Information processing for new generation of clinical decision support systems

Thomas Mazzocco
tma@cs.stir.ac.uk

COSIPRA lab - School of Natural Sciences
University of Stirling, Scotland (UK)

2nd SPLab Workshop
Brno, 25 October 2012

Outline

Introduction

- Many pilot predictive models for clinical have been successfully developed over the last decades
- Huge amount of information is nowadays collected in healthcare domain (patients' clinical history, diagnostic test results, etc.)
- A key challenge is how to mine the plethora of information in order to effectively help clinicians to make decisions
- Machine learning techniques can be employed to develop and improve such models

Clinical Decision Support Systems

- Tools to help health professional in (optimal) decision making for improved health care
- Knowledge-based or non-knowledge-based CDSS
- Desirable features: integrated in the clinical workflow, usability, transparency, electronic-based, recommendations provided, etc.
- Aimed at supporting the clinical processes and use of medical knowledge (e.g., diagnosis, investigation, treatment, short- and long-term care, etc.)
- Traditional vs. new

CDSS development

CDSS development (I)

The development of a logistic regression based model to aid the diagnosis of early dementia

Thomas Mazzocco, Amir Hussain; "Novel logistic regression models to aid the diagnosis of dementia", *Expert Systems with Applications*, 39(3), pp.3356-3361, 2012, Elsevier

Prototype at `http://www.cs.stir.ac.uk/~tma/`

CDSS: early dementia diagnosis

- A dataset of 164 patients suspected of dementia is considered
- For each patient 14 variables about their clinical history are recorded along with physician's diagnosis
- A logistic regression model is used to associate to each patient the probability of developing a dementia condition

	benchmark model	our model
Technique	Bayesian belief network	logistic regression
Variables	expert driven	expert or data driven

CDSS: early dementia diagnosis

The performance of our logistic regression model (both using expert driven and data driven variables selection) is reported in this table along with the benchmark (Bayesian belief network model)

	previous model	our model	our model
auROCc	0.764	0.783	0.879
R^2	n/a	0.371 - 0.602	0.365 - 0.601
Variables selection	expert driven	expert driven	data driven

UNIVERSITY OF STIRLING

SCHOOL OF NATURAL SCIENCES

Stirling Dementia Risk Calculator

Risk model for patients suspected of having dementia

---

1. Is the patient showing impairment in domestic activities of daily living (ability to carry out activities such as shopping, housekeeping, finance management, food preparation and transportation)?

○ Severely ○ Mildly ◉ None

2. Is the patient showing impairment in personal activities of daily living (ability to carry out activities such as dressing, eating, ambulating and hygiene)?

○ Severely ○ Mildly ◉ None

3. Overall, is the patient showing impairment in current functioning, i.e. in general activities of daily living?

○ Severely ○ Mildly ◉ None

4. Overall, how would you rate the global severity of impairment?

○ Severe ○ Mild ◉ None

5. Is the patient experiencing tremors?

○ Yes ◉ No

6. How long has the patient been showing symptoms for?

◉ Short period ○ Medium period ○ Long period

7. Did the patient show a clear progression in these symptoms?

○ Yes ◉ No

8. Is the patient able to complete the clock drawing test?

○ Yes ◉ No

Stirling Dementia Risk Calculator

Risk model for patients suspected of having dementia

The probability of suffering from dementia is **6%**

Given the information provided the most important factors with respect to the outcome are:

1. the clear progression of symptoms (positive correlation)
2. the presence of tremors (negative correlation)
3. the inability to complete the clock drawing test (positive correlation)

Back

CDSS development (II)

The development of a mortality model to identify AH patients at greatest risk of death

Thomas Mazzocco, Amir Hussain; "A novel mortality model for acute alcoholic hepatitis including variables recorded after 7 days after admission in hospital", submitted to *Computers in Biology and Medicine (Elsevier)*

Prototype at `http://www.cs.stir.ac.uk/~tma/`

CDSS: alcoholic liver disease mortality model

- A dataset of 82 patients with AH is considered
- 45 patients still alive after 28 days of admission, 37 succumbed to various complications
- For each patient, 22 variables about clinical findings and standard laboratory tests at the time of admission are recorded; 4 variables were re-evaluated after 7 days from admission (or at the time of death if patient died within 7 days)
- Our logistic regression model is compared with 3 risk scores currently used in clinical practice

