Pergamon

Mathl. Comput. Modelling Vol. 19, No. 6–8, pp. 273–288, 1994
Copyright©1994 Elsevier Science Ltd
Printed in Great Britain. All rights reserved
0895-7177/94 $7.00 + 0.00

0895-7177(94)E0054-Q

Future Directions in Computational Nursing Sciences*

S. L. MEINTZ
Nursing, University of Nevada, Las Vegas, NV 89154, U.S.A.
E. A. YFANTIS
Computer Science, University of Nevada, Las Vegas, NV 89154, U.S.A.
W. P. GRAEBEL
University of Michigan

Abstract—The advent of the supercomputer and its capabilities for dealing with terabyte-sized data bases has provided a unique opportunity for nursing sciences to enhance and add to its theories. An interdisciplinary team has formed at the University of Nevada, Las Vegas (UNLV), to provide new tools and methodologies for analyzing large-scale data bases. Their first project is a study of infant mortality. The strategy and goals for this project are presented, along with an assessment of the present state of health care data base analysis.

Keywords—Computational nursing, Health care data sets, Interdisciplinary, Nurmetrics, Nursing informatics, Nursing science, Supercomputers.

INTRODUCTION

Nursing has traditionally been a profession whose science was subsumed from other disciplines. Research which originated from nurses was traditionally nonmathematical and noncomputational in nature (as in fact has been true for much of the medical research performed by medical practitioners), often involving the collection of data of small sample sizes with marginal statistical significance. Nevertheless these results were added to the nursing literature. What analysis was performed was typically statistical in nature using packaged statistical programs with no special adaptation of the statistical program to the case in hand. Within the last few years, this mode of operation has been gradually changing, with a shift toward the development and utilization of mathematical tools suited directly to the needs of nursing within the general area of computational health care.

In recent years nursing science has broken out of the traditional mold and has added a strong scientific and theoretical base. In doing so, it has recognized the value of using mathematical form for graphically representing abstract conceptualizations, particularly for describing, explaining and predicting nursing practice. The theory of nursing knowledge can be expressed by the mathematical equation,

$$NF + M + NE + DI = NK \tag{1}$$

the terms in this expression are defined as follows:

NF stands for the Nursing Foundation, a combination of knowledge from both the sciences and humanities. The science contribution includes the empirical knowledge base of nursing

This research was supported by Cray Research Inc., Silicon Graphics Inc., National Center for Health Statistics and Advanced Computer Laboratory at Los Alamos National Laboratories.

*This paper was typeset in $\mathcal{A}\mathcal{M}\mathcal{S}$-TEX and galley proofs to authors provided from the offices of the Guest Editor.

Typeset by $\mathcal{A}\mathcal{M}\mathcal{S}$-TEX

practice, as well as, contributions from associated disciplines such as biology, the physical sciences, medical science, and chemistry. The humanities contribute knowledge from the philosophical and cultural studies that investigate human constructs and concerns.

M represents Methodology applied towards problem solving. The appropriate methodology depends on the context of the problems and the practice domain in which it resides. In clinical practice what is appropriate is the methodology of nursing practice, in nursing research the scientific method is appropriate, in nursing administration it is strategic thinking methodology, and in nursing education taxonomy of education is appropriate.

NE Nursing Essence, represents the evolution of nursing as a profession. Nursing essence is defined according to the practice domain, and includes among others parameters such as: the principles of nursing science; legal parameters; and the definition of person, environment, health, and nursing.

DI Disciplined Inquiry, refers to investigation or experimentation aimed at the discovery and interpretation of facts, revision of accepted theories of law in the light of new data, and the practical application of new or revised theories or laws.

NK Nursing Knowledge, is the understanding of nursing as a science and an art. It included the nursing foundation, essence, methodology, and disciplined inquiry.

The application of nursing theory to practice can be represented by the formula

$$(NK)(I) = P. \tag{2}$$

Here I represents the nurse's integration and synthesis of nursing knowledge through the cognitive, psychomotor, affective/spiritual domain of self. P, the depth and breadth of practice, increases or decreases according to the application of nursing knowledge to this integration.

