VPH2: Extracting Knowledge from Integrated Data

Michele CARENINI[1], Dimitris GATSIOS[2], Oberdan PARODI[3], and Cristiano QUERZÈ[1]

[1]NoemaLife Spa, Bologna, [2] Computer Technology Institute and Press "Diophantus", Patras, [3]Institute of Clinical Physiology, National Research Council, Pisa

Correspondence: mcarenini@noemalife.com; 0039 051 4193911. 52 via Gobetti, 40129, Bologna, Italy

Introduction

Heart failure accounts for almost a quarter of all hospital admissions due to cardiovascular events, has a high mortality, and places a great burden on healthcare systems. VPH2 allows to build *subject-specific multi-scale models of the heart (from tissue to genes) and valves*. The task of VPH2 is the design and development of a platform aiming at: defining the severity and extent of disease in patients with post-ischemic LVD, and in particular at integrating clinical, biological and genetic data retrieved from medical records and laboratories research results; supporting cardiac surgeons with simulation and modelling tools to be applied in order to obtain better selection and simulation of specific surgical procedures; providing a new framework and multimodal approach that will be usefully applied to other clinical scenarios of cardiology and cardiac-surgery. This paper focuses on the first aforementioned aspect, i.e. the integration of data coming from heterogeneous sources, and the extraction of clinical and biological knowledge through data mining.

Materials and Methods: Data Harmonization and Data Mining

In order to extract knowledge from heterogeneous (clinical, biochemical, genetic, imaging) categories of data, the project built a core DB starting from four different data sets. From the *GISSI Prevenzione* study more than 1,000 patients enrolled in a randomized controlled trial on the efficacy of unsaturated fatty acids in preventing mortality after myocardial infarction (MI) were analyzed. The data set consisted of biochemical, behavioral and demographics data, drugs, stress test results. Additional interest derived from the availability of patients' genetic data. Patients' genomic DNA was extracted from mononuclear blood cells and screened for genetic variations. Candidate genes and linked variants screened in this study were the result of a multilayer process considering most recent findings in clinical and molecular genetics of cardiovascular dysfunction (CVD) with special respect to their reproducibility.

The *Niguarda* data set included patients from 2005 to 2008 with a clinical diagnosis of acute myocardial infarction (AMI) or chronic ischemic heart disease (IHD), for acute events or planned procedures. Clinical data were retrieved from current hospital databases and manually checked. Patient records with no more than 25% missing values for demographic, clinical, echocardiography, laboratory and drug therapy data were considered for analysis. Two separate data sets based on clinical diagnosis were examined: patients admitted for AMI (974 cases) and patients admitted for chronic IHD (404 cases).

The project had access also to a *CRT (Cardiac resynchronization therapy)* data set. CRT has been shown to reverse the remodelling process by improving ventricular size, shape, and mass, and reducing mitral regurgitation in the short and long term, on average within 6 months from the procedure. We analyzed the clinical (10 variables) and genetic (5 variables) predictors of RR in a matched series of 76 patients who did not reverse remodeling after CRT (RR-) and 80 controls who had shown a decrease > 15% in LVESV after 6 to 12 months with respect to pre-procedural values. Although the population was not large, useful outcomes were obtained.

Another important aspect of VPH2 was *cardiac MRI* and the correlation of features extracted from MRI with clinical, biochemical, and other variables. For that purpose, the Institute of Clinical Physiology of National Research Council has extracted from its clinical DB a data set consisting of 100 retrospective patients with myocardial dysfunction that were enrolled for cardiac MRI.

The Core DB obtained with these data sets is structured following openEHR specifications and is based on specific XML Archetypes. All the Clinical Data sets required by the clinical part of the consortium have been mapped in several Clinical Archetypes, and refined during the project. Core DB includes information about demographics, visits, behavioural habits, clinical and biochemical data, coronary angiographies, drugs, echocardiographic and electrocardiographic exams, genetics, MRI functional data, procedural data, radiology exams, stress test and all outcomes from the various VPH2 modules. During the data mining work with the four data sets described above always the same steps were followed as indicated by the adopted CRISP-DM standard. Namely:

- *Problem definition – setting goals*: After extensive collaboration with clinical partners from CNR for each data set both the problem and the desired outcomes were clearly defined: the patient will or will not develop late onset heart failure based on the classifiers extracted from GISSI dataset, patient's survival status based on the classifiers extracted from Niguarda retrospective dataset, patient's heart reverse re-modelling after CRT based on the CRT dataset.
- *Preparation of data*: data cleansing (different ranges, different names, etc.), handling missing values (SMOTE, ignoring variables with more than 25% of missing values, etc.), changing data formats in order to be used for mining, etc.
- *Modelling the problem - Evaluation of the proposed solution*: Restricted data sets were defined (firstly automatically with wrappers and filters, then manually in order to use medical knowledge as background information) and many different methodologies (from C4.5 to Bayes) with different parameters were applied in order to compare their accuracies, keeping in mind at the same time that clinicians were asking for transparent methods, i.e. methods that were user friendly (rules, trees etc.) and whose results could be clinically interpreted. 10 fold cross validation was used for the validation of the machine learning results. In order to evaluate the statistical differences of the classifiers with the highest accuracies McNemar testing was also performed.
- *Building the model*: After the comparison of the accuracies and the clinical interpretation of the most transparent and accurate methods (decision trees, decision tables, partial decision trees, random forests), a rule-based method (Partial decision Trees – PART) was adopted as the best choice for the Decision Support Module. This method can be edited by adding, removing, or reordering rules and these functionalities allow clinicians to build customized classifiers based on their own knowledge. Moreover, PART classifiers are automatically updated each time new datasets become available, ensuring that the provided support is based on up-to-date patient data.
- *Deployment of the solution*: Classifiers built with the PART methodology for each data set were reviewed by experts, and the ones considered useful for decision support were imported in the Decision Support Module.

