is when the system predicted label is ”congestive heart failure”; partial match is when some part of it is recognized, e.g. ”heart failure”.

Figure 3 shows the consistency in the performance of the CRF model. Best results are with Test 12. The best F-score is 0.76. There is 6% increase for this feature combination. The precision is above 0.9 except for Test 1.

7 Discussion

In the previous sections, we showed the results from CRFs and SVMs as applied to our task. Here, we present some discussion points.

Test 2 and Test 3 involve the addition of BOW features. There is a marked increase in performance with these features using both CRFs and SVMs. Such an increase indicates the discriminative nature of context for NEs as BOWs are the words surrounding NEs. From a psycho-linguistic perspective, the importance of surrounding con-text could be explained by our tendency to convey ideas in a patterned manner with lexical usage falling within contextual natural classes. This facilitates both language production and comprehension. Pragmatically, new information is always accompanied by some related old information. But why is smaller window size better than larger window size? Probably, due to the informal style of clinical notes where sentences tend to be shorter and information concentrates in shorter contexts. Of note, in the 12 experiments, BOW features are defined in terms of distance and orientation.

Next, we track the contribution of POS tag features. There is only a small increase in the F-score for test 4 and again an increase for test 5 for IOB CRFs and a decrease for Test4 in all other experiments, including SVMs. Apparently, purely grammatical roles are not as important as context. Once again, a small window size yields better results than a larger window size. The recall decreases from 0.648 in Test 3 to 0.645 in Test 4 with CRFs. Precision, on the other hand, increases from 0.896 to 0.926. In absolute numbers, there is no increase in correct NEs (301 in both Test 3 and Test 4) and two more correct NEs are recognized when compared with Test 2 (both of them have window size 5). Similarly, the corresponding recall decrease is from 0.599 to 0.547 and precision increase from 0.809 to 0.847 for SVMs. This indicates that a small window size is more discriminative. As indicated above, clinical notes are unstructured free text with short sentences. If a larger window size is used, many words will share the similar POS features.

As we see from the above charts, IOB CRFs outperforms IOB SVMs in this task (about 0.1 in differences) except test 1. CRFs are in essence a finite state machine derived from HMMs which can naturally consider state-to-state dependences and feature-to-state dependences. They also allow both dependencies of feature vectors and of labels. On the other hand, SVMs do not consider such dependencies. SVMs separate the data into categories via a kernel function. They implement this by mapping the data points onto an optimal linear separating hyperplane. This can explain why only dictionary features bring SVMs better results than CRFs. As stated in section 4, in order for SVMs to catch some dependencies between neighboring words, we add previous predicated label as features for the current feature vector. This way, in fact, improves SVMs' performance to a large degree.

Besides, SVMs do not behave well for large number of feature values. For large number of feature values, it would be more difficult to find discriminative lines to categorize the labels. Attributes with larger numerical ranges could potentially dominate the ones with smaller numerical ranges. Meanwhile, there may be a large number of over-lapping values. The kernel values depend on the inner products of the feature vectors. In our project, BOW features involve 5761 possible attributes. Namely, in our datasets, there are 5761 unique words. With such a large range, attributes are unavoidably overlapping and each label may share the same attributes. LibSVM provides tools to scale data with a large number of attributes which we used to compress the feature attributes into a range between -1 and 1. Even so, the high density of feature values presented classification challenges for the SVMs and results did not improve.

Our next question relates to the contribution of each feature to the SVM model. To evaluate that, we trained one SVM model with all other features except the dictionary features. The results are 0.18 for F-score, 0.12 for recall and 0.36 for precision. This shows that the other features are contributing; however not to the extent the dictionary features are. A possible explanation is that LibSVMs focus on the most discriminative features. We performed a similar experiment with the CRFs. If dictionary features were excluded, the best results with the CRFs were 0.38, 0.27 and 0.67 for the F-score, recall and precision respectively. If these same features are added to the dictionary features, the results improve unlike the trend exhibited by the SVMs. This clearly demonstrates that CRFs can handle various unrelated features without conflicting effects. The last question which may arise is why we only tune the parameters for SVMs but not for CRFs. This is due to the nature of the two models. For parameter estimation, CRFs, as a finite state machine, mainly consider the number of orders and the prior distribution of the data while SVMs, a model based on a kernel function, need to consider both the penalty parameter C and the kernel parameter λ. Maximum likelihood is the way we train a CRF model while maximum margin a SVM model. The typical feature of the former is the consistency for parameter estimation and that of the latter is the consistency for classification. Therefore, parameter tuning is not so significant for CRFs. Besides that, the informal nature of clinic texts favors a small window size and also the nearest dependence (hence, first order used). In general, a Gaussian distribution with a 0.5 variance is good enough for the prior distribution. One more reason is that our datasets are not large either. For a maximum margin model, both the penalty parameter and the kernel parameter are unknown beforehand and various in nature. Consequently, parameter search plays an important role for getting a better model and thus better results of predication. Without doubt, due to the greedy nature of a maximum likelihood model, we cannot guarantee its performance on various data. For larger data, it is still necessary to do cross-validation in training the model.

8 Conclusion and Future Work

In this paper, we reported the results of applying two machine learning techniques, CRFs and SVMs, to the task of disorder NER in clinical free-text. Our results show that, in general, CRFs performed better. We demonstrated that well-chosen features along with dictionary-based features improve the performance of both models. In general, the precision results of the CRF models are high. It is the recall that lags behind. In the future, we are planning to focus on improving recall. In our current work, we considered dictionary look-up features, contextual features, capitalization features and some external features. We are planning to experiment with other features; especially morphological markers, since many medical terms are derived from Latin. In addition, we are planning to investigate the application of the two machine learning approaches to the discovery of anaphoric and co-referring expressions. We also plan to experiment with different SVM models. A possible alternative is SVM-HMMs [Altun *et al.*, 2003] which consider both the label dependency and feature interaction. Further, GSNE experiments will also go to SVMs once we get more granular feature vectors for them. Then, a more comprehensive contrast can be made between the two models.

Acknowledgments

The work was partially supported by a scholarship by the Biomedical Informatics and Computational Biology graduate program at the University of Minnesota. We thank James Masanz, Vinod Kaggal and Philip Ogren for programmatic help.

References

[Altun *et al.*, 2003] Yasemin Altun, Ioannis Tsochantaridis, and Thomas Horfmann. Hiddend markov support vector machines. In *International Conference on Machine Learning*, pages 119–128, 2003.

[Chang and Lin, 2001] Chih-Chung Chang and Chih-Jen Lin. Libsvm: a library for support vector machine. In *Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm*, 2001.

[Cortes and Vapnik, 1997] Corinna Cortes and Vladimir Vapnik. Support-vector network. In *Machine Learning*, pages 20:273–297, Cambridge, Massachusetts, 1997. MIT Press.

[Ji *et al.*, 2002] Kazama Ji, Maino T, Ohta Y, and Tsujii Ji. Tuning support vector machines for biomedical named entity recognition. In *Association for Computation Linguistics Workshop on Natural Language Processing in the Biomedical Domain*, Philadelphia, 2002. ACL.

[Lafferty *et al.*, 2001] John Lafferty, Andrew McCallum, and Fernando Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, 2001.

[Leaman and Gonzalez, 2008] Robert Leaman and Graciela Gonzalez. Banner: an executable survey of advances in biomedical named entity recognition. In *Pacific Symposium on Biocomputing*, pages 13:652–663, 2008.

[McCallum and Li, 2003] Andrew McCallum and Wei Li. Early results for name entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *Pacific Symposium on Biocomputing*, pages 13:652–663, 2003.

[McCallum, 2002] Andrew McCallum. Mallet: A machine learning for language toolkit. In *Software available at http://mallet.cs.umass.edu.*, 2002.

[Ogren *et al.*, 2008] Philip Ogren, Guergana Savova, and CG Chute. Constructing evaluation corpora for automated clinical named entity recognition. Marrakesh, Morocco, 2008.

[Savova *et al.*, 2008] Guergana Savova, Karin Kipper-Schuler, Jim D Buntrock, and CG Chute. Uimabased clinical information extraction system. Marrakesh, Morocco, 2008.

[Tsuruoka *et al.*, 2007] Yoshimasa Tsuruoka, John McNaught, Junichi Tsujii, and Sophia Ananiadou. Learning string similarity measures for gene/protein name dictionary look-up using logistic regression. *Bioinformatics*, pages 2768–74, 2007.

Exploring semantic locality patterns in the biomedical literature

Alexa T. McCray, Kyungjoon Lee
Center for Biomedical Informatics
Harvard Medical School
10 Shattuck Street, Boston, Massachusetts 02115, USA
{alexa_mccray, joon_lee}@hms.harvard.edu

Abstract

We developed fully automated methods and tools for identifying trends and patterns in scientific discovery through investigation and visualization of MEDLINE data. We illustrate our general methods through the detailed analysis of one concept, 'autistic disorder'. We analyzed MeSH topic co-occurrence data in progressively higher levels of semantic aggregation and showed trends from 1962 to 2007. The 2,518 topics that co-occurred with 'autistic disorder' in 8,356 citation records are unevenly distributed across all MeSH categories, across 116 UMLS semantic types, and all 15 UMLS semantic groups. We present a number of views of the data. Our results closely mirror key events and accepted knowledge about this topic. Our methods should prove useful for others as they study evolving knowledge in a field of interest.

1 Introduction

The National Library of Medicine's Medical Subject Headings (MeSH) not only provide an important means to access the biomedical literature, but, taken in their entirety, they also provide a complex and rich window into the world of biomedical scientific discovery. MeSH is a thesaurus of descriptors that consists of sixteen top level nodes with up to eleven levels of embedding. MeSH is regularly updated and currently comprises almost 25,000 descriptors with almost four times as many additional entry terms [National Library of Medicine, 2008].

The work reported here has been done in the context of our broader bibliome mining research. We have created a web-based automated bibliome mining system, called Medvane. Medvane summarizes past research history and shows research trends for a given bibliome. Bibliome views can be constructed in any arbitrary way: a single disease such as autism, multiple diseases such as mental diseases, drugs, a researcher, an institution, or a country. Users can easily search and view the interrelationships between these aspects and their evolution over time. Medvane has been released as one of the first modules in Harvard's Clinical Translational Science Center (CTSC) system [Harvard Catalyst, 2008].

1.1 Semantic Locality

In this study, we have been interested in the "semantic locality", or semantic neighborhood, of concepts in biomedicine [Nelson *et al*., 1991; Bodenreider, 2001].

In particular, we have chosen one evolving concept and have investigated its place in the biomedical literature since the time this concept first came into the MeSH system in 1962. Specifically, we investigated semantic locality through the MeSH terms that have co-occurred with the MeSH topic 'Autistic Disorder' in MEDLINE citation records. Others have studied co-occurrence information with the goal of making explicit the implicit relationships between MeSH co-occurring terms (e.g., [Srinivasan and Rindflesch, 2002; Srinivasan and Hristovski, 2004; Zhu *et al*., 2003]), between UMLS (Unified Medical Language System) co-occurring concepts [Burgun and Bodenreider, 2001], and between text words and other keywords in MEDLINE records (e.g. [Stapley and Benoit, 2000; Alako *et al*., 2005, Urbančič *et al.,* 2007]. We have been concerned with automated methods for characterizing the semantic space of any arbitrary concept - in this case we chose the concept 'Autistic Disorder' - and in understanding how that semantic space has evolved over time.

1.2 Autism

Autism is one of a spectrum of disorders that encompass a range of highly heritable, complex neurodevelopmental disorders diagnosed in large numbers of children, usually before the age of three, from all socioeconomic, ethnic, and cultural backgrounds [Folstein and Rosen-Sheidley, 2001; Fombonne, 2003]. Autism spectrum disorders are four times more likely to occur in boys than girls, and family studies show a strong, though not yet well-defined, genetic contribution to this condition.

The Diagnostic and Statistical Manual of Mental Disorders (DSM) is the standard classification of mental disorders used by mental health professionals in the United States [American Psychiatric Association, 2000]. The DSM-IV-TR recognizes five pervasive developmental disorders: autistic disorder, Rett's disorder, childhood disintegrative disorder, Asperger's disorder, and pervasive developmental disorder, not otherwise specified. For each of these conditions, DSM lists extensive diagnostic criteria. For autistic disorder there are three major criteria: 1) qualitative impairment in social interaction, 2) qualitative impairments in communication, and 3) restricted repetitive and stereotyped patterns of behavior, interests, and activities. Autistic disorders last throughout a lifetime and are often associated with other co-morbid conditions.

2 Methods and Materials

2.1 Materials

We downloaded all MEDLINE citation records indexed with 'Autistic Disorder' as a MeSH major topic. For each MeSH major topic that co-occurred with ‘Autistic Disorder’, we extracted from the 2008 MeSH Descriptors XML file both the MeSH tree numbers and the UMLS semantic types that had been assigned to those descriptors. We additionally downloaded from the NLM the UMLS semantic groups, a construct that aggregates the semantic types into higher level categories [McCray *et al.,* 2001]. We used existing open-source visualization tools to present high-level views of the data [Shneiderman, 1992; yWorks, 2008].

2.2 Methods

Our methods grouped the data in progressively higher levels of semantic aggregation. First, our algorithms calculated the number of specific MeSH topics that co-occurred with the concept of interest. This overall frequency was compared with the frequency during roughly ten year time periods from 1962 (when ‘Autistic Disorder’ was first used as an index term) to 2007. For each of these periods, our algorithms normalized according to the total number of articles in MEDLINE, so as not to over-estimate the growth of a particular topic over time. That is, if a MeSH term such as ‘Epilepsy’ co-occurred with ‘Autistic Disorder’ with increasing frequency over the decades, we wanted to be able to distinguish whether this was because there was increased observed incidence of epilepsy in the context of autism or whether it was simply an artifact of there being more articles overall in the MEDLINE database.

Second, our algorithms aggregated according to high level MeSH categories (e.g., A (Anatomy), F (Psychiatry and Psychology), etc.) and subcategories (e.g., A01 (Body Regions), F01 (Behavior and Behavior Mechanisms), etc.). This would allow us to demonstrate, for example, that our concept of interest appeared more heavily with, for instance, psychiatry and psychology terms than with anatomy terms. In the aggregation process, a descriptor that is found in multiple MeSH categories or subcategories is distributed evenly across each of them, so as not to over-estimate the contribution of one semantic (MeSH category) area over another.

Third, our algorithms aggregated according to UMLS semantic types and semantic groups. We were interested in comparing whether semantic locality could be better expressed at this higher level of semantic aggregation than at the level of MeSH categories and subcategories. Again, for this step, any descriptor that was assigned to multiple semantic types or semantic groups was evenly distributed.

The previous steps all analyzed pair-wise co-occurrences which then led to overall views of semantic locality. The final step investigated local co-occurrence patterns in each of the citation records and then generalized further from these patterns. Specifically, for each citation record we calculated all co-occurrence tuples and created a highly interconnected graph of these concepts.

3 Results

For the period we considered in this study (1962-2007), there are over 16 million citation records in MEDLINE. A total of 8,356 of these are indexed with the MeSH major topic ‘Autistic Disorder’.

3.1 Autistic Disorder as Major Topic

Table 1 shows the total number of citation records for each of the periods we analyzed .

Time Period	**[a]Number of articles**	**[b]Autistic Disorder as Major Topic**	**b / a × 10,000**
1962-1967	970,004	156	1.6
1968-1977	2,299,892	580	2.5
1978-1987	3,034,552	1,133	3.7
1988-1997	4,099,059	1,876	4.6
1998-2007	5,837,437	4,611	7.9
Total	16,240,944	8,356	5.2

Table 1. MEDLINE citation counts for 1962-2007

Note that there has been a steady increase in the number of articles about autism over the years, measured both in actual numbers and also as a relative proportion of the entire MEDLINE database. Of the total 8,356 autism articles, 7,622 articles are co-indexed with one or more of 2,518 unique major topics. The remaining 734 articles are indexed with only one major topic, ‘Autistic Disorder’.

3.2 Topics Co-occurring with Autistic Disorder

Figure 1 shows the 10 most highly frequent topics that co-occurred with ‘Autistic Disorder’ over time. We show

All Years	**1962-1967**	**1968-1977**
Mental Retardation	Child	Schizophrenia, Childhood
Brain	Schizophrenia, Childhood	Behavior Therapy
Behavior Therapy	Psychotic Disorders	Mental Retardation
Social Behavior	Child Behavior Disorders	Language Disorders
Language Development Disorders	Psychotherapy	Language Development
Asperger Syndrome	Infant	Education, Special
Attention	Mental Retardation	Verbal Behavior
Child Development Disorders, Pervasive	Parent-Child Relations	Conditioning, Operant
Cognition Disorders	Schizophrenic Psychology	Attention
Parents	Adolescent	Parents
1978-1987	**1988-1997**	**1998-2007**
Mental Retardation	Mental Retardation	Brain
Behavior Therapy	Language Development Disorders	Asperger Syndrome
Sex Chromosome Aberrations	Behavior Therapy	Social Behavior
Brain	Child Development Disorders, Pervasive	Mental Retardation
Fenfluramine	Attention	Cognition Disorders
Fragile X Syndrome	Brain	Measles-Mumps-Rubella Vaccine
Language Development Disorders	Social Behavior	Behavior Therapy
Stereotyped Behavior	Fragile X Syndrome	Child Development Disorders, Pervasive
Communication	Delirium, Dementia, Amnestic, Cognitive Disorders	Language Development Disorders
Child Development	Communication	Attention

Figure 1. Ten most highly frequent topics that co-occurred with 'Autistic Disorder' over time.

the topics according to individual time periods as well as for the entire period under study, 1962-2007.

Table 2 shows the distribution across MeSH high level categories of the 100 most frequent major topics co-occurring with 'Autistic Disorder'.

MeSH Category	MeSH Tree Number	# of Co-occuring MeSH Topics	Examples
Anatomy	A	6	Face; Brain; Cerebellum
Diseases	C	17	Tuberous Sclerosis; Epilepsy; Fragile X Syndrome; Language Disorders
Chemicals & Drugs	D	7	Fenfluramine; Serotonin; Risperidone; Secretin
Techniques & Equipment	E	7	Magnetic Resonance Imaging; Communication Aids for Disabled; Language Therapy
Psychiatry & Psychology	F	56	F01: Interpersonal Relations; Emotions; Intelligence F02: Cognition; Object Attachment; Aptitude F03: Schizophrenia Child Behavior Disorders; Developmental Disabilities; F04: Psychotherapy; Personality Assessment
Education & Social Phenomena	I	5	Social Environment; Education, Special; Play and Playthings
Named Groups	M	1	Child
Health Care	N	1	Early Intervention (Education)
Total		100	

Table 2. Distribution across MeSH categories of top 100 topics co-occurring with 'Autistic Disorder'.

3.3 Semantic Type Aggregation

The 2,518 topics co-occurring with 'Autistic Disorder' are distributed across 116 semantic types of the 135 available and across all 15 semantic groups. The most frequent semantic types of these co-occurring concepts are 'Mental or Behavioral Dysfunction', 'Mental Process', 'Disease or Syndrome', and 'Social Behavior'.

3.4 Views of Co-occurring Concepts

Figures 2 and 3 present global views of the co-occurring concepts. Figure 4 shows local co-occurrence patterns, and Figure 5 shows the distribution of the co-occurring concepts over time.

4 Discussion

Figure 1 shows the 10 most highly frequent topics that co-occurred with 'Autistic Disorder' over the total time period and in each of the segments we considered. Notice that over the last 46 years, and in each of the individual time periods, 'Mental Retardation' appears in the top ten list. This is consistent with the observed rate of co-morbidity in the autism population [Folstein and Rosen-Sheidley, 2001; Fombonne 2003]. Notice further that in the early time period (1962-1967) there was a strong focus on the population ('Child', 'Infant', 'Adolescent') in which autism was being diagnosed and on the attempt at an explanation for this phenomenon ('Parent-Child Relations'). In fact, in the early years, mothers of autistic children were often blamed by the medical establishment for the social reticence of their children and were labeled "refrigerator moms" [Folstein and Rosen-Sheidley, 2001].

