The use of Data Mining Techniques in Antibiotic Resistance Surveillance.

Joseph Papaparaskevas,[1] Yannis Batistakis,[2] Maria Halkidi,[2] Christos Amanatidis,[2] Maria Kanellopoulou,[3] Michalis Vazirgiannis,[2] Evangelos Papafragas,[3] and Alkiviadis C. Vatopoulos[1].

[1]Department of Hygiene and Epidemiology, Medical School, National and Kapodistrian University of Athens,

[2]Department of Informatics, Athens University of Economics and Business,

[3]Clinical Microbiology Laboratory, "Sismanoglion" General Hospital, Athens.

Address for correspondence:
Dr A.C. Vatopoulos
Dept. of Hygiene and Epidemiology, Medical School,
National and Kapodistrian University of Athens
Mikras Asias 75, GR 115 27, Athens, Greece.
Phone: ++30.10.7462071
Fax: ++30.10.7462080
E-mail: avatopou@med.uoa.gr

Key words: WHONET, data mining, Greece, antibiotic resistance, surveillance.

Abstract

In the present study, antibiotic resistance data generated in Greece by the WHONET Network were further analyzed by the use of data mining techniques. More specifically association rules were extracted among data collected in the Microbiology Dept. of "Sismanoglion" General Hospital, a 500-bed general hospital, in Athens, Greece. The data studied were the susceptibility results, as well as data concerning the patient's wards, the day of isolation and the type of clinical specimen, of a total of 20,794 bacterial isolates collected in the period January 1st 1996 to December 31st 2000,. The factors used to measure the importance of each association rule were its strength (confidence), its support, its coverage, its leverage and its lift. Two main rule categories were generated, one associating clinical specimen, time and ward of isolation, with bacterial species and the second one associating the same attributes with resistant phenotypes. The factors most often used to compare and evaluate different rules were leverage and lift. The system generated association rules in an unsupervised automatic way and revealed pieces of knowledge not easily available with standard supervised procedures of analysis, thus making it very useful in an automated public health surveillance system.

INTRODUCTION

A National Electronic Network for the continuous monitoring of antibiotic resistance is being in operation in Greece for the last five years (1) (see also *www.mednet.gr/whonet*). The network, which currently involves 20 Hospitals throughout Greece, analyses the routine data generated in the respective microbiology laboratories through the use of WHONET software, which is distributed by World Health Organization (WHO) (2). The network tabulates and reports basic annual or semi-annual summary statistics, mainly consisting of multi hospital antibiogram summaries by important bacterial species and type of ward, and resistant phenotypes by species and type of ward. Intra hospital comparisons of percent resistances are also reported.

The network has been established as an important surveillance system in Greece (1, 3), is funded by the Greek Ministry of Health (Hellenic Centre for Infectious Disease Control - KEEL) and the generated data are considered very helpful for the control of antibiotic resistance. Nevertheless, the difficulty of the system to perform analyses in real time, an important function of any surveillance network, (4) as well as to automatically identify emerging complex patterns, trends and associations of public health importance was soon recognized. Thus, the need for a new generation of tools and techniques for intelligent data analysis was evident (5).

Data mining is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data repositories (6). It is considered as the main step in the Knowledge Discovery process concerning with methodologies for knowledge extraction from large data repositories.

There are many data mining methods that are available and discussed in the literature. These methods, accomplishing a limited set of tasks, produce a particular enumeration of patterns over datasets. The main tasks, according to well-established data mining process are: i) Clustering, ii) Classification, iii) Association Rules Extraction, iv) Estimation & Prediction, v) Regression. More specifically, Association Rules Extraction techniques are among the most widely known data mining methods in order to extract interesting patterns of knowledge form large amounts of data.

In the present communication we report a feasibility study of further analyzing the data generated by the WHONET system by the use of data mining techniques.

Materials and methods.

A. The WHONET Network.

The network, which currently involves more than 20 Hospitals throughout Greece, is based on the assumption that the routine results of the antibiotic sensitivity tests performed in the hospital laboratories are the major resource for antibiotic resistance surveillance. The network operates through the establishment of a common quality control procedure and the adaptation of an electronic code and data format in all hospitals through the use of the WHONET software. The WHONET software was originally developed by the WHO Collaborating Centre for Surveillance of Antibiotic Resistance in Boston USA and further developed in the Division of Emerging and other Communicable Diseases Surveillance and Control, (EMC), WHO, Geneva, Switzerland. Details about the software are revised elsewhere (7). In each hospital all data are entered into the programme on a daily or weekly basis, either manually for hospitals using the disk diffusion method, or by automatic download with the aid of the BACLINK software (also developed in WHO/EMC), for hospitals using automatic sensitivity systems.

