

Research paper

The use of Data Mining in the categorization of patients with Azoospermia

Themistoklis Mikos¹, Konstantinos Pantazis¹, Dimitrios G. Goulis¹, Nikolaos Maglaveras², John N. Bontis¹, John Papadimas¹

¹Unit of Reproductive Endocrinology, 1st Department of Obstetrics & Gynaecology, Aristotle University of Thessaloniki, Greece, ²Laboratory of Medical Informatics, Aristotle University of Thessaloniki, Greece

ABSTRACT

OBJECTIVE: Data Mining is a relatively new field of Medical Informatics. The aim of this study was to compare Data Mining diagnosis with clinical diagnosis by applying a Data Miner (DM) to a clinical dataset of infertile men with azoospermia. **DESIGN:** One hundred and forty-seven azoospermic men were clinically classified into four groups: a) obstructive azoospermia (n=63), b) non-obstructive azoospermia (n=71), c) hypergonadotropic hypogonadism (n=2), and d) hypogonadotropic hypogonadism (n=11). The DM (IBM's DB2/Intelligent Miner for Data® 6.1) was asked to reproduce a four-cluster model. **RESULTS:** DM formed four groups of patients: a) eugonadal men with normal testicular volume and normal FSH levels (n=86), b) eugonadal men with significantly reduced testicular volume (median 6.5 cm³) and very high FSH levels (n=29), c) eugonadal men with moderately reduced testicular volume (median 14.5 cm³) and raised FSH levels (n=20), and d) hypogonadal men (n=12). Overall DM concordance rate in hypogonadal men was 92%, in obstructive azoospermia 73%, and in non-obstructive azoospermia 69%. **CONCLUSIONS:** Data Mining produces clinically meaningful results but different from those of the clinical diagnosis. It is possible that the use of large sets of structured and formalised data and continuous evaluation of DM results will generate a useful methodology for the Clinician.

Key words: Azoospermia, Data Mining, Decision support systems, Knowledge discovery in databases, Male infertility

INTRODUCTION

Approximately 15% of couples are unable to

Address all correspondence to:

Dr. Dimitrios G. Goulis, 1st Department of Obstetrics & Gynaecology, "Papageorgiou" General Hospital, Periferiaki Odos, Nea Efkarpia, 564 03, Thessaloniki, Greece, Tel.: (+30) 2310 693.384, Fax: (+30) 2310 991.510, e-mail dgg30@otenet.gr

Received 21-09-05, Revised 20-10-05, Accepted 27-10-05

conceive after one year of unprotected intercourse. A male factor is solely responsible in about 20% of infertile couples and contributory in another 30-40%¹. The World Health Organisation suggests specific diagnostic algorithms for the classification and treatment of male infertility². Men with azoospermia constitute a special group of male infertility. Azoospermia is defined as the complete absence of sperm from the ejaculate, even after centrifugation.

It is a laboratory finding and not a diagnosis; therefore it can be the result of a wide spectrum of pathophysiological conditions that call for different therapeutic approaches.

Knowledge discovery in databases is the computing process of finding useful hidden information in large datasets³⁻⁵. Data Mining is one step during the process of knowledge discovery. It consists of mining algorithms which reproduce a set of patterns that can be further applied to other databases. Statistics, machine learning, database techniques, pattern recognition, and optimization techniques are the main tools of Data Mining³⁻⁵.

The aim of this study was to compare Data Mining diagnosis with clinical diagnosis by applying a Data Miner (DM) to a clinical dataset of infertile men with azoospermia.

SUBJECTS AND METHODOLOGY

Clinical notes of 3000 infertile men evaluated at the outpatient clinic of the Unit of Reproductive Endocrinology from 1990 to 2004 were studied retrospectively. Of these subjects, 204 men with azoospermia were identified (6.8%). Full history was retrieved from 147 of these men. A database consisting of eleven clinical and laboratory parameters was constructed: 1) age, 2) referral reason (presenting complaint), 3) personal andrologic history, 4) presence of secondary sex characteristics, 5) mean testicular volume, 6) presence of varicocele, 7) follicular stimulating hormone (FSH) serum levels, 8) luteinizing hormone (LH) serum levels, 9) testosterone (T) serum levels, 10) sperm volume and 11) cytopathology findings from testicular fine needle aspiration biopsies (FNA).