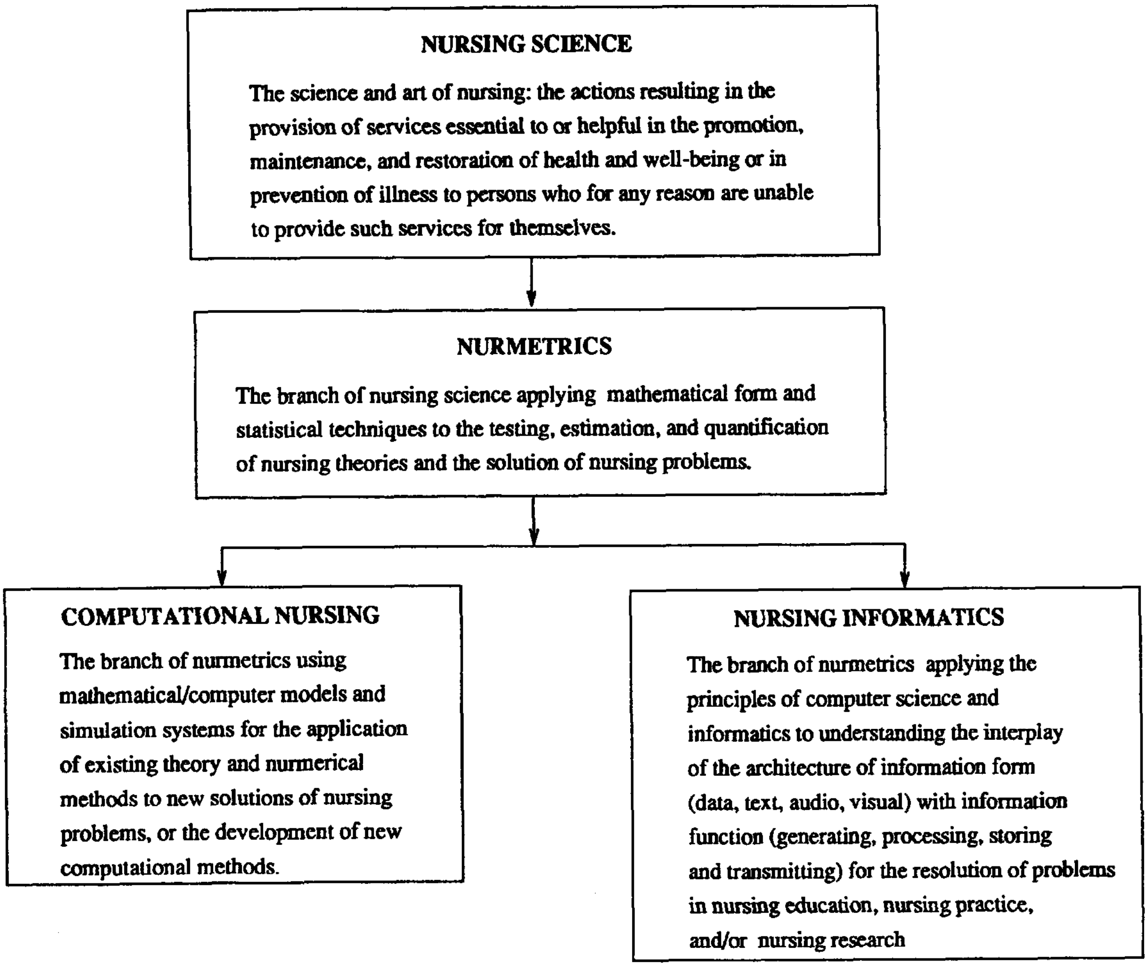
As nursing has become a more scientific profession, several new branches of nursing science have developed. *Nurmetrics* is a new specialty area within the field of computational health sciences. It is the aspect of nursing science which applies mathematical formulations and statistical techniques to the testing, estimation, and quantification of theories and solutions of nursing problems. It can use mathematical/computer models and simulations to test existing health care theories to produce either new solutions for nursing problems, or to develop new computer models or methodologies for nursing. This aspect of Nurmetrics has been designated *Computation Nursing*. Alternately, it can apply the principles of computer science and informatics to understand the interplay between the structural form of the information (text, raw data, audio, visual) and the function of the information (generating, processing, storing, transmitting), to see how the architecture of the information available aids in solving the problems encountered in nursing administration, education, practice, and research. This aspect of nurmetrics has been termed *Nursing Informatice*. Table 1 summarizes the relationship between these branches.

