Discovering of Health Risks and Case-Based Forecasting of Epidemics in a Health Surveillance System *

M. Bull G. Kundt L. Gierl

University of Rostock, Department for Medical Informatics and Biometry
Rembrandtstr. 16/17, D-18055 Rostock, Germany
{mathias.bull|guenther.kundt|lothar.gierl}@medizin.uni-rostock.de

Abstract. In this paper we present the methodology and the architecture of an early warning system which fulfills the following tasks. (1) discovering of health risks, (2) forecasting of the temporal and spatial spread of epidemics and (3) estimating of the consequences of an epidemic w.r.t. the personnel load and costs of the public health service. To cope this three task methods from knowledge discovery and data mining, case-based reasoning, and statistics are applied.

Keywords: knowledge discovery and data mining, case-based reasoning and forecasting,

1 Introduction

1.1 Need of Health Surveillance Systems

During the cholera epidemic in London in 1854, the English physician John Snow discovered that cholera is transmitted by contaminated water. Snow mapped all of the locations of cholera deaths, saw a cluster of the victims in the neighbourhood of a street and found out that nearly all had drunk water from community pump. The handle of the pump was removed, at Snow's insistence, and the epidemic ended in five days (cf. [3]).

Nowaday, the fight against diseases is often more complicated. Many foodborne and waterborn disease outbreaks go unrecognized or are detected late; the magnitude of the problem of antimicrobial drug resistance is unknown; and global disease surveillance is fragmentary. There is reason to believe that the number and incidence of emerging diseases and the risk of reemerging diseases are all increasing. Factors responsible for the increase include such social changes as mass population movements, rural-to-urban migration and accelerated urbanization, population growth, rapid transport, new food technologies, and new life styles as well as environmental changes that increase the risk of exposure to zoonotic or vector-borne infections, such as altered land use patterns and irrigation. The slides offered by the CDC (see http://www.cdc.gov/ncidod/publications

* supported by the DFN-Verein (German society for the national research net) and the AOK Mecklenburg-Vorpommern (General health insurance company)

/slides/slides.htm) illustrate the dramatic situation of emerging and re-emerging infectious diseases, health risks and the resistant pathogens.

For ensuring a high level of disease prevention physicians and pharmacists need information on current health risks as well as on outbreaks and spreads of communicable diseases. Contrary to this information need, the current information systems to monitor human and animal infectious diseases domestically and internationally are inadequate to confront the present and future challenges of emerging infections. Consequently, a new intelligent information tool on current health situation of a population based on a well-designed network is needed which discovers new illnesses which are unknown up to now, warns of epidemics of emerging or re-emerging communicable diseases as soon as possible, forecasts the temporal and spatial spread of epidemics, detects drug resistance of pathogens, estimates the consequences of an epidemic according to the personnel load and costs of the public health service, for instance.

For an introduction in the problem of health surveillance systems we refer to the book "Geomedical Systems" by R. Thomas (see [19]).

1.2 Review of Health Surveillance Systems

The ProMED-Proposals (see http://www.healthnet.org/promed.html), which was circulated for comment over the past year to more than 300 experts around the world, characterizes in a clear and sharp form the need and the technical demands of such a information system. Within this proposal, the following estimation is given:

> There is presently no functional system anywhere for monitoring emerging diseases. Current surveillance capabilities are fragmentary, lack coordination, and are geared toward established diseases. Moreover, they have mostly fallen into neglect. This is true both at the national level and internationally.

Presently, there are some health surveillance networks as singular solutions for observing special chosen or notifiable communicable diseases, which can be classified in three groups.

Sentinel-General-Practitioner-Networks consist of general practitioners scattered thourghout the region report each new case of a notifiable disease to a data processing center, where the collected data are evaluated. On the one hand such networks work rapidly, the methods of biostatistics are well applicable. On the other hand a small sector of diseases are only observed. The French communicable diseaes computer network "le Réseau National Téléinformatique de surveilance et d' information sur les Maladies Transmissbles" equipped with the workstation for an epidemiologist "Station de travail pour l' epidémiologiste" (see http://www.b3e.jussieu.fr:80 /sentiweb and [20],[6]) is an advanced examples for a such information system.