The cohort of azoospermic men was clinically classified into four subgroups: a) eugonadal men with obstructive azoospermia (n=63, 42.8%), b) eugonadal men with non-obstructive azoospermia (n=71, 48.3%), c) hypergonadotropic hypogonadal men (n=2, 1.4%) and d) hypogonadotropic hypogonadal men (n=11, 7.5%) (Table 1).

Secondary sexual characteristics, mean testicular volume, serum levels of FSH, LH and T were clinically considered as primary factors for Data

Table 1. Clinical classification of azoospermia.

	n	%
Hypogonadism	13	8.9
Hypergonadotropic (primary testicular failure)	2	1.4
Hypogonadotropic (secondary testicular failure)	11	7.5
Eugonadism	134	91.1
Non-obstructive azoospermia (impaired spermatogenesis)	71	48.3
Obstructive azoospermia	63	42.8
Total	147	100.0

Mining, whereas the rest of the parameters were considered as secondary factors for the computing process. Cytopathology findings were excluded from further analysis because of the significant number of missing data (34/147, 23%).

The data from the patient files were manually entered in a computer database (Microsoft Excel for Windows®, Microsoft Office 2003). Finally, the database was transferred from Excel to a DataBase2 file so that it be compatible with the DM (DB2/Intelligent Miner for Data®, IBM, version 6.1).

DM was specifically programmed to form four subgroups of men (clustering model with limited number of produced groups). By this means, comparison with the four clinical groups would be more feasible.

RESULTS

The Data Mining results are presented in Table 2. The four-cluster model produced four groups of azoospermic men.

The first group (Group 1, n=86, 58.5% of the study population) consisted of eugonadal men with normal testicular volume (median 22.0 cm³) and normal serum levels of FSH, LH and T (8.1 IU/l, 1.5 IU/l, and 17.0 nmol/l, respectively). During the Data Mining, 59 men with obstructive azoospermia (93% of the total men clinically categorized as obstructive azoospermia), 26 men with non-obstructive azoospermia, and one man with hypogonadot-

Table 2. Intelligent Miner for Data Mining results: The four-clustering model.

	Prediction from Data Mining				Total
	Group 1	Group 2	Group 3	Group 4	
Parameters studied*					
Mean testicular volume (ml)	22.0 (6.0-25.0)	6.5 (0.0-12.0)	14.5 (7.0-23.0)	7.0 (1.0-25.0)	
FSH (IU/l)	8.1 (0.3-60.6)	26.4 (1.7-80.0)	18.6 (3.4-42.0)	6.3 (0.4-31.0)	
LH (IU/l)	1.5 (0.0-4.0)	15.3 (5.0-32.0)	12.0 (0.7-65.0)	7.2 (0.1-44.5)	
Testosterone (nmol/l)	17.0 (2.0-30.8)	13.6 (1.0-31.0)	19.2 (5.5-31.9)	4.6 (0.2-19.1)	
Sperm volume (ml)	2.8 (0.2-7.0)	3.3 (0.4-5.0)	3.9 (1.8-6.5)	2.1 (1.0-3.5)	
Clinical classification**					
Obstructive azoospermia	59	0	4	0	63
Non-obstructive azoospermia	26	29	16	0	71
Hypergonadotropic hypogonadism	0	0	0	2	2
Hypogonadotropic hypogonadism	1	0	0	10	11
Total	86	29	20	12	147

* Data are given as median (range). To convert serum testosterone levels from nmol/l to ng/ml multiply by 0.228.

** Absolute number of patients allocated to each "clinical classification" or "four-clustering model" group.

ropic hypogonadism were allocated in this group.

