Exploitation of data for risk assessment

Niel Hens[1]

[1]UHasselt, Diepenbeek

1. Introduction

In the narrow sense 'database management' can be understood as activities associated with storing and processing available information while in the broader sense it can be understood as all activities associated with collecting, storing, processing and validating the information. In this paper focus goes to the broader meaning and it's relation to risk assessment. Risk assessment is often defined as 'the scientific assessment of risks and potential risks that occur in a specific context'. To this aim, risk models describing the relationship among several entities and the risk are developed. A large caveat for the usability of a risk model is often the lack of (accurate) data and the lack of assessment of the associated variability. Next to summarizing the basic concepts of a database management system, it is shown how the statistical world answers to the inevitable complications that arise when collecting data and conducting surveys.

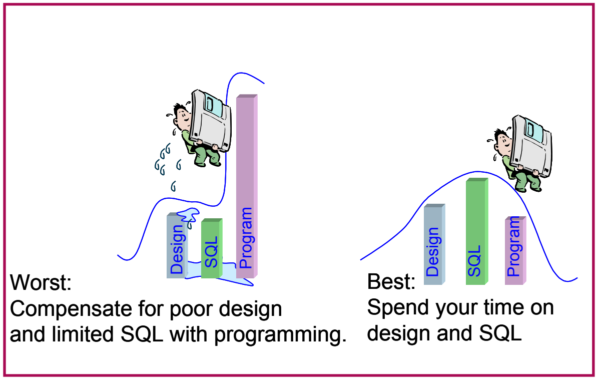
2. State of the art

Originally, databases were set up as rectangular, spreadsheet-like structures, closely followed by the development of hierarchical database structures and network structures.

The associated inefficiency of storing data in this way resulted in the development of relational databases and finally objected-oriented databases. The latter database structure is for example used for SANITEL, the Belgian herd identification and registration database. In this way data are stored with minimal data redundancy, maximal data consistency, maximal integration and sharing of data, enforced standards, ease of application development, uniform security, privacy and integrity. In terms of software, the most popular and recent database management software packages include Oracle, Ingres, Informix, DB2, SQL Server and Microsoft Access[1].

3. Database management

A database management system (DBMS) consists of three different components: the design, the standard query language (SQL) and the programming component. It is emphasized that the excellence of a DBMS is a result of spending time on design and SQL rather than on programming (Figure 1)[1].



Worst:
Compensate for poor design
and limited SQL with programming.

Best:
Spend your time on
design and SQL

Figure 1. *The trade-off between design/SQL and programming.*

Design

The design of a database is the first important step in the development of a DBMS. It is the essential core ingredient to construct an efficient and comprehensive DBMS and consists of six construction steps: 1) the identification of the exact goals of the system; 2) the identification of the basic data collection forms and reports; 3) the identification of the data items to be stored; 4) the design of the classes (tables) and relationships; 5) the identification of the so-called 'business rules', i.e. the context-specific rules; and 6) the verification of the design matching the 'business rules'.

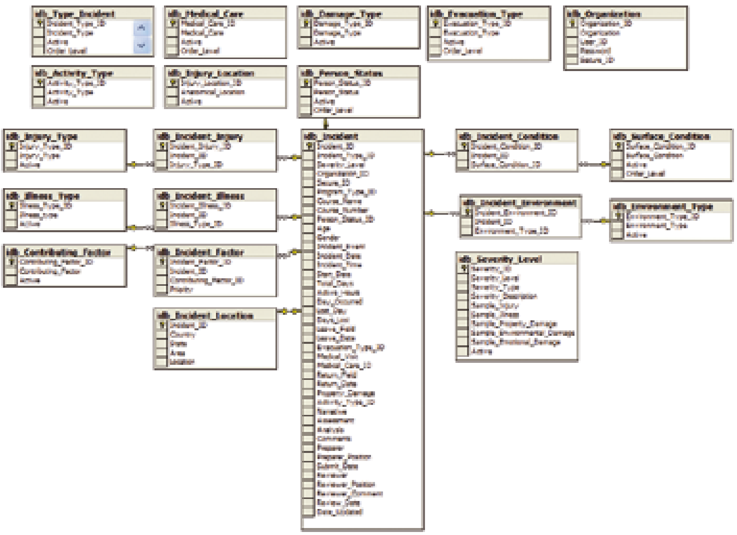


Figure 2. *The Wilderness Risk Management Committee Incident Reporting database (© Outdoor Ed LLC (www.outdoored.com/safety/incidents/design/rationale.aspx)).*