CDSS: alcoholic liver disease mortality model

A logistic regression model has been developed and coefficients are tabulated.

	coefficient	**std. err.**	**p value**
creatinine	-0.022	0.010	0.033
creatinine @7d	0.046	0.013	<0.001
prothrombin time @7d	0.159	0.070	0.023
encephalopathy	1.390	0.670	0.038
constant	-6.303	1.630	<0.001

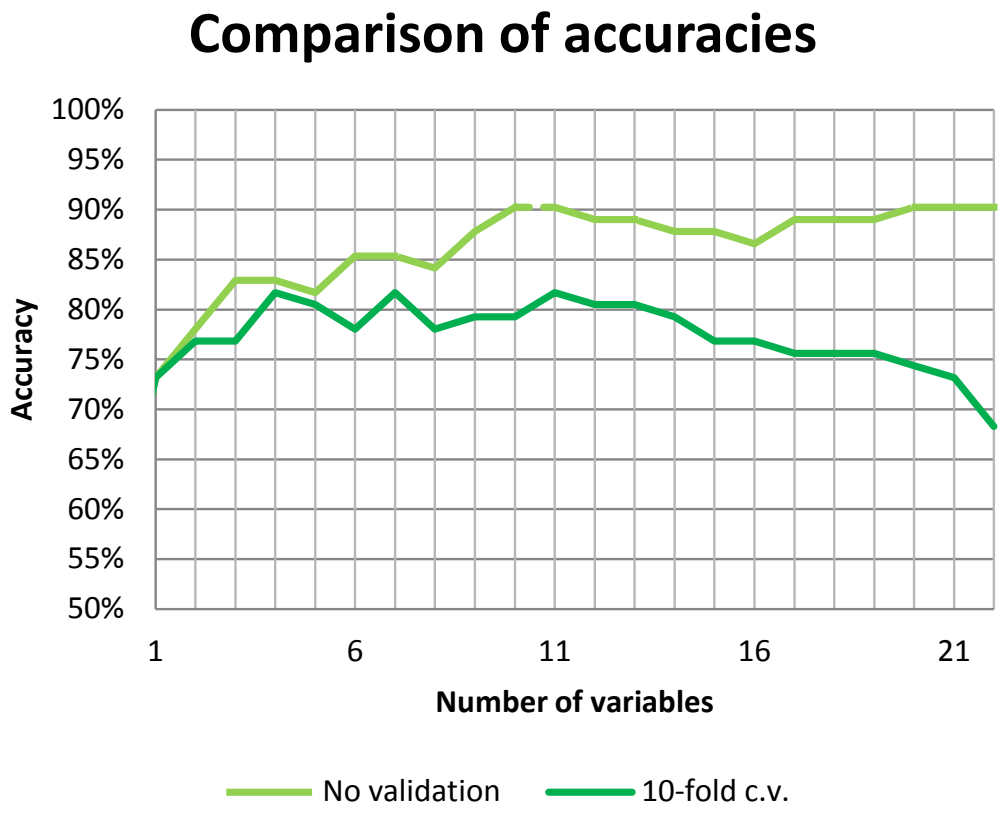
CDSS: alcoholic liver disease mortality model

The performance of our logistic regression model is reported in the confusion matrix below.

		Prediction		
		died	survived	% correct
Real outcome	died	28	9	75.7%
	survived	6	39	86.7%
	Overall			81.7%

CDSS: alcoholic liver disease mortality model

The optimal (minimum) number of variables has been selected in order to maximize performance.