The second group (Group 2, n=29, 19.7% of the study population) consisted of eugonadal men with decreased testicular volume (median 6.5 cm³), raised FSH and LH serum levels, and normal serum T levels (26.4 IU/l, 15.3 IU/l, and 13.6 nmol/l, respectively). All men allocated in this group had non-obstructive azoospermia according to the clinical classification.

The third group (Group 3, n=20, 13.6% of the study population) consisted of eugonadal men with moderately decreased testicular volume (median 14.5 cm³), slightly raised FSH and LH serum levels, and normal serum T levels (18.6 IU/l, 12.0 IU/l and 19.2 nmol/l, respectively). Four men clinically classified as obstructive azoospermia and 16 men as non-obstructive azoospermia were allocated into this group.

Finally, the fourth group (Group 4, n=12, 8.2% of the study population) consisted of hypogonadal men with decreased testicular volume (median 7.0 cm³), normal FSH and LH serum levels, and low serum T levels (6.3 IU/l, 7.2 IU/l and 4.6 nmol/l, respectively). Ten out of the 11 clinically classified as hypogonadotropic hypogonadal men and the two hypergonadotropic hypogonadal men were allocat-

ed into this group.

Overall, DM concordance rate in hypogonadal men was 92% (clinical: 13 patients vs. DM - Group 4: 12 patients), in obstructive azoospermia 73% (clinical: 63 patients vs. DM - Group 1: 87 patients), and in non-obstructive azoospermia 69% (clinical: 71 patients vs. DM - Groups 2 and 3: 49 patients).

DISCUSSION

The overall performance of the DM on the andrology database was satisfactory from a technical point of view. The DM ran flawlessly over the database and reproduced a four-cluster model that could be compared with the clinical classification of the azoospermic men⁶. A successful Data Mining procedure would classify the azoospermic men in a way similar to that of the clinical diagnosis. The DM successfully formed the groups of men with hypergonadotropic hypogonadism (Group 4) and obstructive azoospermia (Group 1). Regarding the men with non-obstructive azoospermia (Groups 2 and 3), there was only partial agreement of the Data Mining results with the clinical classification. Moreover, the DM's classification was generally characterised by low sensitivity as compared to the clinical classification, with the exception of Group 4.

An instance of significant success of the DM would be the exclusive inclusion of eugonadal men with non-obstructive azoospermia in Group 2. Nevertheless, less than 40% (29 of 71) of the total eugonadal men with non-obstructive azoospermia were included in this group; thus the classification was characterised by high sensitivity but low specificity.

In Group 3 of the Data Mining included four eugonadal men with obstructive azoospermia and 16 eugonadal men with non-obstructive azoospermia. The inclusion of these two subgroups of men into the same DM group has no clinical relevance. The same also applies to Group 1.

Finally, Group 4 of the Data Mining included two hypergonadotropic hypogonadal men and 10 of the 11 hypogonadotropic hypogonadal men. This group of patients was fairly similar to the clinical grouping of the hypogonadal men. The error of the Data Mining was the allocation of both the hypergonadotropic and the hypogonadotropic hypogonadal men into the same group.

In spite the above mentioned discrepancies, the ability of the Intelligent Miner for Data to reproduce clinically justifiable results is obvious. Each one of the groups that were formed by the Data Miner consisted of men with similar clinical features. A likely explanation for the differences between the clinical and DM classification is the weakness of the mining model to conclusively use the secondary clinical parameters included in the database.

It is true, however, that the database that was used for the Data Mining process was rather small. Commercial Data Miners extract useful new information from datasets of thousands or millions of records⁷. The relatively small size of the cohort that was used in this study was definitely a limitation for the reliable interpretation of the results. The overall impression however is that the application of the DM in a larger dataset would further improve its performance and data interpretation.

The most important step during Data Mining in databases is the analysis and the interpretation of the results. DMs produce numerous algorithms, models, correlations, prognostics, and decision trees, all of which have to be clinically assessed and vali-

dated. The clinician should initially go through this time-consuming evaluation and then choose the application that would be functional in the clinical situation one has to deal with⁸.