To this respect 'data normalization' has been developed as the method to efficiently store data coming from different sources. To ensure minimal data redundancy, a specific set of rules is followed when collecting and/or consulting data. In general, one distinguishes between 1st, 2nd, 3rd and 4th normal form. Once data have been normalized and thus minimal data redundancy is ensured, referential integrity is more easily verified and more easily validated. An example of a design is presented in Figure 2 for the Wilderness Risk Management Committee Incident Reporting database.

Standard query language

In a sense, normalizing databases leads to the construction of small datasets linked to each other by identifying relations between them (Figure 2). Information is thus spread over different sub-datasets. When interested in a specific scientific question, this information is collected from the database by merging different entities from these different sub-tables of the dataset using queries. To avoid the creation of ad hoc query languages, several proposals to construct a standard query language have been made among which SQL is the most commonly used one.

Programming

The final step in the development of a DBMS to achieve an easy-to-use application is the 'programming' step. This step essentially finalizes the development of the DBMS to answer user-defined needs. Depending on the time spent on design and SQL, the programming step ideally covers only a minor part of the DBMS (Figure 1). Programming cannot solve deficiencies created in the design stage of the DBMS development. It is therefore, that whenever user-specific needs change in time, a re-evaluation of the DBMS should take place and whenever necessary a new DBMS should be developed. To avoid creating new DBMS over and over again, developing a DBMS goes hand in hand with forecasting future needs, obviously a very difficult task in a rapidly changing information society.

4. On the use of different databases in risk assessment

Building up risk models often relies on constructing mathematical/stochastic models as for example compartmental models. In such compartmental models, different parameters are used to describe the rate of transition between compartments (Figure 3). These parameters are often determined using specific databases as for instance infection data, demographic data… There is often no link between these different datasets and thus consistency, especially with respect to space and time, should be checked.

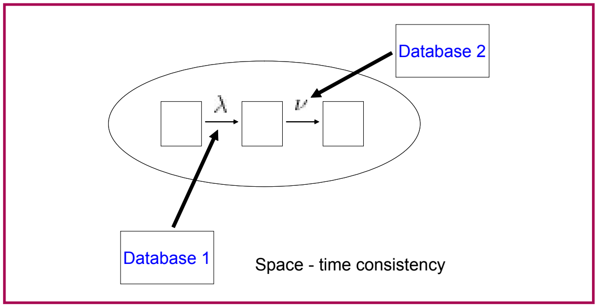


Figure 3. *An example of the compartmental model with two parameters as estimated from two different databases.*

Whenever a link is available, different databases can be merged to one. However, by doing so there is a likely occurrence of creating structural incompleteness in the non-common entities (Figure 4). While incompleteness often results in bias in terms of modelling results, this is not necessarily the case for structural incomplete data. A formal sensitivity analysis could show the effect of potential bias on the modelling results and provides a more honest analysis.

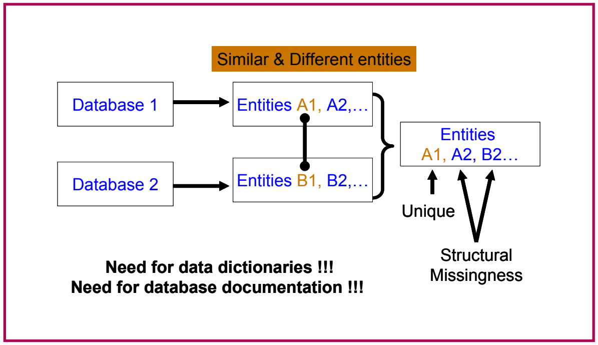


Figure 4. Database merging scheme.