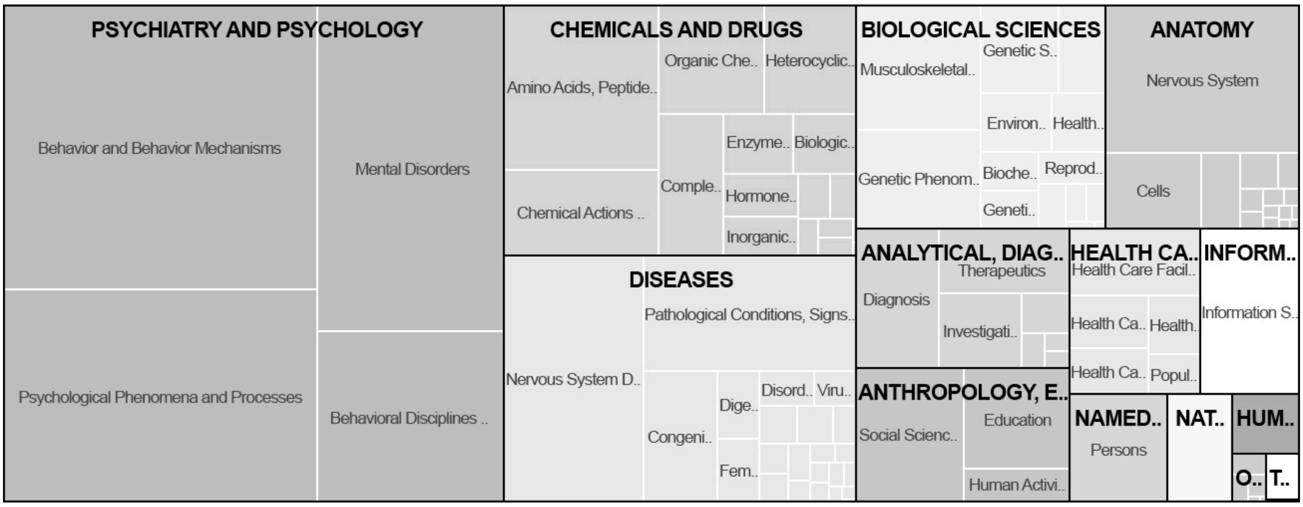


Figure 2. Treemap of all MeSH major topics co-occurring with 'Autistic Disorder', organized according to MeSH categories and subcategories.

The 1968-1977 time period reflects a number of important events during that era. First, the so-called TEACCH program was founded in 1971 to improve the treatment and therapy of children with autism. This is reflected by the high frequency of the topics 'Behavioral Therapy' and 'Education, Special'. In addition, according to Singh *et al*., comparative studies were being conducted during this time in order to distinguish this condition from childhood schizophrenia [Singh *et al*., 2007]. In the most recent time frame 1998-2007 brain and behavior research has been heavily funded, and this is reflected in the most highly co-occurring term. Other trends become evident in studying the timeline data, but one additional area of research was spurred by a 1998 study that suggested a causative effect of vaccines on autism [Singh *et al*., 2007:157]. The study has had enormous impact, and though there have been multiple subsequent studies finding no evidence of this effect, the controversy continues to this day.

Table 2 shows a view of the most frequently co-occurring topics with the concept of autism. What is interesting to note is that a relatively few number of unique terms in each MeSH category form the semantic space of this concept. We can see known co-morbid conditions (e.g., 'Epilepsy', 'Fragile X Syndrome'), a small number of drug therapies (e.g., 'Risperidone'), a small number of diagnostic and therapeutic techniques (e.g., 'Language Therapy), and an even smaller number of anatomical terms (e.g., 'Brain').

While the MeSH categories are useful for thinking about the semantic space, notice that terms that, arguably, belong together actually cross MeSH high level categories. For example, in Table 2, terms in both 'C' and 'F03' refer to disorders; terms in both 'E' and 'F04' refer to diagnostic and therapeutic procedures, etc.

It is just this sort of issue for which the UMLS semantic types have proved useful. Using treemaps we are able to see at a glance the difference between aggregating all co-occurring concepts at the MeSH category level (Figure 2) and aggregating at the semantic group level (Figure 3).

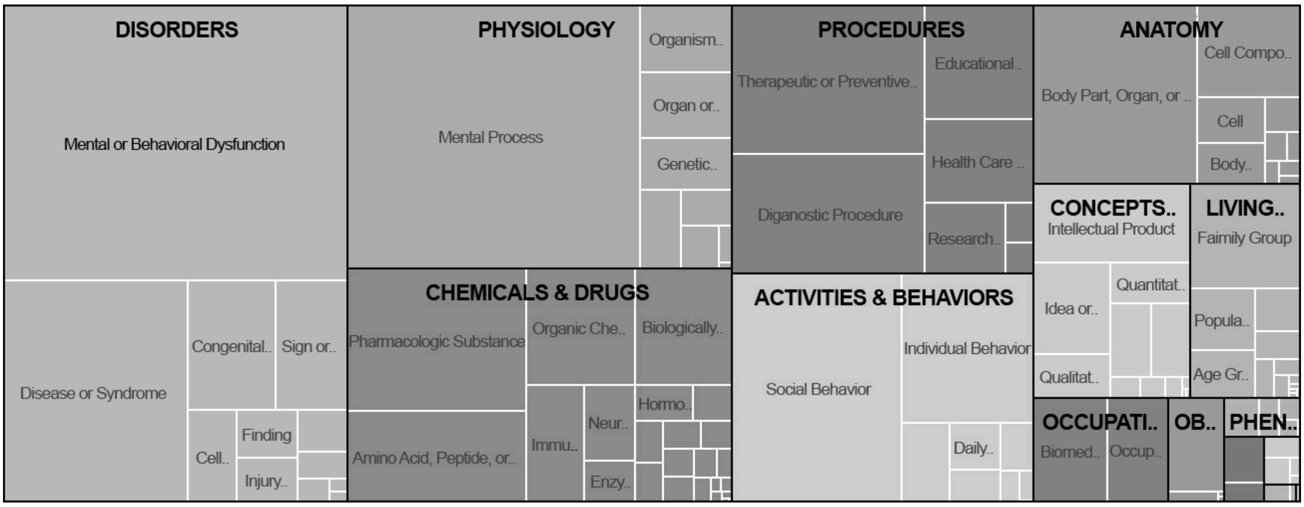


Figure 3. Treemap of all MeSH major topics co-occurring with 'Autistic Disorder', organized according to UMLS semantic groups.

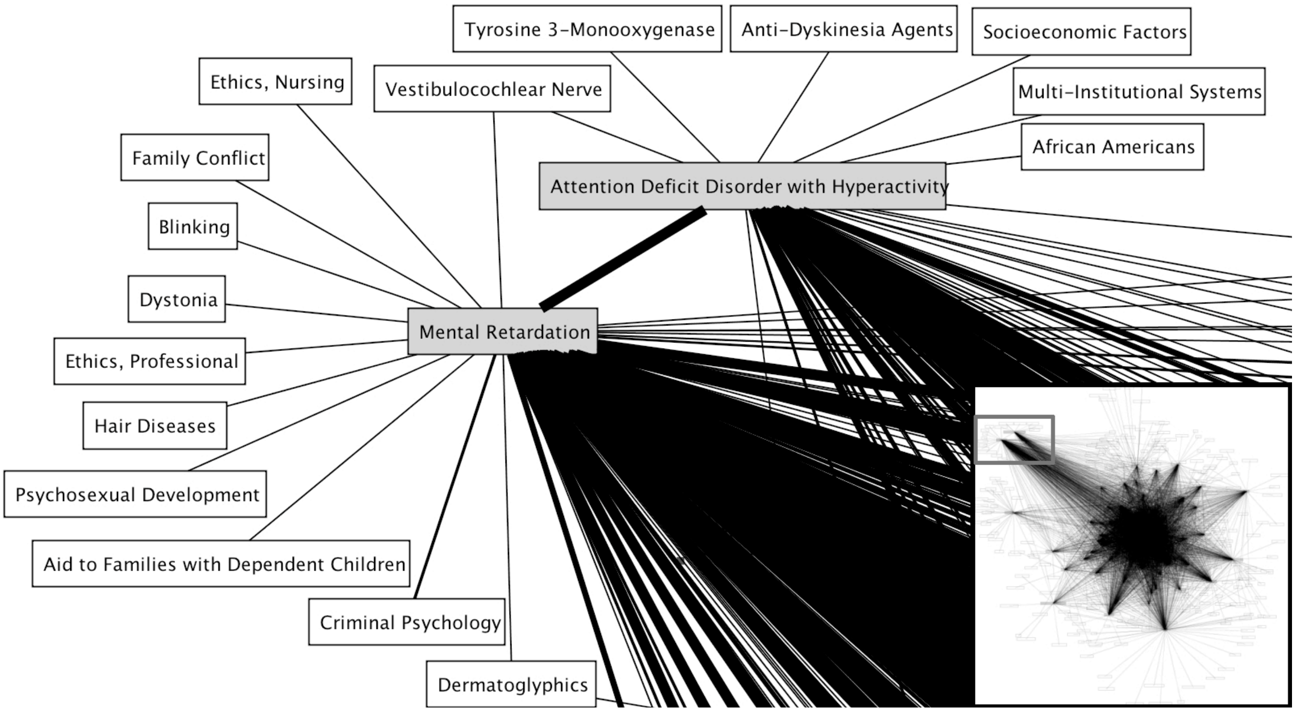


Figure 4. Graph of local co-occurrence patterns of top 100 concepts, with partial interactive view.

Notice that compared with Figure 2, Figure 3 groups all of the "Disorders' together (drawn from MeSH categories C and F) and all diagnostic and therapeutic techniques (drawn from MeSH categories E and F) are grouped together in the 'Procedures" semantic group.

The results of our analysis of local co-occurrence patterns (tuples of MeSH major topics that co-occur in a given citation record), allows a view of the data that clusters and then interrelates topics according to their local neighborhood. Figure 4 illustrates the graph of local co-occurrence patterns for the top 100 MeSH major topics that co-occur with 'Autistic Disorder'. The box in the right hand corner shows a global view (with "drill-down" capabilities) of all of the interconnections for the top 100 topics. In the "birds-eye" view we see that 'Mental Retardation', a highly co-occurring term also co-occurs with terms such as 'Family Conflict', 'Ethics, Nursing', and 'Vestibulocochlear Nerve', etc. This latter term is further connected to 'Attention Deficit Disorder with Hyperactivity', and, of course, to 'Autistic Disorder' (not shown).

Figure 5 shows a changing profile over the years. Autistic disorders continue to be heavily studied, together with either co-morbid conditions, or with other conditions, such as schizophrenia, which share some behavioral similarities but from which they need to be distinguished [Singh *et al.,* 2007]. In the last decade, the increase in

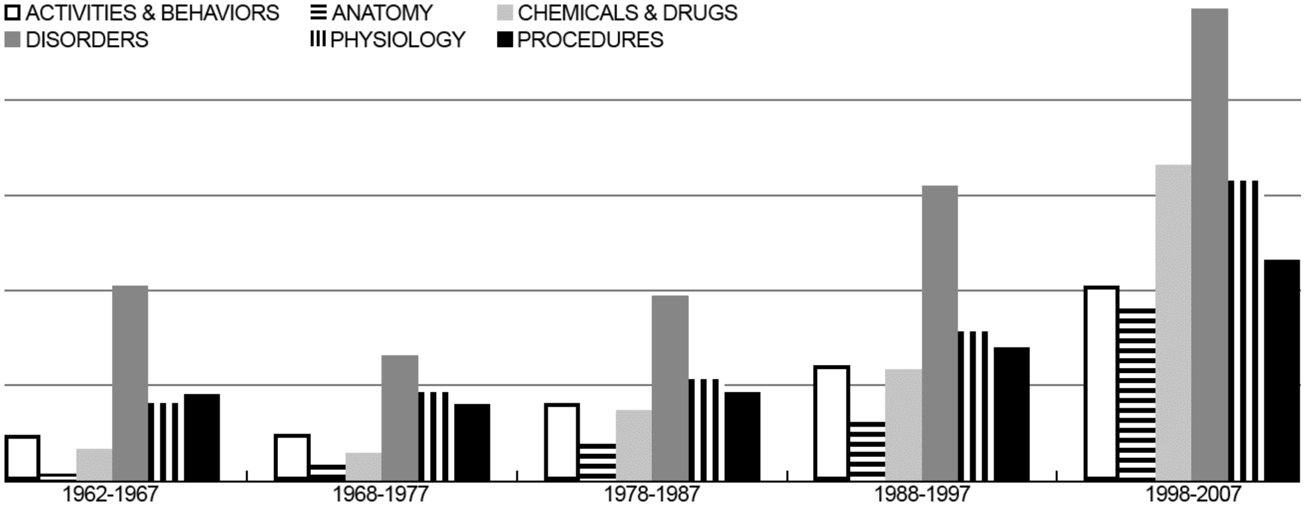


Figure 5. Distribution of co-occurring terms over time, displayed according to UMLS semantic groups.

both anatomical terms (heavily brain research oriented) as well as physiology terms (including genetics), may well be accounted for by increased funding in these areas.

5 Conclusions

We have been concerned with developing automated methods for identifying trends and patterns in scientific discovery. In this study we have been especially interested in the semantic locality of biomedical concepts, as this is evidenced in the biomedical literature. We developed methods to investigate semantic locality at several levels of semantic aggregation. We illustrated these general methods through the detailed investigation of one concept, 'Autistic Disorder'. Our results were evaluated and validated by their consistency with key events in the evolution of autism research.

Our methods can be applied to any biomedical concept, giving insight not only into the specific concept under study, but also into the entire semantic neighborhood of that concept, either at a specific point in time or across the years. Our methods are general and should prove useful to others as they study evolving knowledge in a field of interest.

Acknowledgments

The first author is a member of the Autism Consortium and is partially funded by them. We thank the National Library of Medicine for access to the MEDLINE, MeSH and UMLS data files.

References

[Alako *et al.,* 2005] Alako BT, Veldhoven A, van Baal S, et al. CoPub Mapper: mining MEDLINE based on search term co-publication. BMC Bioinformatics. 2005;6:51.

[American Psychiatric Association, 2000] Diagnostic and Statistical Manual of Mental Disorders, DSM-IV-TR, Washington, DC: American Psychiatric Association, 2000.

[Bodenreider, 2001] Bodenreider O. An object-oriented model for representing semantic locality in the UMLS. Medinfo. 2001;10(Pt1):161-5.

[Burgun and Bodenreider, 2001] Burgun A, Bodenreider O. Methods for exploring the semantics of the relationships between co-occurring UMLS concepts. Medinfo. 2001;10(Pt 1):171-5.

[Folstein and Rosen-Sheidley, 2001] Folstein SE, Rosen-Sheidley B. Genetics of autism: Complex aetiology for a heterogeneous disorder. Nature Rev Genet. 2001;2(12):943-55.

[Fombonne E , 2003] Fombonne E. Epidemiological surveys of autism and other pervasive developmental disorders: An update. J Aut Dev Disord. 2003;33(4):365-82.

[Harvard Catalyst, 2008] The Harvard Clinical and Translational Science Center. http://catalyst.harvard.edu/.

[McCray *et al.*, 1992] McCray AT, Burgun A, Bodenreider O. Aggregating UMLS semantic types for reducing conceptual complexity. Medinfo. 2001;10:216-20.

[National Library of Medicine, 2008] Medical Subject Headings. http://www.nlm.nih.gov/pubs/factsheets/mesh.html.

[Nelson *et al.*, 1991] Nelson SJ, Tuttle MS, Cole WG, et al. From meaning to term: semantic locality in the UMLS Metathesaurus. Proc Annu Symp Comput Appl Med Care. 1991;:209-13.

[Shneiderman B, 1992] Shneiderman B. Tree visualization with tree-maps: 2-d space-filling approach. ACM Trans Graph. 1992;11:92-99.

[Singh *et al.*, 2007] Singh J, Hallmayer J, Illes J. Interacting and paradoxical forces in neuroscience and society. Nature Rev Neurosci. 2007;8(2):153-60.

[Srinivasan and Rindflesch, 2002] Srinivasan P, Rindflesch T. Exploring text mining from MEDLINE. Proc AMIA Symp. 2002;:722-6.

[Srinivasan and Hristovski, 2004] Srinivasan P, Hristovski D. Distilling conceptual connections from MeSH co-occurrences. Medinfo. 2004;11(Pt 2):808-12.

[Stapley and Benoit, 2000] Stapley BJ, Benoit G. Biobibliometrics: information retrieval and visualization from co-occurrences of gene names in Medline abstracts. Pac Symp Biocomput. 2000;:529-40.

[Urbančič *et al.,* 2007] Urbančič T, Petrič I, Macedoni-Lukšič M. Literature mining: Towards better understanding of autism. In Belazzi R, Abu-Hanna A, Hunter J (eds.): AIME 2007, LNAI 4594, 2007:217-226.

[yWorks, 2008] yEd – Java Graph Editor. http://www.yworks.com.

[Zhu *et al.*, 2003] Zhu AL, Li J, Leong TY. Automated knowledge extraction for decision model construction: a data mining approach. AMIA Annu Symp Proc. 2003;:758-62.

An Improved Methodology for Pathway Based Microarray Analysis Based on Identification of Individual Pathways Responsible for Gene Regulation

Stelios Pavlidis*, Annette Payne, Stephen Swift
School of Information Systems, Computing and Mathematics
Brunel University
Uxbridge, UB8 3PH, London, UK
*Contact e-mail stelios.pavlidis@brunel.ac.uk

Abstract

After a decade characterised by intense research in the field of Gene expression analysis there is a considerable shift in interest towards the integration of distinct types of data in a holistic analytical approach, in an attempt to facilitate the large amounts of biological data more efficiently. In this paper we propose a methodology for pathway based microarray data analysis, based on the observation that a number of genes can be simultaneously involved in distinct biochemical processes. We attempt to decipher the functional state of biochemical pathways, under given experimental conditions, to assist the identification of pathways that have been stimulated or repressed.

1 Introduction

Pathway based microarray analysis aims to benefit from the integration of biochemical pathway data and microarray technology. Instead of concentrating on the often subtle change occurring in the expression of each individual gene, between two conditions of interest, we are now looking for a coordinated change among a set of genes.

Ideally, given a microarray experiment, a biologist would want to simply and efficiently identify those biochemical pathways that have been stimulated along with pathways that have been suppressed by the experimental conditions. Many available tools leave this to the biologist to decide by providing a useful visualisation of biochemical pathways. Some attempt to statistically verify the validity of the obtained data, but often, the state of each individual gene, meaning up- or down-regulation, is not taken into account, with the analysis considering only two types of genes, i.e. affected (up- and down-regulated) and unaffected (stable).

We propose a method that may facilitate the identification of the state of biochemical pathways of interest, by taking into consideration the state of the genes forming each pathway. We show that the configurations produced by our method are consistent, by comparing the results of multiple runs of our algorithms and explore three different starting configurations, discussing their advantages and disadvantages.

In Section 2 we provide a more detailed analysis of the background and the motivation behind this study. We discuss the rationale of our approach in section 3, while section 4 describes the applied algorithms. In section 5 we evaluate our hypothesis and present the results of the application of our method to some real life microarray data sets. Finally, in section 6 we conclude with some discussion of our approach and give some suggestions for future work.

2 Background

The study of biochemical pathways is of great significance in biological and medical research, notably in our effort to design efficient drugs that can target molecules of interest. Currently, a number of pathway lists and other meaningful gene sets are available in public databases, including KEGG [Kanehisa *et al.*, 2000], Reactome [Joshi-Tope *et al.*, 2005], SABIO-RK [Rojas *et al.*, 2007] and others. Characteristic examples of tools developed for visualisation of a metabolic pathway's behaviour, based on microarray data, include GenMapp [Dahlquist *et al.*,2002], Cytoscape [Shannon *et al.*, 2003], Pathfinder [Goesmann *et al.*, 2002], GeneNet [Kolpakov *et al.*, 1998] and Eu.Gene [Cavalieri *et al.*, 2007]. These tools do not attempt to give definitive answer regarding the state of biochemical pathways. They refrain themselves to graphically portraying the expression of each gene of the pathway, leaving it to the user to make a judgment about the state of the pathway. At most, tools like Eu.Gene provide some statistical indication regarding the reliability of the examined data, as discussed in Section 5.5.

Regardless of each individual approach, existing methods take into account the expression of all the genes forming a pathway, thus all methods are based on some type of averaging. However, it is a well established fact that not all genes in a biochemical pathway show similar behaviour in terms of RNA production. To some extent, this is due to gene diversity, given that genes forming a particular pathway can encode proteins of very different functionality [Stryer *et al.*, 2002]. Additionally, since there are many different levels of regulation, including translation, protein maturation and degradation rate [Quadroni *et al.*, 1999] and so on, it is not unusual to observe that genes respond differently in terms RNA production.

