Towards a clinical decision support system for pregnancies of unknown location

Ben Van Calster and Sabine Van Huffel
Dept. Electrical Engineering, K.U.Leuven
Kasteelpark Arenberg 10, B-3001 Leuven, Belgium
{bvancals, vanhuffe}@esat.kuleuven.be

Dirk Timmerman
University Hospitals K.U.Leuven
Herestraat 49, B-3000 Leuven,Belgium
dirk.timmerman@uz.kuleuven.ac.be

Emma Kirk and Tom Bourne
St. George's Hospital Medical School
University of London
Cranmer Terrace, London SW17 0RE, UK

George Condous
Nepean Clinical School
University of Sydney
NSW 2006 Australia

Abstract

This paper presents our activities towards the development of a clinical decision support (CDS) system for the early diagnosis of pregnancies of unknown location (PULs). PULs are pregnancies that are not visualized on ultrasound. Generally, they turn out to be failing, normal intra-uterine, or ectopic pregnancies. Using 856 PULs collected at a London hospital, we developed probabilistic multi-class models. At the moment, a large multi-center data set is being collected. Data are entered using stand-alone software, and are sent to the database server by ftp. These data will be used to prospectively test existing models, and to build new, hopefully more robust, models. The ultimate aim is to select one model for implementation in a CDS system that is easy to use in routine clinical practice.

1. Introduction

A pregnancy of unknown location (PUL) describes a situation in which there is a positive pregnancy test without signs of an intra- or extra-uterine pregnancy on transvaginal sonography (TVS) nor remnants of miscarriage [2]. Generally, PULs can turn out to be one of three types of pregnancy. Most PULs are failing pregnancies (FPs) that are never seen on TVS. A second, large group of PULs represents early intra-uterine pregnancies (IUPs). The third group of PULs, which is least common, represents ectopic pregnancies (EPs). Rarely, the type of pregnancy remains unknown (persisting PULs). An ectopic pregnancy can rupture, such that it is potentially dangerous for the mother. Although a wait-and-see approach is known to be safe for women with a PUL that turns out to be ectopic [2], this approach can put serious pressure on the parents. This stresses the need for early diagnosis of PULs.

This paper concisely presents our activities towards the aim of implementing a robust mathematical diagnostic model for PULs into gynecological software for use in routine clinical practice. Using a large single center data set, we have developed diagnostic models based on various statistical and machine learning algorithms (Section 2). We focused on probabilistic multi-class models because diagnostic uncertainty is a relevant factor in medical decision making. Finally, we have initiated a multi-center study for prospective model evaluation and robust model development (Section 3).

2. Diagnostic models

2.1. Probabilistic multi-class models

The aim was to develop multi-class classifiers using various algorithms in order to compare their performance [3]. The data at our disposal were collected at St. George's Hospital (London, UK), involving 856 women (460 FPs, 330 IUPs, 66 EPs; 508 training cases) and 14 variables. We applied multi-class logistic regression (MLR) with manual checks for interaction effects and the linearity in the logit assumption, a combination of binary logistic regression models with similar checks (LR-PC), a Bayesian multi-layer perceptron (BMLP) and a Bayesian perceptron (BPER) using the evidence framework, combinations of binary Bayesian least squares support vector machines (LS-SVMs) using either the linear or RBF kernel (BLSSVMlin, BLSSVMrbf), and LS-SVM-based multi-class kernel logistic regression using the RBF kernel (MKLR). The binary logistic regression models for LR-PC were combined using pairwise coupling. Binary LS-SVM models were combined using either pairwise coupling or ensembles of nested dichotomies (END). Thus, we implemented nine methods: MLR, LR-PC, BMLP, BPER, BLSSVMlin-PC, BLSSVMlin-END, BLSSVMrbf-PC, BLSSVMrbf-END, and MKLR. Input selection was performed separately for the different models. Remember that MLR and LR-PC involved manual checks, possibly resulting in adaptations of the input set (e.g. transformations, interaction or polynomial effects). For the MLR model and LR-PC's binary models, the selection of an input set was mainly based on the average validation AUC for each class based on 20 runs of five-fold cross-validation. For the BMLP, BPER, and the binary BLSSVMs, automatic relevance determination was used as a ready extension of these models' Bayesian framework. For the MKLR model, all inputs selected for any of the binary BLSSVM models were used.

The `hcgratio` (ratio of the hCG levels at presentation and 48 hours later), `lhcgmean` (log of the average hCG level), `lprogmean` (log of the average progesterone level), and `bleeding` (level of vaginal bleeding) were selected for use in all methods. `Age` and `prevect` (history of ectopic pregnancy) were selected often. When applying the developed classifiers to the test set, the most important measures to evaluate performance are the test set AUCs for each class. The AUCs for FP and IUP were very high with minor variations between classifiers: they ranged from 0.985 to 0.990 for FP, and from 0.984 to 0.990 for IUP. For EP, larger differences were found. Models were compared using a ranking analysis based on the AUC for EP. Using 1000 bootstrap samples of the test set, the average rank (AR) is computed, as well as the proportion of times the classifier was ranked among the best three ($Pm(3)$). The results are shown in Table 1. LR-PC and MLR are ranked at the top, with good values for AR and $Pm(3)$. MKLR comes in third, closely followed by BLSSVMrbf-END. Using nonparametric regression, we observed that the models' predicted probabilities for test set cases were fairly well calibrated: predicted probabilities were fairly similar to true probabilities. To obtain a black-and-white prediction using LR-PC, we predicted EP if the probability of EP was at least 0.1. Else, IUP was predicted if its probability was at least 0.5. These thresholds were derived using the training data. On the test set, 82.5% of FPs, 79.6% of IUPs, and 83.3% of EPs were correctly predicted.