In *Sentinel-Microbiological-Laboratory-Networks* the results of the analysis of pathogens are collected and evaluated by a central computer center. Consequently, the monitoring of resistant pathogons is possible. By means of such a network the top of the health situation of population is only observable since in laboratories there are only analysed specimens of particians with a severe diseases. In the USA there work the networks "Public Health Laboratory Information System" and "Laboratory Information Tracking System" (see [18]).

Electronic-Mailing-Lists, where all subscribers can inform and discuss any disease or symptoms of a sickness, are the simplest form of a warning and communication system. Unfortunately, the informal form of an e-mail allows no evaluation by computers. ProMED is worldwide one of the greatest e-mail-conferences to monitor emerging diseases. (see http://www.healthnet.org/promed.html)

In all health surveillance systems mentioned above the information gathering, the information flow, the information processing and the publication of the monitoring results are separate from the offical report channels. Consequently, these systems do not improve the information management within the networks of the offical report channels.

Information Gathering. The data sources which are gathered already on machine-readable data mediums are not often used. The data streams are not continuously, often only sporadally. Partial aspects of the health monitoring are considered.

Information Processing. The computer-aided evaluation of the gathered data are often weak developed and there are no abilities for automatic derivation.

Publication and Presentation of the Monitoring Results. The printed report of the international and national health organizations are not suffient because of the delay of publication as well as the temporal and spatial inaccurancy of the monitor results. The information channels do not reach all user groups and keep not all user groups informed about current health risks.

1.3 Outline of our Approach

In this paper we present an architecture of an health surveillance system for regional health risks which we establish in the country Mecklenburg-Vorpommern (north-east part of Germany, cf. [2]).

Information Gathering. The crucial point with the analysis of the time-spatial pattern of diseases and pathogens is the reliability, the validity and, for an early warning system, the timeliness of the data.

In Germany, the hospitals have to collect patient records on machine-readable mediums (§301 SGB V, §301-agreement on the transmission of medical data) and have to send these records to the health insurance companies. Moreover, both patient and general practitioner have to send the medical certificate within three days to the health insurance companies. We recieve daily all these data as anonymous patient records via net from the computing center of general health insurance company "AOK Mecklenburg-Vorpommern". The "AOK Mecklenburg-Vorpommern" insures 50% the population. Furthermore, the public health office "Landeshygieneinstitut Mecklenburg-Vorpommern" sends weekly information about notifiable diseases and about pathogens identified by the microbiological laboratories.

Information Processing. The public health surveillance is a complex problem of multiparametric times courses of diseases, pathogens, resistance, health services etc. in a geographical region.

In [13] R. J. Marshall discussed various methods for analysis of geographical distribution of diseases. In his interesting review, he point out that statistical methods are partially applicable to partial problems of disease clusterings either in time or in space but not appropriate to examine the time-spatial dynamics of communicable diseases. Further, he mentioned that there are observable time-spatial patterns of the spread of diseases which are complicate to describe by means of statistics (especially see [13], section 5). G. F. Pyle [15] compared the diffusion paths of influenza epidemics and found out that there exist similar epidemic waves according to the time-spatial dynamic.

All successful applications used case-based techniques for monitoring complex processes in time and space on other domains. Case-based reasoning (CBR) is a cyclical AI paradigm for solving problems by analogical reasoning and for learning new problem-solution pairs through experiences (see [1]). The CBR-cycle [1] consists of the following four stages:
retrieve the most similar case or cases;
reuse the retrieved case or cases to solve the problem by analogical reasoning;
revise the proposed solution;
retain the parts of this experience likely to be useful for future problem solving.
The MetVUW Workbench [9] is a system for intelligent retrieval and display of historical meteorological data. An interesting combination of statistics and CBR is implemented in the expert system Air Quality Predictor [12] developed to predict air pollution levels in Athens, Greece. For a more statistical application we refer the reader to [11]. The problem of managing and retrieval of large case bases is discussed in [10]. In [17] we made some steps using data and time abstraction to cope with the problem of multiparametric time courses. Our system ICONS prognoses possible kidney failure of patients in an intensive care unit based on former, similar courses.

Following the argumentation given above, we combine methods of statistics and artifical intelligence (AI), namely case-based reasoning, knowledge discovery and data mining (KDD) to cope with the problem of health surveillance.

Publication and Presentation of the Monitoring Results. For a rapid publication of the monitoring results we develop a visualization tool. It provides a means to present knowledge on the current epidemic situation in a region via INTERNET.

