Acta Genet Med Gemellol 43: 83-88 (1994)

International Congress
40th Anniversary of the Mendel Institute

Knowledge Data Base System for Twins Study

S. Reina, F. Miozza

Informatics Laboratory, The Gregor Mendel Institute, Rome, Italy

Abstract. The medical research on twins, carried out at the Gregor Mendel Institute for Medical Genetics and Twin Study in Rome over the past four decades, has resulted in a vast collection of clinical paper records. A challenge was presented by the need for a more secure method of storage to preserve this enormously valuable historical and scientific patrimony and to render its contents more easily accessible for research purposes. We met the challenge by planning and developing the computerization of this material. New concepts, currently being explored in biomedical informatics, were applied to build a Knowledge Data Base System, using a fourth-generation language (SQL). This architecturally innovative computer system enables its users to manipulate data supplied, rather than just simply storing it. Based on heuristic relational criteria between variables and parameters, the system is employed to solve problems of sibling design analysis typically arising from twins' records, but is also equipped to meet future data base requirements. Another feature of the system is its users' ability to pull off data in the form of regular automated reports, which are distributed through a Local Area Network (LAN). Through a Bulletin Board System (BBS) and modem, any scientist (outside as well as within the Institute) is thus able to access data and exchange scientific information.

Key words: Informatics, Computerized twins research, KDBMS

INTRODUCTION

Informatics and computer science have been used extensively to transfer scientific information from paper to magnetic media [2, 18-20, 25]. Thanks to the speed and power of computers today, the creation of an Informatic Laboratory (ILM) made it feasible to convert to electronic form the huge amount of data on twins held by the Mendel Institute. Our primary goal consisted in developing a common data-structure to link the different areas of scientific activity within the Institute, while, at the same time provid-

https://doi.org/10.1017/S0001566000002981 Published online by Cambridge University Press

ing a central reference bank to be shared by all [11-17, 22, 23]. Up until a few years ago, the more common method of storing and evaluating scientific data was based on the simple creation of electronic data bases. Such data banks were a response to the need of having information sorted according to one or more taxonomic criterion (indexed tables) [1, 3, 4, 7, 10, 17]. More recently, however, a new technique called "B-tree indexing" was introduced, which allowed scientists to interrogate a database rapidly, using a formal computer language [6] called Structured Query Language (SQL). This language is based on the algebraic theory of J. Boole, and enables logical operators like AND, OR, and NOR to search coherent clusters and sub-clusters within a large group of related variable lists, namely the Relational Data Base Management System, R-DBMS [5, 8, 9, 24]. The latest development in these concepts is the Knowledge-DBMS, which allows for qualitative information such as, in our case, twins' genealogy, anamneses and follow-ups to be stored inside the data base in terms of knowledge rules. In

Table

Page	Section	Information units
1	General case information	Name, surname, dates of birth of: twins, mother, father. Zygosity determination criteria Medical pathological case reports in the parental and sibling genealogies.
2	Ethnic origin of the parents and pregnancy data	Geographical location and origin of parents Pharmacological treatment during pregnancy Previous Premature interruption of pregnancy Blood relation between parents
3	Pregnancy profile/assisted fecundation (AF)	Pregnancy duration and delivery modality Fetus presentation Placenta and sac anomalies Pharmacological treatments Infertility treatment or assisted fecundation techniques Location and outcome of AF
4	Siblings	Names, ages, blood groups and karyotypes of brothers/sisters
5	Twin anamneses	Height, weight, lactation and behaviour Follow-up data of twin
6	Twin update	Residence, jobs, habits, growth malformations and pathological events, update of the twins
7	Pathology profile follow-up	Incidence of illness during the follow-up Smoking habits, drug and/or alcohol consumption
8	Psychological profiles	Twins' similarities and peculiarities Interviews with parents

https://doi.org/10.1017/S0001566000002981 Published online by Cambridge University Press

developing its twin data archive, the Mendel Institute adopted "B-tree indexing" and the latest concepts deriving from it. This paper schematically describes the technical engineering aspects of the K-DBMS created, placing special emphasis on the strategy used for categorizing data relating to the pathologies found in twins' siblings.

MATERIALS AND METHODS

Hardware. Four AT-386 and 486 computers were connected in a LAN by using a 10 megabyte Ethernet communication protocol based on a 50 Ohms coaxial cable. One of the machines was equipped with a 1.5 Gigabytes SCSI hard-disk and used as server for central heavy-processing calculations. Another machine was connected to a Hayes compatible Modem (2400 buds UART) to function as a door/bridge between the Informatic Laboratory and the external consultant. One central High-Resolution laser printer (HP LaserJet 4™) was installed to be shared by all the LAN nodes.