CDSS: alcoholic liver disease mortality model

Comparison of the scoring systems in patients with AH

Score	Patient alive after 28 days	Patient died within 28 days	p value	auROCc
mDF	37.2 ± 26.2	67.5 ± 56.9	< 0.01	0.705
CPS	10.2 ± 1.6	11.8 ± 1.4	< 0.01	0.681
GAHS	7.6 ± 1.6	8.7 ± 1.6	< 0.01	0.687
Our model	24.5 ± 23.7	70.3 ± 30.7	$<$**0.001**	**0.873**

UNIVERSITY OF
STIRLING

SCHOOL OF
NATURAL SCIENCES

Stirling ALD Mortality Predictor (SAMP)

Mortality risk model after 28 days from admission in hospital for patients suffering from alcoholic liver disease during acute hepatitis

Severe form of alcoholic hepatitis in patients with alcoholic liver disease is associated with high mortality; it is therefore vital to identify patients at greatest risk of mortality as they may benefit from aggressive intervention. This new predictive model, which uses four statistically significant predictiors, could be used in clinical practice to identify such patients. The comparison with the available predictive scores showed an increase of 25% predictive power, demonstrating increased accuracy in identifying these sick patients with alcoholic hepatitis.

---

1. Level of creatinine at admission: 12 micromoles per litre

2. Level of creatinine on 7th day: 12 micromoles per litre

3. Prothrombin time: 12 seconds

4. Is the patient suffering from encephalopathy?

○ Yes ◉ No

Calculate score

UNIVERSITY OF
STIRLING

SCHOOL OF
NATURAL SCIENCES

Stirling ALD Mortality Predictor (SAMP)

Mortality risk model after 28 days from admission in hospital for patients suffering from alcoholic liver disease during acute hepatitis

---

SAMP score: **-4.104**

The probability of death within 28 days from admission is about **2%**

A score of -4.6 or 4.6 corresponds respectively to a probability of death of about 1% or 99%.
A score of -2.9 or 2.9 corresponds respectively to a probability of death of about 5% or 95%.
A score of -2.2 or 2.2 corresponds respectively to a probability of death of about 10% or 90%.

Back

CDSS development (III)

The development of a side-effects mapping model in patients with lung, colorectal and breast cancer receiving chemotherapy

Mazzocco, Thomas; Hussain, Amir; "A side-effects mapping model in patients with lung, colorectal and breast cancer receiving chemotherapy", *13th IEEE International Conference on e-Health Networking Applications and Services (Healthcom), 2011*, pp.34-39, 13-15 June 2011

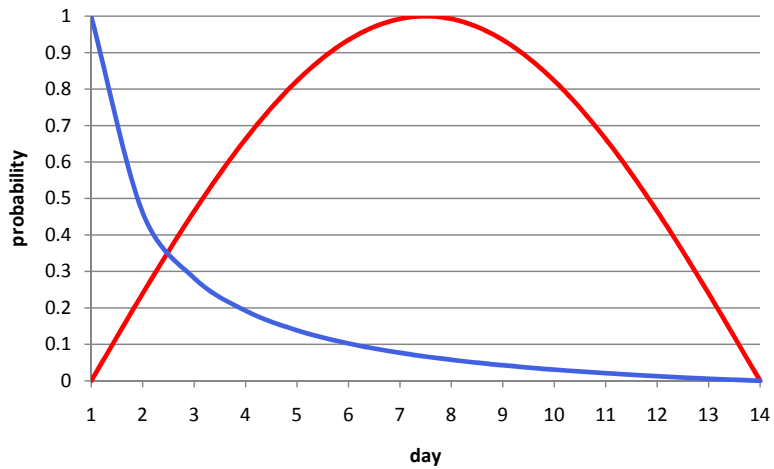
CDSS: register, monitor, predict chemotherapy side-effects

	Benchmark (ASyMS©)	our model
Dataset	24 patients	56 patients
Conditions	breast cancer	breast, colorectal, lung cancer
Model	different for each symptom	same for all symptoms

- In both models, data about symptoms were collected over 4 cycles of chemotherapy, each lasting 14 days
- 5 selected symptoms

CDSS: register, monitor, predict chemotherapy side-effects

- Two main time-dependant tendencies over time were outlined: the 'peak effect' and the 'inverted U-shape effect'



$$S(d) = sin\left(\frac{d-1}{d_{max}-1}\pi\right) = sin\left(\frac{d-1}{13}\pi\right)$$

$$H(d) = \frac{d_{max}}{d_{max}-1}\left(\frac{1}{d} - \frac{1}{d_{max}}\right) = \frac{14}{13}\left(\frac{1}{d} - \frac{1}{14}\right)$$