Results and Discussion: Biologists and Clinicians' Feedback

The most important outcome from data mining in VPH2 was of course the knowledge the experts extracted from the produced classifiers. Firstly we present the biologists' feedback. During the genetic study in the GISSI population three genetic variants (rs4291, rs5443 and rs4646994) were shown to be associated with late onset HF (all p-values < 0.05). More precisely, we identified a significant association for two genes within the study population. One gene encodes for the angiotensin I-converting enzyme (ACE), the other for the guanine nucleotide-binding protein (GNB3). Two genetic variations positioned in ACE, termed rs4291_a=1 and rs4646994_INS=6 and one positioned in GNB3, termed rs5443_b=2 marked the two identified genes. The three alleles of the identified variants, namely rs4291=1, rs4646994=6 and rs5443=2 are predictors for late-onset HF in the study population used. The other alleles rs4291=4, rs4646994=5 and rs5443=4 are not associated with late-onset HF. Neither of the variants used are predictors for MI since they were not associated with MI. It has to be underlined that the functionality of the variants identified has not been experimentally proven. Some findings may not be in agreement with common knowledge, since genetic data is combined with biochemical and other markers. Discrepancies might point towards underlying unknown mechanisms and present potential starting points for selective research activities. Some indicative (even controversial) rules and the knowledge extracted from them are presented in table 1.

Rule	Class	Samples following the Rule	Correct	Wrong	Rule Accuracy
AMI = Anterior	Did not develop late onset HF	503	467	36	92.84%
AMI = Anterior and rs4291_b=4 and Diabetes = No	Did not develop late onset HF	199	190	9	95.48%
AMI = Anterior and rs5443_b=4 and rs4646994=6 and Diabetes = No	Did not develop late onset HF	62	58	4	93.55%
AMI = Anterior and rs5443_b=4 and rs4646994=5 and rs4291_b=4 and rs4291_a=1and Diabetes = No	Did not develop late onset HF	62	61	1	98.39%

Table 1. Classification of patients (and prediction for the case being assessed with VPH2 DSS) according to whether or not they have developed late on set heart failure with the application of PART methodology in a restricted (defined by the cardiologists) subset from GISSI dataset

Looking at these four rules we identify again that patients who had suffered from anterior AMI were not readmitted to the hospital. The rule accuracy is raised to 95.5% by adding rs4291=4 and Diabetes=0, while diabetes was identified not to be a risk predictor for late-onset HF and vice versa. Combining this rule with rs4646994=6, which is associated with late-onset HF, lowers the rule accuracy to 93.5%. Combining all "protective" alleles of the three genetic variants in one rule raises rule accuracy to 98.4%, which is a difference of 5.6%, even if rs4291 is heterozygous (rs4291=4 and rs4291=1). This is a good example that combination of genetic variants can remarkably increase accuracy of outcome prediction.

Data mining methods have been applied also in order to derive rules that may improve clinicians' decision-making; this has been accomplished mainly on Niguarda data set (consisting of retrospectively enrolled IHD patients). Results are consistent in general with indications from the literature even in the reperfusion and statin era. Here follow two examples:

Rule	Class	Samples following the Rule	Correct	Wrong	Rule Accuracy
Statins_Lipid_Lowering=YES AND Pre-existing_Vascular_Disease= NO	Patient survived	697	653	44	93.69%
Atrial_Fibrillation_history= NO AND STENT=YES AND Haemoglobin_blood>11.548035	Patient survived	345	333	12	96.52%

Table 2. Classification of patients (and prediction for the case being assessed with VPH2 DSS) according to their survival status with the application of PART methodology in a restricted (defined by the cardiologists) subset from Niguarda retrospective dataset

The high accuracy of the first rule in the prediction of a good outcome in this wide population subset confirms results from RCT on secondary prevention with statins in patients who do not have coexistent vascular disease in district other than the coronary one. The second rule also confirms results of previous studies (see [1, 2]). The negative prognostic impact of hypertension had been previously described in the classical Cox model from the GISSI Prevenzione data set (see [3]), and is confirmed by VPH2 results. Although the predictive role of clinical HF on presentation and atrial fibrillation are well established in AMI, the combination with other predictors is novel and interesting; in particular prescription of beta-blockers in this subset when still unstable is suggested by the negative impact of this class of drugs of proven efficacy in HF.

Conclusion

VPH2 developed the technology required to generate complex models composed by multiple sub-models, each simulating one of the many processes that may be observed at various dimensional scales, and that collectively produce the physiological or pathological manifestations observed in the clinical practice.

VPH2 conjugates the need for feasible and reliable patient-specific models with the timing of clinical decision making. Additionally, VPH2 technology represents an intelligent medical simulation environment for surgery training, planning and interventions. The multi-scale heart model simulates disease progression and allows for testing different therapeutic strategies in a specific simulated anatomical, functional and clinical scenario.

Acknowledgements

The VPH2 ("Virtual Pathological Heart of the Virtual Physiological Human") Project has been supported by the European Commission – FP7, DG-INFSO, Project No. 224635.

References

1. Schmitt J, 2009 DOI:10.1093/eurheartj/ehn579.

2. Saczynski, J S, 2009 DOI:10.1016/j.amjcard.2009.03.011.

3. Macchia A, 2005 DOI:10.1016/S1885-5857(06)60413-1.