Informatic environment. The LAN software configuration was based on a Peer-to-peer design, allowing each workstation to access data and send instructions to any other PC in the LAN without passing through the server station. The Operating system used was Microsoft DOS 6.2™ and the graphical interface for all running applications was Windows for Workgroup 3.11.

Data Base System. The Relational Data Base System used to implement the twin record information was Open Access IV™ (Software Product International, San Diego, Ca, USA).

Programming language. The X-BASE 4th-generation language was adopted within the Open Access IV DBMS to programme for specific management functions such as special queries, automated tabulation, cross table etc. Using this language it was also possible to produce basic statistical contingencies and frequency counts for all the variables (fields) of the records. The programming language was particularly suitable for developing exporting procedures to create external data tabulation formats to share information with other systems, by phone line or simply via floppy disk exchange. Our data base can also be accessed by Digital™ Mainframe, Unix™ System and Apple™ computers.

Statistical package. Besides standard statistical functions, such as average, median and canonical counts, which were included in the DBMS, non-parametric statistical standards, Statgraphics 2.1™ (STSC, Rockville, MD, USA) and SAS™ (SAS, MD, USA) were also used.

Communication Software. The commercial products Cosession™ 6.2 and PCBoard™ 15.0 were used to compute at a distance, and in programming the BBS interface of the twin data base.

DISCUSSION

Since 1950, some 20,000 cases of twins have been clinically studied by The Gregor Mendel Institute and their progress monitored in several follow-up visits. For each twin pair studied, their anamnesis, parental genealogy, sibling and clinical profiles were recorded in a total of 230 information units (fields). Because of the complexity and multiformity of the qualitative information relating to twins' histories, each data base record was divided into several sections to contain specific itemized data. Each section of the electronic archive of the Institute could be viewed on the computer screen as pages of a mask whose frames contained the specific data relating to that section. For example, the third page of the mask might show the section relating to the mother of the twins, and its frames contain information on her pregnancy, labour and delivery as well as any infertility treatment or assisted fecundation programme followed. The Table briefly summarizes the informational sections used in the main twin data base, although many other data bases not described here were used in parallel to facilitate record compilation and to eliminate any possible keyboarding errors.

The specific itemized data within a section of the electronic archive were divided into quantitative and qualitative information, by defining variables and parameters; for instance, within the category of number, we distinguished the number of placentae and sacs from physical parameters such as weight, age, etc., by a logical correlation between the first two values and the variable 'zygosity' instead of the variable 'karyotype'. On the other hand, such variables as pharmacological treatments during pregnancy had a priority score-correlation with regard to numerically related parameters such as weight, age, and number of malformations. Once the architecture of the data base was designed, and the data-entry performed with a mask system, an automated procedure was installed that created a glossary of pathologies intended to speed up searches of particular case-reports. In a simulated trial, the programme was able to select 4 cases of a Hodgkin malignant lymphoma out of 50, 000 total records contained in the Institute's twin data base in 1.2 seconds in a local station. For an external request (RS-232) connection, the time taken was 3.4 seconds. The design of the multi-task and multi-user software allows a large number of people to access the data base to introduce or call up data simultaneously. Moreover, another feature of the the software architecture allows the automatic updating of information introduced into the data base.

CONCLUSION

Our preliminary experience with the new technology on electronic informative media, leads us to the conclusion that a large amount of scientific information can be reinterpreted in a more useful way, thereby allowing for new kinds of epidemiological studies to be undertaken. New, hitherto undiscovered correlations in the investigation of twinning can now be discovered with computer science, if it is applied correctly. Moreover, informatics appeared to be the most suitable and flexible tool to accumulate knowledge that can redirect an entire line of research ergonomically, without rearrangement of the database structure. Lastly, it should be noted that the adoption of informatic media, has

https://doi.org/10.1017/S0001566000002981 Published online by Cambridge University Press

enormously increased the scope for the exchange of data betweeen different disciplines and different laboratories.