The hospital WHONET data files are in *.dbf* format and include one isolate (the first) per patient. More specifically, they include the following information, which are the attributes characterising each isolate: the bacterial species, the patient's ward, the date of isolation, the type of the clinical specimen, and the results of the susceptibility test.

In the present study, the data set of the "Sismanoglion" General Hospital, a 500 bed general hospital which participates in the WHONET network since January 1996, was used. Bacterial identification and susceptibility testing was performed by the PASCO MIC/ID semi-automated system (Difco Laboratories, Detroit, MI, USA). The resulting database, covering the years 1996 to 2000, consisted of 20,794 isolates. A description of this database is given on tables 1-3.

B. The Data mining background – Definitions.

An association rule reveals underlying interactions between attributes and attribute values in the data set. In our case the data set consisted of the clinical bacterial isolates included in the WHONET files.

An association rule is of the form of A, B, C,…. → Z, where A, B, C etc. are the attribute-value sets that consist the Left Hand Side (LHS) of the equation, and Z is the attribute-values set included in the Right Hand Side (RHS) of it.

The aim of association rules is to represent interesting relationships among the data under consideration. A number of approaches and methods have been proposed for rules extraction and their main idea of these techniques is based on the concept of frequent itemsets. Moreover five different "interestingness measures" are used in order to evaluate the importance of each association rule. These are:

1. *The association rule* ***strength*** *(confidence).* The strength of an association rule is the proportion of the isolates that fulfil the LHS of the rule that are also fulfilling the RHS.
2. *The association rule* ***support****.* The support of an association rule is the proportion of isolates fulfilling both LHS and RHS among the total number of isolates.
3. *The association rule* ***coverage****.* The coverage of an association rule is the proportion of isolates in the data that have the attribute values or items specified on the LHS of the rule.
4. *The association rule* ***leverage****.* The leverage of an association rule is the proportion of additional isolates covered by both the LHS and RHS above those expected if the LHS and RHS were independent of each other.
5. *The association rule* ***lift****.* The lift of an association rule is the strength divided by the proportion of all isolates that are covered by the RHS.

C. The Data Mining Procedure.

1. *Pre-processing step*.

For the purpose of the present study, the WHONET files were send from "Sismanoglion" General Hospital to the coordinating center, where they were extracted via the WHONET software in EXCEL format, and were submitted into a cleaning and correcting process. In that process entries with mistakes or missing attributes were omitted (in our case this was about 2% of the whole database). Moreover, in an attempt to categorize time into time periods useful for analysis, date of isolation was transformed into the number of weeks, one month periods, two month periods and three month periods from January 1st, 1995, which represents the conventional date the WHONET network was launched in Greece. These numbers were incorporated into the data set as new attributes. Lastly, each isolate of the bacterial species with known clinical importance (*Staphylococcus aureus, Klebsiella pneumoniae, Enterobacter* spp*, Enterococcous* spp*, Pseudomonas aeruginosa* and *Acinetobacter baumanii)* was allocated into a subgroup according to its resistant pattern and a new attribute called Phenotype was generated and incorporated in the data set (Table 4).

Wards, specimen types and time periods were included in the LHS of the model, whereas species or phenotypes were allocated into the RHS of it.

2. *Rules extraction.*

For the association rules extraction step the techniques supported by MAGNUM OPUS V1.1 software (G.I. Webb & Assoc, RuleQuest Research Pty Ltd, 30 Athena Avenue, St Ives NSW 2075, Australia) were used. The software discovers rules based on user defined thresholds with a defined minimum for coverage (≥ 0.0), leverage (≥ 0.001) and lift (≥ 1.0). The output is a text file in which the extracted rules and their related interestingness measures are presented. Based on a PERL script (*www.training.perl.com*), the text files were converted and imported into EXCEL software (Microsoft Corporation, One Microsoft

Way, 98052, Redmond, WA, USA), sorted by leverage in descending order and evaluated for possible novel and interesting patterns and trends. The *autoFilter* function of the EXCEL software was used for quick grouping and reviewing. The results of this process were also integrated with other application analysis results and previous knowledge in order to draw the right conclusions. Trivial associations are discarded by “Magnum Opus”. In the appendix the fundamental concepts of the rule mining process are discussed in more detail.

