# Web-based Data Collection for Uterine Adnexal Tumors: A Case Study

Stein Aerts Peter Antal Dirk Timmerman* Bart De Moor
Yves Moreau

Department of Electrical Engineering (ESAT-SCD)
Katholieke Universiteit Leuven, Belgium
stein.aerts@esat.kuleuven.ac.be
http://www.esat.kuleuven.ac.be/~dna/BioI/

*Department of Obstetrics and Gynecology
University Hospitals Leuven, Belgium
dirk.timmerman@uz.kuleuven.ac.be

**Abstract**

*We have developed a web application for the collection of EPR's (Electronic Patient Records) from uterine adnexal masses pre-operatively examined with transvaginal ultrasonography. The application has been used intensively since November 2000 by 9 of the 19 international centers that joined the International Ovarian Tumor Analysis (IOTA) consortium. At the moment of writing, the IOTA database contained 68 parameters for 1150 masses. Here we report the design and implementation of the generic web-based clinical data entry system and describe the advantages and drawbacks that we have experienced while developing, using, and maintaining the system. The data model, the user interface, the help system, the constraints (mandatory/optional), and the quality checking were all based on the medical protocol created by the IOTA consortium. The data collection system has become an open and transparent implementation of the formalized protocol. It covers the complete path of the patient data from the clinical situation to the finalized database. This approach provides new types of possibilities for the data analysis since all aspects of the data collection are documented and formally available to the data analyst. The IOTA website can be found at https://www.iota-group.org which also serves as the entry point for the secure EPR application.*

## 1. Introduction

The ability to characterize persistent cystic ovarian tumors is important because of the interest in the conservative management of benign ovarian cysts, as well as in the selection of patients for less invasive surgical techniques. With the combined use of transvaginal ultrasonography (TVS) and color Doppler imaging, it is possible to identify neovascularization in new tissue growth. If other available information (for example, medical history) is included, the accuracy of the subjective assessment increases even more [14]. Subjective evaluation of an ovarian mass can however only be an accurate method for discriminating

between benign and malignant masses if the ultrasonographer is experienced. The use of mathematical approaches to predict malignancy can therefore be of value in helping less experienced operators to obtain better overall diagnostic accuracy [14]. Several studies have investigated the use of morphological scoring systems, multiple logistic regression, neural networks, and Bayesian networks [3, 13, 6, 8, 10, 11, 16]. However, a detailed review of the literature revealed considerable variation in the diagnostic accuracy of test procedures [12] and diagnostic algorithms derived from the retrospective analysis of data in one center do not necessarily produce comparable results when used prospectively in another center [4]. Both findings might be explained, at least in part, by differences in the interpretation and usage of terms and definitions of the diagnostic end-points. Consequently, a new initiative was started to address the problem, which involved the participation of researchers from different centers; the participants comprise the International Ovarian Tumor Analysis (IOTA) group [15]. The consensus opinion (i.e., the IOTA protocol) that arose is used in the multicentric study to collect a large amount of data of adnexal masses (currently 68 parameters including TVS, color Doppler imaging results, medical history, CA 125 values, and subjective assessment) to develop more accurate mathematical models for the pre-operative classification of malignant and benign ovarian tumors.

The paper is organized as follows: Section 2 describes how the formalization of the protocol connects data entry and data analysis; in Section 3 describes all steps of the data flow together with data quality issues; Section 4 covers the technical implementation of the system. Section 5 presents possible analysis approaches. The last section is a conclusion of this work.

## 2. Formalized protocol links data collection and data analysis

The 25-pages protocol [15] contains the following information: (1) general information about the project, (2) inclusion and exclusion criteria for patient records, (3) a description of each variable with its format, its value list, mandatory/optional constraints, and possible intervariable dependency rules, (4) the grouping of variables into sections (family history, medical history, ultrasonography, serum tumor marker, subjective assessment, histopathology and staging), and (5) the diagnostic methods for all tumor variables together with self-explaining figures. This protocol served as a user requirement specification for the development of the EPR system: the backbone of the graphical user interface (GUI), the input formats of the variables, the data model, and the validation rules were all derived from the protocol. The data collection and the data analysis are brought closer together by using an integrated architecture. This led to an efficient implementation and maintenance of the data collection system and to an increased openness of all aspects of the data collection.