REFERENCES

1. Angluin D. Learning regular sets from queries and counterexsamples. Information and Compuatation.
2. Antman EM, Lau J, Kupelnick B, Mosteller F, Chalmers TC (1992): A comparison of results of meta-analyses of randomized control trials and recommendations of clinical experts. JAMA 268:240-48.
3. Appelt D. (1987): Bidirectional grammars and the design of natural language generation systems. In Theoretical Issues in natural Language Processing (TINLAP 3).
4. Bandelloni R, Alabiso G and Reina S (1988): Computer-assisted identification methods for gram-negative marine bacteria. Proceedings of the 5th European Congress. MIS La Nuova Italia Editrice, pp. 199-222.
5. Berliner HJ (1977): Search and Knowledge. In Proceedings IJCAI.77.
6. Berliner HJ (1979): The B-tree search algorithm: a best-first proof procedure. Artificial intelligence 21(1).
7. Bertone S, Casareto L, Reina S, Berti R, Sandri C, Calegari L (1992): Acidi Grassi Cellulari totali: una chiave chemotassonomica computer assistita per la classificazione dei batteri marini appartenenti al genere Alteromonas. Atti del 24° Congresso Nazionale Società Italiana di Microbiologia, Genova, pp. 181-182.
8. Brachman RJ, Levcesque HJ eds (1985): Readings in Knowledge Representation. Los Altos, CA: Morgan Kaufmann.
9. Buchanan BG, Shortliffe (1984): Rule based Expert System: The MYCIN experiments of the Stanford Heuristic Systems. Tech. Rep. 3453. Boston MA: Bolt Beranek and Newman.
10. Casareto L, Bertne S. Reina S, Sandri C, Berti R, Calegari L (1992): Caratterizzazione tassonomica dei batteri del genere deleya mediante confronti computer assistiti dei profili gascromatografici degli acidi grassi. Atti del 24° Congresso Nazionale Società Italiana di Microbiologia, Genova, p. 182.
11. Chalmers I (1991): Improving the quality and dissemination of reviews of clinical research. In Lock S (ed): The Future of Medical Journals, London: BMJ, pp. 127-146.
12. Cochrane AL (1972): Effectiveness and Efficiency. Random Reflections on Health Services. London: Nuffield Provincial Hospitals Trust. (Reprinted in 1989 in association with the BMJ).
13. Cochrane AL (1989): Foreword. In Chalmers I, Enkin M, Keirse MJNC (eds): Effective Care in Pregnancy and Childbirth. Oxford: Oxford University Press.
14. Cochrane AL (1979): 1931-1971: a critical review, with particular reference to the medical profession. In: Medicines for the Year 2000, London: Office of Health Economics, pp. 1-11.
15. Cochrane AL (1993): Cochrane Pregnancy and Childbirth Database. Cochrane Updates on Disk, Issue I. Oxford: Update Software.
16. Enkin M, Kerise MJNC, Renfrew MJ, Neilson JP (1994): A Guide to Effective Care in Pregnancy and Childbirth. 2nd Edn. Oxford: Oxford University Press.
17. Haynes RB (1991): How Clinical Journals could Serve Clinician Readers Better. In Lock S (ed): The Future of Medical Journals. London: BMJ, pp. 116-126.
18. L'Abb KA, Detsky AS, O'Rourke K (1987): Meta-analysis in clinical research. Ann Int Med 107:224-232.
19. Mulrow CD (1987): The medical review article: state of the science. Ann Int Med 106:485-88.
20. Oxman AD, Guyatt GH (1988): Guidelines for reading literature reviews. Can Med Assoc J 138:697-703.
21. Ruggiero C, Giacomini M, Reina S, Gaglio S (1993): A qualitative process theory based model

https://doi.org/10.1017/S0001566000002981 Published online by Cambridge University Press

of the HIV-1 Virus-Celle interaction. Proceedings of Medical Informatics Europe, pp. 147-150. Israel.

22. Silagy C (1993): Developing a register of randomised controlled trials in primary care. BMJ 306:897-900.
23. Sinclair JC, Bracken MB, eds (1992): Effective Care of the Newborn Infant. Oxford: Oxford University Press.
24. Stefanelli M (1988): Sistemi Esperti in medicina: le metodologie, i progetti di ricerca e i prodotti commerciali, MIS La Nuova Italia Editrice, pp. 199-222.
25. Williamson JW, German PS, Weiss R, Skinner EA, Bowes F (1989): Health science information management and continuing education of physicians. Ann Int Med 110:151-160.

Correspondence: Dr. S. Reina, Informatics Laboratory, The Gregor Mendel Institute, Piazza Galeno 5, 00162 Rome, Italy.

https://doi.org/10.1017/S0001566000002981 Published online by Cambridge University Press