RESULTS: EXTRACTING RULES - AN EXPERIMENTAL STUDY

Preliminary experiments using the various time intervals revealed that the two month period was the optimal period to observe secular differences as well as differences among wards and specimens. Lesser time intervals included too few cases for associations to be detected, whereas longer time periods were less sensitive for prompt recovery of new patterns (data not shown).

Two main experiments, one with bacterial species and the second with phenotypes on the RHS of the rules were performed and a total of 269 and 291 association rules were recovered, respectively. The prevalence of isolates satisfying these rules in the database varied considerably from a support of 0.1% (20 isolates) up to 15% (3000 isolates). Moreover, the leverage ranged from 20 up to 1600 additional cases (data not shown).

Although and as expected, most of these rules revealed obvious associations with no added value from the public health point of view, interesting patterns were also detected. The assessment of these rules was based on their leverage in association with coverage, strength and support.

In table 5, examples of rules associating lower respiratory tract specimens (LRT), respiratory wards and the time period (two-month period) on the LHS, with bacterial species on the RHS, are shown. The first rule of the table associates LRT specimens from the 1st Respiratory ward of the hospital with isolation of *Stenotrophomonas maltophilia.* The various factors of this rule are its coverage (0.040), indicating that 4% of the total database consists of LRT specimens from 1st Respiratory ward (834 cases), of which 8.2% are *S. maltophilia* (strength). The whole association rule (*S. maltophilia* isolates coming from LRT specimens from the 1st Respiratory ward) represents the 0.32% of the whole database (support). Both these factors indicate that *S. malophilia* is rather rare in the database. On the other hand, the leverage of this rule, which is the highest among rules associating LRT specimens from the 1st Respiratory ward with any species, is 0.21 corresponding to 44 cases. This figure indicates the additional cases of *S. maltophilia* isolated from LRT specimens, in the 1st Respiratory ward, than those expected if there were no association these attributes, revealing that the rate of isolation of this species in this ward should be further evaluated. Moreover, lift is also a measure of the importance of this association rule, indicating that *S. malophilia* is 2.87 times more often found in LRT specimens coming from the 1st Respiratory ward, than in the whole data base.

Interestingly, although *Klebsiella oxytoca* is recovered from LRT specimens in the 1st Respiratory ward more often than expected (Table 5, rule 5, leverage 0.0011, 22 additional cases), but in a lesser extend than *S. maltophilia* in the previous rule, the risk of isolating *K. oxytoca* in LRT specimens in the 1st Respiratory ward is higher than that for *S. maltophilia* (lift 3.25).

In addition differences were detected between the three main Respiratory wards of the hospital (1st, 5th and 7th), with respect to the species most often associated, which were *S. maltophilia, P. aeruginosa* and *Pseudomonas* spp., respectively (Table 5, rules 1, 7 and 16, respectively). In the 5th and 7th Respiratory wards, the only enterobacterium present in association rules are *Serratia marcescens* and *Enterobacter cloacae*, respectively (Table 5, rules 11 and 19), whilst in 1st Respiratory ward *K. pneumoniae, K. oxytoca* and *E. cloacae* are detected (Table 5, rules 4,5 and 6, respectively). Furthermore, only in the 5th Respiratory ward a gram-positive species (*Staphylococcus aureus*) is recovered in association rules (Table 5, rule 9). Interestingly in 5th Respiratory ward, rules associating *P. aeruginosa* with LRT specimens were generated for the two-month periods January - February 1999 and May – June, July – August and September - October 2000 (Table 5, rules 12 – 15, respectively). These associations may reveale a possible epidemic increase in the isolation rate of *P. aeruginosa*.

In table 6, rules associating urinary tract (UT) specimens from the three Internal Medicine wards are displayed. The very low coverage in the Academic Ward (Table 6, rule 9) is due to the fact that this ward was established in the hospital only two years ago. Understandably *E. coli* is present in rules in all internal medicine wards (Table 6, rules 1, 5 and 9). Interestingly although *Enterococcus faecalis* is present in rules from all wards (Table 6, rules 2 and 6), *E. faecium* is associated with UT specimens coming only from the 1st (Table 6, rule 3) but not from the 2nd ward. Similarly *K. pneumoniae*, is present in rules in the 2nd (Table 6, rule 8) but not in 1st ward. Additionally *P. mirabilis*, although present in both wards (Table 6, rules 4 and 7), has a leverage rate in 2nd ward more than twice the one in 1st (0.0031 and 0.0014 respectively).