Extensive archived health care sets have been collected by many sources over the past decades, from both government and nongovernment sources. These data bases include information vital to such applications as the monitoring of health costs and quality, monitoring of vital statistics records for trends, and determination of the informatics portion of such projects as the Human Genome Project. The size of the data sets have progressively grown, in many cases faster than the capabilities for analyzing them. Currently, more than 250 gigabytes of national nursing and health data exists. Trends in data taking of health data indicate that future data sets will be even larger, rapidly approaching in size the terabyte range. Past, and unfortunately, current tendency is to expend funds to generate, transmit, and store this data, but to allocate little or no funds for data analysis.

PRESENT STATE OF AFFAIRS

For several reasons, much of the information in these data bases today still remains untapped. While we are now accustomed to the availability of ever more powerful personal computers and

Table 1. A new branch of nursing science.



NURSING SCIENCE

The science and art of nursing: the actions resulting in the provision of services essential to or helpful in the promotion, maintenance, and restoration of health and well-being or in prevention of illness to persons who for any reason are unable to provide such services for themselves.

NURMETRICS

The branch of nursing science applying mathematical form and statistical techniques to the testing, estimation, and quantification of nursing theories and the solution of nursing problems.

COMPUTATIONAL NURSING

The branch of nurmetrics using mathematical/computer models and simulation systems for the application of existing theory and nurmerical methods to new solutions of nursing problems, or the development of new computational methods.

NURSING INFORMATICS

The branch of nurmetrics applying the principles of computer science and informatics to understanding the interplay of the architecture of information form (data, text, audio, visual) with information function (generating, processing, storing and transmitting) for the resolution of problems in nursing education, nursing practice, and/or nursing research

workstations at ever decreasing cost, looking back at the history of computers makes the present seem like something out of last decade's science fiction. In the mid-1950's the IBM650 was the first mass-produced computer with a sufficient large internal memory system (a rotating magnetic drum) to allow internal storage of the program. Before that, the program existed in punched cards which were read serially by the machine. Looping was done by duplicating that portion of the card deck a sufficient number of times, a cumbersome process at best! The 1960's saw the introduction of the IBM7xx(x) computer family which added other forms of internal memory storage as well as the ability to input large data sets from magnetic tape. In the 1970's supercomputers started to appear, although it would be another decade before their use started to become widely available. There allowed rapid processing of the data, changing the time scale for processing from weeks to hours. And, the early 1980's saw the advent of the personal computer and the computer workstation, which in many instances, dwarfed the capabilities of the main frame computers of the 1970's and before. The late 1980's was perhaps the first time that computational capabilities were finally capable of handing the vast size and complexity of the health care data bases.

The development of software and computational techniques naturally follows somewhat behind the development of the computer hardware. The primary concern of much of the techniques still practiced today is to reduce the data so as to select only that portion of the data which is "pertinent" to the study. This has influenced the initial collection of the data, with the survey questions being asked influenced by and being tailored to the computer capabilities available to the investigator [1–4]. In many cases, the reduction process carried out on the date is done by first performing a preliminary analysis of the data set using assumptions such as linearity of interaction of the variables in the data set. Even with the best of intentions biases could well be introduced in the conclusions found from the study. It cannot be stated with certainty that such biases do exist, but neither can it be stated with certainty that these biases do not exist!

NEED FOR AN INTERDISCIPLINARY APPROACH

A vital component of the computational scheme, the nursing and health science analyst, has not developed at the same rate as computer technology and has not kept up with the high-tech advances in computer capabilities. The training of nurses and doctors normally spends little time on mathematics of computational skills. The nonmathematical physics course is more the norm than are semesters if advanced mathematics and programming. Thus, the health care specialist has had to rely on statisticians, mathematicians, and programmers to do the analysis of health care data for them. In many cases, this has been done in institutes which are more social science oriented than health care oriented. While many goals are shared by these two groups, there are important gaps in the information output which has resulted in a filtering process, to the degradation of information available to the health care practitioner. What appears to be needed is an interdisciplinary team approach, suited to the needs of the health care practitioner.