It seems that certain prerequisites must be fulfilled for a DM to be able to offer clinical assistance. A DM can successfully operate only in complete and accurate patient databases. The clinician could then use the DM to get help in decision support systems, research projects, or risk management issues (i.e. in the discovery of associations that could assist in the prevention of health-professional errors)^{3,5,9}.

To date, Data Mining has been sparsely used in the field of Andrology. Dzeroski et al reported a database consisting of clinical details from infertile men that was analysed with the use of a DM^{10,11}. The researchers investigated 382 infertile men by using 177 gene markers, in order to detect Y-chromosome microdeletions and establish a correlation between genotype and phenotype. Clustering and decision-tree making were the main Data Mining tools that were used in this study. Data Mining suggested a group of 13 gene markers that were responsible for more than 90% of the infertile male phenotype. The authors, however, did not try to compare clinical with Data Mining findings¹⁰. In the present study, the main aim was the comparison of the Data Mining findings with the clinical classification.

To the best of our knowledge, this is the first attempt to clinically validate the use of Data Mining in azoospermia. Our data suggest that, given the Data Mining tools we used in this study, the clinical and laboratory data available are already sufficient for the experienced andrologist to reach a diagnosis, and further assistance from a computer does not optimize diagnostic ability. Nevertheless, DM could have a potential role in education and in supporting decisions of trainees.

In this dataset, making a diagnosis without human involvement did not appear feasible. Given the poor performance of the DM in the learning set, any further examination on a new group of patients would most likely give results of equally low value. Finally, a potential application of the DM would be to unravel a hidden layer of data, i.e. overlooked pathophysiological connections. However, this ap-

plication does not appear achievable, as the computer algorithm is proprietary and hidden from the user. Further testings required to determine the clinical significance of the four computer generated groups.

In conclusion, in this study a database of azoospermic men was initially built and DM was used subsequently to extract useful clinical information. The Data Mining results were compared to the clinical classification of the patients. The comparison suggests that DM in its present form is not currently capable of consistently helping the clinician in this particular entity. The small sample size may be an important limitation. It is quite possible that Andrology, Assisted Reproduction Techniques (ART), gene therapy, and molecular biology are fields that could benefit from the use of DMs. It remains to be proven, however, whether these hi-tech products will in the future secure a place in everyday clinical practice.

REFERENCES

1. The Male Infertility Best Practice Policy Committee of the American Urological Association and The Practice Committee of the American Society for Reproductive Medicine. 2004 Report on evaluation of the azoospermic male. *Fertility & Sterility* 82: Suppl 1: S131-S136.
2. WHO Manual for the standardized investigation, diagnosis and management of the infertile male, 2000 Cambridge University Press.
3. Cios KJ, 2000 Medical data mining and knowledge discovery, *IEEE Eng Med Boil Mag* 19: 15-16.
4. Stilou S, Bamidis PD, Maglaveras N, Pappas C, 2001 Mining association rules from clinical databases: an intelligent diagnostic process in healthcare. *Medinfo* 10: 1399-1403.
5. van Bommel JH, Musen MA 1997 Handbook of Medical Informatics. In: Springer (ed); pp, 362.
6. Papadimas J, Papadopoulou F, Ioannidis S, et al, 1996 Azoospermia: clinical, hormonal, and biochemical investigation. *Arch Androl* 37: 97-102.
7. DB2 Intelligent Miner for Data, IBM. Web site: <http://www-306.ibm.com/software/data/iminer/>. Accessed on 27/10/2005.
8. Babic A, 1999 Knowledge discovery for advanced clinical data management and analysis. *Stud Health Technol Inform* 68: 409-413.
9. Greenes RA, 2002 Future of medical knowledge management and decision support. *Stud Health Technol Inform* 80: 29-44.
10. Dzeroski S, Hristovski D, Kunej T, Peterlin B. 2000 A data mining approach to the development of a diagnostic test for male infertility. *Stud Health Technol Informa* 77: 779-783.
11. Dzeroski S, Hristovski D, Peterlin B. 2000 Using data mining and OLAP to discover patterns in a database of patients with Y-chromosome deletions. *Proc AMIA Symp* pp, 215-219.