CDSS: register, monitor, predict chemotherapy side-effects

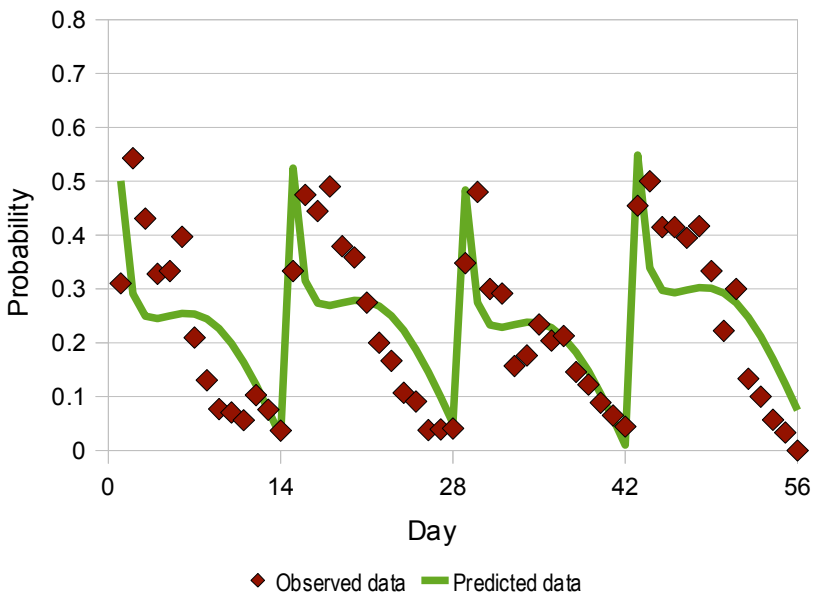
- The same formal model for all symptoms combining two effects (inverted U-shape and peak) and possible differences between cycles
- Three groups: lung, colorectal and breast cancer
- Coefficients determined using regression

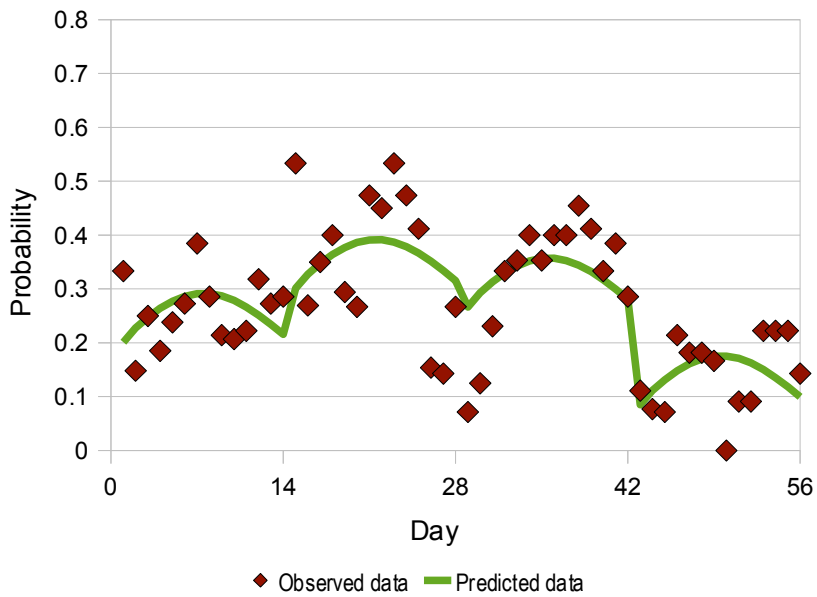
$$P(d) = a \cdot S(d) + b \cdot H(d) + \sum_{n=1}^{4} c_n \cdot D_n$$

$P(d)$	probability of experiencing the symptom on day d
a, b, c_n	coefficients determined using regression
D_n	dummy variable to identify the n-th cycle