The open and transparent implementation of the data collection system results in formal descriptions that can be used as a reference for the data analysts. It means that the medical description, the details of submission, and even the checking of the variables are efficiently available in the data analysis phase. In other words, the use of open methods for formalizing the protocol allows the data entry application and the data analysis to rely on the same underlying medical expert knowledge and on the same technical and practical constraints.

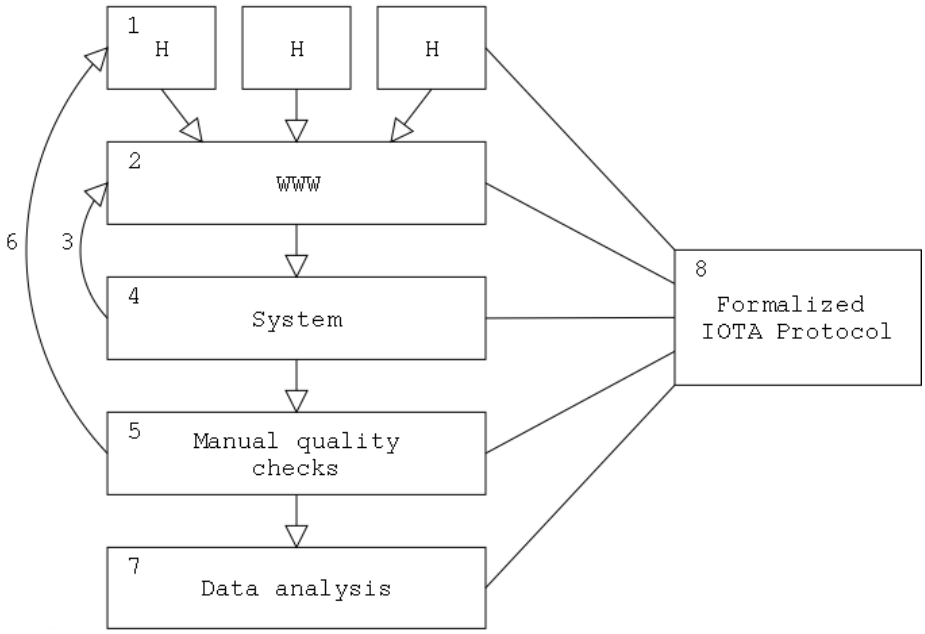


**Figure 1. Data flow from hospital (H) to the system and further to the data analysts. Loops ensure data quality. See text for the description.**

## 3. Data flow and data quality

Figure 1 represents the complete data flow. Clinicians of the participating hospitals use an internet browser to enter patient data into a HTML form (nr 1 on Figure 1) [1]. Compared to paper-based records, electronic record systems provide many advantages including global access, fast interaction, increased availability, improved legibility, long-term accessibility, greater completeness, data encoding, and automated decision support and analysis [9]. These advantages have lead to the choice of a WWW-based system for the IOTA data collection (nr 2). Based on the identifier of the center, the patient birth date, the date of ultrasound scan, and the follow-up number, a unique patient ID is created. These identifiers allow the retrieval of the patient data back from the database to update the data or to print a report. The user can request information on each variable and can consult an extended manual. Without recording the name of the patient, all data is encrypted (using a 56-bit SSL certificate) and submitted to the server. If some of the mandatory variables are missing or if the dependency rules are not fulfilled, the user is requested to correct his/her mistakes (nr 3). Only a fully compliant patient record that is confirmed by the user is entered into the database (nr 4). Although participating clinicians were initially sceptic about this restriction, the constraints appeared not too tight in the end. Even extra constraints were added in production phase on request of several users.