Recreating a DBMS from merging two or more databases involves extended coding, renormalization, re-evaluating queries and conducting consistency checks. Whenever links are broken, they are often lost forever with an associated increased uncertainty. In that case, pragmatic solutions are often used. This stresses the use of data dictionaries and full data documentation on alterations in the database.

5. From databases to statistics

In any statistical package the optimal statistical database structure is of a rectangular format where the consensus is often made that entries in rows correspond to subjects while columns correspond to different factors measured on those subjects.

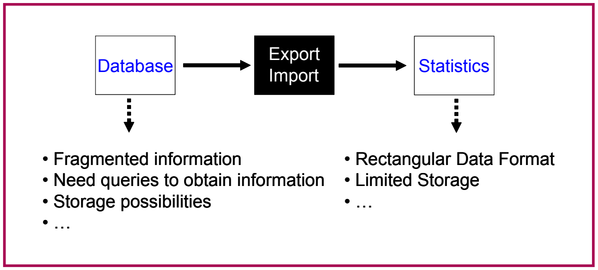


Figure 5. A graphical scheme of the relationship between databases and statistics.

The export/import function often available in any DBMS or statistical program provides the link between both. However, the transition from a DBMS to a statistical program is often associated with writing the corresponding SQL code and limiting the information to the essential since statistical software often have limited storage capability (Figure 5).

Exporting/importing data is ideally done only once and as such the relevant information has to be identified and retained from the DBMS using queries. This is accompanied with the identification of potential risk factors, a potentially complex task in many applications.

Once data have been stored into a rectangular data format, statistical analyses are easy to conduct although different complex data structures require different statistical techniques to be employed. In the following section, a flavour of modern statistical techniques developed to handle specific complex data structures such as clustering and incomplete data, are briefly discussed.

6. Modern statistics for complex data

The development of new statistical methodology is mostly motivated from practical applications. In the past few decades, several new statistical techniques have been developed to deal with clustering, incomplete data and other complications that arise when collecting and analyzing data. Let us give an overview of a few of these domains that can help to ease the burden associated with collecting data and conducting surveys. Although these techniques can deal with a lot of these complications, they still inevitably rely on assumptions and the quality of data collection. The information loss associated with sloppy data can never be rectified.

Data mining

In terms of risk assessment it is important to identify risk factors. Data mining is the principle of sorting through large amounts of data and picking out relevant information. It is increasingly used in the sciences to extract information from the enormous data sets generated by modern experimental and observational methods. Often the information is scarce and the identification of a sole or a limited set of risk factors is the scope of mining the data. An excellent reference to this respect is the work of Hastie et al.[2].

Incomplete data

Collecting survey data is inevitably associated with data loss, potentially due to several different causes. Analyzing incomplete data with inappropriate methods leads to serious bias and false results. There exists a whole taxonomy on different incomplete data mechanisms and techniques to handle them (see e.g. Little and Rubin[3]; Schafer[4]). In essence, two components are now modelled simultaneously, i.e. the measurement part and the missingness part, the latter referring to modelling the underlying mechanism generating the missingness, while the measurement part is modelled in the same way as if the data would have been completely observed. A sensitivity analyses with respect to the model describing the missingness mechanism is needed since often untestable assumptions need to be made.

Diagnostic uncertainty

In the risk assessment context, data are often collected using diagnostic tests. These test results are merely markers but are often treated as if they were the true result. This could lead to bias but more surely to overestimated precision. Complications as detection limits and diagnostic uncertainty need to be taken into account. Censoring techniques and mixture modelling are two approaches that have been proven worthwhile.

Whenever a test result is lower or higher than a certain detection limit, the observation is said to be censored. Appropriate techniques to handle censoring exist[5]. Mixture modelling deals with diagnostic uncertainty in that test-positives and test-negatives are classified as true positives and true negatives by directly using the marker of infection[6]. Note that other techniques relying on for example receiver operator characteristic curves exist but are often considered to be outdated.