CDSS: register, monitor, predict chemotherapy side-effects

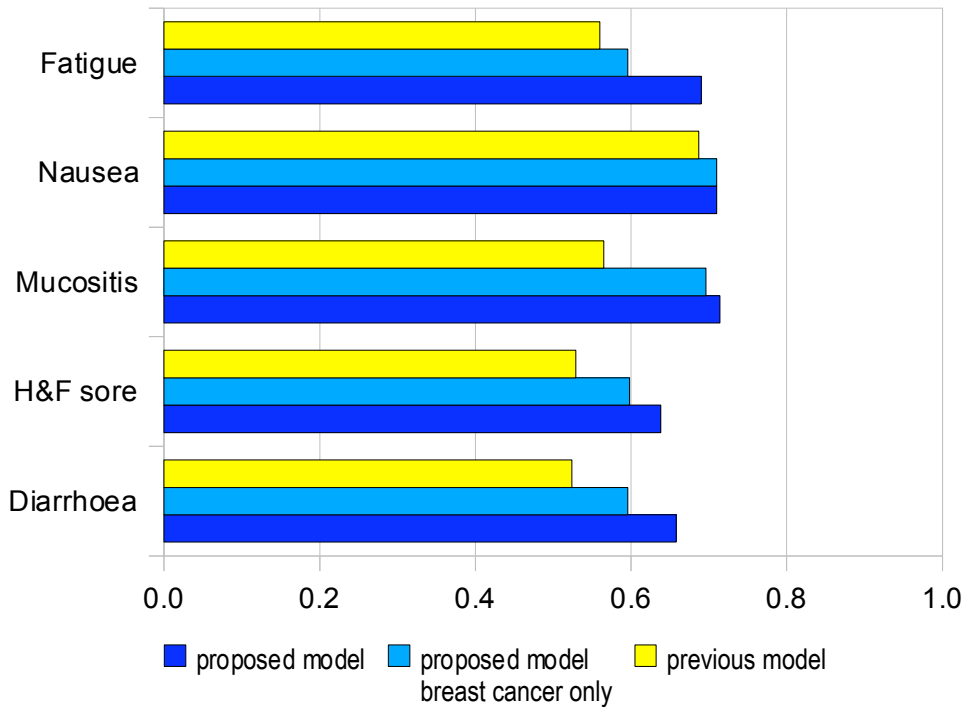
Observed versus predicted data for nausea in breast cancer (left) and for mucositis in lung cancer (right)





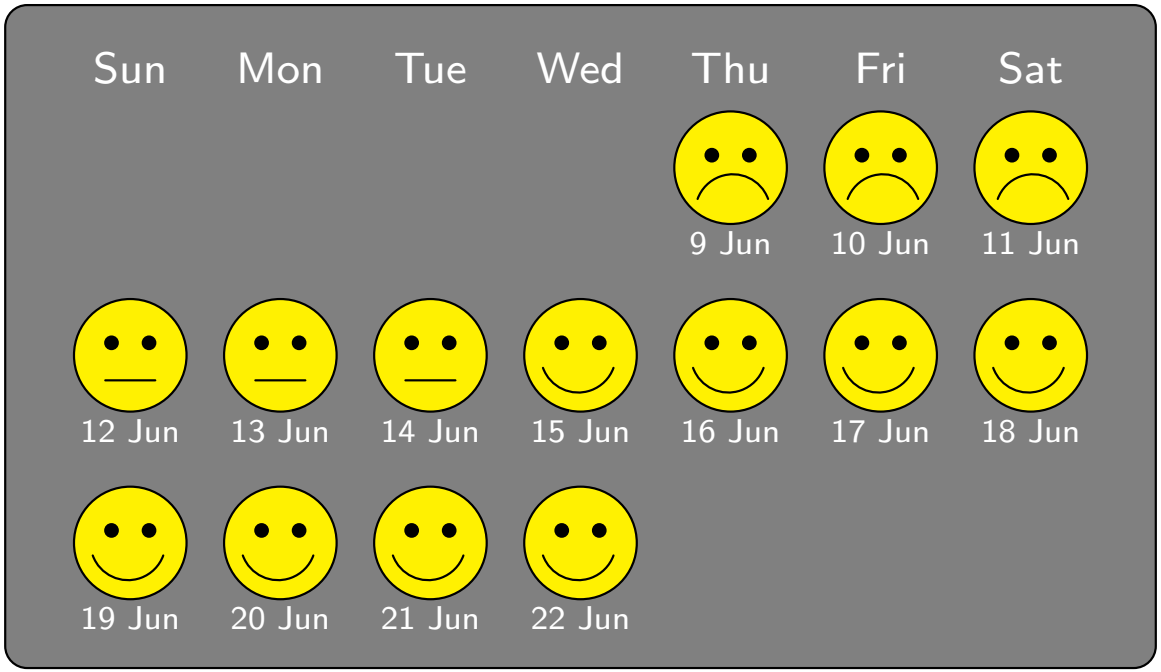
CDSS: register, monitor, predict chemotherapy side-effects