Furthermore, as previous work in the field [Panteris *et al.*, 2007] has already suggested, not all genes in a pathway are representative of the pathway's behaviour, partly due to the fact that a number of genes can participate in more than one biochemical pathways.

It can be inferred that there are two types of genes, meaning those that are members of only one pathway, to which we will hence forth refer as single-membership genes, and those that are members of two or more pathways, to which we will hence forth refer as multi-membership genes. Consequentially, up-regulation of a multi-membership gene that is a member of two pathways can be due to its contribution to the activity of either or both pathways. In that sense, the intensity value for any multi-membership gene, extracted from a microarray chip, is the net effect of the contribution of that gene to all biochemical pathways it is a member of. To our knowledge the multi-membership nature of genes, has not been extensively examined in pathway based microarray analysis. We believe that this omission may lead to misleading conclusions, as it is likely that genes that participate in more than one biochemical pathways can mask the true pathway behaviour and confer analysis trivial.

For example let's assume a hypothetical pathway consisting of genes A, B, C, D, E and F, responsible for the production of molecule X, and a situation where the cell requires increased amounts of molecule X. Additionally, let's assume that genes D, E and F are also involved in other pathways where there is a parallel need for substantial decrease of activity. When examining RNA levels on a microarray corresponding to these conditions, we might observe increased expression of molecules A, B and C but decreased expression for molecules D, E and F. If we do not take into account the multi-membership nature of genes D, E and F, but instead solely concentrate on the path consisting of genes A to F, producing biochemical product X, we may find the results confusing. However, the controversy is removed, if we take into consideration that genes D, E and F are also involved in down-regulated pathways.

3 Rationale

We assume that given a certain pathway and a certain state, e.g. up-regulation, genes constituting that pathway, that fluctuate in their expression, follow the trend and show up-regulation. That is, they produce more RNA to contribute to the pathway function. Thus, we attempt to ascribe any observed down-regulation of genes in this pathway to decreased activity of other pathways of which these genes are also members. We assume that the net effect of the contribution of these genes to pathways of contradicting behaviour may be responsible for the intensity values extracted from the microarray chip.

Escherichia coli data from [Khodursky *et al.*, 2000] available as experiment GSM513 at **G**ene **E**xpression **O**mnibus (**GEO**), can exemplify our line of thought. The experiment examines the cell response to addition of excess tryptophan in the growth medium. Naturally, the cell intensively degrades the amino acid following up-regulation of the tryptophan metabolism pathway, as present in **K**yoto **E**ncyclopedia of **G**enes and **G**enomes (**KEGG**) database. Most tryptophan metabolism genes show subtle to substantial up-regulation except from yqeF which shows significant down-regulation (Table 1). However, according to KEGG gene yqeF is also member of seven other biochemical pathways, some of which involve degradation of amino acids, other than tryptophan. It is biologically meaningful to observe decline in the activity of such pathways given that the cell is presented with excess tryptophan to cover its nutritional needs.

Table 1. Log2 ratios of tryptophan metabolism genes, for experiment GSM513

Gene Symbol	Log2 ratio	Gene Symbol	Log2 ratio
'atoB'	1.1150	'trpS'	5.8490
'yqeF'	-1.8120	'katE'	-0.4370
'fadB'	2.6340	'katG'	1.4110
'sucA'	1.8200	'tynA'	-0.7870
'tnaA'	1.4660		

The Pentose Phosphate pathway in the diauxic shift experiments [DeRisi *et al.*, 1997] discussed in the Results section offers further evidence that taking the multi-membership nature of genes into account may be beneficial. Six genes included in the pathway show up-regulation while another six show down-regulation (Figure 1a). It may be insufficient to just say that 12 genes are affected, in order to make an informative guess about the state of the pathway. However, an examination of the pathway membership of the genes reveals that most up-regulated genes are unique members of the Pentose Phosphate pathway, while all down-regulated genes are involved in one additional pathway, in most cases the Purine metabolism pathway. Given that Purine metabolism is severely down-regulated (data not shown) it may well be responsible for their behaviour.

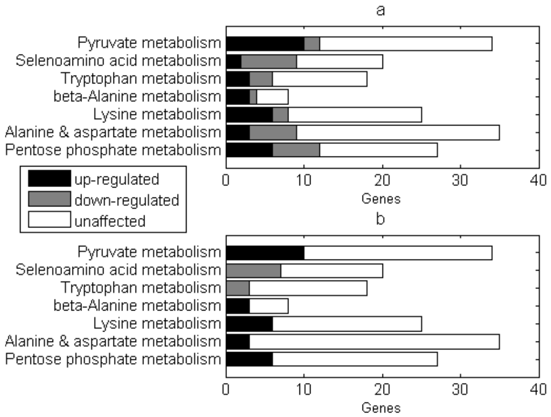


Figure 1. Figure 1a shows gene expression for a set of chosen pathways. Figure 1b reveals the allocation of genes for the same pathways, upon processing with our method.

4 Methods

To render data analysis more comprehensive, each microarray dataset is trimmed to only include genes contained in KEGG pathways. The state of each gene is defined as up-regulated, down-regulated or stable, based on a chosen set of thresholds. Following that, we apply a Hill

Climbing algorithm [Michalewicz *et al.*, 1998] that changes the possible multi-membership gene configuration.

Essentially, we are changing the allocation of multi-membership genes to their constituent pathways to define which pathways are more influential as far as their expression is concerned. Assigning a gene to a pathway, suggests that the gene is involved in the function of that pathway, in the experiment in hand. Similarly, removing a gene from a pathway means that the state of expression of that gene, up- or down-regulation, is not due to its involvement in the activity of that pathway, in the particular experiment. If for example the gene is down-regulated and we remove it from an up-regulated pathway, we are implying that it is not the contribution of that gene to the particular biochemical pathway that is causing it to produce less RNA, in the particular experiment under examination.

4.1 Algorithm

We will first define some notation that is used within our methods and algorithms. Let P be an N row by M column binary matrix, $P \in B^{NxM}$. Let p_{ij} (the element in the ith row and jth column of matrix P) = 1 if gene i is a member of pathway j,and P_{ij} = 0 if gene i is not a member of pathway j. Therefore P represents a snapshot of KEGG membership of genes to pathways for a given species and does not change.

Let $A \in B^{NxM}$ be a binary matrix such that $P\text{-}A \in B^{NxM}$. A represents a potential allocation of genes to pathways and will be used by our method, see algorithm 1, here $a_{ij} = 1$ if gene i is allocated to pathway j and $a_{ij} = 0$ if gene i is not allocated to pathway j.

The restriction $P\text{-}A \in B^{NxM}$ means that A can define pathways to have less genes than originally in P, but can never have genes that contradict P, i.e. we do not allow allocations that would be contrary to that in KEGG.

Let us assume that we have a single set of gene expression data (one experiment) for the N genes called G. We score an allocation on how much each pathway is down or up regulated according to equations (1) to (3), note that the constant c is a threshold parameter.

$$X(i) = \begin{cases} +1, \text{if } G(i) > \text{c} \\ -1, \text{if } G(i) < \text{- c} \\ 0 \text{ ,otherwise} \end{cases} \tag{1}$$

$$F(A) = \sum_{j=1}^{M} \left| \sum_{i=1}^{N} H(a_{ij}) \right| \tag{2}$$

$$H(a_{ij}) = \begin{cases} X(i), \text{if } a_{\text{ij}} = 1 \\ 0 \text{ ,otherwise} \end{cases} \tag{3}$$

X (i) has a value of +1, -1 or 0 if gene i is up-, down-regulated or stable respectively. $F(A)$ is our fitness function, which we aim to maximise by changing the allocation of multi-membership genes to their corresponding pathways. We use equation (3) to define if gene i is a member of pathway j, which is true if a_{ij}=1, and if that is the case to define if the gene is up-, down-regulated or stable. Function $\sum H(a_{ij})$ reveals the difference between the numbers of up- and down-regulated genes in pathway j. Thus, the more genes of similar expression are allocated to pathway j the greater $|\sum H(a_{ij})|$ becomes for that pathway.

We implement three different starting allocations. In the case of single membership allocation, only the single membership genes are assigned to pathways. In the case of full membership, multi-membership genes are assigned to all possible pathways they belong to. In the case of directed membership, we allocate single membership genes to their corresponding pathways, and then go through the pathways that are still empty, to check if they would contain more up- or down-regulated genes upon full allocation. If the full allocation contains more up-regulated genes, we randomly assign one of the up-regulated genes to the corresponding pathway in the starting allocation. If on the other hand the full allocation contains more down-regulated genes we randomly assign one of the down-regulated genes to the pathway, in the starting allocation.

```
ALGORITHM 1: SEARCH ALGORITHM

1)  Input: $a$ = list of gene IDs coupled with their pathway
    IDs
2)  Input: $b$ = Expression vector of log2 ratios (only
    KEGG pathways genes)
3)  Input: $c$ = threshold for up-/down-regulated genes
    Input: allocation_type = one of {single, multiple, directed}
4)  Remove all genes between $+c$ and $-c$
5)  If allocation_type = single then allocate single membership genes to their pathways (thus create $A$)
6)  Elseif allocation_type = multiple then allocate all
    genes to all the pathways they are members of (thus
    create $A = P$)
7)  Elseif allocation_type = directed then Call Algorithm 2
8)  Get fitness $F(A)$, set $F\_old = F(A)$
9)  For $j$ = 1: number of iterations
10)    Save gene configuration
11)    Use $P$ to randomly choose a gene ($i$) with multi-
       membership and randomly choose one of the
       pathways($j$) it belongs to
12)    If according to $A$ gene ($i$) is already present in the
       pathway ($j$) then remove the gene, i.e. set $a_{ij} = 0$
13)    Else if not present, place it in the pathway, i.e. set
       $a_{ij} = 1$
14)    End if
15)    If upon completion of steps (10) to (14) the gene is
       not assigned to at least one pathway, randomly
       choose a pathway and assign the gene to it
16)    Estimate fitness $F(A)$
17)    If $F(A) > F\_old$ set $F\_old = F(A)$
18)    Else if $F(A) < F\_old$ restore gene configuration
       (from step (10))
19) End for
20) Output: $A$
```

Algorithm 2: Set Directed Starting Allocation

```
Allocate single membership genes to their pathways creating $A$
Set Q =  a list of pathways that do not contain single membership genes
For k = 1: length of $Q$
    If $\sum_{i=1}^{N} H\left(a_{iQ_k}\right) > 0$ Then
        Let x = a random up-regulated gene from path way $Q_k$
        Set $a_{xQ_k} = 1$
    End if
    If $\sum_{i=1}^{N} H\left(a_{iQ_k}\right) < 0$ Then
        Let x = a random down-regulated gene from pathway $Q_k$
        Set $a_{xQ_k} = 1$
    End if
```

Each starting allocation has different properties. In the case of single membership starting allocation, the presence of a single membership gene in a pathway will cause the algorithm to fill that pathway with genes of similar behaviour. Hence, if a pathway initially contains a down-regulated single membership gene, the algorithm will keep assigning more down-regulated genes to it. This is a sensible choice, because the behaviour of a single membership gene can be only due to its involvement in that pathway. Thus it is evidence of the pathway's behaviour.

Starting from full membership allocation may also be beneficial. If for example, upon full allocation a certain pathway contains more up-regulated genes, we may argue that the pathway is probably up-regulated and vice versa.

However, there are cases where starting from single membership allocation or full membership allocation may influence the final allocation in opposite directions. To target such issues, we have implemented the directed membership allocation. Here, the presence of single membership genes directs the allocation of genes to the pathways that contain them, while the full allocation directs the filling up of pathways which do not contain single membership genes.

4.2 Comparison of Allocations

The allocation of a gene to pathways, may be represented as a binary string. Thus, the Hamming Distance (*Hamm* below) measure [Hamming et al., 1950] reveals the dissimilarity between two allocations of the same gene. We divide the difference between the number of pathways containing the gene (length of binary string) and the observed Hamming distance by the length of the binary string. Then we add the results for all genes and divide the sum by the number of genes.

Let $D,E \in B^{NxM}$ be binary matrices such that $P\text{-}D \in B^{NxM}$ and $P\text{-}E \in B^{NxM}$, i.e D and E are allocations of genes. Let the similarity between D and E be:

$$S(D,E) = \frac{1}{NM}\sum_{i=1}^{N}(M - Hamm(D_i, E_i)) \qquad (4)$$

where D_i is the *i*th row of D.

5 Results

We first present an informative statistical evaluation of the validity of our hypothesis that unlike single-membership genes, the expression of which is true indication of a pathway's behaviour, multi-membership genes expression is the net effect of their contribution to all their constituent pathways. We discuss some result produced by our method, the convergence of our algorithm, the consistency of the produced allocations and conclude with a comparison of our allocations to the original ones, using a standard statistical approach.

5.1 Hypothesis Validation

We have identified 19 experiments (GSM99081 to 83, GSM99108 to 112, and GSM99171 and GSM99172) from GEO platform GPL3503 that contain a large number of expressed Urea Cycle genes. KEGG Urea Cycle pathway consists of 16 single-membership and 12 multi-membership genes. We divide the intensities for each group and experiment by their sum, to obtain a measure of the contribution of each gene to the behaviour of the pathway. We then compare the correlation between the obtained contribution values of the 12 multi-membership genes and the 16 single-membership genes, throughout the 19 experiments. For both cases we acquire a set of 171 correlation values, and perform a two sample t-test which reveals that the values are significantly different with a *p*-value of $1.3251\text{x}10^{-12}$. Furthermore, in the case of single-membership genes the correlation values are higher with 86.5% of the values being above the level of significant correlation at p=1%. In contrast, for the multi-membership genes only 41.5% of the values exceed the threshold of significance at 1%.

The assumption that multi-membership genes expression is the net effect of their contribution to their constituent pathways is in agreement with our findings. Single membership genes apparently show more consistent behaviour as they only contribute to the functionality of one pathway.

5.2 Data Processing

We have applied our methodology to process data from diauxic shift experiments on *Saccharomyces cerevisiae* [DeRisi *et al*., 1997], using a threshold log2 ratio value of 1 and -1, to consider a gene up- and down-regulated respectively, as suggested by the authors. The data consists of 7 time points and we have chosen to use time point 6 for analysis to demonstrate the utility of our algorithm.

It is evident that all pathways on Figure 1a contain both up- and down-regulated genes. Especially, in the case of the Pentose phosphate pathway, the number of up and down-regulated genes is the same, which makes it difficult to infer the state of that pathway.

Figure 1b shows the new configuration of genes in the very same pathways upon processing of the data with our method. Now, the Pyruvate metabolism pathway contains

10 up-regulated genes and no down-regulated genes. The algorithm has been able to move the down-regulated genes to other pathways, of which these genes are also members, giving us a stronger indication that the Pyruvate metabolism pathway is up-regulated. Similarly, in the case of the Pentose phosphate pathway, the configuration on Figure 1b suggests that the pathway is up-regulated. For the unique down-regulated gene in the beta-alanine pathway, our method implies that the observed down-regulation is due to decreased activity of another pathway, in which this gene participates, rather than the beta-alanine pathway itself. Overall, for the above 7 pathways our method was able to produce a convenient allocation of genes, efficiently removing contradictions from the final results. Each pathway is now filled with genes of similar expression, which we consider to be the most indicative of the pathway's state.

To further examine our methodology we have applied it to *Escherichia coli* K-12 data from experiment GSM513 discussed in Section 3 (data not shown). As already mentioned, Escherichia coli cells were grown in tryptophan enriched medium, leading to increased activity of the tryptophan metabolism pathway. Our method was able to remove down-regulated genes from the latter pathway and ascribe their behaviour to the activity of other amino acid degradation pathways. As discussed in the Section 3, this is biologically meaningful, given that the cell is presented with excess tryptophan to partly cover its nutritional needs. In both cases discussed here, our method produces results that are consistent with the findings of publications accompanying the data, while reducing the number of genes per pathway contradicting their expression and making it easier for a biologist to infer the state of each pathway.

5.3 Convergence

The solid line on Figure 2 represents the convergence, starting from full membership, the dashed line starting from single membership and the dotted line starting from directed membership allocation. The lines represent the average performance for 20 runs of the algorithm. Evidently, the full-membership allocation shows faster convergence while the directed membership allocation is slower but performs better in terms of the final fitness.

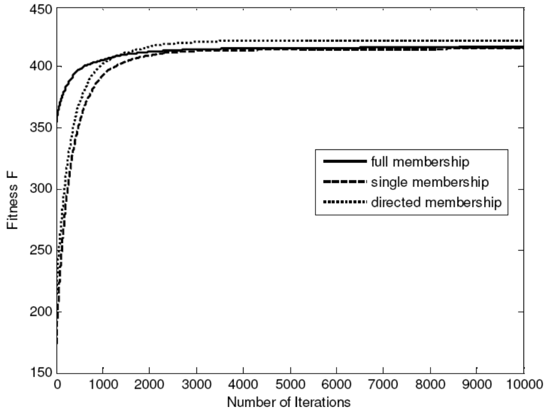


Figure 2. Convergence. Each line shows the mean fitness for 20 runs of the algorithm on data from GSM513.

5.4 Consistency of Results

We have run the algorithm 20 times, on GSM513 data, starting from full, single and directed gene allocation, to examine the consistency of the produced allocations. We perform *(n-1)n/2* comparisons, hence for *n* = 20 runs, we perform 190 comparisons. Figures on Table 2 reveal that the algorithm produces sufficiently consistent results. Especially in the case of directed membership, for 5000 iterations, the two most distinct configurations produced by our method are still 94% similar.

Table 2. Comparison of results produced by 20 separate runs of the algorithm for each separate starting point.

Membership		Full	Single	Directed
1000 Iterations	Max	97.4	92.6	94.9
	Min	86.7	72.8	88.1
	Mean	92.6	83.6	91.1
5000 Iterations	Max	98.2	93.8	99.9
	Min	87.82	78.5	94.0
	Mean	93.4	86.3	96.9
10000 Iterations	Max	97.6	97.7	99.7
	Min	88.7	86.9	92.8
	Mean	93.5	92.7	96.8

5.5 Comparison of Allocations

In pathway based microarray analysis, to validate data quality it is common practice to estimate the probability per pathway of obtaining the results in hand by chance. We have applied this approach to compare the results produced by our directed membership method to the standard full membership allocation. We used a microarray dataset, consisting of experiments from GEO, platform 17, with sufficient number of expressed genes. We have chosen the NBH statistic described in [Swift *et al.*, 2004], which given the overall number of genes and the overall number of affected genes on the array, reveals the probability of obtaining the observed number of affected genes in a pathway of a certain size, purely by chance. Figure 3 shows the mean probability of obtaining the results in hand, per pathway. Evidently, while there is no substantial change in probability, between our and the full membership allocation, our methodology adds an intuitional, biologically meaningful step to the data processing course.

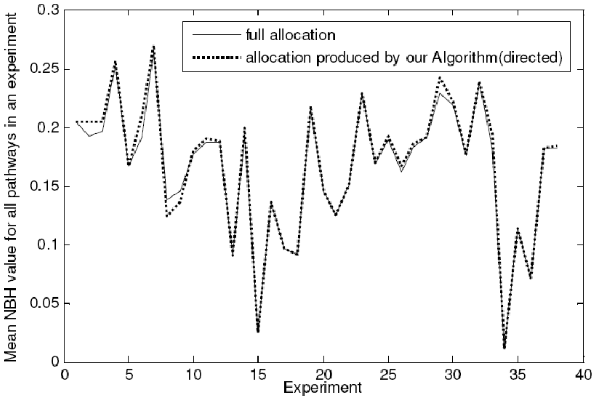


Figure 3. Mean NBH values per experiment, for standard allocation and the one produced by our methodology.

6 Conclusions

We have shown that our algorithm can effectively assign multi-membership genes to their constituent pathways, increasing the level of agreement, in terms of the direction of expression per pathway. This methodology is of potential interest, as it may assist a Biologist to infer the state of individual biochemical pathways, based on microarray data. Given that the multi-membership pathway nature of genes has not been extensively considered in currently used tools for pathway based microarray analysis, we believe that our method suggests an interesting direction for future work.

We acknowledge that for some genes, e.g. repressors, it is expected to observe change in expression that contradicts the up- or down-regulated state of a pathway. We believe that our method can still perform sufficiently well, since this is confined to individual cases. Nevertheless, we plan to tackle this issue by improving our fitness function and facilitating a more detailed pathway categorisation, using the Reactome database.