In table 7 rules associating *K pneumoniae* phenotypes with wards and clinical specimens are presented. It can be seen that *K. pneumoniae* susceptible to tobramycin and ciprofloxacin is associated with UT specimens from 2nd Internal Medicine ward (Table 7, rule 4), and from LRT specimens from the 1st Respiratory ward (Table 7, rule 7). On the other hand multiresistant (resistant to tobramycin and ciprofloxacin) *K. pneumoniae* is associated with LRT, UT and i.v. catheters specimens form the ICU (Table 7, rules 10, 11 and 13, respectively).

Similarly, on table 8, it can be seen that both *E. faecalis* and *E. faecium* susceptible to streptomycin and *E. faecalis* highly resistant to streptomycin are associated with 1st and 2nd Internal Medicine ward (Table 8, rules 3, 5, 6, 8, 21, 12, 13, 15 and 17), whereas streptomycin-resistant *E. faecium* is associated only with ICU (Table 8, rules 23, 25 and 26).

Finally on table 9, the emerging of the various phenotypes during successive time in the ICU is displayed. It can be seen that from May 1996 until October 2000 multiresistant *P. aeruginosa, Enterobacter aerogenes, Acinetobacter baumanii*, and MRSA, are sequentially included in association rules, revealing that during different time periods, patients are in increase risk of acquiring different multi-resistant isolate.

DISCUSSION

Antibiotic resistance is characterized by biological complexity, being the result of the emergence and spread of a multitude of genes on a variety of bacterial vectors. Moreover changes in antibiotic resistance are the result of a series of interactions between reservoirs of different genomes (species, plasmids, and transposons) in different ecological niches with constantly changing and evolving selection pressures.

In that respect, for surveillance to be a critical and useful part of any public health strategy in confronting antibiotic resistance, it must be able to monitor changing patterns of resistance in different bacterial species bearing different resistant traits in a multitude of locations and time periods under the influence of a series of external and internal factors. Furthermore this must be done in a timely manner with no significant increase of workload in the hospital laboratory. The prime role of the hospital laboratories is considered to be the diagnostic work, with surveillance generally regarded as an additional task, not always an obligatory part of the job description of the laboratory.

Traditional presentations of susceptibility data are based on the tabulation and comparison of resistant rates by ward and clinical specimen. (8, 9, 10) Additionally, specifically designed epidemiological studies are used for the assessment of associations, and identification of risk factors. (11, 12) These studies are time-consuming, need funding, and more importantly, are designed for the study of specific tasks predefined by the investigators.

An important characteristic of Data Mining is the unsupervised production of association rules (non supervised learning). The unsupervised production of association rules conveys that an algorithm is executed over the entire dataset of interest and all the potential association rules of the form $A \rightarrow B$ are produced. This procedure is quite different to the supervised learning procedures where the system is guided towards a focused portion of a knowledge domain relating specific attributes. The advantage of unsupervised learning is that it reveals all pieces of knowledge, among which several can be unexpected, and potentially useful. So any association among any ward, specimen type, and time period with any kind of species or phenotypes is revealed. This ability of the system to perform unguided identification of potentially important associations is compatible with the requirement of an ideal public health surveillance system to automatically identify, on different time and geographical scales, unusual and interesting patterns from time slices of raw data (5). Additionally this production is automatic, no substantial workload is created and no special design is needed. Further more, the method is user friendly and the output is very easy to be understood by the non-expert. Obviously, the results of the system should be regarded as preliminary and all these associations must be further assessed and analyzed.

Although data mining work focuses mainly on very large data sets that are megabytes to terabytes in size, we experiment with data that are kilobytes in size, since in public health and infection control, a great number of interesting and unusual patterns may exist in kilobyte-size data sets.

In the present study association rules with a coverage ≥ 0.001, representing in our database at least 20 cases at the LHS of the rule, and leverage ≥ 0.001, representing at least a number of 20 cases more than expected, were generated.

As shown on table 9, the addition of a two- month time period in the data bank can reveal a new association rule that referrers to that period with out changing the previous set of rules, a fact compatible with the requirement of the system to detect secular trends in a timely manner. Possible new trends are automatically revealed within this time frame.