Each month a text file is exported from the database and sent to two experts who manually check the data set for consistency (nr 5). If problems have occurred, an e-mail is sent to the clinician requesting an update of the data (again using the web application)(nr 6). For the present 1150 masses in the database, between 50 and 100 e-mails have been sent back to the research centers. The fault-proof data is then available to the data analysts who

again check the data set in a more automated way by performing basic statistical analyses and investigations (nr 7). The protocol is used in each step of the flow: to construct the help system, to construct the HTML entry form, to check the correctness and completeness of the variables, and to define the variables in the data analysis (e.g., nominal values, thresholds ...)(nr 8).

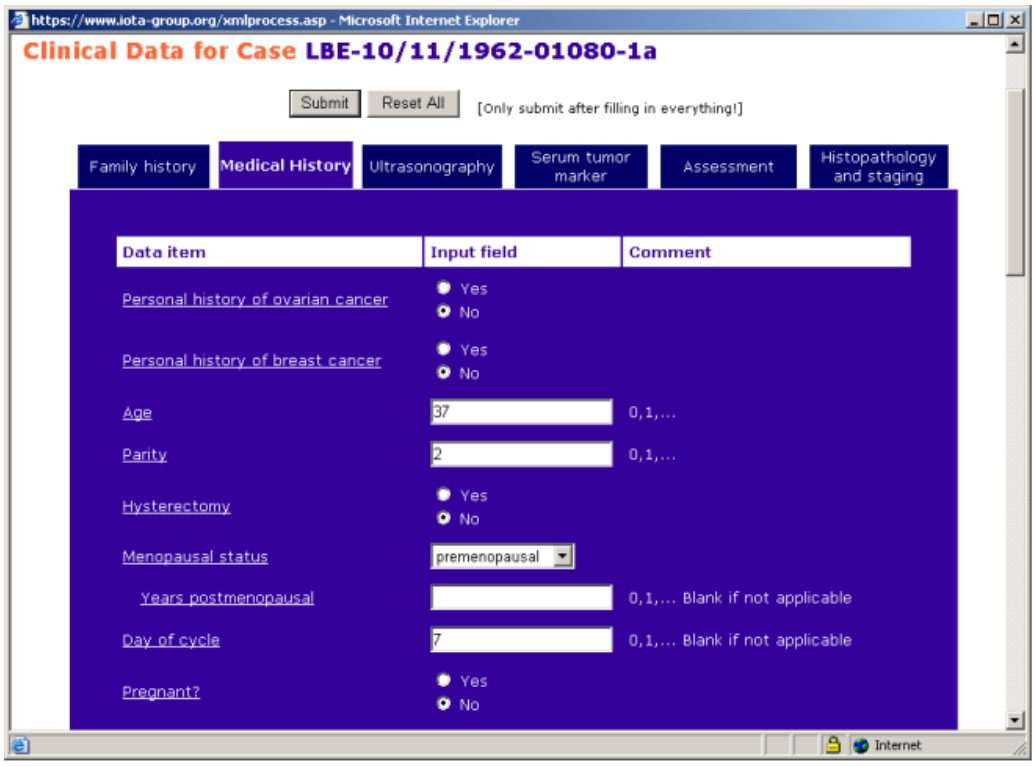


**Figure 2. Screenshot of the EPR data entry system. The HTML form is generated using XML/XSL.**

## 4. Technical implementation of the system

The three-tier EPR application has the following layers:

1. User interface: HTML pages and Javacript in a web browser (client side) and Active Server Pages (ASP) and XML for the presentation logic (server side). The complete HTML form for data entry is encoded in XML containing the type of input (text box, radio button, dropdown, or checkbox), the list of possible values, and a mandatory/optional specification. When the user requests the input form, the XML file is processed on the server using XSL (eXtensible Stylesheet Language) to generate HTML and Javascript that is sent to the client browser (see Figure 2). This design is open (new variables can be added to the XML file) and generic (a new XML file can be generated from another protocol). This allows us to deploy a similar EPR application in another field of medicine with another protocol and other variables, or to deploy several case reporting 'modules' within the same framework.