Furthermore, we intend to develop our search method, using simulated annealing and a genetic algorithm, and compare these approaches, to investigate their performance.

Additionally, we plan to apply our method to more complex organisms, including human data. While *Escherichia coli* and *Saccharomyces cerevisiae* are relatively simple living forms, we speculate that our methodology may produce more dramatic results and have more beneficial impact in cases of organisms with larger, more sophisticated genomes and complicated biochemical networks.

We also wish to investigate the possibility of devising a method to predict the approximate absolute values of gene expression per pathway, that give rise to the net effect values obtained by microarray analysis, discussed in this paper.

Finally, we aim to use the proposed approach on a large dataset, consisting of more than a thousand microarray experiments to explore potential association rules between biochemical pathways and elucidate pathway regulation and interaction.

Acknowledgments

This research was supported by the studentship scheme of the School of Information Systems, Computing and Mathematics, Brunel University.

References

[Cavalieri *et al*., 2007] D Cavalieri, C Castagnini, S Toti, K Maciag, T Kelder, L Gambineri, S Angioli, P Dolara: Eu.Gene Analyzer a tool for integrating gene expression data with pathway databases. *Bioinformatics* 2007, 23(19): 2631-2632

[Dahlquist *et al*.,2002] KD Dahlquist, N Salomonis, K Vranizan, SC Lawlor, BR Conklin: GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways. *Nat Genet* 2002, 31(1): 19–20

[DeRisi *et al*., 1997] JL DeRisi, VR Iyer, PO Brown: Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* 1997, 278: 680–686

[Goesmann *et al*., 2002] A Goesmann, M Haubrock, F Meyer, J Kalinowski, R Giegerich: PathFinder: reconstruction and dynamic visualization of metabolic pathways. *Bioinformatics* 2002, 18: 124–129

[Hamming *et al*., 1950] R Hamming: Error Detecting and Error Correcting Codes. *Bell System Technical Journal* 1950, 26(2): 147-160

[Joshi-Tope *et al*., 2005] G Joshi-Tope, M Gillespie, I Vastrik, P D'Eustachio, E Schmidt, D de Bono, B Jassal, GR Gopinath, GR Wu, L Matthews, S Lewis, E Birney, L Stein: Reactome: a knowledgebase of biological pathways. *Nucl Acid Res* 2005, 33: Database issue D428-D432

[Kanehisa *et al*., 2000] M Kanehisa, S Goto: KEGG: Kyoto encyclopaedia of genes and genomes. *Nucl Acid Res* 2000, 28: 27–30

[Khodursky *et al*., 2000] AB Khodursky, BJ Peter, NR Cozzarelli, D Botstein, PO Brown, C Yanofsky: DNA microarray analysis of gene expression in response to physiological and genetic changes that affect tryptophan metabolism in *Escherichia coli*; *PNAS USA* 2000, 97: 12170–12175

[Kolpakov *et al*., 1998] FA Kolpakov, EA Ananko, GB Kolesov, NA Kolchanov: GeneNet: a gene network database and its automated visualization. *Bioinformatics* 1998, 14: 529–537

[Michalewicz *et al*., 1998] Z Michalewicz, DB Fogel: How to solve it: Modern heuristics. Berlin: Springer; 1998

[Panteris *et al*., 2007] E Panteris, S Swift, A Payne, X Liu: Mining pathway signatures from microarray data and relevant biological knowledge. *Journal of Biomedical Informatics* 2007, 40(6): 698-706

[Quadroni *et al*., 1999] M Quadroni, P James: Proteomics and automation. *Electrophoresis* 1999, 20: 664-677

[Rojas *et al*., 2007] I Rojas, M Golebiewski, R Kania, O Krebs, S Mir, A Weidemann , U Wittig: SABIO-RK: a database for biochemical reactions and their kinetics. *BMC Systems Biology* 2007, 1(Suppl 1): S6

[Shannon *et al*., 2003] P Shannon, A Markiel, O Ozier, NS Baliga, JT Wang, D Ramage, N Amin, B Schwikowski, T Ideker: Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research* 2003, 13(11): 2498-2504

[Stryer *et al*., 2002] L Stryer, MJ Berg, LJ Tymoczko: Biochemistry. 5th ed. W.H. 612 Freeman; 2002.

[Swift *et al*., 2004] S Swift, A Tucker, V Vinciotti, N Martin, C Orengo, X Liu, P Kellam: Consensus clustering and functional interpretation of gene-expression data. *Genome Biology* 2004, 5: R49

Temporal Data Mining for the Analysis of Administrative Healthcare Data

Stefano Concaro[1,2], Lucia Sacchi[1], Carlo Cerra[2], Pietro Fratino[2], Riccardo Bellazzi[1]
[1]Laboratory for BioMedical Informatics, University of Pavia, Pavia, Italy, stefano.concaro@unipv.it
[2] PAC Department, ASL, Pavia, Italy

Abstract

Over the last few years there has been an increasing attention in the collection of data from healthcare institutions. Given that the healthcare delivery process evolves over time, capturing the temporal relationships between interesting events is a crucial point for forecasting and planning operational policies and interventions. In this paper we will present a new algorithm oriented to the mining of Temporal Association Rules which has the main innovative feature of handling both events with a temporal duration and events represented by single time points. The new method will be applied to the analysis of healthcare administrative data of diabetic patients, as recorded in the Pavia local healthcare agency information repository. The analyses performed on these data show the usefulness and potentialities of the proposed framework in Public Health.

1 Introduction

Over the last few years there has been an increasing attention in the collection of data from healthcare institutions such as hospitals and local healthcare agencies (ASL in Italy). Besides clinical data, also administrative information is recorded and used to monitor the evolution and functioning of the healthcare processes. Data of this kind are usually referred to as “process data”. One of the essential components of process data is time. Given that the healthcare delivery process evolves over time, capturing the temporal relationships between interesting events is a crucial issue for forecasting and planning operational policies and interventions.

A descriptive analysis about the processes can be very useful to show how they are working or evolving. In this scenario, the application of Data Mining (DM) techniques can be particularly suitable for the extraction of meaningful information and knowledge from large healthcare databases. However, traditional DM techniques are characterized by a ”static” approach, which is not able to handle the temporal dimension in an explicit way. The exploitation of Temporal Data Mining (TDM) methods able to deal with time and temporal events could be particularly fruitful in this field [Post and Harrison, 2008; Roddick and Spiliopoulou, 2002; Shahar, 1997].

Recently the Local Healthcare Agency (ASL) of Pavia started a research activity aimed at mining its large data warehouse, which includes the administrative records concerning all the main healthcare admissions to the public health system of the population of Pavia area (about 520000 citizens) since 2002. All the records have a temporal structure, that describe the clinical history of the citizens over the last 6 years. From a TDM viewpoint, the main feature of this database is the heterogeneity of the temporal data it contains. It in fact includes time-point like events, as ambulatory visits, short time interval events, as hospital admissions, and longstanding events, like drug delivery associated to pharmacological prescriptions. Up to now, TDM methods have been designed to deal with homogeneous temporal events, namely single time points [Agrawal and Srikant, 1995; Ayres et al., 2002] or intervals [Höppner and Klawonn, 2002; Sacchi et al., 2007]. Considering the different temporal nature of the data that characterize this and related studies, we developed a new algorithm able to handle both events with a time duration (intervals) and events with no duration (time points) at the same time. In this paper we will show the application of the new method to the analysis of the data of the ASL patients suffering from Diabetes, which is one of the chronic diseases with the wider social impact.

2 Methods

In this section we describe a new algorithm devoted to the extraction of Temporal Association Rules on temporal sequences of events. A sequence of events can be defined as a time ordered succession of episodes. Each sequence is related to a subject (for example a specific patient) and to a variable (for example a prescription of a specific drug) and each episode is characterized by its temporal duration within an observation period.

The main new feature of our algorithm stands in its capability to handle heterogeneous temporal events; in particular, considering the time dimension, we divided the represented events as either with a duration (identified by a time interval) or without a duration (identified by a time point). Of course, the latter are instantaneous events when we consider the temporal granularity of the former. For example an event with a duration can be an hospitalization, which may last for few days, while a time-point event could be a blood test, which lasts less then a minute.

In this framework, we will represent each temporal episode *e* through a tuple with five fields:

1. the *start time: e.start*
2. the *end time: e.end*
3. the *subject*
4. the *variable*
5. the variable *group.*

Fields 1 and 2 are temporal attributes of an observation period; they may be expressed in a suitable time unit corresponding to a certain temporal granularity (seconds, hours, days, etc…). In particular, for a temporal episode *e* characterized by a duration *e.start<e.end*, while for an episode without a duration *e.start≡e.end.* Fields 3 and 4 together constitute the distinctive identifier of any temporal sequence: the subject can be for example a specific patient, while the variable can be the assumption of a certain drug. Field 5 defines the class to which the variables may belong, such as hospital admissions or ambulatory visits. This field is used to introduce a taxonomy of the variables, which can be exploited by the algorithm to properly constrain the search only to a set of desired rules. The goal of the algorithm is to find Temporal Association Rules (TARs). A TAR is a relationship defined through a temporal operator which holds between an antecedent, consisting in one or more patterns, and a consequent, consisting in a single pattern. In this case a pattern is defined as the presence of a certain variable. An example of TAR could be: “Cholesterol and Triglycerides tests are performed before a statin prescription”. In this example the variables *cholesterol* and *triglycerides* constitute the antecedent while the variable *statin* is the consequent. The relationships between antecedent and consequent were defined through 7 temporal operators; 6 of them are derived from Allen’s algebra [Allen, 1984] (BEFORE, MEETS, OVERLAPS, FINISHED-BY, EQUALS, STARTS), while the last one is the more general PRECEDES operator [Bellazzi et al.,2005]. Given two episodes *e1* and *e2*, the PRECEDES relationship holds (*e1 PRECEDES e2*) if *e1.start* $\leq$ *e2.start* and *e1.end* $\leq$ *e2.end.* By definition the PRECEDES operator synthesizes all the six mentioned Allen’s operators. Thanks to the definition of three design parameters it is anyway possible to select only a subset of desired relationships. These parameters are [Bellazzi et al.,2005]: the *left shift* (LS), defined as the maximum allowed distance between *e1.start* and *e2.start*, the *gap* (G), defined as the maximum allowed distance between *e1.end* and *e2.start*, and the *right shift* (RS), similarly defined as the maximum allowed distance between *e1.end* and *e2.end.* As an effect of this parameterization, the search of the rules can be properly constrained, because the number of episodes involved in a relationship can be conveniently reduced to those sufficiently close in time (see [Sacchi et al.,2007] for a detailed discussion). All the operators were implemented in order to handle both events with a duration and events without a duration at the same time (except for OVERLAPS which is defined only for temporal intervals) [Vilain, 1982]. If one or both the events involved in a TAR are represented by a single time point, the number of possible mutual positions between the events decreases; in this case the operator PRECEDES synthesizes a reduced subset of temporal relationships. Table 1 and 2 show the definition of the temporal operators in the case of an interval and a time point and of two time points, respectively.

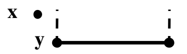
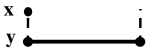
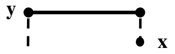
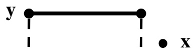
Mutual position	Description
	x *BEFORE* y
	x *MEETS* y x *STARTS* y
	y *FINISHED-BY* x y *MEETS* x
	y *BEFORE* x

Table 1. Mutual position and temporal operator definition when one episode is represented by an interval (y) and the other by a time point (x)

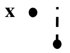
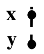
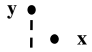
Mutual position	Description
	x *BEFORE* y
	x *EQUALS* y x *MEETS* y
	y *BEFORE* x

Table 2. Mutual position and temporal operator definition when both episodes (x,y) are represented by time points

The method for TAR extraction is based on an Apriori-like strategy, and it is made up of three steps:

1. iterative selection of a variable as consequent of the rule;
2. extraction of the *basic-set* of rules, that is the whole set of rules with single cardinality in the antecedent;
3. extraction of *complex rules*, defined as rules with antecedent of multiple cardinality K obtained through the intersection of the episodes of the antecedents of the rules of cardinality K-1.

In addition to the implementation presented in [Sacchi et al.,2007], the algorithm offers the optional opportunity to select specific target rule types, defining the classes of the variables allowed for the antecedent and the consequent selection respectively. This feature helps to focus the search only on those kind of relationships between classes which the user wants to investigate.

As happens in the Apriori algorithm [Agrawal and Srikant, 1994], the notions of confidence and support play an essential role in the definition of frequent patterns. The original concepts were adapted to the temporal domain and properly extended to handle events of different nature. In particular, the *support* of a rule is defined as the proportion of subjects for which the rule is verified over the total number of subjects involved in the study:

$$support = \frac{size(SR)}{size(S)}$$

The set *SR* is defined as the union between the sets *SRS* and *SRF*, where *SRS* is the set of subjects whose rule episodes meet a minimum *temporal span* threshold, while *SRF* is the set of subjects whose rule episodes meet a minimum *frequency* threshold. Thanks to this definition, episodes supporting a rule can be either episodes with low frequency but long lasting in time, or short episodes (as for events without a duration) but with high frequency. An example of how these two thresholds are used is shown in Figure 1.

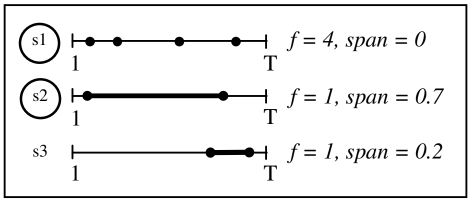


Figure 1. This example shows the hypothetic rule episodes for three subjects s1, s2 and s3. If we set *f_th=2* and *span_th=0.5*, then the rule is supported by s1 and s2 (*support=2/3=0.66*).

The *confidence* is defined as the ratio between the number of episodes of the antecedent involved in the rule (*NAR*) and the total number of episodes of the antecedent (*NA*) during the whole observation period:

$$confidence = \frac{NAR}{NA}$$

According to this definition, the confidence represents the probability to find one episode of the consequent satisfying the temporal relationship given that one episode of the antecedent occurs.

3 Results and Discussion

In this section we show the application of the presented method to the analysis of public health data. The regional healthcare agency (ASL) of Pavia is collecting a large amount of data concerning all the main healthcare expenditures, including hospital admissions, pharmacological prescriptions and ambulatory visits of the population of Pavia area (520000 people).

In particular we will show the results obtained on a sample of diabetic patients. The population involved in this study includes all the patients who had been diagnosed during 2005. For each patient we considered all the healthcare events included in a period of two years after the date of the diagnosis. The reason why we chose this event as the initial point of the period is that it usually marks a change in the treatment of the disease. According to the structure of the presented temporal data mining framework, we can define the temporal episodes as follows:

- *subjects*: diabetic patients
- variable *groups*: hospital admissions (HA), ambulatory visits (AV), drugs (Dr)
- *variables*:
  - HA group: DRG {HA.DRG}, diagnoses {HA.D} and procedures {HA.PR} (ICD9-CM code);
  - AV group: ambulatory visits (Italian National code, DGR n. VIII/5743 - 31/10/2007);
  - Dr group: drug prescriptions (ATC code up to the 3rd level).
- *start/end time*: the temporal location of an event within the 2-years observation period, considering a granularity of one day. Typically,ambulatory visits and day hospital admissions will have *e.start≡e.end*. In the case of ordinary hospital admissions and drug prescriptions the episodes will range over a time interval. The duration of drug assumption has been estimated by the number of days of prescribed *Defined Daily Dose* (DDD).

The selected population includes 2513 diabetic patients, with a total number of 3107 variables. Overall we detected a total number of 263224 healthcare episodes, distributed as shown in Table 3.

Variable Group	N variables	N episodes
DRG (HA.DRG)	291	2187
Diagnosis (HA.D)	1200	6371
Procedure (HA.PR)	660	6868
Ambulatory Visit (AV)	838	183331
Drug (Dr)	118	64467
Total	3107	263224

Table 3. Episodes distribution for the diabetic sample.

In the next two sections we will present two different applications of the proposed method to the described data set. The first set of analyses (Section 3.1) is aimed at finding temporal association rules specific for the diabetic population with respect to a control group. In the second part (Section 3.2) we will instead present an example of how the method can be used as a tool to check the compliance of the assistance path to medical guidelines.

3.1 Analysis of diabetes-specific temporal association rules

A well-known problem of the application of data mining methods to large databases is the management of the great amount of output information. This is true in our case too, especially considering that several ambulatory visits can be prescribed simultaneously (e.g. Cholesterol test together with Triglycerides). The plain application of the algorithm to the diabetic dataset would generate a great amount of TARs, the most part of which would consist in redundant, trivial or already known information.

In order to reduce the number of extracted rules, we focused our attention only on population-specific rules, i.e. the rules related to the diabetic condition or to its concomitant pathologies. To this aim, we considered a “control” sample of non-diabetic patients, paired with the diabetic sample for the variables age, sex and observation period, to generate a “null” benchmark distribution of temporal rules. TARs extracted on the diabetic population undergo to a post-processing step aimed at pruning the rules that are similar in the two populations. The resulting TARs for the diabetic group are thus only the rules which are not present in the control group or those presenting a statistical difference in the support or confidence indexes. The significance of the difference of the indexes is checked through a chi-square test (*p-value=0.005*).

We performed several analyses, each of them characterized by a different target rule type, in order to investigate all the relationships among the four variable groups. The settings of the analyses which generated the most meaningful results are reported in Table 4; AID is the analysis identifier.

For the analyses on the diabetic sample some of the design parameters were kept constant for all the configurations: *f_th=1*, *span_th=1*, *LS=RS=731 days*, *minconf=0,3*. The chosen temporal operator was "BEFORE". The remaining parameters, i.e. the minimum support (*minsupp*) and the *gap*, were instead conveniently tuned for each case, according to the domain knowledge on the healthcare processes underlying the different target rule type.

Table 5 shows a representative subset of TARs obtained from the five considered analysis configurations. These rules were selected by a ASL clinician within the total set of rules extracted by the algorithm (Table 4). All the rules are characterized by a significant statistical difference between the two samples (diabetic vs non-diabetic), being more frequent in the diabetic one. For each rule we report the rule identifier (RID), the analysis configuration identifier (AID), the antecedent, the consequent and the values for confidence and support in the diabetic sample (D). Besides the information herein reported, the user has also the possibility to know the average duration of the antecedents, of the gap and of the consequents.

The first analysis configuration (AID=1) is related to hospital admissions. Rules 1, 2 and 3 show that hospitalizations are most frequently related to cardiovascular diseases in the diabetic population, as shown by the statistical differences for both support and confidence.

The second analysis configuration investigates the relationships between hospitalizations and ambulatory visits. Rules 4 and 6 highlight the close correlation of diabetes with an associated condition, respectively kidney and liver disorders. Rule 5 offers the opportunity to explicitly compare the role of diabetes related hypertension with respect to the risk of kidney disorders. The rule is more frequent in the diabetic population (higher support value), and the correlation between hypertension and risk of kidney disorders is stronger (higher confidence value).

AID	Rule type	Parameters_D	Parameters_{ND}	N rule

1	A={HA}
	C={HA}
		Minsupp=0,02
		Gap=365
			Minsupp=0,01
			Minconf=0,16
				23

2	A={HA}
	C={AV}
		Minsupp=0,04
		Gap=90
			Minsupp=0,025
			Minconf=0,2
				44

3	A={HA}
	C={Dr}
		Minsupp=0,02
		Gap=60
			Minsupp=0,01
			Minconf=0,16
				35

4	A={AV}
	C={Dr}
		Minsupp=0,02
		Gap=30
			Minsupp=0,01
			Minconf=0,16
				4

5	A={Dr}
	C={Dr}
		Minsupp=0,02
		Gap=60
			Minsupp=0,01
			Minconf=0,16
				56

Table 4. Each analysis configuration (identified by AID) is characterized by a different rule type and a particular selection of the parameters for the two samples (D: Diabetic, ND: Non-Diabetic). Herein A stands for Antecedent, C for Consequent; the acronyms in the second column refer to the variable groups.