A paradigm for an interdisciplinary approach might be the field of bioengineering, which started in the 1960's. There began a gradual and continuing contact between medical doctors and engineering researchers, prompted by the success of devices such as the first mechanical heart pumps. Research performed by such teams typically would start out being performed in an iterative manner. The medical doctor would propose a medical problem which he thought deserved attention and development towards a solution. The engineer would then translate the problem into his or her field of expertise, and then present a first step in the solution of the problem in the form of what seemed possible at that time. Criticism, comments, and tests, with much interaction between the medical doctor and the engineering researcher, led to further development towards the goal, and to a broadening of the understanding between the members of the team and respect for their mutual abilities. The medical practitioner, with his training and practice could best formulate the needs for the research and its eventual absorption into medical practice. The engineering researcher, trained in abstract analysis in a wide range of physical disciplines, could cast the problem into either a physical or mathematical mode, and provide a quantitative analysis which could improve present practices. While strictly mechanical problems such as mechanical hearts and artificial joints are the obvious results of such interdisciplinary approaches, better knowledge of pulmonary and blood circulation and flow, understanding of the structural behaviour of the spine and limbs, optima; administration of drugs, and hospital administration and utilization all have benefited from such an approach. It would appear that such an interdisciplinary model, so successfully applied to the benefit of the medical doctor and hospital administrator, could equally well benefit the nurse health care practitioner.

The need for this new direction in nursing research is particularly timely now. The new administration in Washington has focused on health care costs as being a major concern of the United States economy. One of the factors with the largest percentage of the national health care costs is hospital costs, which have grown at a rate well above the rate of inflation of the overall economy. A major deliverer of health care in hospitals is the nursing staff, whose procedures have grown from a number of sources over time, but with relatively little input from nursing researchers. For nursing research to be recognized as being of an equal stature with other health care research, it is imperative that it be equal in scientific sophistication. Since biological and health care research has changed greatly in direction toward more use of mathematics, the computer, and mathematical/computational simulation [5], the need for nursing research to move in this direction is clear.

As a start towards such an interdisciplinary team, in May of 1991 the project for Nursing and Health Data Research (PHNDR) of the University of Nevada, Las Vegas (UNLV) received access to massive nursing and health data sets from both government and nongovernment sources. In many cases, this data had been archived without analysis because many agencies neither had access to the type of computers needed to analyze data sets of this size, nor to the computing and statistics professionals needed for processing of the data [6]. For example, the data collected for Medicare in just one year documents 10 million cases equaling 15 reels of data, or 1.8 gigabytes. To complete a five-year comparative or trend analysis, a computing system capable of handling 9 gigabytes of data is required. Analyzing trends over the length of the average human life would require computing resources several orders of magnitude greater.

The research involved in PHNDR's study of these data bases included the following items:

(1) Conceptualization of complex problem resolution,
(2) Design of product and/or process,
(3) Development of product and/or process,
(4) Dissemination of preliminary results and/or prototypes,
(5) Alpha sites application for testing and debugging,
(6) Beta site application for refinement and further debugging,
(7) Dissemination of product and/or process through software, manuals, educational programs, publications and/or presentations.

To perform this R & D effort, a framework was developed to validate the application of supercomputers to nursing and health data research. The case study utilized a manageable data base of 17,000 cases which were analyzed with the Statistical Package for Social Scientists (SPSS). Computer platforms including personal computers, Sun workstations, and Cray Y-MP supercomputer were used to analyze the data. The Cray Y-MP was able to process the data in approximately 127 seconds. The Sun workstation was several orders of magnitude slower than this, while the personal computer was unable to process the data. The work resulted in a multilevel approach for supercomputing application to establishing the foundational parameters of health through gigabyte/terabyte data base analysis. Table 2 summarizes the issues involved in the phase one study.

The initial conceptualization for complex problem resolution resulted in identification of barriers to supercomputer application to gigabyte-sized data base analysis in nursing and health. The research focused on removal of barriers to allow successful analysis of gigabyte/terabyte-size data sets. To remove these barriers, a statistical analysis process termed GATES (for global analysis terabyte exploratory statistics) was designed with the aid of researchers from the Los Alamos National Laboratory. The prototype for GATES was developed using 249 megabytes of data from the Hispanic Health and Nutrition Examination Survey, which was obtained from the National Center of Health Statistics.