Association rule discovery in hospital infection data was used by investigators in different studies (13, 14) Their approach was based on the detection of increases in coverage of a certain association rule between different time periods. In addition, in other studies (15) with similar methodology and the same software (DMSS), incidences of certain events were calculated and compared (incidence of multi resistant infections etc).

Our experience is that leverage (and to a lesser extend lift) is a very convenient indicator of the public health importance of an association rule, both in cross sectional and in longitudinal perspective, since it focus on the significance of an association and not on its magnitude: In public Health and more specifically in surveillance of antibiotic resistance it is important to discover new associations and patterns before they become widely spread in the hospital or in the region. It is a well-recognised fact that once a new resistance mechanism is established, it will decline very slowly if at all.

It must be underlined that an important presupposition for the successful application of data mining or any other technique for knowledge discovery in epidemiology, is the (biological and epidemiological) quality of the data and also the time frame of data entry into the system. In that respect it is important that Hospital Laboratory Information Systems should be developed in a way to be able to handle data accordingly (16).

In conclusion we believe that the application of data mining in the analysis of public health data and more specifically in antibiotic resistance surveillance has potentials that must be further explored.

Acknowledgments

- This work has been sponsored by a grand (PENED 1999, EPET II) from the General Secretariat of Research and Development (GSRT) of the Greek Ministry of Development.
- The Greek National Electronic Network for the Surveillance of Antibiotic Resistance is funded by the Hellenic Centre for Infectious Disease Control (KEEL) of the Greek Ministry of Health.

APPENDIX

Association rules definition.

Association rules reveal underlying interactions between the attributes in the data set. These interactions can be presented with the following form: *LHS*→*RHS*, where *LHS*, *RHS* refer to sets of attributes in underlying data. The symbolism *RHS* and *LHS* stands for the attributes present on the *Right Hand Side* and on the *Left Hand Side*, respectively. More specifically, LHS and RHS are selected so as to be frequent item sets. A *frequent item set* is a set of attributes' value, found together in at least T records in a dataset, where T is a user-defined threshold.

Taking that into account, we may describe the problem of mining association rules as follows: (17)

> *Given a set of transactions* $S=\{A_1, \ldots, A_n\}$*, where* A_i*,* $i=1, 2, \ldots, n$*, are the attributes of the data set, generate all the association rules LHS → RHS, where LHS and RHS are frequent sets of attributes,* $LHS \subset S$*,* $RHS \subset S$ *and* $LHS \cap RHS \neq \emptyset$*.*

The intuitive meaning of such a rule is that records in the data set, containing the attributes in LHS, tend also to contain the attributes in RHS (18). It must be noted also that the extracting rules have to satisfy some user-defined thresholds related with association rules measures (such as *support, confidence, leverage, lift*).

These measures give an indication of the association rules' importance and confidence. They may represent the predictive advantage of a rule so as to help to identify interesting patterns of knowledge in data and make decisions.

Rules interestingness measures. The related measures are: (19, 20)

Association rule <u>*strength*</u> *(confidence).*

The strength of an association rule is the proportion of the isolates that are covered by the LHS of the rule that are also covered by the RHS {Strength =n(RHS ∪ LHS)/n(LHS) and it takes values inside [0,1]}**.** The symbolism n(LHS) stands for the number of isolates covered by the Left Hand Side. Values of strength near value 1 are expected for an important association rule.

Association rule *support.*

The support of an association rule is the proportion of isolates covered by LHS and RHS among the total number of isolates {Support = n(RHS ∪ LHS)/N and it takes values inside [0,1]}. N stands for the total number of isolates used in the analysis. Support includes both strength and coverage: **support = strength * coverage**, because strength = n(RHS ∪ LHS)/n(LHS) = (n(RHS ∪ LHS)/N)/(n(LHS)/N) = support / coverage. Values of support near value 1 are expected for an important association rule. Support can be considered as an indication of how often a rule occurs in a data set and as a consequence how significant is a rule.

Association rule *coverage.*

The *coverage* of an association rule is the proportion of isolates in the data that have the attribute values or items specified on the Left Hand Side of the rule. The total number of isolates that this represents is indicated in brackets, by the software, following the coverage {Coverage = n(LHS)/N = P(LHS) and it takes values inside [0,1]}. The symbolism P(LHS) means the proportion of isolates covered by the LHS. Values of coverage near value 1 are expected for an important association rule.