2 Problem Description

In the following we sketch the basic terminology and describe our approach to the problem of monitoring the public health situation in a region.

We divide the considered geographical region in a finite number of disjoint geographical units, here called *locations*. Thus a region is a non-empty, finite set L. For our project we have chosen the ZIP-code districts because this dissection includes implicitly facts on the demography and the infrastructure. Further, we choose a reasonable period τ (e.g. $\tau=$ one week) and divide the time scale in equidistant time steps. A finite number of consecutive time steps (of the period τ) is said to be a *time interval.*

By a *scenario* we understand a concept which describes the public health situation and the load of the health service in the considered region during a period. More precisely, a scenario s includes the following data for each location l in L which arise during a period τ: (i) each new case of illness given by an anonymous patient record, (ii) the load of the health service, and (iii) contextual information (weather and season, holiday seasons). Such data are gathered on machine-readable mediums by health insurance companies and hospitals.

Let S be the set of all scenarios. A *scenario sequence* σ is a finite sequence of scenarios. By S^* we denote the set of all scenario sequences over S. We also denote a scenario sequence by the concatenation of its elements, e.g. $\sigma = s_0 s_1 \dots s_n \in S^*$ with $s_i \in S$ for $i \in \{0, 1, \dots, n\}$. So we are able to describe the course of any epidemic by a scenario sequence since it keeps all information on the public health situation during a time interval. Furthermore, we define inductively a similarity metric *sim* on S^*, i.e. at first we define similarity metrics on all elementary sets and then similarity metrics on sets of structured objects by using the metrics of its components.

By a *health risk* we understand an observable changing of the public health situation which may cause an epidemic. So we have the following main problems:

Discovering of Health Risks
Given: σ current scenario sequence
Question: Does there exists a health risk ?
Forecasting a Scenario
Given: σ current scenario sequence
Question: How does the successor scenario of σ looks like ?
Estimating the Consequences for Health Care Resources
Given: σ current scenario sequence
Question: What are the consequences for health care workers, pharmaceutical industry and over all costs ?

In the following sections we denote the case base by $\Sigma \subseteq S^*$ which contains some scenario sequences describing the course of past epidemics. For the case base Σ it holds $(s\sigma \in \Sigma \Rightarrow \sigma \in \Sigma) \wedge (\sigma s \in \Sigma \Rightarrow \sigma \in \Sigma)$. Let $\sigma = s_0 s_1 \dots s_m$ be the current scenario sequence.

2.1 Discovering of Health Risks

We want to detect a health risk as soon as possible *to warn* whereas in epidemiological studies the aim is *to prove* a causal relationship of disease and exposition. Thus, the discovery of a health risk is based on the question "Has something changed in a suspicious manner?" Here, conspicious changes in the health situation of the inhabitants of a region could be an accumulation of an illness of undetermined origin within a certain time interval, a suspect accumulation of resistance to antibiotics in a certain area or a sickness which occurs frequently within a period of time or a certain area. Remark that an accumulation can also occur with a very low incidence rate of an event. Therefore, it is nescessary to determine each minor changes of the health situation. However, a frequent false-positive reaction of the system must be avoided (artifacts). It must be possible to set the sensitivity and specificity of the system.

Although we have a large and rich data sources, we have to notice that these data can not be considered as a representative sample within the meaning of statistics. The social structure as well as the age distribution of the insurants and of the whole population do not correspond. Consequently, parametric statistical methods are not applicable. Therefore we combine case-based techniques and permutation tests. Permutation tests (see [8], [4]) are special distribution-free statistical tests which require only one or two relatively weak assumptions about the distribution of the variables, e.g. the underlying distribution is symmetric.

Discover conspicuous changes in the current scenario sequence
By using methods of KDD (cf. [5], [14]) the current scenario sequence σ are explored. To detect disease distribution pattern we apply the so called G-test by A. Getis and J.K. Ord (cf. [7]), which based on a permutation test. The G-statistic measures the concentration or the lack of concentration of the sum of considered values in the region. This statistic is a proportion of the sum of all values that are within a certain neighbourhood of a location to the total sum of the region. The null hypothese is that the set of all values within a certain neighbourhood of a location is a random sample drawn without replacement from the set of all values in the region, i.e. spatial independence of the considered values. If we reject the null hypothese then we obtain positive or negative spatial association.
Moreover, existing techniques of Scientific Visualisation can be applied to show temporal and spatial dependencies and correlations of the data contained in the current scenario sequence σ.
Retrieve the scenario sequences with similar onsets and epidemic courses
The system looks for a scenario sequence σ' which starts with a subsequence similar to the current scenario sequence σ and terminates in a dangerous scenario. For that purpose each onset of epidemic scenario sequence in the case base are indexed by the disease and by the charaterization of the outbreak locations.
Reuse the retrieved scenario sequences to warn
If there is such a scenario sequence σ' then the system has to warn or alarm of an epidemic depending on severity degree of the detected health risk.