Table 1. PUL classifier ranking results.

Classifier	AUC_{EP}	AR	$Pm(3)$
LR-PC	.924	2.1	.89
MLR	.920	2.6	.74
MKLR	.917	3.5	.56
BLSSVMrbf-END	.914	3.7	.36
BMLP	.910	4.5	.23
BLSSVMrbf-PC	.907	4.9	.20
BLSSVMlin-PC	.896	6.8	.01
BLSSVMlin-END	.875	8.3	.00
BPER	.866	8.6	.00

3. The IPULA study group

Convened by the International Society of Ultrasound in Obstetrics and Gynecology (ISUOG), a panel of gynecologists with expertise in early pregnancy and gynecological ultrasound agreed on a consensus statement regarding the diagnosis and management of PULs [2]. The experts acknowledged the need for multi-center studies to evaluate the performance of diagnostic models in different populations. Furthermore, they stress that future research should focus on minimizing follow-up. This is an issue where diagnostic models can be of use.

Therefore, the International Pregnancy of Unknown Location Analysis (IPULA) study group was established in 2007. The aim of the IPULA study group was to collect a large multi-center data set in order to prospectively test existing diagnostic models and to retrospectively develop new models using PULs from various populations and clinical centers. Seven centers from around the world contribute to the IPULA data set: Leuven (Belgium), Cagliari (Italy), London (UK), Toronto (Canada), Philadelphia (USA), and two centers from Sydney (Australia).

The multi-centric nature of IPULA is important because populations may have different behavioral patterns such that pregnant couples may consult their gynecologist earlier in one population versus another. Also, the rate of PULs varies between centers [2]. This may be related to the experience of the gynecologist. Less experienced gynecologists could for example overlook ectopic pregnancies that are localized by more experienced colleagues. Thus, the PUL population may differ between centers.

Data is collected using software from Astraia GMBH (www.astraia.de). Patient information is entered locally, and is sent by ftp to a secured database server at the Katholieke Universiteit Leuven (Belgium). Collected data include patient information such as age and presenting complaints, patient history including risk factors for ectopic pregnancy, ultrasound data, and hormone levels (see Figures 1 and 2). At the end, the gynecologist has to give his/her personal diagnosis of the PUL (FP, IUP, or EP), along with his/her degree of certainty (cf. bottom of Figure 2). Finally, the true outcome of the PUL is recorded, together with the time until diagnosis, number of blood samples, and number of ultrasound scans.

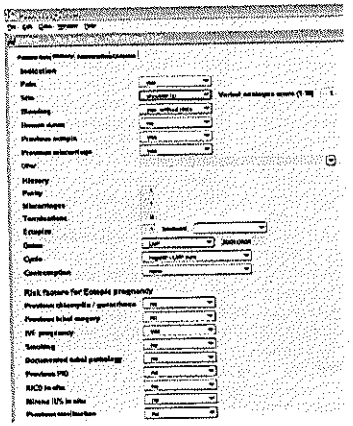


Figure 1. IPULA data collection.

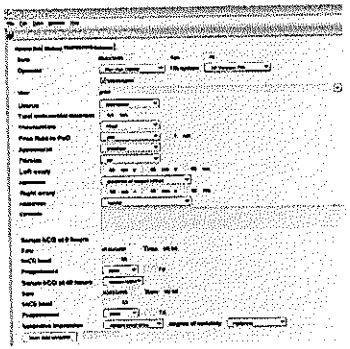


Figure 2. IPULA data collection (cont'd).

Using the same software, gynecologists can be informed of the probabilistic predictions made by a diagnostic model such as those presented in section 2.1 (see Figure 3). In the IPULA study, this feature was omitted to avoid any influence of the diagnostic model on the gynecologist. It will be used, however, in future studies assessing the tangible effect of a model on everyday clinical practice by investigating if the model can help to reduce follow-up time, the number of blood samples taken, and the number of ultrasound scans performed. When using this feature to compare diagnostic models to the personal diagnosis of the gynecologist, it is important to give the model's predictions after the gynecologist's prediction.

4. Conclusions

A lot of work has already been done in order to develop a CDS system for the early diagnosis of PULs. Yet, a long way still lies ahead. For example, new markers for PUL diagnosis are being investigated. Once the IPULA data set is ready, prospective model testing can start, and new models can be built. Next, new studies are required to prospectively test IPULA models, and to assess the effect of the models when applied in everyday clinical practice. Throughout the whole process, two aspects are of utmost importance to us: a continuous in-depth collaboration between statisticians and gynecologists to ensure that the diagnostic models meet the requirements set by clinicians, and a rigorous performance evaluation of the models, both in terms of diagnostic performance and tangible clinical benefits [1, 4].

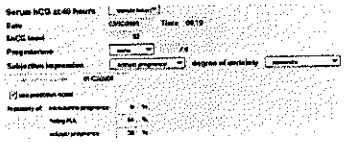


Figure 3. Implementation of a diagnostic model.

References

[1] D. Bates, G. Kuperman, S. Wang, et al. Ten commandments for effective clinical decision support: making the practice of evidence-based medicine a reality. *J Am Med Inform Assoc*, 10:523–530, 2003.

[2] G. Condous, D. Timmerman, S. Goldstein, et al. Pregnancies of unknown location: consensus statement. *Ultrasound Obstet Gynecol*, 28:121–122, 2006.

[3] B. Van Calster. *Predictive diagnostic models for gynecologic applications with focus on multi-class classification*. PhD thesis, Dept. of Electrical Engineering, Katholieke Universiteit Leuven, Belgium, 2008.

[4] J. Wyatt and D. Altman. Prognostic models: clinically useful or quickly forgotten? *BMJ*, 311:1539–1541, 1995.