RID	AID	Antecedent	Consequent	Conf_D	Sup_D
1	1	HA.PR: Electrographic monitoring	HA.PR: Diagnostic ultrasound of heart	0,605	0,021
2	1	HA.PR: Diagnostic ultrasound of heart	HA.D: Diabetes II uncomplicated	0,375	0,050
3	1	HA.D: Essential benign hypertension	HA.D: Essential benign hypertension	0,311	0,043
4	2	HA.PR: Diagnostic ultrasound of abdomen and retroperitoneum	AV: Creatinine test	0,341	0,045
5	2	HA.D: Essential benign hypertension	AV: Creatinine test	0,322	0,065
6	2	HA.PR: Diagnostic ultrasound of abdomen and retroperitoneum	AV: Alanine transaminase (ALT-GPT)	0,306	0,041
7	3	HA.DRG: Diabetes over 35	Dr: Oral hypoglycaemic agents	0,556	0,031
8	3	HA.PR: Diagnostic ultrasound of heart	Dr: Beta blocking agents	0,465	0,060
9	3	HA.PR: Computerized axial tomography of head	Dr: Antithrombotic agents	0,448	0,036
10	3	HA.PR: Diagnostic ultrasound of abdomen and retroperitoneum	Dr: Drugs for peptic ulcer and gastro-oesophageal reflux disease	0,315	0,040
11	4	{AV: Kirby-Bauer antibiotic testing; AV: Urine culture test}	Dr: Quinolone antibacterials	0,425	0,027
12	5	Dr: Beta blocking agents	Dr: Beta blocking agents	0,744	0,214
13	5	Dr: Antiglaucoma preparations and miotics	Dr: Antiglaucoma preparations and miotics	0,671	0,057
14	5	Dr: Thyroid preparations	Dr: Thyroid preparations	0,607	0,037
15	5	Dr: Diuretics and potassium-sparing agents in combination	Dr: Diuretics and potassium-sparing agents in combination	0,601	0,064

Table 5. A representative set of TARs defined by the "BEFORE" operator. Herein RID is a rule identifier, while AID is the analysis identifier (Table4).

The third analysis is focused on drug prescription performed just after an hospitalization. Rule 7 reveals that the probability to be treated with oral hypoglycemic just after an hospitalization due to diabetes corresponds to 55%. Rules 8 and 9 are related to cardio-cerebral-vascular disorders, probably identifying patients who underwent an heart failure or an ictus respectively. Rule 10 shows that peptic ulcer is more frequent in the population when associated to diabetes.

The fourth configuration evaluates the relationship between drug prescriptions and ambulatory visits. An example of these relationships is given by rule 11 that highlights another specific issue of the diabetic population, that is the increased frequency of bacterial infections compared with the non-diabetic sample.

The fifth configuration is devoted to the analysis of the relationships between drug prescriptions. The majority of the extracted rules (12, 13, 14, 15) identify prescriptions which are repeated according to a cyclic pattern during the observation period; this is due to the fact that diabetic patients often suffer from chronic pathologies. Moreover, such rules give an hint on specific concomitant diseases, as hypertension, cardiovascular, ocular, thyroid and stomach disorders.

3.2 Analysis of the Compliance to Guidelines Recommendations

This section shows how the method can be suitably used to check the compliance of the extracted patterns to the guidelines that GPs and primary care are supposed to follow, with particular attention to the temporal dimension. We considered the prescription of the HbA1c test (glycated hemoglobin) as an example of the application of the method. According to the ASL recommendations, the HbA1c test should be prescribed every three months (~90 days) for diabetic patients. Considering the observation period of two years after diabetes diagnosis in the sample, the expected optimal frequency would correspond to 8 prescriptions.

The compliance to this recommendation was investigated by studying the support index in the target rule "*HbA1c BEFORE HbA1c*", which is supposed to show a cyclic behavior over the observation period. For this analysis the parameters were set as follows: *gap=135 days*, *span_th=1*, *LS=RS=731 days*, *minconf=0.1, minsupp=0.02*. The value for the gap was chosen considering the recommended time within two tests (90 days) increased by the 50% to allow more flexible prescription intervals. The chosen temporal operator was "BEFORE". The analyses were in this case performed by varying the frequency threshold parameter (*f_th*). The compliance to the recommendation is then expressed by the behavior of the support index as a function of the frequency threshold (Fig. 2). Since one episode of the rule includes two episodes of the HbA1c test (one in the antecedent and one in the consequent), if we consider a cyclic chain of HbA1c episodes, the number of correct prescriptions of the test corresponds to the value *cp=f_th+1*. We evaluated the performance of the algorithm using 5 different values for *f_th* going from 1 (at least one rule during the observation period) to 7 (the expected frequency according to the recommendation).

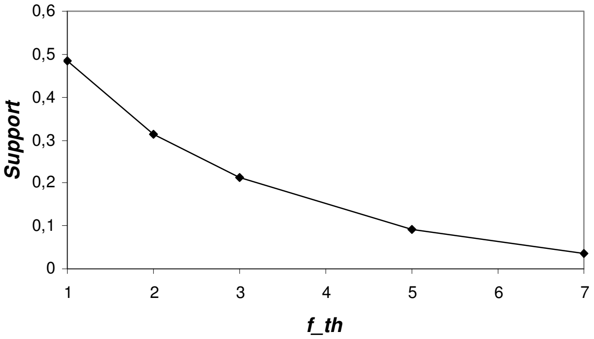


Figure 2. Trend of the support index of the rule "*HbA1c BEFORE HbA1c*" as a function of the frequency threshold. The number of correct prescriptions of the test correspond to the value *cp=f_th+1*.

A complete adherence to the recommendation is reached only for the 3.4% of the diabetic sample, obtained for *f_th=7*. Moreover, the support value reaches only the 48.5% for the minimum value of frequency. This means that only the 48.5% of the diabetic patients undergo to at least two HbA1c tests in a period of 135 days. These results give an hint on the low compliance of the physicians' practice in treating the diabetic patients with respect to the ASL recommendations.

4 Conclusions

In this paper we presented a new algorithm oriented to the mining of Temporal Association Rules on temporal sequences of events. The main innovative feature of the algorithm is its ability to handle both events with a temporal duration and events represented by single time points. We applied the new method to the healthcare administrative data of diabetic patients, as recorded in the Pavia ASL information repository, which mainly includes hospital admissions, pharmacological prescriptions and ambulatory visits. The analyses performed on this data set show how the proposed framework could be used as a general instrument in Public Health. It in fact represents a useful method to observe frequent health care temporal patterns in a population, with the opportunity to monitor how the healthcare processes are working and to evaluate the compliance of the processes with respect to the recommended medical guidelines.

As a future direction of the work, the main challenge will be the integration of prior knowledge into the temporal association rule extraction process. This could be done for example by exploiting the intrinsic hierarchical structure which underlies the encoding system for DRG, diagnoses and procedures, drug prescriptions, etc., or by developing an online strategy to check the compliance of the patterns with respect to formalized reference guidelines [Raj et al., 2007]. This method would allow to define an interestingness score able to quantify the adherence of the rule to the formalized knowledge. According to this score we may filter out the trivial or already known patterns, and consequently perform a deeper analysis (lower support threshold) in order to search for highly interesting patterns with low frequency in the population or even to discover new knowledge which can be used for decision support or to define sub-populations identified by particular diagnostic or therapeutic paths.

Finally, we are currently planning the implementation of a software tool for the local healthcare agency of Pavia. Such tool will have to be characterized by an high flexibility to allow users to mine TARs over the wide set of application fields which concern ASL management (e.g. hypertension, cardiovascular risk, adverse drug reactions).

References

[Agrawal and Srikant, 1994] Rakesh Agrawal, Ramakrishnan Srikant. Fast algorithms for mining association rules in large databases. In *Proceedings of the International Conference on Very Large Databases*. Morgan Kaufmann, pp.478-499, 1994.

[Agrawal and Srikant, 1995] Rakesh Agrawal, Ramakrishnan Srikant. Mining Sequential Patterns. *In Proceedings of the 11th Intl Conf on Data Engineering*, 3-14, 1995.

[Allen, 1984] James F. Allen. Towards a general theory of action and time. *Artificial Intelligence*, 23:123-154, 1984.

[Ayres et al., 2002] Jay Ayres, Johannes Gehrke, Tomi Yiu, Jason Flannick. Sequential Pattern Mining Using a Bitmap Representation. In *Proceedings of the 8th ACM SIGKDD Intl Conf on Knowledge Discovery and Data Mining*, 429-435, 2002.

[Bellazzi et al., 2005] Riccardo Bellazzi, Cristiana Larizza, Paolo Magni, and Roberto Bellazzi. Temporal data mining for the quality assessment of hemodialysis services. *Artificial Intelligence in Medicine*, 34(1):25-39, 2005.

[Höppner and Klawonn, 2002] Frank Höppner, Frank Klawonn. Finding Informative Rules in Interval Sequences. *Intelligent Data Analysis - An International Journal*, 6(3):237-256, 2002.

[Post and Harrison, 2008] Andrew R Post, James H Harrison. Temporal Data Mining. *Clinics in Laboratory Medicine*, 28:83-100, 2008.

[Raj et al., 2007] Rashmi Raj, Martin J. O'Connor, Amar Das. An Ontology-Driven Method for Hierarchical Mining of Temporal Patterns: Application to HIV Drug Resistance Research. In *Proceedings of AMIA 2007 Annual Symposium*, 2007.

[Roddick and Spiliopoulou, 2002] John F Roddick, Myra Spiliopoulou. A survey of temporal knowledge discovery paradigms and methods. *IEEE Transactions on Knowledge and Data Engineering*, 14(4): 750-767,2002.

[Sacchi et al., 2007] Lucia Sacchi, Cristiana Larizza, Carlo Combi, Riccardo Bellazzi. Data mining with Temporal Abstractions: learning rules from time series. *Data Mining and Knowledge Discovery*. 15(2): 217-247, 2007.

[Shahar, 1997] Yuval Shahar. A framework for knowledge-based temporal abstraction. *Artificial Intelligence*, 90:79-133, 1997.

[Vilain, 1982] Marc B. Vilain. A system for reasoning about time. In *Proceedings of AAAI-82*, 197-201,1982.

Data-Driven Hybrid Method for Synthetic Electronic Medical Records Generation

Anna L. Buczak, PhD[1], Linda J. Moniz, PhD[1], John Copeland, MS[2] , Henry Rolka RN, MPH,MS[2], Joseph Lombardo, MS[1], Steven Babin, MD, PhD[1], Brian Feighner, MD, MPH[1]

[1]JHU/APL, Laurel, MD; [2]Centers for Disease Control and Prevention, Atlanta, GA
Anna.Buczak@jhuapl.edu

Abstract

In this research we develop a disease injection scenario and produce fictional disease victims. We use data mining techniques to analyze electronic medical records to extract the care models of existing patients and to find the best match to the health states of the victims. This method uncovers the projection of disease dynamics onto the medical records. We use this reconstructed projection to synthesize electronic medical records for the disease injection victims.

1 Introduction

Recent advances in the use and utility of disease surveillance systems require equally sophisticated data sets with which to test the systems. Luckily, there have been few significant outbreaks of disease in the digital age. However, this means that realistic testing is at a particular disadvantage. The lack of controllable, realistic data for testing of disease surveillance on electronic medical records significantly hinders the ability to evaluate these systems.

1.1 Dynamics and Projection

A population's health states are a complicated equation that is at best finite-dimensional and at worst the infinite dimensional solution of a complicated partial differential equation. At any time we do not know the true health states of the population. The projection from health state to electronic medical records includes noise that arises from many factors, including but not limited to the health-care seeking behavior of the population, the mapping from health state to chief complaint or ICD-9 code, differences in care and treatment protocols from one provider to the next, and differences in language and its interpretation. Another source of noise is clerical and omission errors during the coding of the records. However, with sufficient data the reconstruction of the dynamics of illness from the electronic medical records can give a picture of the true health states.

1.2 Current Models

Historical healthcare data often has not been (or could not have been) analyzed for a ground truth determination of "normal" background disease levels and epidemics or outbreaks. However, testing for sensitivity and specificity of biosurveillance algorithms requires specific knowledge of the duration and extent of an outbreak. For this reason, fictitious outbreak data is often injected into historical data for testing purposes. Several models currently exist for injection of cases. The naïve model produces numbers of cases per day according to a lognormal epidemic curve, and disperses the cases according to a spatial model (e.g., plume, point-source, etc.) [Barrett *et al.,* 2005, Meltzer *et al.,* 2001]. The entire electronic medical record is usually not involved. Some recent models [Daley *et al.,* 1999] incorporate social models for both healthcare seeking behavior and for chief complaint or other measures of injections. These models tend to be very flexible—all that is required for most is an epidemic curve and a spatial dispersion/demographic model—but not very realistic.

The most recent and realistic models employ sophisticated social networking models [Breban *et al.,* 2005, Frasca *et al.,* 2006] with a great deal of detail. Some models include geospatial data as well as detailed social data [MIDAS]. These models are in general tailored to a particular disease injection scenario, and frequently take a long time to run. They are not as flexible as the naïve models. We note that none of these models produce, at this time, entire electronic medical records for the injected victims.

Hybrid models, which inject cases on top of real background disease data, incorporate both methods. They are more realistic than the naïve models, and can include as much detail as desired in the modeling of social and epidemic effects.

2 The Data-Driven Hybrid Approach

The approach we will describe here is a novel method that goes well beyond the hybrid model and uses the background data as a model for injection of the entire electronic medical record. The method splits the problem into three parts—firstly, producing victim identities and health states; secondly, extracting the care models of the existing patients from the background data; and thirdly, projecting those health states into existing background data. A synopsis of the method is given in Figure 1.

The approach has the following steps:

1. The fictitious victims, their symptoms and demographics are generated ("Injected Disease/Victim Generation Model").

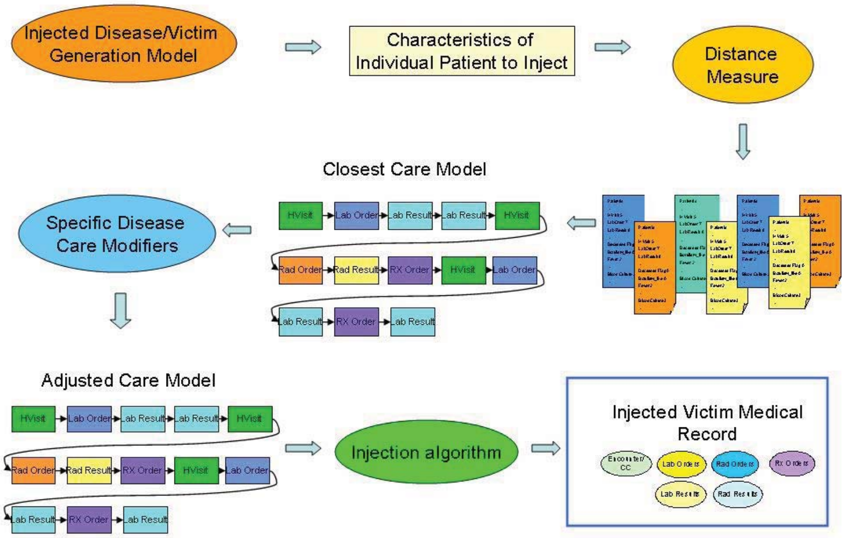


Figure 1. Steps in Creation of Synthetic Patient Medical Records.

2. The characteristics of the generated victims are made into inject descriptors ("Characteristics of Individual Patients to Inject").
3. A distance measure is computed between the inject descriptors and existing patient care descriptors. The few closest descriptors are identified ("Distance Measure").
4. The patient care models corresponding to the Closest Patient Care Descriptors are selected from the background data ("Closest Care Model").
5. The closest care models are altered by the Specific Disease Care Modifiers to reflect the characteristics of the injected disease producing the Adjusted Care Models. ("Specific Disease Modifiers").
6. The injected victims are projected into the existing data base using the Adjusted Care Models ("Injection Algorithm").

The advantage of this approach is that the analysis of the electronic medical records can be done separately from any modeling of victim health states. The result of the analysis is a projection function from the victim health states to an electronic medical record. Any victim generation model, from the simple (for straightforward injection of non-infectious disease or localized toxic agents) to the sophisticated (for infectious disease spread or widespread bioterrorism) can be used to model the victim health states. Here, we use a simple model for generation of victim health states.

2.1 Generation of Injected Victim Health States

The health states of injection victims form one of the inputs to the algorithm that produces the medical records. We apply the disease starting with healthy people. We use information from the literature concerning the possible manifestations of the injected disease to change the victims' health states to values that reflect the disease progression.

Consider, for example an injected illness of weaponized Tularemia to a small population. The symptoms and timeline of the disease, from latent period to initial symptoms to various stages of infection, are tracked for each individual victim, based on initial bacterial dose. We assign symptoms and severity to each injected patient based on values from various literature studies regarding the timeline of the illness [Ellis *et al.,* 2002, Dennis *et al.,* 2001, Evans *et al.,* 1985, Arikan *et al.,* 2003]. Thus, the manifestation of disease from an influenza-like illness in the prodrome to pneumonia and severe infection for later stages are present in the injected victims. Although all victims are exposed at the same time, the number of infecting organisms is varied and the time in which severe illness is manifested is based on a well-studied epidemic curve and related to the number of initial infecting organisms.

Victims are chosen according to a demographic model that matches the exposure or illness scenario. In this weaponized Tularemia example, the exposure is modeled to be at the restrooms in the luxury boxes at a baseball game. The demographic distributions of gender and age for the population segment present at the game are used to "choose" fictitious victims with particular ages, genders and zip codes. If ethnicity or race data are present in the background data, additional census/demographic distribution information is used to assign a race and ethnicity to each victim.

Each individual victim is assigned a set of initial symptoms and the healthcare seeking behavior for initial symptoms is set to a percentage of the affected population. The disease is modeled to progress to all of the victims until symptoms become severe enough that the exposed and infected population seeks care. However, the symptoms and severity of the manifest illness vary. For example, some victims will have fever only, others fever and cough, and some victims will develop difficulty breathing or septicemia and bacteremia. After the health states and symptoms of the injected illness are modeled in the victims, the analysis of the background data provides the projection function from the health states of the victims to the electronic medical records.

2.2 De-identification of Background Records

In order to discover disease patterns in the electronic medical records' background data without compromising patient confidentiality or hospital demographic information, the data first must be rendered anonymous and indiscernible by facility. Any patient-identifying factors such as zip code and name must be modified (e.g., zip3's used in lieu of zip codes) or removed (name or personal information). In addition, a date shuffle, for example, the RBNR algorithm [Burr *et al.,* 2005] is performed so patient identification is not possible through symptoms and timeline. Any records for rare, familial, or hereditary disorders must be removed from the data set; those records could possibly be traced to a person. Since injection scenarios focus on illnesses from exposure to toxins or to pathogens, the rare or hereditary disease information is not necessary for analysis.

In order to protect the identity and propriety information for a facility, the facility identifiers are removed or changed to random codes. We note that with the data shuffle, any records associated with a visit are shuffled with the same time period so that the timeline information for a visit is intact. The result of the shuffling and facility removal is a set of data with reassigned identification fields but with intact medical record information.

3 Identification of Patient Care Models

The data mining approach that we developed for identifying Patient Care Models that exist in the data has the following major steps (Figure 2):

1. Identification of a subset of patients of interest (the "patients of interest" have a diagnosis or some symptoms of the injected disease).
2. Building sequences of care events for each of these patients, i.e., individual patient care models (Figure 1).
3. Building descriptors of the individual patient care models**.**

The whole data set that we are dealing with has over 400,000 patients. From this data set we extract the records concerning patients that have symptoms similar to those that patients with any stage of Tularemia could have. Typically, these patients have fever and some combination of cough, shortness of breath, painful respiration, malaise / fatigue, sometimes diarrhea or nausea and vomiting; in later stages they could have hemoptysis, respiratory failure, and enlarged lymph nodes. In step 1, a query is performed on the whole data set to retrieve patients who have combinations of the symptoms described above. This query identified a subset of over 10,000 patients. The rest of the operations are performed on that subset.

The second step consists of extracting for each of the 10,000 patients care models from the different data tables. A care model (bottom of Figure 2) is a time sequence of the analysis visits (*HVisit*) that the patient had, laboratory orders (*Lab Order*), laboratory results (*Lab Result*), radiology orders (*Rad Order*), radiology results (*Rad Result*), and finally, the medications prescribed to the patient (*Rx Order*). The care events in the sequence are ordered by date and time they happened.

Inside the lab orders there is information on each of the labs that were ordered, including date and time. In the lab results there is information on each of the labs, with bacteria identified or information that the test was negative. Individual patient care models are of different length (depending on the number of visits and specific information in lab orders, lab results, rad orders, rad results and Rx orders). People who came only once and did not have any labs or rad orders, have patient care models with one record. People who came many times and had a lot of lab or rad orders and results, have very lengthy care models (hundreds of records).