2. Business logic: ASP for the general business logic and XML for the validation rules. The ASP code forms the backbone of the application and performs the database access. All dependencies among variables are encoded in XML and parsed with ASP.
3. Database: The data model is generic as well since the clinical variables are entered as key-value pairs. An extra variable would result in extra records, not in the adjustment of the database tables. For moderately complex cases like this one, the key-value approach is still feasible and again results in the possibility to extend the application towards multiple clinical modules. The data model has been implemented using MS Access.

The IOTA web application approaches security from two levels: (1) network security is implemented using standard SSL (Secure Socket Layers) and (2) access to the application is restricted by user/password authentication.

## 5. Data analysis

After collecting a significant amount of validated patient data, several mathematical models can be applied on it. The question whether existing models based on logistic regression [13], neural networks [16], and Bayesian networks [3] will perform well on this upcoming huge data set is intriguing because of the participation of multiple centers. The use of Bayesian Networks [7], and specifically the Annotated Bayesian Networks (ABNs) [2], are particularly well suited for the integration of the data collection process, the incorporation of domain knowledge and the statistical data analysis in the same consistent framework defined by the IOTA protocol. Note, that these complex statistical models can be similarly described in XML-based formats [5] that is used to describe the data model and the collection process/interface itself.

## 6. Discussion and conclusion

The number of EPR applications that are web-based and that are used in an intensive multi-center study such as the IOTA study is still limited. However, the necessary effort to develop the application resulted in a reduced manual effort to collect, inspect, and encode the data. Secondly, participating clinicians are fairly satisfied with the ease and speed of data entry. Thirdly, the data quality we have achieved would be hard to achieve using paper records in an international multi-center setup and potential further improvements (checks) can be easily integrated in the system. The feedback on the IOTA web application has thus been largely positive. The main objection was from centers where an own EPR is already in use. The complaints were sporadic unavailability and the fact that certain redundant data of a patient has to be entered again for a second tumor mass of the same patient. The preliminary analysis of the submitted data indicates that the selected architecture covering the path of the data from the clinical environment to the final data provides new possibilities for the data analysis. For example, the formal, semantic data model of the collection system allows the semi-automated definition of certain statistical models. In the data analysis phase it also provides detailed information about the factors that influenced the submitted data.

## 7. Acknowledgements

Dr. Bart De Moor is a full professor at the Katholieke Universiteit Leuven, Belgium. Stein Aerts is a Research Assistants with the K.U.Leuven. Supported by grants from: Research Council KUL: Concerted Research Action GOA-Mefisto 666 (Math.Eng.), IDO (IOTA Oncology, Genetic networks), Flemish Gvt.: Fund for Scientific Research Flanders (projects G.0256.97, G.0115.01, G.0240.99, G.0197.02, G.0407.02 research communities ICCoS, ANMMM), AWI (Bil. Int. Collaboration Hungary BIL2000/19 ), IWT (Soft4s), STWW-Genprom,GBOU-McKnow Eureka-Impact, Eureka-FLiTE,); Belgian Fed. Gvt: DWTC (IUAP IV-02 (1996-2001) and IUAP V-10-29 (2002-2006)), Program Sustainable Development PODO-II (CP-TR-18); Direct contract research: Verhaert, Electrabel, Elia, Data4s, IPCOS.

## References

[1] S. Aerts, D. Timmerman, and Y. Moreau. Official website of the IOTA project with general information and access to the IOTA data entry system. *Internet Web site: http://www.iota-group.org*, November 2000.

[2] P. Antal, T. Meszaros, B. De Moor, and T. Dobrowiecki. Annotated bayesian networks: a tool to integrate textual and probabilistic medical knowledge. In *Proc. of the 13th IEEE Symp. on Comp.-Based Med.Sys. (CBMS01)*, pages 177–182, 2001. Bethesda, MD.