Receiver-operating characteristic (ROC) curve has been used to evaluate the model's performance: areas under curve (AUC) are here compared



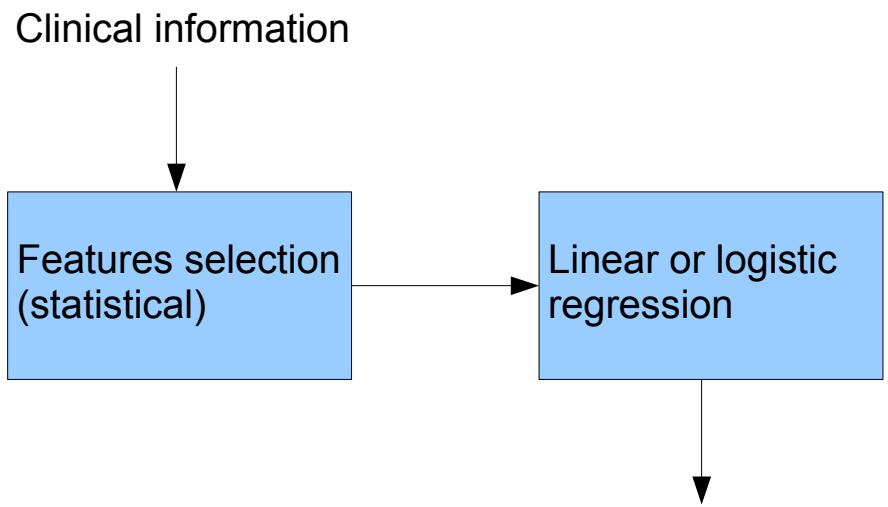
CDSS: register, monitor, predict chemotherapy side-effects

A diary is presented on patients mobile phones where, for each day, a smiley, sad or neutral face is used to depict the overall side-effects situation predicted for that particular day



The framework so far

The framework so far



Advantage: easy and transparent model (i.e. easy to identify how much each feature is influencing the output)