The third step is to build descriptors that summarize the patient care models. For each patient care model, one patient care descriptor is computed. Each descriptor has attributes describing the number of analysis visits, overall lab orders, specific lab orders (e.g., blood culture, respiratory culture, urine culture, aerobic culture/smear, platelet auto AB), microorganisms identified (e.g., Enterobacter Cloacae, Pseudomonas Aeruginosa, Staphylococcus Aureus), types of radiology orders (e.g., DX chest, PX chest, DX Abdomen), syndromes (e.g., fever, gastrointestinal, severe illness or death), subsyndromes (e.g., malaise and fatigue, myalgia, upper respiratory infections). The value of all the attributes mentioned so far are integer numbers, depicting how many times such a given syndrome / subsyndrome / lab test occurred in the patient care model. For a given patient many attributes have a value of zero. Additional attributes in the descriptor include the patient's race, ethnic group, and age group (0-3, 4-11, 12-19, 20-49, 50+) and they are textual attributes.

The process of building descriptors is completely data driven: if there are *n* different microorganisms identified in the data set, there will be *n* corresponding fields in the descriptor; if there are *m* different types of laboratory tests in the data set, there will be *m* corresponding attributes in the descriptor. For a given data set, each patient descriptor has exactly the same attributes (in our case about 700 of them) making it easy to perform subsequent computations.

4 Distance Measure

Once the descriptors of all patient care models are computed, a distance measure is used to identify the closest (min distance) patient care descriptor to the desired inject. Gower Similarity [Gower, 1971] and several Euclidean distances were investigated. The distance measure needs to be well tuned to the illness of interest since its goal is to identify the patients who have attributes similar to those of the inject. Certain attributes are not related to Tularemia, and therefore their presence or absence is inconsequential (therefore they will not be used in the distance). An example of such an attribute is *Cardiac dysrhythmias* (whether or not a person has *Cardiac dysrhythmias* subsyndrome, this person can have

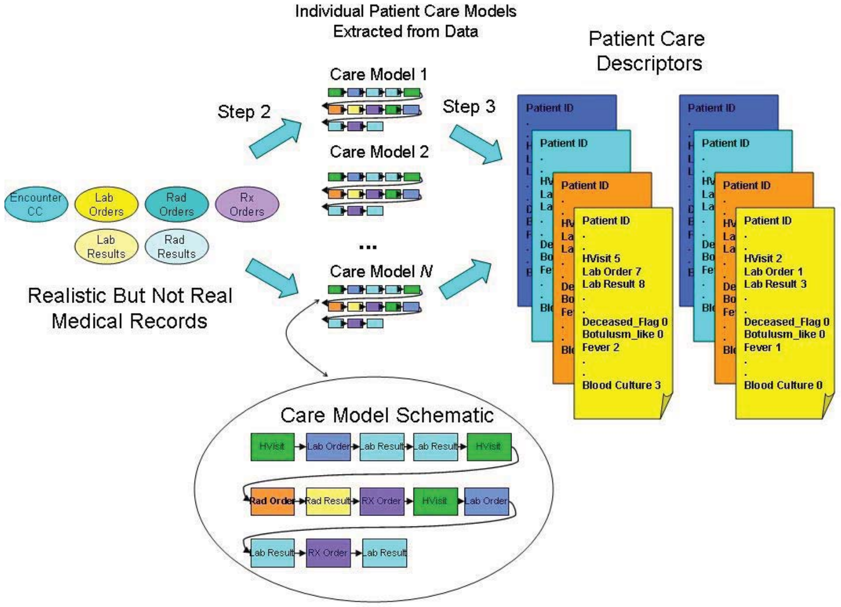


Figure 2. Data Mining Approach to Identify Existing Patient Care Models.

symptoms similar to those of Tularemia and the treatment will also be similar).

The Euclidean distance used operates on the following 23 attributes. Syndrome attributes: Fever, Gastrointestinal, Rash, Respiratory; and Subsyndrome attributes: Abdominal pain, Alteration of consciousness, Chest pain, Convulsions, Cough, Diarrhea, Dyspnea, Headache, Hemoptysis, Hemorrhage, Influenza-like illness, Lymphadenopathy, Neoplasms, Malaise and fatigue, Nausea and vomiting, Respiratory failure, Septicemia and bacteremia, Upper respiratory infections, Severe Illness or Death.

The Euclidean distance used is:

$$Distance = \sum_{i=1}^{23} w_i \cdot A_i$$

Where $w_i = 1$ for $1 \le i \le 22$, and $w_{23} = 0.1$ and A_i signifies the value of a given attribute (from the attributes specified in the previous paragraph). The attribute for which the weight of 0.1 is used is *Severe Illness or Death.*

Ten nearest neighbors in terms of the Euclidean distance from the same age group as the inject are retrieved. These nearest neighbors are processed by the Specific Disease Care Modifier.

4.1 Specific Disease Care Modifier

If the diagnosis is present in the data, the care models will match the care model expected for symptoms and severity of the injected illness. It may be the case that the diagnosis is present in the care models, but the age of the patient in the care model does not match the age of the injected victim. In this case, the care model will be modified to reflect age-appropriate treatment protocols.

In the case that the diagnosis is not present in the data, the patient care models that will be used for the synthetic medical records for injection victims must be modified to reflect the results of tests and radiographs that the victim population is likely to exhibit. Some examples of this for the Tularemia injection are the following:

1. For patient models with an influenza-like illness, the results of an influenza test are negative.
2. For patient models with chest x-rays, the results are modified to include the clinical manifestations of disease than have been observed with pneumonic Tularemia. For example, patients with Tularemia will exhibit bilateral pleural effusions, parenchymal infiltrates, lobal infiltrates, multi-lobar opacities and other similar characteristics in their chest x ray results.
3. Patient models for patients with non-Tularemia bacterial infections will have laboratory results modified to reflect the fact that *f. tularensis* is not usually able to be cultured without specialized culture media. Thus, culture results will be negative.
4. Patients models with rapid strep tests will have negative results on those tests.

5 Injection Algorithm

Once an injection time is determined for a victim, the synthesis of the medical record follows the patient care model identified by the data mining of the background data (with adjustments as above). The timeline for the patient care model for the injected patients is implicit in the synthesis of the records. However, it follows that of the patient care model in the following way. The visit date and time of the patient care model are noted as time 0. Then any subsequent laboratory, radiology, or clinical activity records *i* occur at time t_i after the visit date/time. The injected patient will have visit date/time *T* given by the disease injection function, and any laboratory, radiology, or clinical activity records *i* will have timestamp $T + t_i +$ *(random time < 1 hour)*.

The visit type—emergency, outpatient, or inpatient—follows the patient care model. Similarly, patients who succumb to the injected illness will match those in the background data who do not survive. Their death dates/times will match the timeline in the patient model.

6 Results

The results of this data synthesis method was a complete set of electronic medical records for over 200 fictitious victims of an airborne Tularemia attack interwoven into the de-identified medical records of the background population. Thus it would be impractical to present all the results here. However, in Figure 3 we offer selected results from some of the victim records. With suitable safeguards, the data can be distributed to the research facility that requested the injects via the proposed bioinformatics GRID [Kratz *et al.,* 2007].

The two records in the examples are for a female victim, age 49 and a male victim, age 57. The injection date for the Tularemia outbreak in this example (the date the pathogens were introduced) was 7/29/2006. The usual incubation period for Tularemia is 3-5 days. The female injected victim sought care on 8/5/2006, presenting with prodrome symptoms of a flu-like illness and fever. The care model for this patient then followed the care model for a flulike illness patient – clinical observations included a chief complaint of fever and influenza like symptoms, and a group A strep test was ordered. This victim was discharged with a flu-like illness as the diagnosis; the care model for the 8/5 visit did not include diagnosis of a bacterial illness. The male injected patient did not seek care until 8/19/2006 and consequently the illness for this victim was more advanced. Clinical observations included pulmonary congestion, cough and acute bronchitis. Because the care model for many patients of this age and gender included observations of hypertension, this victim also has this observation as a final diagnosis. This victim, more severely ill than the first, had a radiology order for a two-view chest x-ray. There were two preliminary radiology result records (this was the norm in this data set) and one final radiology result record, all indicating evidence of Tularemia. Unfortunately, our background data set did not include prescription or medication orders for this time period so prescription orders could not be included in any injected victim records.

The set of 200 injected victim records was reviewed by an ER physician for both content and accuracy; it was determined that the victim records were in line with the manifestations of pneumonic Tularemia and the records included resultant tests and observations of patients with this disease.

Because this method is entirely new, we have yet to develop statistical validation criteria. This is necessary to ensure that the injected data both reflect the injected illness and exhibit the same statistical profile as the background data. We are currently developing validation tests.

Analysis Visit Table.

VisitID	PatientID	VisitDate	Patient Class	Age	Birthdate	Sex	CASummary	Summary	SubSyndrome	Syndrome
584447	210321	05AUG2006	E	49	14JAN1956	F	FEVER OTHER FLULIKE..	|Emergency-Chief.. |	|Fever|Influenza-like…|	|Fever|Respiratory|
502621	215649	19AUG2006	O	57	27JAN1949	M	401.9Acute Bronchitis ….|	|Outpatient-Final…|	|Cough|Dyspnea|Bro…|	|Respiratory|

Clinical Activity Table

VisitID	PatientID	VisitDate	Activity	Category	SubSynText	SynText
584447	210321	05AUG2006	FEVER OTHER FLULIKE SYMPTOMS…	Chief Complaint	|Fever|Influenza-like Illness|	|Fever|Respiratory|
502621	215649	19AUG2006	CONGESTION IN CHEST COUGH DIFF BR…	Reason for Ad..	|Cough|Dyspnea|	|Respiratory|
502621	215649	19AUG2006	466.0 Acute Bronchitis	Final Diagnosis	|Bronchitis and Bron..|	|Respiratory|
502621	215649	19AUG2006	401.9 Hypertension NOS	Final Diagnosis		
502621	215649	19AUG2006	514 Pulm congest/Hypostasis	Working Diagnosis		|Respiratory|

Laboratory Order Table

Visit ID	PatientID	Visit Date	Diagnostic Service	Ordered Test Code Local	Ordered Test Name
584447	21031	18AUG2006	IMM	RStrep	Strep Group A Rapid (RST)

Radiology Order Table

Visit ID	PatientID	Visit Date	Ordered Test Code Local	Ordered Test Name Local	Reason for Test
502621	215649	19AUG2006	15488699	DX CHEST 2 VIEW	COUGH

Radiology Result Table

Visit ID	PatientID	Visit Date	Ordered Test Name	Reason	Status		Impressions
502621	215649	19AUG2006	DX CHEST 2 VIEW	COUGH	P	CHEST XRAY, TWO VIEWS	Clinical History: Cough, shortness of breath. Findings: Multiple parenchymal infiltrates seen bilaterally. There are bilateral pleural effusions.
502621	215649	19AUG2006	DX CHEST 2 VIEW	COUGH	P	CHEST XRAY, TWO VIEWS	Clinical History: Cough, shortness of breath. Findings: Multiple parenchymal infiltrates seen bilaterally. There are bilateral pleural effusions.
502621	215649	19AUG2006	DX CHEST 2 VIEW	COUGH	F	CHEST XRAY, TWO VIEWS	Clinical History: Cough, shortness of breath. Findings: Multiple parenchymal infiltrates seen bilaterally. There are bilateral pleural effusions.

Figure 3. Synthetic Data Record Example.

7 Conclusions

The Data-Drive Hybrid approach is a flexible model that will facilitate placing realistic electronic medical records with injected illnesses on a research grid for testing and evaluation of detection, fusion, and clustering algorithms. These medical records will mimic the data quality and quantity of the background records by using the background records as the model. The background records will be stripped of records for rare or familial illnesses, date-modified, rendered anonymous, and geographically randomized so that identification of either individuals or facilities from the data set is not possible. Future enhancements include randomization (based on distributions found in the background records) in the assignment of clinical observations, laboratory orders and results, radiology orders and results, and prescription orders. Another enhancement is improved automation in the analysis and injection phases as well as the ability, with expert knowledge, to select other injected diseases for the synthetic records.

We note, however, that the injection of additional diseases is non-trivial. Injection of infectious disease would require a simulation of disease transmission in the injected population. Even injection of non-infectious disease or toxic agents requires knowledge of the incubation period, timeline, and clinical manifestations of illness both to construct adequate descriptors (should the injected disease not be present in the background data, as in this example) and in the modification of the care models to reflect peculiarities of the illness.

Acknowledgments

This research was supported by Grant Number P01 HK000028-02 from the Centers for Disease Control and Prevention. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of CDC.

References

[Arikan *et al.,* 2003] Osman Kürşat Arikan, Can Koç, Önder Bozdoğan. Tularemia Presenting as tonsillopharyngitis and cervical lymphadenitis: a case report and review of the literature. *Eur Arch Otorhinolaryngology* 260: p. 298-300 (2003).

[Barrett *et al.,* 2005] Chris L. Barrett, Steven G. Eubank, James P. Smith. If Smallpox strikes Portland… *Sci Am* (March 2005).

[Breban *et al.,* 2005] Romulus Breban, Raffaele Vardavas, Sally Blower. Linking population-level models with growing networks: a class of epidemic models. *Physical Review* E **72**, 046110 (2005).

[Buckeridge *et al.,* 2004] David L. Buckeridge, Howard Burkom, Andrew Moore, Julie Pavlin, Protagoras Cutchis, William Hogan. Evaluation of Syndromic Surveillance Systems – Design of an Epidemic Simulation Model, *MMWR*, Vol. 53/Supplemental, Sept. 24, 2004.

[Burr *et al.,* 2005] Tom Burr, Richard Klamann, Sarah Michalak, Richard Picard. Generation of Synthetic Bio-Sense Data. *Los Alamos National Laboratory Report* LA-UR-05-7841 (2005).

[Daley *et al.,* 1999] Daryl Daley, Joe Gani. Epidemic Modelling: An introduction (Cambridge Studies in Mathematical Biology, 14) New York: *Cambridge University Press* 1999.

[Dennis *et al.,* 2001] David W. Dennis, Thomas V. Inglesby, Donald A. Henderson, John G. Bartlett *et al.,* . Tularemia as a Biological Weapon. *JAMA* 285(21) p. 2763-2773 (2001).

[Ellis *et al.,* 2002] Jill Ellis, Petra C.F. Oyston, Michael Green and Richard W. Titball. Tularemia. *Clinical Microbiology Reviews* 15(4), p. 631-646 (2002).

[Evans *et al.,* 1985] Martin E. Evans, David W.Gregory, William Schaffner, Zell A. McGee Tularemia: A 30-year Experience with 88 cases. *Medicine* 64(4), p 251-269 (1985).

[Frasca *et al.,* 2006] Mattia Frasca, Arturo Buscarino, Alessandro Rizzo, Luigi Fortuna, Stefano Boccaletti. Dynamical network model of infective mobile agents. *Physical Review* E. **74**, 036110 (2006).

[Gower, 1971] Gower, J. A General Coefficient of Similarity and Some of its Properties, *Biometrics*, 27, pp. 857-874, (1971).

[Harrell and Whitaker, 1985] Robert E. Harrell and Gary R. Whitaker. Tularemia: Emergency Department Presentation of an Infrequently Recognized Disease. *Am J Emerg Med* 3: p. 415-418 (1985).

[Klompas, 2007] Michael Klompas, Gillian Haney, Daniel Church, Ross Lazarus, Xianlin Hou, Richard Platt. Automated identification of acute hepatitis B using electronic medical record data to facilitate public health surveillance. *Advances in disease surveillance* 4: 52 (2007).

[Kratz *et al.,* 2007] Mary Kratz, Jonathan Silverstein, Parvati Dev, editors, Telemedicine and Advanced Technology Research Center Integrated Research Team Final Report: Health Grid: Grid Technologies for Biomedicine. (Nov. 2007).

[Lombardo and Buckeridge, 2007] Joseph S. Lombardo and David L. Buckeridge, *Disease Surveillance: A Public Health Informatics Approach*. Wiley Interscience, 2007.

[Meltzer *et al.,* 2001] Martin I. Meltzer, Inger Damon, James W. LeDuc, J. Donald Millar. Modeling Potential Responses to Smallpox as a Bioterrorist Weapon," *Emerging Infectious Disease* **7**(6) 959-969 (2001).

[MIDAS] MIDAS Models of Infections Disease Agent Study http://www.nigms.nih.gov/Initiatives/MIDAS

Improving a Knowledge Base for Use in Proteomic Data Analysis

Jonathan L. Lustgarten[1], Vanathi Gopalakrishnan[1,2], William R. Hogan[1,3], Shyam Visweswaran[1,2]
[1]Department of Biomedical Informatics and the [2]Intelligent Systems Program
University of Pittsburgh
200 Meyran Ave, M-183, Pittsburgh, PA 15260, USA
[3]University of Pittsburgh Medical Center
Pittsburgh, PA, USA
JLL: jll47@pitt.edu, VG: vanathi@pitt.edu, WRH: wrh9@pitt.edu, SV: shv3@pitt.edu

Abstract

We have developed a knowledge base containing information linking experimentally validated *m/z* ratios to proteins that a user can query to retrieve candidate proteins corresponding to a *m/z* ratio of interest. In this paper, we describe an improvement to the knowledge base that allows a user to query for *m/z* ratios associated with a protein of interest.

1 Introduction

Several significant challenges remain in the analysis of high-dimensional data generated in mass spectrometry–based proteomic experiments. One goal of such experiments that utilize mass spectrometry data is protein identification. Given a spectrum of mass-to-charge ratios (*m*/*z*) the aim is to identify those proteins and peptides that are present with high intensity in the data. The chief challenge in the assignment of putative protein identification to an experimentally identified *m*/*z* ratio stems from the variability of that value. Typically, the *m*/*z* ratio corresponding to a protein varies due to several factors: the biofluid from which the sample is obtained (e.g., plasma, cerebrospinal fluid, tissue lysate, etc.) [Gopalakrishnan, *et al.*, 2004], the mass spectrometry platform, and the occurrence of different forms of a protein resulting from post-translational modifications, alternative splicing and the existence of isoforms [Lustgarten, *et al.*, 2008].

The researcher who needs to identify candidate proteins corresponding to a *m/z* ratio, typically, either searches the literature or online knowledge bases such as the UniProt [The UniProt, 2007] and the Expasy [Gasteiger, *et al.*, 2003]. We have recently developed a knowledge base called the EPO-KB (Empirical Proteomic Ontology Knowledge Base) [Lustgarten, *et al.*, 2008] containing information linking experimentally validated *m/z* ratios to proteins that was curated from the biomedical literature. The EPO-KB can be queried via a Web interface (http://www.dbmi.pitt.edu/EPO-KB) to retrieve proteins that have experimentally determined *m/z* ratios that are close (based on a distance score) to the user-specified *m/z* ratio. Thus, EPO-KB enables identification of candidate proteins corresponding to a *m/z* ratio of interest that can then be evaluated further by experimental methods such as immunochemistry.

A second purpose of generating proteomic data is to identify unique proteomic patterns such as a set of *m*/*z* ratios that discriminates well disease samples from healthy samples. Such analyses may be assisted by the identification and removal from the data of those *m/z* ratios that represent non-specific inflammatory proteins such as serum amyloid A. We have now added the ability to perform reverse look-ups in EPO-KB, i.e., identification of all *m/z* ratios that have been linked to a particular protein in the literature. We further describe this functionality below.