This prototype was selected for demonstration as a Research Exhibit at the Supercomputer'92 Conference. Even with GATES, additional questions related to statistical analysis of gigabyte/ terabyte-size bases have arisen, providing a direction for further methodology, conceptualization and design, and development of both process and product.

The original concept of using high performance computing for the analysis of nursing and health gigabyte/terabyte-size data base was initially perceived as a simple application process. However, the complexity of gigabyte-size data base analysis has resulted in a significant challenge for the computational health sciences. Before gigabyte-sized data analysis could be used to establish the foundational parameters of health, removal of barriers to supercomputer application was necessary. Actual and potential barriers exist related to theory, statistics, software, hardware, data base management, data access, supercomputing resources, learning curves for nurse researchers in high performance computing and communication, and establishing a new interdisciplinary team incorporating nurse researchers, engineers, computer scientists, programmers, computational mathematicians, etc. [7].

Several separate areas evolved from this preliminary research investigation. These functional areas provide organization structure for the present proposed research effort. Each of the areas has independent but interrelated functions. The functional areas include the following:

(1) Complex Problems Resolution/Future Directions Function: Responsible for the conceptualization of complex problem resolution, determining criteria for design, establishing future development, proving theoretical function, and providing technology transfer.

(2) Scientific Validation/Research Advisor Function: Responsible for selection of data sets for analysis, development of validation procedure for statistical methodologies, identification of existing statistical problems with analysis of complex survey design data, advise or mentor users group.

(3) Methodology/Software Function: Responsible for design and development of software, expert systems interface, graphic user interface, computer model simulations linking or interfacing existing data sets, and determining new simulation linking or interfacing existing data sets, and determining new statistical methodologies for gigabyte/terabyte-size data base analysis.

(4) Operation/Hardware Function: Responsible for user productivity, transparent systems application, and machine peak performance.

(5) Interactive Archival/Research User Function: Responsible for collecting and archiving data sets from government, nongovernment and international sources, identify research needs, service enrolled research users, monitor system effectiveness, refer research user problems with gigabyte/terabyte-size base analysis to the Scientific Validation/Research Advisory Group.

(6) Information Dissemination/Education Function: Responsible for development of user's manuals, education programs, marketing products (software) and process, and demonstrations, publications and presentations.

The research which has been performed in these areas has heavily influenced the conception of the present project.