[3] P. Antal, H. Verrelst, D. Timmerman, S. Van Huffel, B. De Moor, and I. Vergote. Bayesian networks in ovarian cancer diagnosis: Potentials and limitations. *13th IEEE Symposium on Computer-Based Medical Systems*, pages 103–108, June 2000.

[4] N. Aslam, S. Banerjee, J.V. Carr, M. Savvas, R. Hooper, and D. Jurkovic. Prospective evaluation of logistic regression models for the diagnosis of ovarian cancer. *Obstet Gynecol*, 96:75–80, 2000.

[5] R. Cover. Xml belief network file format. *Internet Web site: http://www.oasis-open.org/cover/xbn.html and http://research.microsoft.com/dtas/bnformat/default.htm*, April 1999.

[6] S. Granberg, M. Wikland, and I. Jansson. Macroscopic characterization of ovarian tumors and the relation to the histological diagnosis: criteria to be used for ultrasound evaluation. *Gynecol Oncol*, 35:139–144, 1989.

[7] D. Heckerman. Learning bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20:197–243, 1995.

[8] I. Jacobs, D. Oram, J. Fairbanks, J. Turner, C. Frost, and J.G. Grudzinskas. A risk of malignancy index incorporating ca 125, ultrasound and menopausal status for the accurate preoperative diagnosis of ovarian cancer. *Br J Obsted Gynaecol*, 97:922–929, 1990.

[9] S.N. Luxenberg, D.D. Dubois, C.G. Fraley, R.R. Hamburgh, X.L Huang, and P.D. Clayton. Electronic forms: benefits and drawbacks of a www-based approach to data entry. *Proc AMIA Annu Fall Symp*, pages 804–808, 1997.

[10] A.M. Sassone, I.E. Timor-Tritsch, A. Artner, C. Westhoff, and W.B. Warren. Transvaginal sonographic characterization of ovarian disease: evaluation of a new scoring system to predict ovarian malignancy. *Obstet Gynecol*, 78:70–76, 1991.

[11] A. Tailor, D. Jurkovic, T.H. Bourne, W.P. Collins, and S. Campbell. Sonographic prediction of malignancy in adnexal masses using multivariate logistic regression analysis. *Ultrasound Obstet Gynecol*, 10:41–47, 1997.

[12] D. Timmerman. *Ultrasonography in the assessment of ovarian and tamoxifen-associated endometrial pathology.* PhD thesis, Leuven, Belgium, 1997.

[13] D. Timmerman, T.H. Bourne, A. Tailor, W.P. Collins, H. Verrelst, K. Vandenberghe, and I. Vergote. A comparison of methods for the pre-operative discrimination between benign and malignant adnexal masses: the development of a new logistic regression model. *Am J Obstet Gynecol*, 181:57–65, 1999.

[14] D. Timmerman, P. Schwärzler, W.P. Collins, F. Claerhout, M. Coenen, F. Amant, and I. Vergote. Subjective assessment of adnexal masses with the use of ultrasonography: an analysis of interobserver variability and experience. *Ultrasound Obstet Gynecol*, 13:11–16, 1999.

[15] D. Timmerman, L. Valentin, T.H. Bourne, W.P. Collins, H. Verrelst, and I. Vergote. Terms, definitions and measurements to describe the sonographic features of adnexal tumors: a consensus opinion from the International Ovarian Tumor Analysis (IOTA) Group. *Ultrasound Obstet Gynecol*, 16:500–505, 2000.

[16] D. Timmerman, H. Verrelst, T.H. Bourne, B. De Moor, W.P. Collins, I. Vergote, and J. Vandewalle. Artificial neural network models for the preoperative discrimination between malignant and benign adnexal masses. *Ultrasound Obstet Gynecol*, 13:17–25, 1999.