$$\boxed{\begin{array}{ll}
 & \hfill\text{epidemic scenario} \\
\text{retrieved scenario sequence} & \hfill\swarrow \\
\text{from the case base } \Sigma & \longrightarrow \sigma' = s'_0 s'_1 \ldots s'_m s'_{m+1} \cdots s'_{m+k} \\
 & \qquad\qquad \Downarrow \textit{warning} \\
\text{current scenario sequence} & \longrightarrow \sigma \;= \underbrace{s_0 s_1 \ldots s_m}_{\text{similar subsequences}} \boxed{!}
\end{array}}$$

The severity degree of a health risk is characterized by the following items:
1. incidence, e.g. how much new cases of illness occur.
2. population density of a location, e.g. whether the outbreak occurs in a city or village;
3. quality of a prevention, e.g whether there is a vaccine;
4. quality of a therapy, e.g. how successful will be a medical treatment;
5. course of the illness.

Revise the warning

The quality of the discovery of health risks is evaluated by monitoring of the succeeding scenarios.

Retain the useful scenario sequences

If a scenario sequence which led to an epidemic and has not been discovered by the system then this sequences will be stored in the case base.

2.2 Forecasting of a Scenario

Forecasting of a regional health situation means to predict a future scenario. We use the following steps.

Retrieve the most similar scenario sequences

The system retrieves all scenario sequences $\sigma' s'_{m+1} = s'_0 s'_1 \dots s'_m s'_{m+1}$ from the case base Σ so that the differences between σ' and the current scenario sequence σ are minimal, i.e. for all $\sigma'' s'' \in \Sigma$ it holds $sim(\sigma, \sigma'') \leq sim(\sigma, \sigma')$. Here, all epidemic scenario sequences are indexed by the diseases, by the outbreak locations and by the diffusion paths of the epidemics.

Reuse the retrieved scenario sequences to forcast a scenario

By using the retrieved sequences $\sigma' s'_{m+1}$ the system adapts the scenario s'_{m+1} to a forecast scenario $\tilde{s}$.

$$\begin{array}{lll} \text{similar scenario sequence} & & \\ \quad\text{from the case base } \Sigma & \longrightarrow \sigma' s'_{m+1} & = s'_0 s'_1 \dots s'_m s'_{m+1} \\ & & \qquad\qquad\quad \Downarrow \textit{adaptation} \\ \text{current scenario sequence} & \longrightarrow \sigma & = s_0 s_1 \dots s_m \boxed{?} \\ & & \qquad\qquad\quad \nwarrow \\ & & \qquad\qquad \text{forecast scenario } \tilde{s} \end{array}$$

Here, we use background knowledge about demographic structures, several statistical models of epidemics (see [16], [13], [19]) and the defined threshold values of epidemical levels which are based on the epidemiological studies. The resulting data will be stored in a case based fashion and will be available for graphical presentation on the user interface.

Revise the forecast scenario

After a time period τ we know the scenario s_{m+1}. In comparing the forecast scenario $\tilde{s}$ with the scenario s_{m+1} the system evaluates the forecast mechanism.

Retain the useful scenario sequences

If the difference of these scenarios is too large then the system has to learn this new scenario sequence σs_{m+1} or the inference mechanism must be changed. In the first case, the current scenario sequence is integrated in the case base. In the latter, an expert (e.g. epidemiologist, biostastician) has to execute a working cycle with the system. This work cycle contains the modifying of the threshold values and of the metrics, the visualizing of the multiparametric data of the current scenario sequence and those in the case base and the forecasting of a scenario (as a simulation step). This cycle is executed until the difference of the current scenario and the forecast senario is within a tolerance range.