Association rule *leverage.*

The leverage of an association rule is the proportion of additional isolates covered by both the LHS and RHS above those expected if the LHS and RHS were independent of each other. This is a measure of the importance of the association that includes both the strength and the coverage of the rule. The total number of isolates that this represents is presented in brackets, by the software, following the leverage {leverage = P(RHS $\cup$ LHS) – (P(LHS) $*$ P(RHS) and it takes values inside [-1,1] }, where P(RHS $\cup$ LHS) = n(RHS $\cup$ LHS) / N = Support. It is also referred in literature as rule-interest function. (21) Values of leverage equal or under value 0, indicate a strong independence between LHS and RHS. Those associations would not be extracted from "Magnum Opus" because the program does not permit changing "minimum leverage" under value 1. On the other hand values of leverage near value 1 are expected for an important association rule.

Association rule lift.

The lift of an association rule is the strength divided by the proportion of all isolates that are covered by the RHS. This is a measure of the importance of the association that is independent of coverage. {lift = strength / P(RHS) and it takes values inside $\Re_+$ (the space of the real positive numbers)}.

As for the values of lift there are some conditions to be considered:

1. Lift=1 means that RHS and LHS are independent, which indicates that the rule is not important. They are independent because: *Lift* = 1 $\Leftrightarrow$ P(RHS $\cup$ LHS) / P(RHS) $*$ P(LHS) =1 $\Leftrightarrow$ P(RHS $\cup$ LHS) = P(RHS) $*$ P(LHS) which is the definition of independency between two facts. This is also what happens, if we have values of Lift near 1.
2. Lift values close to $+\propto$ means two things:
   - RHS $\subseteq$ LHS or LHS $\subseteq$ RHS, which is not true, due to the definition of the Association Rule, in which we find the restriction LHS $\cap$ RHS $=\varnothing$.
   - P(RHS) is close to 0 or P(RHS $\cup$ LHS) is close to 1, also this first conclusion (P(RHS) $\rightarrow$ 0) means that the rule is not important. But the second conclusion is a very good indicator that the rule is a good one. Now what is urgent to

consider is leverage measure. If leverage takes values close to 1 we have the second of the two options, if leverage is close to –1 then we have the first of the two options.

3. Lift = 0 means that P(RHS $\cup$ LHS) = 0, which indicates that the rule is not important.

For example, suppose a data-base consists of 100 isolates, and an association rule covers with the LHS 100 isolates, with the RHS 200 isolates, and with the RHS 50 isolates that are also covered by the LHS (Figure 1). Strength is 50/100=0.5, support is 50/1000=0.05, coverage is 100/1000=0.1, number of isolates of coverage is 100 (0.1*1000), leverage is (50/1000)-(100/1000*200/1000)=0.03, number of isolates of leverage is 30 (0.03*1000), and lift is (50/100)/(200/1000)=2.5.

Trivial Associations.

An association *x* is *trivial* if there exists another association *y* that is identical, except that the LHS of *y* contains a subset of the LHS of *x* and both associations cover exactly the same cases in the data file. The extra attribute value(s) in the LHS of the trivial association provide no additional lift or leverage.

REFERENCES

1. Vatopoulos AC, Kalapothaki V, Legakis NJ & the Greek Network for the Surveillance of Antimicrobial Resistance: An Electronic Network for the Surveillance of Antimicrobial Resistance in Bacterial Nosocomial Isolates in Greece. *WHO Bulletin*, 1999; 77:595-601.
2. O'Brien TF & Stelling JM. WHONET: an information system for monitoring antimicrobial resistance. *Emerging Infectious Diseases*, 1995; 1: 66-__.
3. Vatopoulos AC, Kalapothaki V, The Greek Network for the Surveillance of Antimicrobial Resistance and Legakis NJ: Trends in bacterial resistance to ciprofloxacin in Greece. Results from the Greek National Electronic Surveillance System. *Emerging Infectious Diseases* 1999; 5: 471-476
4. Gaynes RP. Surveillance of nosocomial infections: a fundamental ingredient for quality. *Infectious Control and Hospital Epidemiology*, 1997; 18: 475-478.
5. Dean AG, Fagan RF & Panter-Conner PJ. 1994. Computerising public health surveillance Systems. PP200-217. In: SM Teutch and RE Churchill (eds), Principles and Practice of public health surveillance. Oxford University Press, New York, NY.
6. Fayyad UM, Piatetsky-Shapiro G & Smyth P. From data mining to knowledge discovery: an overview. In: Fayyad UM, Piatetsky-Shapiro G, Smyth P, Uthurusamy R (eds). Advances in Knowledge Discovery and Data Mining. Cambridge, Mass.: MIT Press, 1996:1 –36.
7. Stelling JM & O'Brien TF. Surveillance of antimicrobial resistance: The WHONET program. *Clinical Infectious Diseases* 1997, 24 (suppl 1): S157-168.
8. McDonald LC, Banerjee SN & Jarvis WR. Seasonal Variation of Acinetobacter Infections: 1987 - 1996. National Nosocomial Infections Surveillance System. *Clinical Infectious Diseases* 1999; 29 (5): 1133–37.

