EJSO (2005) **31**, 115–116

Available online at www.sciencedirect.com

SCIENCE DIRECT®

EJSO
the Journal of Cancer Surgery
www.ejso.com

EDITORIAL

Research databases and data sets in cancer surgery

Clinical data management presents major challenges and opportunities in cancer treatment. Each patient under care provides an experiment in treatment and a lesson in outcomes. The mass of data from all patients treated successfully and unsuccessfully for cancer produces a mountain of observational and experiential information, which if properly mined, should provide grand insights into the benefits and limitations of treatment strategies.

Professional practice presents a continuous stream of patients of different ages, states of health, origins, specific diagnoses, stages of treatment, social and geographic mobility and clinical outcomes, often requiring multiple interventions. Electronic data collection necessarily takes second place to the immediate needs of clinical care. Well resourced cancer units capture their clinical activity and results on a range of computerized databases of various capabilities and vintages. Unfortunately, the limitations of existing data systems and their connectivity severely constrain the amount of useful insights that can be gleaned.

Modern digital technologies offer the capacity for the rapid and efficient capture, mass storage, manipulation, analysis, mining, transmission and graphical representation of massive data sets. Most clinical databases are limited in their analytical capabilities. They represent dated, legacy systems of various vintages and designs, including self help and commercial systems. Database technology has generally grown out of commercial and government needs for processes such as archiving, inventory, customer and client profiling, rather than to meet clinical design specifications. Most current clinical databases are quite adequate for the task of patient registration and the recording of serial interventions, but they are not generally useful for the interpretration of multiple interventions over various timescales which contribute to clinical outcomes in large and heterogenous populations of cancer patients.

Consider a data set of thousands of breast cancer patients collected over many years. Outside the tightly controlled and selected conditions of clinical trials, most patients will have been treated with different combinations of surgery, chemotherapy, radiotherapy and endocrine manipulations at different times. They will also undergo concurrent treatments for other conditions, and may develop recurrent disease or complications of treatment in different ways. If it were possible to ask complex, multifactorial and multidimensional questions of these data sets, it might well be possible to determine the benefits or disadvantages of each and every component of treatment. For example, do we really know what the global and specific benefits of chemotherapy and the results are in our own local populations of cancer patients of various ages, comorbidities and interventions. Outside clinical trials, treatments run a variable course and may be terminated or compounded by complications and treatment intolerance. We certainly know the results of ideally controlled clinical trials, but do we know how closely these reflect reality in the generality of practice? Try asking such questions of your own proprietary databases.

Clinical data collection has a significant cost. Most medical record keeping is still based on folders of written notes. Data research and input takes expensive professional and clerical time, and the larger the number of fields in each record, the greater the expense. Excessive data collection for contingencies and opportunistic future analysis may waste substantial resources, while the rapid evolution of generations of software, hardware and storage devices incurs costs and losses in data transport to new platforms. Poor design and

0748-7983/$ - see front matter
doi:10.1016/j.ejso.2004.09.016

inadequate or misdirected data collection risk missing insights hidden in the data. We need affordable, carefully thought out data analysis systems which provide new levels of insight in local systems and large public registries.

A number of new digital tools merit closer inspection. For example, technology helps rapid, cost effective data capture through bar code readers, swipe cards, optical character recognition, voice recognition and patient assist data entry systems. The Internet and the use of standard transmission protocols allows more efficient data sharing, subject to the constraints of data security and protection.

There have been substantial advances in the visual presentation of complex data sets. At the 'front end' of interaction between the user and the computer screen, there have been advances in graphical tools, maps, trees, colour enhanced data presentation tools, and interactive screen based devices. These include pointers, clickable icons, movable gates and sliders. These tools allow the selection and analysis of subsets of large data bases and continuous variables in multiple dimensions in visually effective and easily understood formats.

The representation of time dependent variables in current generations of cancer databases is a particular problem. For example, the lifespan of a patient with cancer may cover many years from diagnosis, and involve a whole series of interventions, each of which will be subtly different in timing and type from those in other patients. How can we analyse the significance of the sequencing of adjuvant treatments over large data sets? There are no simple tools for such analyses, which are thus rarely if ever attempted. Newer graphical tools based on electronic calendar concepts, and project management software, which allow the plotting of many parallel, serial and interrelated events along a time line, may help.

Data is the raw material of scientific and clinical enquiry. It does not reveal insights unless specific questions are conceived and asked of it, and unless the tools exist to interrogate the data and present the findings in intelligible ways. Data Warehousing and Data Mining, tools essential to modern large retail commercial enterprises such as supermarket chains, allow large volumes of data to be interrogated for trends and abnormal events. Some of these techniques of analysis are logical and systematic. Other methods, such as neural network analysis, have a touch of alchemy about them, sometimes referred to as the black box approach, wherein the machine generated outputs cannot readily be linked by the human observer to the data inputs.

Such systems may also improve the management of patients in the clinic. On Line Transaction Processing allows real time decision-making based upon complex variables and drawing on experience and data held elsewhere. Thus for example, treatment selection can be optimized for individual cancer patients with complex variables of age, treatment history, concurrent illness from a computer terminal in the consulting room.

The digital technology wave rolls on. Units and organisations will be wary of investments in new generations of software systems, but the limitations of current cancer databases in the context of the achievable are now apparent. Better clinical data analysis might obviate the need for many clinical trials. The development of such systems is work for the best brains and best resourced organisations in computer software and systems analysis, but the rewards are considerable in economic, clinical and scientific terms.

We hope to further the understanding of the problems and solutions to data management in the pages of the EJSO, through the publication of papers and the stimulation of debate. We would welcome contributions in the subject area both from clinicians and computer specialists, and we hope that you will feel prompted to engage your friendly neighbourhood software brains to focus on the problems and contribute to our pages.

D.A. Rew*
Editor, EJSO

*Tel.: +44 2380 767823; fax: +44 2380 825148.