2.3 Estimating of the Consequences for the Health Care Resources

Based on the discovered risk and on the forecast scenario, monitrary and medical consequences will be estimated. Furthermore, the demands on intensive care beds, the required nursing care or amount of vaccines can be concluded. Length of stay, possible surgery, letal cases and further consequences can also be deduced. From consumables, departmental care rates etc. the expected costs for a case induced by the epidemic can be estimated.

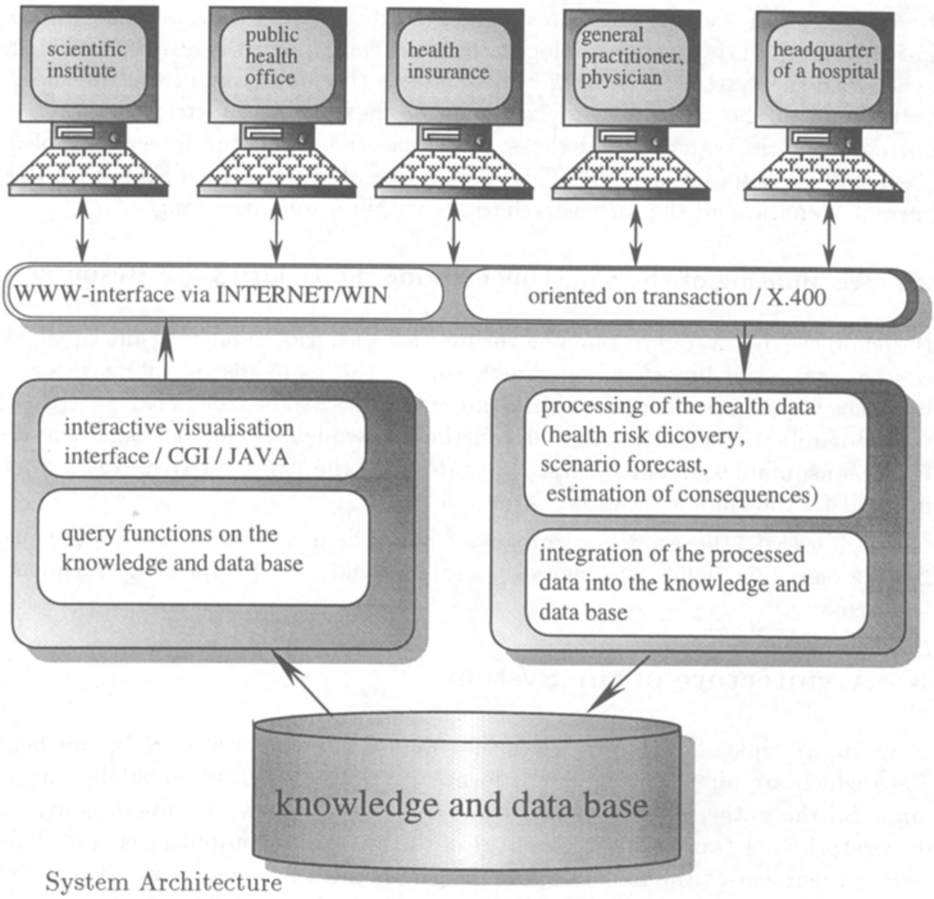
To illustrate the economic impact of an epidemic, we notice that the influenza epidemic 1995/1996 caused 900 million DM only for medical treatment in Germany.

3 Architecture of the System

A main principle of our approach is that we use various data source of medical data which are already available as formatted data on machine-readable mediums. So the gathered data of hospitals, health insurances and microbiological laboratories are transmitted to a server of the university computing center. This server preprocesses the raw data and integrates them into a scenario. Then, the above described procedures starts. The results are integrated in the knowledge and data base. The visulization tool generates graphical representations of the data which shows the forecast of health risks and the spread of communicable diseases as well as the temporal and spatial dependencies and correlations of the data.

In the next future, the functionality of the visualization system will be used via the net without having the actual visualization system implemented on the user's terminal or network. The computer language JAVA offers the possibility to write platform and operating system independent applets which can be run on any computer operating system.

The general architecture of our health surveillance system is shown in the figure given below.


scientific institute
public health office
health insurance
general practitioner, physician
headquarter of a hospital
WWW-interface via INTERNET/WIN
oriented on transaction / X.400
interactive visualisation interface / CGI / JAVA
query functions on the knowledge and data base
processing of the health data (health risk dicovery, scenario forecast, estimation of consequences)
integration of the processed data into the knowledge and data base
knowledge and data base

System Architecture

4 Conclusion

The problem of emerging and re-emerging diseases will increase dramatically in the near future.