9. Fluit AC, Jones ME, Schmitz FJ, Acar J, Gupta R, Verhoef J, and the SENTRY Participants Group. Antimicrobial susceptibility and frequency of occurrence of clinical blood isolates in Europe from the SENTRY Antimicrobial Surveillance Program, 1997 and 1998. *Clinical Infectious Diseases* 2000; 30 (3): 454–60.
10. Hanberger H, Garcia-Rodriguez JA, Gobernado M, Goossens H, Nilsson LE, Struelens MJ, and the French and Portuguese ICU Study Groups. Antibiotic susceptibility among aerobic gram-negative bacilli in intensive care units in 5 European Countries. *The Journal of the American Medical Association*, 1999, Jan 6; 281 (1): 67-71.
11. Mekontso-Dessap A, Kirsch M, Brun-Buisson C, & Loisance D. Poststernotomy mediastinitis due to *Staphylococcus aureus:* Comparison of methicillin-resistant and methicillin-susceptible cases. *Clinical Infectious Diseases* 2001; 32 (6): 877-883.
12. Cheong HJ, Yoo CW, Sohn JW, Kim WJ, Kim MJ, & Park SC. Bacteremia due to quinolone-resistant *Escherichia coli* in a teaching hospital in South Korea. *Clinical Infectious Diseases* 2001; 33 (1):48-53.
13. Brossette SE, Sprague AP, Hardin JM, Waites KB, Jones WT & Moser SA. Data Mining in Hospital Infection Control and Public Health Surveillance. *Journal of the American Medical Informatics Association*, 1998; 5: 373 - 381.
14. Brossette SE, Sprague AP, Jones WT & Moser SA. A Data Mining System for Infection Control Surveillance. *Methods of Information in Medicine* 2000; 39: 303–10.
15. Moser SA, Jones WT & Brossette SE. Application of Data Mining to Intensive Care Unit Microbiologic Data. *Emerging Infectious Diseases* 1999; 5: 454-457
16. Martin SM & Bean NH. Data management issues for emerging diseases and new tools for managing surveillance and laboratory data. *Emerging Infectious Diseases* 1995; 1: 124-128

17. Agrawal R & Srikant R. (1994). Fast Algorithms for Mining Association Rules. [p: 487] In: Proceedings of the 20th Very Large Data Bases Conference, Santiago de Chile, Chile, 1994.

18. Srikant R & Agrawal R. Mining Generalized Association Rules. [p: 407] In: Proceedings of the 21st Very Large Data Bases Conference, Zurich, Switzerland, 1995.

19. Han J & Kamber M. Data Mining: Concepts and Techniques. In: Han J & Kamber M. Data Mining: Concepts and Techniques. [p: 23] Academic Press, 2001.

20. Berry M & Linoff G. Data Mining Techniques: For Marketing, Sales, and Customer Support. In: Berry M & Linoff G. Data Mining Techniques: For Marketing, Sales, and Customer Support. John Willey, 1997.

21. Piatesky-Shapiro G. Discovery, analysis and presentation of strong rules. In Knowledge Discovery in Databases. In: Piatesky-Shapiro G. Discovery, analysis and presentation of strong rules. In Knowledge Discovery in Databases. [p: 229 – 248]. MIT Press, 1991.

22. Bayardo RJ Jr, Agrawal R, & Gunopoulos D. Constraint-based rule mining in large, dense databases. [p: 188]. In: Proceedings of the 5th International Conference on Data Engineering, Sydney, Australia, 1999.