2 Enabling protein to *m/z* ratios search

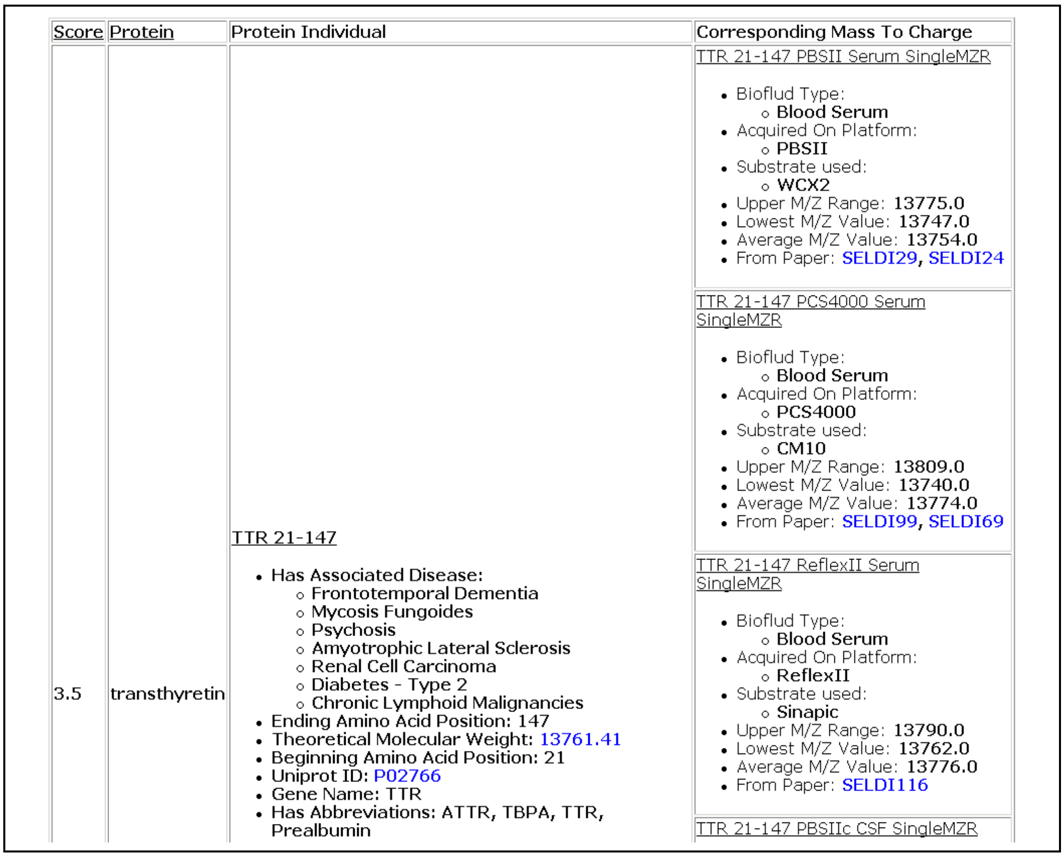
To allow querying of EPO-KB for proteins, we needed a standardized nomenclature to identify proteins uniquely. We chose the UniProt knowledge base [The UniProt, 2007] which provides a unique identifier for each protein and also contains information such as the sequence, possible modifications, and the gene that produces the protein.

For reverse look-ups, EPO-KB can be queried by a user-specified UniProt identifier. This retrieves all *m/z* ratios in the knowledge base associated with the protein and includes *m/z* ratios associated with fragments, mutations, and post-translational modifications of the protein. The retrieved *m/z* ratios can further be filtered on the basis of mass spectrometry techniques and the sample biofluid.

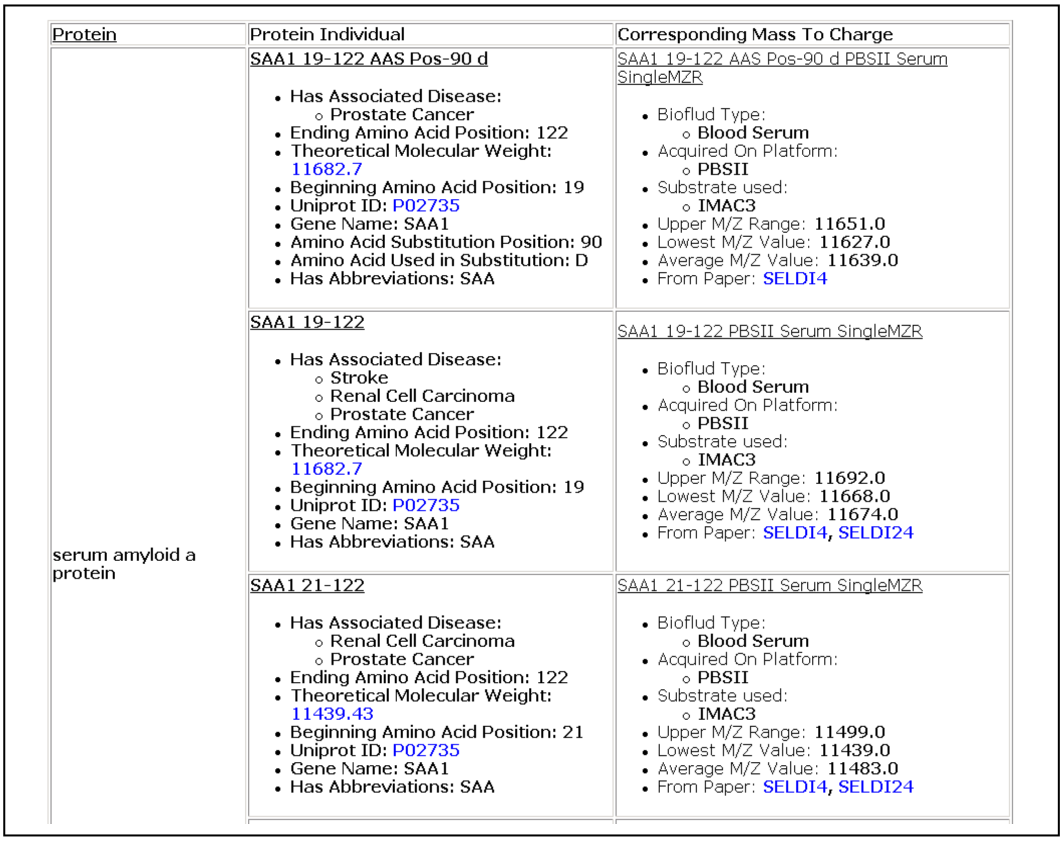
3 Analysis of proteomic data using EPO-KB

We now illustrate with examples the utility of the two methods of querying EPO-KB. A primary use of EPO-KB is for biomarker (*m/z* ratio) identification in experimental spectra. The researcher may use the EPO-KB to retrieve candidate proteins for *m/z* ratios of interest or confirm the identity of already identified *m/z* ratios. For example, the top panel in the Figure shows a partial list of candidate proteins identified by EPO-KB for *m/z* ratio 13750.

Another use of EPO-KB is in the identification of *m/z* ratios associated with a protein of interest. As an example, a researcher may want to analyze known disease specific proteins; however, identifying which *m/z* ratios are linked to those proteins is difficult. In another example, a researcher may want to identify and filter out *m/z* ratios in



Score	Protein	Protein Individual	Corresponding Mass To Charge
3.5	transthyretin	TTR 21-147 - Has Associated Disease: ○ Frontotemporal Dementia ○ Mycosis Fungoides ○ Psychosis ○ Amyotrophic Lateral Sclerosis ○ Renal Cell Carcinoma ○ Diabetes - Type 2 ○ Chronic Lymphoid Malignancies - Ending Amino Acid Position: 147 - Theoretical Molecular Weight: 13761.41 - Beginning Amino Acid Position: 21 - Uniprot ID: P02766 - Gene Name: TTR - Has Abbreviations: ATTR, TBPA, TTR, Prealbumin	TTR 21-147 PBSII Serum SingleMZR - Bioflud Type: ○ Blood Serum - Acquired On Platform: ○ PBSII - Substrate used: ○ WCX2 - Upper M/Z Range: 13775.0 - Lowest M/Z Value: 13747.0 - Average M/Z Value: 13754.0 - From Paper: SELDI29, SELDI24 TTR 21-147 PCS4000 Serum SingleMZR - Bioflud Type: ○ Blood Serum - Acquired On Platform: ○ PCS4000 - Substrate used: ○ CM10 - Upper M/Z Range: 13809.0 - Lowest M/Z Value: 13740.0 - Average M/Z Value: 13774.0 - From Paper: SELDI99, SELDI69 TTR 21-147 ReflexII Serum SingleMZR - Bioflud Type: ○ Blood Serum - Acquired On Platform: ○ ReflexII - Substrate used: ○ Sinapic - Upper M/Z Range: 13790.0 - Lowest M/Z Value: 13762.0 - Average M/Z Value: 13776.0 - From Paper: SELDI116 TTR 21-147 PBSIIc CSF SingleMZR



Protein	Protein Individual	Corresponding Mass To Charge
serum amyloid a protein	SAA1 19-122 AAS Pos-90 d - Has Associated Disease: ○ Prostate Cancer - Ending Amino Acid Position: 122 - Theoretical Molecular Weight: 11682.7 - Beginning Amino Acid Position: 19 - Uniprot ID: P02735 - Gene Name: SAA1 - Amino Acid Substitution Position: 90 - Amino Acid Used in Substitution: D - Has Abbreviations: SAA	SAA1 19-122 AAS Pos-90 d PBSII Serum SingleMZR - Bioflud Type: ○ Blood Serum - Acquired On Platform: ○ PBSII - Substrate used: ○ IMAC3 - Upper M/Z Range: 11651.0 - Lowest M/Z Value: 11627.0 - Average M/Z Value: 11639.0 - From Paper: SELDI4
	SAA1 19-122 - Has Associated Disease: ○ Stroke ○ Renal Cell Carcinoma ○ Prostate Cancer - Ending Amino Acid Position: 122 - Theoretical Molecular Weight: 11682.7 - Beginning Amino Acid Position: 19 - Uniprot ID: P02735 - Gene Name: SAA1 - Has Abbreviations: SAA	SAA1 19-122 PBSII Serum SingleMZR - Bioflud Type: ○ Blood Serum - Acquired On Platform: ○ PBSII - Substrate used: ○ IMAC3 - Upper M/Z Range: 11692.0 - Lowest M/Z Value: 11668.0 - Average M/Z Value: 11674.0 - From Paper: SELDI4, SELDI24
	SAA1 21-122 - Has Associated Disease: ○ Renal Cell Carcinoma ○ Prostate Cancer - Ending Amino Acid Position: 122 - Theoretical Molecular Weight: 11439.43 - Beginning Amino Acid Position: 21 - Uniprot ID: P02735 - Gene Name: SAA1 - Has Abbreviations: SAA	SAA1 21-122 PBSII Serum SingleMZR - Bioflud Type: ○ Blood Serum - Acquired On Platform: ○ PBSII - Substrate used: ○ IMAC3 - Upper M/Z Range: 11499.0 - Lowest M/Z Value: 11439.0 - Average M/Z Value: 11483.0 - From Paper: SELDI4, SELDI24

Figure. Top: screenshot showing partial list of candidate proteins identified by EPO-KB for *m/z* ratio 13750. Bottom: screenshot showing partial list of *m/z* ratios identified by EPO-KB for serum amyloid A (UniProt identifier P02735).

the data that correspond to proteins that are known to be increased in concentration in the sample but are non-specific to the disease process being examined. In such cases, the researcher may use the EPO-KB to retrieve all *m/z* ratios in the knowledge base that correspond to a protein of interest. For example, the bottom panel in the Figure shows a partial list of *m/z* ratios identified by EPO-KB for the protein serum amyloid A.

4 Future directions

We plan to expand the EPO-KB in several ways including the addition of the ability to search for diseases that will utilize a disease ontology with unique disease identifiers. In addition, we plan to add associated microarray and single nucleotide polymorphism data to the knowledge base.

Acknowledgments

We thank Henrik Ryberg and Chad Kimmel for their assistance in populating the knowledge base, and Gary Garvin for his assistance in programming the website. We thank the Department of Biomedical Informatics for providing server space for the EPO-KB. This work was funded by a training grant from the National Institutes of Health (#5 T15 LM007059 NLM).

References

[Gopalakrishnan, *et al.*, 2004] Vanathi Gopalakrishnan, Eric Williams, Srikanth Ranganathan, Robert Bowser, Merit E. Cudkowic, Max Novelli *et al.* Proteomic data mining challenges in identification of disease-specific biomarkers from variable resolution mass spectra. Proceedings of SIAM Bioinformatics Workshop 2004:1-10.

[Lustgarten, *et al.*, 2008] Jonathan L. Lustgarten, Chad Kimmel, Henrik Ryberg, William Hogan. EPO-KB: A searchable knowledge base of biomarker to protein links. Bioinformatics 2008;24(11):1418-1419.

[The UniProt, 2007] Consortium The UniProt. The Universal Protein Resource (UniProt). Nucl. Acids Res. 2007;35(suppl_1):D193-197.

[Gasteiger, *et al.*, 2003] E. Gasteiger, A. Gattiker, C. Hoogland, I. Ivanyi, R. D. Appel, A. Bairoch. Expasy: The proteomics server for in-depth protein knowledge and analysis. Nucleic Acids Res 2003;31(13):3784-3788.

[Tolson, *et al.*, 2004] Jonathan Tolson, Ralf Bogumil, Elke Brunst, Hermann Beck, Raimund Elsner, Andreas Humeny *et al.* Serum protein profiling by SELDIi mass spectrometry: Detection of multiple variants of serum amyloid alpha in renal cancer patients. Lab Invest 2004;84(7):845-856.

Construction and Annotation of a UMLS/SNOMED-based Drug Ontology for Observational Pharmacovigilance

Gary H. Merrill, Patrick B. Ryan, Jeffery L. Painter
GlaxoSmithKline, Research Triangle Park, North Carolina

Abstract

The primary goal of the SafetyWorks project has been the development of an integrated set of methodologies enabling the use of large observational data sources in monitoring and assessing drug safety concerns. To support its analytical and exploratory capabilities, SafetyWorks makes use of two large hierarchically structured ontologies – one for medical conditions, and one for drugs. In this paper we focus on the drug ontology employed in SafetyWorks and on its construction and annotation based on the SNOMED CT and RxNorm subsets of the Unified Medical Language System Metathesaurus. The result is a case study illustrating the value of SNOMED and its integration with UMLS and RxNorm in a critical and innovative drug safety application. We expose sufficient details of our methods to enable others to make use of these methods and to encourage the expanded use of both SNOMED and the UMLS in data exploration and analysis applications, particularly in the area of improving approaches to drug safety. [1]

1 Introduction

FDA "Guidance for Industry Good Pharmacovigilance Practices and Pharmacoepidemiologic Assessment" [FDA, 2005] describes pharmacovigilance as "all scientific and data gathering activities relating to the detection, assessment, and understanding of adverse events." While a drug is in development, one of the primary sources of safety information is clinical trials, but most trials suffer from insufficient sample size and lack of external validity to reliably estimate the risk of any potential safety concerns for the target population. Once a medicine has been approved, spontaneous adverse event reporting becomes an increasingly important tool for safety evaluation. Case review remains a key component of the ongoing surveillance of medicines, and the application of disproportionality analysis tools on spontaneous adverse event databases has greatly enhanced the signal detection process. Unfortunately, these spontaneous reporting systems have several limitations that make causal assessments difficult ([Almenoff *et al.*, 2005; Hauben *et al.*, 2005]): voluntary reporting suffers from chronic underreporting and maturation bias, and the unknown nature of underlying populations make true reporting rates difficult to obtain and use for comparisons. Several recent safety issues have received significant public attention ([Furberg *et al.*, 2006]), resulting in heightened awareness of the challenges of the current safety review process and increased demand for improved methods for understanding the effects of medicines and ensuring patient safety.

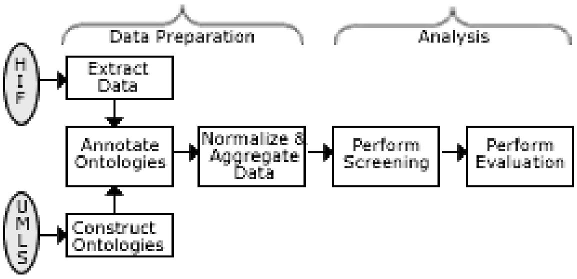


Figure 1: The SafetyWorks Process

SafetyWorks is an integrated system for leveraging observational data in support of the identification and evaluation of potential safety concerns of medicines. This system encompasses a data processing procedure that transforms disparate data sources into a common framework that enables normalized analyses across sources and the integration of automated methods for observational screening and observational evaluation. Figure 1 illustrates how raw data is extracted from the GlaxoSmithKline Healthcare Information Factory (a repository of large databases), normalized and aggregated with the help of annotated medical condition and drug ontologies constructed from the data

[1] All references to the Unified Medical Language System, the UMLS Metathesaurus, RxNorm, and the UMLS Lexical Tools are accessible through [NLM, 2008]. The SafetyWorks project began in the spring of 2005 and most of the ontology work was developed on the basis of the 2005-2006 releases of the UMLS and its documentation. However, we have continually updated our ontology as new releases have appeared.

An extended argument for the use of multiple observational databases in pharmacoepidemiology and how the methods described here may play a central role in this can be found in [Ryan, 2008]. Some additional details and related work may be found in [Painter *et al.*, 2006], [Ryan *et al.*, 2008], [Ryan and Powell, 2008], [Merrill *et al.*, 2008], and [Painter, 2008].

and UMLS, and then used in observational screening and observational evaluation to assess drug safety. Combined, this "observational pharmacovigilance" approach provides a systematic solution to supplement – rather than replace – current practices, enabling more proactive monitoring and better informed decision making.

The data processing procedure involves extracting key elements from each data source into a common relational model. While the representations (codes or strings) of individual elements (medical conditions and their treatments) in each source may be different, both contain common concepts of persons with drug utilization and condition incidence. We construct *drug eras* to represent periods of time where the data suggest a person may be persistently taking a medicine based on prescriptions written, prescriptions filled, or medication history provided to the physician. Similarly, we construct *condition eras* to represent common episodes of care for the same medical condition, aggregating related diagnostic codes that occur within a persistence window. Analytical methods are then applied to these eras to discover drug/condition associations and to evaluate the strength of such associations. Biomedical ontologies play an instrumental role in facilitating the normalization and aggregation of similar drug and condition concepts, and their application is the focus of this paper.

Observational screening applies an unmatched cohort design to facilitate comparisons of incidence rates of all outcomes across two populations of interest, both pre- and post-exposure. It provides an exploratory context (including information concerning patient demographics, comorbidities and concomitant medications) that can be used to understand and compare drugs, their uses, and their effects. Observational screening analyses should be considered exploratory and hypothesis-generating in nature, and should facilitate the identification and prioritization of drug-condition pairs that warrant further evaluation.

Observational evaluation (or "risk estimation") is a targeted analysis aimed at providing a robust estimate of the strength of a drug/condition association within the population of interest by systematically assessing the temporal association between a specific drug and a specific condition within the observational data sources. It models the specific exposure-outcome relationship using multivariate Poisson regression within a propensity score matched cohort design, adjusting for important covariates related to both exposure and outcome. Observational evaluation can be one mechanism to assess the hypotheses generated within observational screening by constructing cohorts that are comparable (adjusted for confounding) and representative of the population of interest.

Two classes of observational data that hold promise in this domain are administrative claims databases and electronic health records. Each type of data has its own advantages and limitations, and specific data sources may have unique features that need to be well understood and carefully considered when conducting observational analyses and interpreting results. SafetyWorks currently makes use of one instance of each type of observational data.

The administrative claims database contains health information for over 74 million persons with an average 24 months of coverage. Drug utilization is extracted from over 833 million pharmacy claims of prescriptions filled. Conditions are captured from diagnosis codes on inpatient and outpatient medical billing claims; 5.6 billion distinct diagnoses were aggregated into 1 billion condition eras. Insurance claims data has the advantage of very large sample size, and generally comprehensive summaries of health-related activities during enrollment. However, claims are also susceptible to misclassification bias, and may not adequately capture symptoms or other important aspects of the patients' medical histories. The database represents an employed, privately insured population which may not be generalizable to other populations of interest.

The electronic health record (EHR) database provides health information for 5.8 million patients. Drug utilization is extracted from prescriptions written by the provider and medication history lists to create 58 million drug eras, averaging 101 days of exposure. Condition eras are constructed from a problem list of diagnoses, symptoms, and other components of medical history, resulting in 32 million condition eras.

One key opportunity in observational pharmacovigilance lies in enabling the systematic use of disparate observational databases for a more comprehensive review of the utilization and effects of medicines in populations. By establishing a common conceptual framework to structure observational data and to normalize references to drugs and conditions, analyses can be conducted consistently across sources, thereby enabling direct comparison of otherwise disparate results. Formal biomedical ontologies provide us with the mechanism for achieving this goal.

2 Methods

In choosing a drug ontology for SafetyWorks, we were guided by several criteria. The ontology must provide a correct and uniform classification of drugs and drug categories. It must be comprehensive relative to the relevant data – which is to say that it must exhibit a sufficiently high granularity of categories to which drug references in our data (and anticipated future data) could be annotated. It must contain categories for branded drugs as well as generics. It must exhibit a hierarchical structure in terms of individual drugs, their generic forms, and various levels of drug classes; and this hierarchical structure must adequately represent the relations of drug products to multiple ingredients that they contain.