Physians, hospitals and health authorities do not have access to an appropriate instrument which provides the information needed for a timely decision making from the huge body of medical data avaiable. Our approach meets the requirements of timely handling of the very large amount of data to identify health risks in a region. Intelligent visualization is an indispenable tool to present this informations for a broad class of uses.

As mentioned in section 1.3 and following, our system approach is an intelligent combination of advanced techniques from AI, statistics and visualization. This approach will generate a knowledge base for modern health surveillance systems.

References

1. Aamodt, A. ; Plaza, P. : *Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches.* AI Communications 7(1) 1994, 39-59
2. Bull, M.; Gierl, L.: *Architecture of an Early Warning System for Regional Health Risk.* Rostocker Informatik Berichte (1996)19, 5-18
3. Cliff, A.D.; Haggett, P.: *Atlas of Disease Distribution: Analytical Approaches to Epidemiological Data.* Blackwell, Oxford 1988
4. Edgington, E.S.: *Randomization Tests.* Marcel Dekker, Inc., NewYork 1987
5. Fayyad, U. M. ; Piatetsky-Shapiro, G.; Smyth,P. ; Uthurusamy, R.: *Advances in Knowledge Discovery and Data Mining.* MIT Press, 1996
6. Garnerin, P.; Valleron, A.J. : *The French Communicable Diseases Computer Network : a technical view.* Comput Biol Med 1992; 22,3:189- 200.
7. Getis,A.; Ord, J.K.: *The Analysis of Spatial Association by Use of Distance Statistics.* Geographical Analysis, Vol 24, N0.3 (July 1992), 189-205
8. Good, Ph.: *Permutation Tests.* Springer-Verlag, Berlin 1986
9. Jones, E.K. ; Roydhouse, A. : *Iterative Design of Case Retrieval Systems.* Victoria University of Welligton, New Zealand, Technical Report CS-TR-94/6, see also: Proc.of the AAAI-94, Workshop on Case Based Reasoning, Seattle, Washington, 1994
10. Kitano, H.; Shimazu, H.; Shibata, A.: *A Methodology for Building Large-Scale Case-Based Systems.* in: Proceedings of the AAAI-93, Washington, 1993
11. Lee, J.K.; Oh, S.B.; Shin, J.C.: *Unik-FCST: Knowledge-assisted adjustment of statistical forecast.* Expert Systems with Applications, Vol. 1, 39-49
12. Lekkas, G.P.; Arouris, N.M.; Viras, L.L.: *Case-Based Reasoning in Environmental Monitoring Applications.* Applied Artificial Intelligence, Vol. 8, 1994, 349-376
13. Marshall, R.J.: *A Review of the Statistical Analysis of Spatial Patterns of Disease.* J. R. Statist. Soc. A (1991) 154, Part 3, 421-441
14. Piatetsky-Shapiro, G.; Frawley, W.J. (Eds.): *Knowledge Discovery in Databases.* AAAI Press, Menlo Park, 1991
15. Pyle, G.F.: *The Diffusion of Influenza.* Rowman & Littlefield, Totowa NJ 1986
16. Rothman, K.J.: *Modern Epidemiology.* Little Brown, Boston, Toronto 1986
17. Schmidt, R.; Heindl, B.; Pollwein, B.; Gierl, L.: *Abstraction of Data and Time for Prognoses of the Kidney Function in a Case-Based Reasoning System.* in J. Brender, J.P. Christensen et al.(eds.): Proc. of Medical Informatics Europe '96, IOS Press, Technology and Informatics, Vol. 34, Part A, 570-574
18. Stanley, M. M.; Bean, N.H.: *Data Management Issues for Emerging Diseases and New Tools for Managing Surveillance and Laboratory Data.* EID Volume 1, Number 4, October-December 1995
19. Thomas, R.J.: *Geomedical Systems.* London, New York 1992
20. Toubiana, L.; Vibert, J.F.; Garnerin, P.; Valleron, A.J. : *A health Care Workstation Integration Architecture for Epidemiologists.* Comput Biomed Res 1995;28:100- 15