Current work: overview

Current work: overview

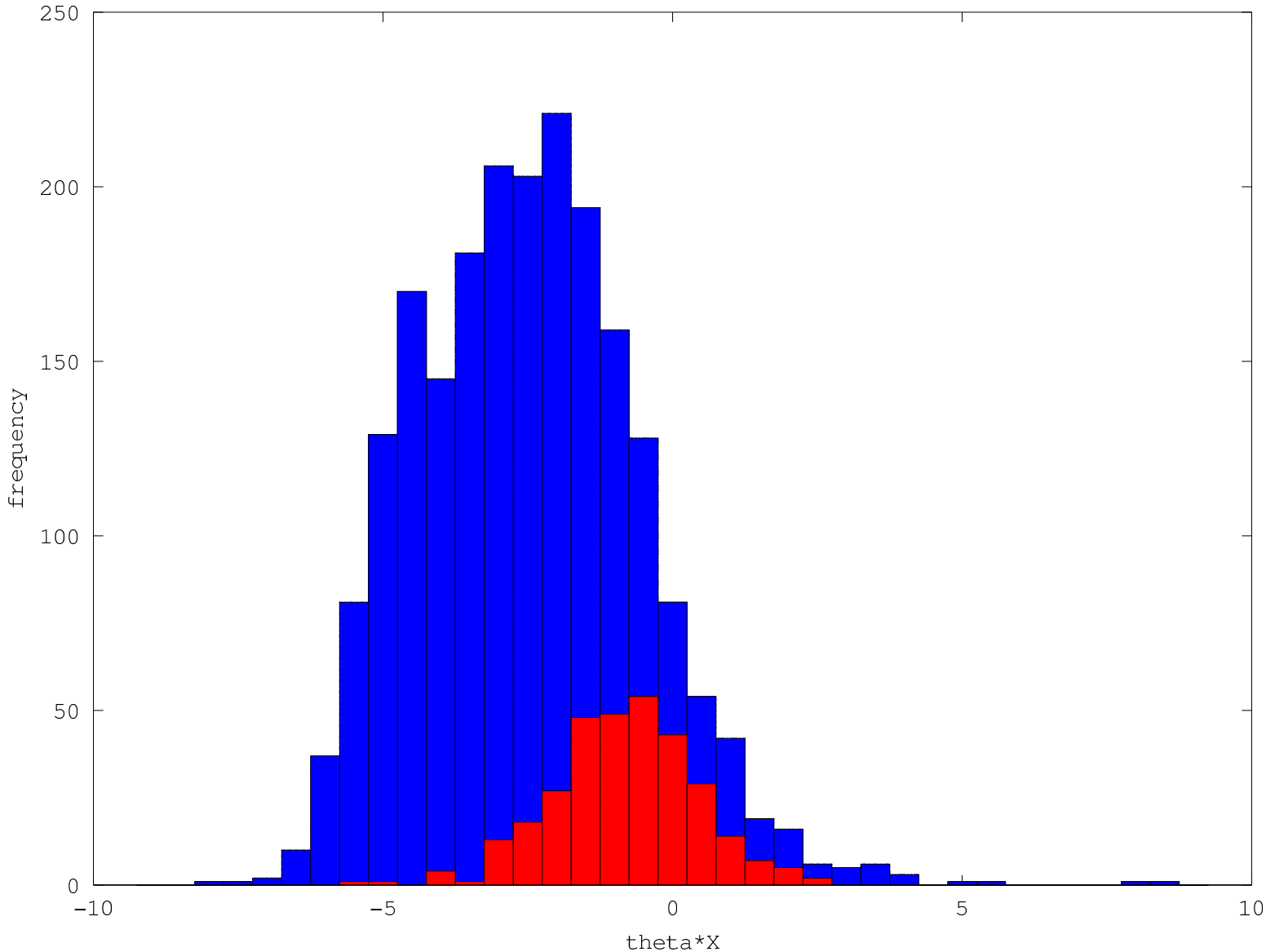
- MIT dataset for mortality prediction in Intensive Care Units
- 5 datasets for predicting mortality at day 28 based on data collected on the 1^{st}, 2^{nd}, ... 5^{th} day after admission
- Original model based on features selection and logistic regression
- For this experiment, we used 5^{th} day dataset, with 2,471 patients and 727 features (reduced to 13 in the MIT model)
- Collaborative work with Dr Hicham Atassi, Brno University of Technology

Current work: results

Performance:

- whole dataset (N=2,471): balanced accuracy = 73%

Current work: analysis of misclassifications

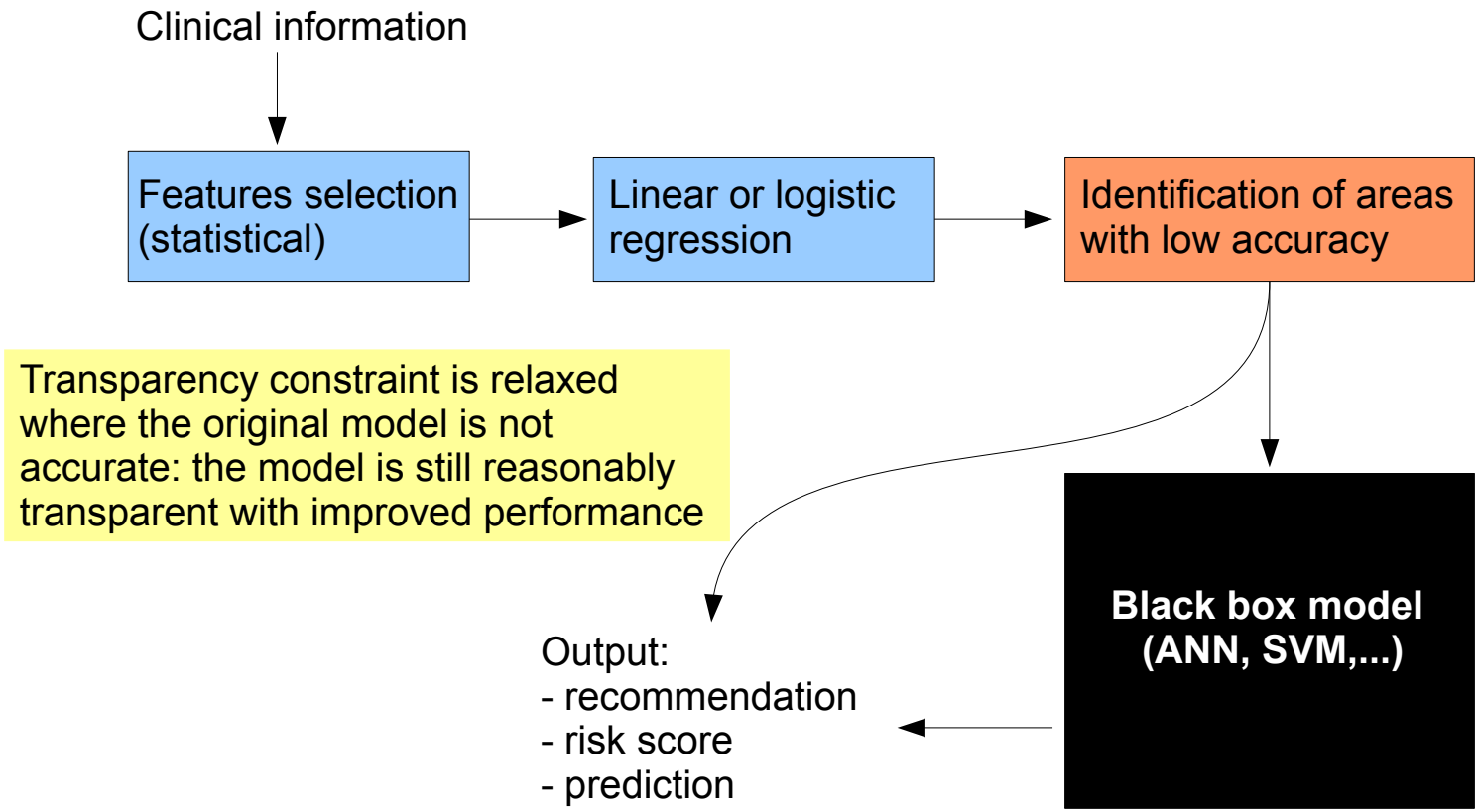


Current work: results

Performance:

- whole dataset (N=2,471): balanced accuracy = 73%
- central region (N=1,231): balanced accuracy = 60%
- outside central region (N=1,240): balanced accuracy = 90%

Current work: an alternative framework



Current work: preliminary results

Performance:

- whole dataset (N=2,471): balanced accuracy = 73%
- central region (N=1,231): balanced accuracy = **60%**
- outside central region (N=1,240): balanced accuracy = 90%

Relax the transparency constraint and apply a different classifier and/or features selection technique to improve results:

- central region (N=1,231): balanced accuracy = **66%** (Bayesian classifier with 10 features)

Conclusions

Conclusions

- A systematic and effective use of patients information will be crucial for delivering better healthcare in the future
- Statistical and machine learning techniques are applied to help medical staff to take appropriate decisions
- Key features which ensure successful deployment of CDSSs into clinical practice will need to look beyond their accuracy
- Proposed extensions to the commonly used framework (including intelligent analysis of misclassifications and subsequent data processing) will help to reduce the misclassifications, while trying to keep the models as transparent as possible