Beyond these purely formal or structural constraints, we also felt it necessary to impose constraints of usability since the ontology would be employed in a graphic and interactive manner by drug safety scientists. Accordingly, it is necessary that the ontology exhibit categories and a structure of some familiarity to such users, and it must be easily navigable and searchable by them. Finally, prior research (see, for example, [Painter *et al.*, 2006]) had convinced us that the UMLS comprised a powerful resource in the areas of drug discovery, coding scheme translation, and broader areas of biomedical informatics, and we were committed to taking advantage of the richness of the relations provided in the Metathesaurus across such domains as medical conditions, diagnoses, symptoms, and drugs. As a consequence, we sought an ontology that was represented among the UMLS Metathesaurus sources.

We therefore settled on SNOMED CT as the basis of the SafetyWorks drug ontology which is constructed and annotated in a sequence of steps:

- The *Drug or medicament* sub-hierarchy of SNOMED CT is extracted from the UMLS Metathesaurus.
- RxNorm is used to extend this hierarchy by grafting leaf nodes to it for branded drugs.
- The extended ontology is annotated with drug references from the observational data sources.
- The annotated ontology is simplified by applying several transformations to its hierarchical structure.
- The resulting ontology is then emitted as a set of files suitable for importation into a relational database for use by the SafetyWorks methodologies.

The first of these steps is accomplished straightforwardly through use of the MRCONSO.RRF and MRHIER.RRF files of a Metathesaurus subset containing the SNOMED CT source. The hierarchy is represented as a set of "nodes" identified by their UMLS Atom Unique Identifiers (AUIs), or extensions of these, and is extracted simply as the *isa* hierarchy with *Drug or medicament* (AUI A6938913) as its root.

Adding Branded Drugs

As part of our data extraction process, we created unique drug product reference identifiers as product-name/strength strings (such as "Zantac 150 Mg"), and associated with these may be additional information (varying with the data source being used) in the form of codes from a variety of coding schemes. Ideally, each drug product reference would be such a string consisting only of a drug name and a strength. However, actual drug product references in the reference file and in the data sometimes contain additional information as well ("tablet", "syringe", etc.); and this makes identifying the drug and correctly annotating it to the ontology more challenging.

Occasionally it is important to distinguish between the occurrence of a drug product reference in the drug reference file (where each drug product has only a single reference) and occurrences of a drug product reference in the observational data itself (where there may be millions of references to a particular drug or drug product). In the latter case we will then refer to *instances* (in the data) of the drug reference or drug product reference.

Unfortunately, the otherwise quite satisfactory *Drug and medicament* hierarchy extracted from the UMLS lacks categories for branded drugs such as "Wellbutrin", "Zyban", etc. While for the most part the interest of drug safety scientists is focused on generic forms, we felt it necessary to achieve the granularity of branded drugs for the sake of completeness and because there are circumstances in which drug/condition associations may occur with one specific drug product and not with another. Our immediate challenge was to extend SNOMED CT with branded drug categories, and RxNorm provided us with a mechanism to meet this challenge.

The goal, then, is to take each branded drug in RxNorm (term type TTY = BN) and find the set of generic categories in the *Drug or medicament* hierarchy of SNOMED CT that represent the ingredients of that drug. The branded drug (represented by its RxNorm AUI) is then grafted to each such category as a child node in the hierarchy. In turn, this requires first finding the CUI (Concept Unique Identifier)representing the drug's "concept" and then finding the set of AUIs (Atom Unique Identifiers) *in our hierarchy* that "realize" that concept. This goal is facilitated by the RxNorm relations *tradename_of*, *ingredient_of* (and *has_ingredient*), *consists_of*, and *form_of*. These relations allow us to construct a mapping from CUIs for brand names in RxNorm to the desired sets of AUIs in our SNOMED sub-hierarchy. In fact, we restrict this mapping to only those BNs in RxNorm that have the semantic type of *Organic Chemical* since experience has shown us that this is the class of entities that most closely approximates what are intuitively regarded as the "normal" set of branded drug products.

In virtually all cases it is possible to map directly from a drug's brand name through the *tradename_of* relation to its ingredient(s). An example of this is the branded drug name (BN) "Wellbutrin" (C0085934) which in RxNorm is a tradename of the ingredient (IN) Bupropion whose CUI is C0085208) and this in turn is realized in the SNOMED CT *Drug or medicament* hierarchy as AUI A2879308. Thus we can attach the a category for the branded drug Wellbutrin as a child of the *Bupropion* category in our extended hierarchy.

There is some question as to whether, and to what degree, the coverage and accuracy of our annotation could be enhanced by the use of other information the data might contain – such as associations with codes from various coding schemes. This is still something of an open question, but at one stage of the project substantial effort was put into making use of NDC codes in the data and their occurrences within several sources (NCI, NDFRT, NDDF, and VANDF) in the Metathesaurus. After a careful and thoughtful implementation, it was determined that the use of this approach yielded not a single enhancement to our lexically-based heuristic approach, and so it was removed from the annotation component.

396 of the brand names in RxNorm did not map to ingredients (i.e., these were BNs that had no corresponding INs) and consequently were not added as categories to our hierarchy. A single case (Meclomen) failed to map by means of the *tradename_of* relation. However, in our experimental approach to mapping arbitrary drug names into UMLS sources, we had developed a set of sophisticated algorithms involving relations among semantic clinical drugs (SCD), semantic clinical drug components (SCDC), and semantic branded drug components (SBDC); and the Meclomen case fell to these.

The *Drug or medicament* hierarchy extracted directly from the Metathesaurus contained 6,800 categories, and adding categories for branded drugs raised this count to 15,159.

Annotating Drug Data References to the Ontology

The goal of annotation is to associate each drug reference in our data with one or more categories in the drug ontology. Our fundamental approach to annotating the drug ontology with such drug references is then to match the string representation of the product-names to category names in the drug ontology, and we employ a number of algorithms and heuristics in this pursuit.

Matching of this sort requires a careful approach to string normalization, and initially our approach was to depend on the UMLS normalization of strings found in the MRXNS_ENG.RRF file of our SNOMED CT and RxNorm subsets, and couple this with the use of the UMLS Lexical Tools *norm* utility. However, for a variety of reasons we cannot detail here, we ultimately abandoned this approach in favor of developing our own string normalizer which is tuned more specifically to the needs of a clinical drug vocabulary.

The fundamental concept supporting our annotation of the drug ontology with drug references from our data is that of *Product Instance*. A Product Instance represents a single "drug product", distinguished by the product name, and also associates with this an "expanded" version of that name, a set of generic ingredients (if known), and potentially other information as well (such as codes from a variety of coding schemes, if these are known and might be useful). We annotate the ontology with drug references from each data source in turn, and the first step in annotation is to construct a Product Instance Table comprising Product Instances for each distinct drug product referenced in the given data source. Once the Product Instance Table is created, we then consider each Product Instance in turn and attempt to annotate it to the ontology.

The high level heuristic we follow in attempting to annotate a Product Instance to the ontology is a sequence of steps, each of which is tried if the previous ones have failed to produce a successful annotation. Table 1 illustrates what proportion of drug reference matches are captured by each method in the case of annotating our data to our SNOMED-based and enhanced drug ontology.

	Drug Coverage	
Annotation by matching normalized form of	**Claims**	**EHR**
The exact product name	51.60%	45.40%
An expanded form of product name	3.16%	1.40%
The generic name(s)	44.89%	52.13%
A variant of the product name	0.35%	1.07%
A variant of the expanded product name	0.00%	0.00%

Table 1: Drug Reference Coverage

The relatively high percentage of generic matches as compared to product name matches reflects our current conservative strategy of preferring a generic match over a match to a product name that has been modified in ways that might render a resulting match inaccurate. This is a challenging problem in the case of various forms of products (such as "extended release", "flu", "nighttime", "cold/cough" vs. "cold/allergy" variants, etc.) where the variant may contain significantly different additional ingredients than the base product; and tuning our matching heuristics is an ongoing research project.

To consider some examples:

- A reference to "Wellbutrin" succeeds as a direct match.
- "Aber-Tuss HC" fails to match (it does not occur in RxNorm). Its generic in the EHR data source appears as "PHENLYEPH-CHLORPHEN-HYDROCOD" which also fails to match. "CHLORPHEN" does match a SNOMED category, but neither of "PHENLYEPH" or "HYDROCOD" do. However, expansions of these do match, and so drug product references involving "Aber-Tuss HC" in the EHR data are annotated to each of the categories *Phenlyephrine*, *Chlorphen*, and *Hydrocodone*.
- "Dextrose in water" fails to match, but its variant "Dextrose" succeeds and so "Dextrose in water" is annotated to *Dextrose*.
- The product "Haleys M-O" fails to match, as does its generic "Mag hydroxide in mineral oil" and the expansion "Magnesium hydroxide in mineral oil". But the simplified variant of the expanded generic, "Magnesium hydroxide", succeeds and so "Haleys M-O" is annotated to the category *Magnesium hydroxide*.

Any drug reference that fails to be annotated to a drug category is annotated to an *unclassified substances* category added to the ontology for this purpose, and the consequence of this is that no purported drug reference in the data is ever lost to analysis.

2.1 Simplifying the Ontology

Initial attempts to use the annotated ontology as described in the previous section showed us that it exhibited some unfortunate features relative to our criteria and our plans for using it to support the SafetyWorks analytical methodologies.

Our last stage in the construction of the drug ontology is then to perform a series of refinements in which we

- Prune unnecessary "forms" of drugs from the hierarchy.
- Ensure that no drug reference annotates both a node and an ancestor of that node.
- Create "generic product" nodes to ensure that only hierarchy nodes at the lowest level are annotated.

Note that no hierarchy pruning or restructuring should take place until annotation is complete in order to maximize the degree and accuracy of the annotations.

Pruning forms

If we look at the ontology immediately after annotating it from the data sources, we will see a number of categories that serve no useful purpose and are something of a hazard to efficient navigation. These categories represent "forms" of a drug and are of no interest in the drug safety context within which we are working. The most common example of such categories are salts of substances such as *Fluvastatin sodium* which in RxNorm is the *tradename_of* the branded drug *Lescol*. But *Fluvastatin sodium* is a direct child of *Fluvastatin* in the extended hierarchy.

The RxNorm documentation describes the *form_of* relation as holding "between a base ingredient and a precise ingredient", and in our analytical and exploratory context such "precise" ingredients are unnecessary. In addition, to facilitate application of some of the SafetyWorks analytical methodologies and to provide more meaningful results to our users, we adopt the principle of annotating only branded drug categories or generic drug categories. Leaving forms (such as salts) in the hierarchy yields a structure that essentially contains (non-uniformly, since only at some places) two levels of generic drug categories. As a consequence, we choose to eliminate these unnecessary intermediate categories and prune the ontology of them by making use of the RxNorm *form_of* relation. As this pruning takes place, it is necessary to move any annotations attached to these categories to the higher level (base ingredient) category that remains.

Eliminating ancestral annotations

It is possible that a data reference has been annotated to multiple ontology categories, and we have seen examples in previous sections where this makes perfect sense (as in the case where a drug has multiple components). However, it is also possible that a data reference has been annotated to a category and also to an *ancestor* of that category. How does this happen?

The answer to this question lies in what are often slight incommensurabilities between the concept structure of UMLS and the content or structure of the specific ontology we are annotating (in this case a subset of SNOMED CT). Recall that we accomplish annotation of the ontology by first finding the UMLS concept (CUI) that represents a data reference and then "projecting" that concept into SNOMED CT to find the UMLS AUIs that realize the concept in that source.

It is true that a given AUI will be associated with only one CUI, but a given CUI may be associated with (realized by) multiple AUIs *in the same source*. This is simply a feature of how CUIs, AUIs, and their relationships have been implemented in the Metathesaurus. After all, the UMLS concept structure is simply yet another ontology (though it is intended to be a very general one). So it should not be at all surprising that mappings of Metathesaurus concepts into Metathesaurus sources will not, in some cases, be structure-preserving. A simple example of this is found in the case of the concept C0028040 which is realized in SNOMED CT by both A2877800 (*Nicotine*) and A3581984 (*Nicotine Agent*), where the latter is in fact a child (in the SNOMED CT hierarchy) of the former. And as a consequence of this, our data reference of "Nicotrol NS 1-Wk 10MG/ML" ends up being annotated to both of these categories.

For our purposes of data analysis (and also from the perspective of a user attempting to navigate the ontology) such redundant higher level annotations are confusing and can be computationally problematic. We therefore modify the annotated ontology to ensure that if a drug reference has been annotated to a node and also to one or more ancestors of that node, then the annotation is detached from the ancestor nodes.

Restricting annotation to the lowest level

Up to this point, we have allowed annotations to attach to both branded drug categories and to generic drug categories. Thus, for example, "Zantac 15MG/ML" annotates the *Zantac* category while "Ranitidine 15MG/ML" annotates the *Ranitidine* category. But this means that annotations are being made to two distinct levels of the hierarchy: branded drugs and their generics.

Again, this may complicate certain computations and it can be confusing to users navigating the hierarchy or searching for annotations. For these reasons we decided to annotate drug data references to only the lowest-level categories of the hierarchy. In order to do this coherently and uniformly we introduce the concept of an "NOC" (not otherwise characterized) category. An NOC category is grafted to the hierarchy at the same level as branded drug categories (i.e., as a child of a generic), and is used to hold annotations which would otherwise annotate the parent generic node. In the context of our Zantac example, then, we introduce the *Ranitidine NOC* category, make it a child of *Ranitidine*, and move any annotations from the *Ranitidine* category to the new lowest-level *Ranitidine NOC* category (which is a sibling of *Zantac*). Another way of thinking of Ranitidine NOC is as "unbranded Ranitidine product". More generally, an NOC category is expected to be annotated with unbranded *products* (or otherwise unknown/unrecognized branded products) of its parent generic category. Thus generic categories are never directly annotated, and annotations apply only to the leaves (lowest levels) of the hierarchy graph.

At this point our drug ontology (now with 16,100 categories) is a modified extended version of the SNOMED CT *Drug or medicament* hierarchy and is complete for our purposes. As shown in Figure 1, we then make use of the annotated ontology to "normalize" all drug references in each source into ontology categories, and on the basis of these normalized drug references we then create the aggregated drug eras described in the *Overview* section. That (together with a similar process involving our medical conditions ontology and the creation of condition eras) completes the SafetyWorks *Data Preparation* process, and the normalized and aggregated data is then used to perform observational screening and observational analysis for drugs and medical conditions of interest within the drug safety monitoring domain.

3 Discussion

Our claims data drug reference set contains 73,553 distinct drug product (product-name/strength) references, from which we can identify 15,848 distinct drug references (distinct product names, independent of strength). The EHR drug reference set contains 38,723 distinct drug product references and 22,851 distinct drug references. After annotating the ontology from both sources, we achieve a drug reference coverage (drug names from our drug reference set annotated to generic or brand name categories) of 90.16% for the claims drug references and 69.16% for the EHR drug references. Two questions that immediately arise are: "What explains the discrepancy in the coverage?" and "What explains the failure rate?".

As we have hinted in earlier sections, the data we are dealing with (in the form of drug product names) is not completely "clean". In fact, many of the purported "drug references" in the data are not references to drugs at all but to medical devices (syringes, braces, ice bags, lancets, etc.), eyecare products, bathroom products, chemicals and minerals, and herbal remedies. The most humorous example of such items occurring in the drug column of our data is "Contour fitted sheets", but others abound. This is the nature of observational data.

Given this, the coverage of over 90% for the claims drug names is quite impressive, and the much lower rate for the EHR drug names is explained by a much higher density of non-drug items that are referenced in that data. In addition, references in the EHR data to generics is of much lower quality than in the claims data. Strings in that data that are supposed to represent drugs sometimes contain a non-drug substring such as the strength, formulation, or delivery system – which makes parsing out the drug name more difficult. Another difference in the quality of data between the two sources is that the claims data represents generic components individually in separate columns while the EHR data represents generics in a single column with complex generic combinations such as "PSEUDOEPH-CHLORPHEN-HYDROCOD" which must be tokenized correctly and then matched through the application of more complicated heuristics.

The coverage of actual drug reference *instances* in the data is even better. While we have successfully classified only about 70% of the EHR drug data references, this represents a successful annotation of 95.6% of the 57.7 million *instances* of drug utilization observed in the EHR data. And of the 494 million *instances* of drug utilization observed in the claims data, 98.2% are successfully annotated.

Although we have not done a formal expert-based analysis of the accuracy of our annotations, any misclassifications appear to be extremely infrequent and for the most part these occur when a fairly specific (e.g., brand) reference is annotated to its generic. An example of this is where "Zyban" is annotated to *Bupropion NOC*, and the reason in this case is that while "Zyban SR" occurs in RxNorm, "Zyban" itself does not. This raises, once again, difficulties in correctly classifying different forms of drug products, and we plan to more fully address these issues in future work. In general, the approach we take to drug name matching can be expected to be highly accurate since the heuristics it employs make use of partial matching of normalized strings but do not make use of any form of probabilistic matching (which we have found to substantially increase the chance of misclassification in both drug and medical condition ontologies). However, our work in the future will focus on improving these methods and assessing their validity.

The drug ontology described here forms an integral part of the SafetyWorks methodologies and has been used in test cases of those methodologies in the assessment of known drug/condition associations. We continue to improve and test these methodologies as SafetyWorks moves from a prototype application to a production-level application that can be used with confidence by drug safety scientists and epidemiologists. And we hope to have exposed enough details of our approach to make it usable by others.

4 Acknowledgements

The authors would like to thank their colleagues Robertino Mera, Kathleen Beach, Hoa Le, and Gregory Powell for contributing medical and drug safety knowledge in evaluating the SafetyWorks drug ontology and its application. Thanks are due Henry Krzywy for his work in understanding the data sources and in developing the technology for extracting the raw data for use in SafetyWorks.

References

[Almenoff *et al.*, 2005] June Almenoff, Joseph Tonning, A. Lawrence, Ana Szarfman, et al. Perspectives on the use of data mining in pharmacovigilance. *Drug Safety*, 28(11):981–1007, 2005.

[FDA, 2005] FDA. *Guidance for industry: good pharmacovigilance practices and pharmacoepidemiologic assessment*. U.S. Food and Drug Administration, March 2005. Available at `http://www.fda.gov/cder/guidance/6359OCC.htm`.

[Furberg *et al.*, 2006] D. Furberg, A. Levin, P. Gross, R. Shapiro R, and B. Strom. The fda and drug safety: A proposal for sweeping changes. *Arch Intern Med.*, 166:1938–1942, 2006.

[Hauben *et al.*, 2005] M. Hauben, D. Madigan, C. Gerrits, et al. The role of data mining in pharmacovigilance. *Expert Opin Drug Saf*, 4(5):929–948, 2005.

[Merrill *et al.*, 2008] Gary H. Merrill, Patrick B. Ryan., and Jeffery L. Painter. *Using SNOMED to Normalize and Aggregate Drug References in the SafetyWorks Observational Pharmacovigilance Project*. KR-MED, Phoenix, AZ, USA, 2008. (Poster session.).

[NLM, 2008] NLM. *The Unified Medical Language System*. U.S. National Library of Medicine, 2008. Available at `http://www.nlm.nih.gov/research/umls/`.

[Painter *et al.*, 2006] Jeffery L. Painter, Kristoph Kleiner, and Gary H. Merrill. Inter-translation of biomedical coding schemes using umls. Technical Report FS-06-06, American Association for Artificial Intelligence, Washington, DC, 2006. Fall Symposium on Semantic Web For Collaborative Knowledge Acquisition.

[Painter, 2008] Jeffery L. Painter. *A Mapping Between SNOMED-CT and the OXMIS Coding Scheme*. KR-MED, Phoenix, AZ, USA, 2008. (Poster session.).

[Ryan and Powell, 2008] Patrick B. Ryan and Gregory E. Powell. *Exploring Candidate Differences Between Drug Cohorts Prior To Exposure: A Systematic Approach Using Multiple Observational Databases*. International Society of Pharmacoeconomics and Outcomes Research, Toronto, 2008.

[Ryan *et al.*, 2008] Patrick B. Ryan, Gary H. Merrill, and Jeffery L. Painter. *Defining medical conditions by mapping ICD-9 to MedDRA: A systematic approach to integrating disparate observational data sources for enabling enhanced pharmacovigilance analyses*. Drug Information Association, Boston, 2008. (Poster session.).

[Ryan, 2008] Patrick B. Ryan. A call to action: Emerging opportunities for pharmacoepidemiology to advance the understanding of the effects of medicines. *The PharmacoEpi Newsletter*, 2008. Forthcoming.