Bayesian data analysis

Next to the well known likelihood framework, the Bayesian paradigm exists[6]. In a Bayesian analysis a postulated model is fitted to the data incorporating prior knowledge on the parameters of that model. In this way, information from several sources can be combined to produce a comprehensive model. Of course this is not without disadvantages. The dependency of the results on the prior distribution makes it necessary to perform sensitivity analysis with respect to their choice. If no prior knowledge is available, uninformative priors are used, implying that the data drives the model and as such the method is comparable to the likelihood approach. Often, the likelihood and the Bayesian paradigm are contrasted to one another instead of considering them to be complementary.

7. Prospects

Current practice in risk assessment should continue aiming at bringing researchers from different fields together. Interdisciplinary research between chemists, epidemiologists, data managers, microbiologists, statisticians, toxicologists and others is of utmost importance to achieve a high standard in risk assessment.

8. Summary and Conclusion

The process from data collection to the models that constitute risk assessment is not a straightforward process. It all starts with the collection of accurate data. To efficiently exploit data, database management systems need to be installed, maintained and made available to researchers. The quality of such a DBMS is determined by the time spent on its design. A specific set of normalization rules should be followed in order to avoid data redundancy and to store data in an efficient way. Future needs have to be foreseen and if necessary the DBMS has to be re-evaluated and re-constructed. Once the DBMS is fully operational with all control facilities (privacy law), it can be exploited by the risk assessor using the appropriate techniques.

9. Samenvatting

Het proces van het verzamelen van gegevens tot het ontwikkelen van modellen als basis voor risicoevaluatie is geen eenvoudig proces. Het begint allemaal met het verzamelen van accurate gegevens. Om data efficiënt te benutten, dienen database management systemen (DBMS) geïnstalleerd, onderhouden en ter beschikking van onderzoekers gesteld worden. De kwaliteit van dergelijke DBMS wordt bepaald door de tijd die besteed wordt aan het ontwerp. Een specifieke set van normalisatieregels dient gevolgd te worden opdat overtolligheid van gegevens voorkomen wordt en gegevens op een efficiënte manier opgeslagen worden. Toekomstige behoeften moeten worden voorzien en, indien nodig, dient de DBMS gereëvalueerd en opnieuw geconstrueerd te worden. Zodra de DBMS, rekening houdend met alle beperkingen (privacy wet), volledig operationeel is, kan deze met behulp van de juiste technieken geëxploiteerd worden door de risicoevaluator.

10. Résumé

Le processus de la collecte des données jusqu'à l'élaboration des modèles comme base pour l'évaluation des risques n'est pas un processus simple. Tout commence avec la collecte de données correctes. Afin d'exploiter efficacement les données, des systèmes de gestion de bases de données (SGBD) doivent être installés, entretenus et mis à la disposition des chercheurs. La qualité d'un tel SGBD est déterminée par le temps consacré à son développement. Un ensemble spécifique de règles de normalisation doit être suivi afin d'éviter la redondance des données et d'enregistrer les données d'une manière efficace. Des besoins futurs doivent être prévus et, si nécessaire, le SGBD doit être ré-évalué et re-construit. Dès que le SGBD est pleinement opérationnel, en tenant compte de toutes les restrictions (loi relative à la protection de la vie privée), celui-ci peut être exploité par l'évaluateur du risque à l'aide des techniques appropriées.

11. References

1. Post, 2007. Database Management Systems: designing and building business applications, McGraw-Hill.
2. Hastie et al., 2001. The elements of statistical learning. Springer.
3. Little and Rubin, 1987. Statistical analysis with missing data. Wiley.
4. Schafer, 1997. Analysis of incomplete multivariate data. Chapman and Hall.
5. Bernoulli, 1766. Essai d'une nouvelle analyse de la mortalité causée par la petite vérole. Mem Math Phy Acad Roy Sci Paris.
6. Gelman et al., 1995. Bayesian data analysis. Chapman and Hall.