THE UNLV INTERDISCIPLINARY TEAM

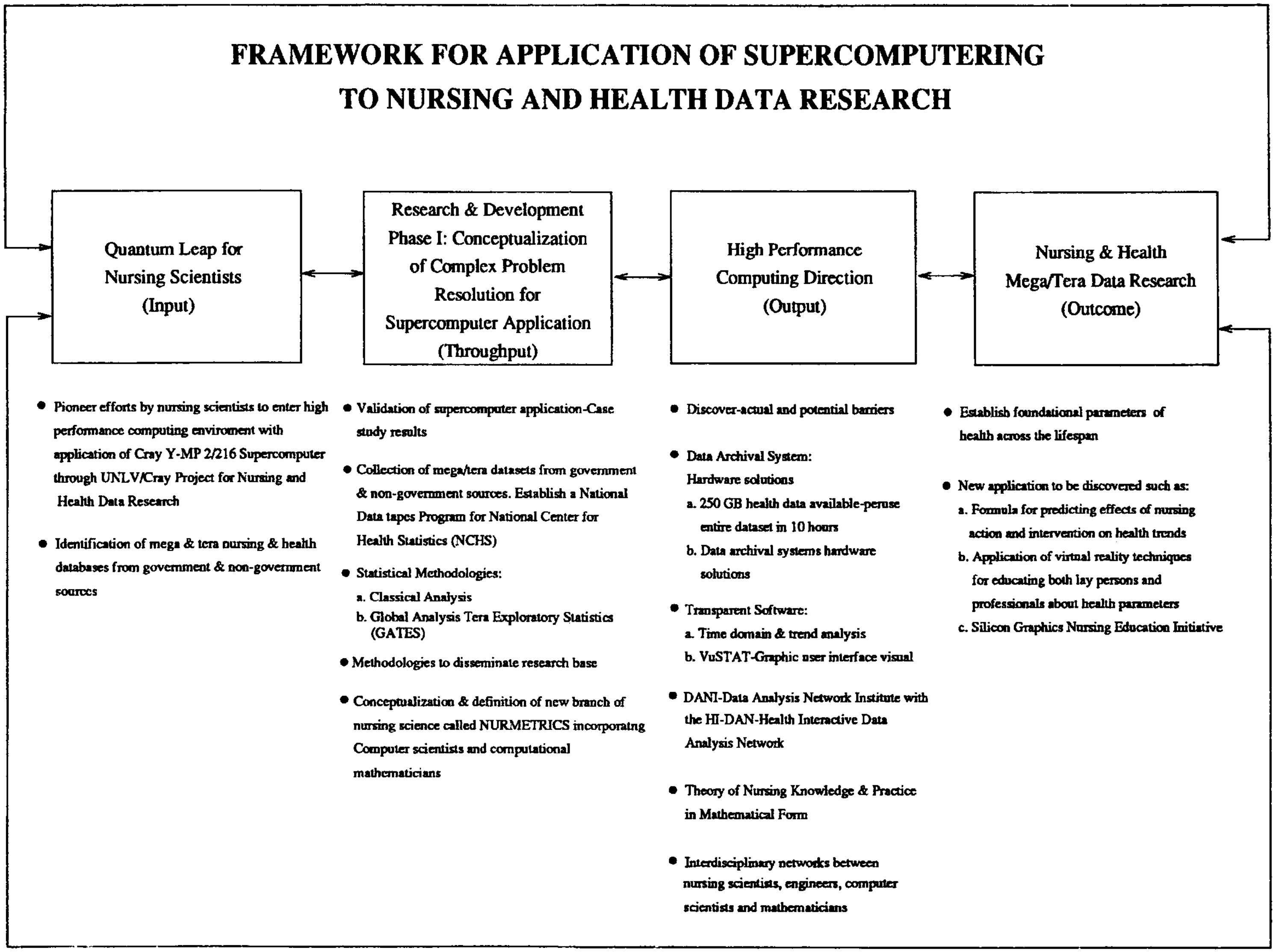
As a continuation of this effort, an interdisciplinary research team has formed at UNLV with the express purpose of advancing the statistical analysis of the gigabyte-size data bases which are presently available. The team consists of individuals with research interests in nursing, computer science, and engineering, and having a strong knowledge of applied mathematics and computational skills, together with problem-solving skills acquired from analysis and simulation in a broad range of physical problems in field including medical research. Collaboration has been initiated between this team and researchers at both Los Alamos National Laboratory. The Cray Y-MP at the National Supercomputing Center for Energy and the Environment at UNLV provides the necessary computer support.

INITIAL PROJECT

As a first research project for this team, funding is being sought to analyze sets of health data presently archived by government and nongovernment resources to study infant mortality (IM). The object of the analysis will be to determine which variables in existing data bases correlate strong with infant mortality, and to assess the importance of these variables in their contribution to infant mortality. Analysis of the large samples of the population represented in these data bases could support or nullify current theories supporting medical practice and the foundations of health care for infants.

Current methods of analysis of these data bases to establish deterministic parameters affecting health, illness, and death are not sufficient to obtain maximum information from such massive

Table 2.



FRAMEWORK FOR APPLICATION OF SUPERCOMPUTERING TO NURSING AND HEALTH DATA RESEARCH

Quantum Leap for Nursing Scientists (Input)

- Pioneer efforts by nursing scientists to enter high performance computing enviroment with application of Cray Y-MP 2/216 Supercomputer through UNLV/Cray Project for Nursing and Health Data Research
- Identification of mega & tera nursing & health databases from government & non-government sources

Research & Development Phase I: Conceptualization of Complex Problem Resolution for Supercomputer Application (Throughput)

- Validation of supercomputer application-Case study results
- Collection of mega/tera datasets from government & non-government sources. Establish a National Data tapes Program for National Center for Health Statistics (NCHS)
- Statistical Methodologies:
  a. Classical Analysis
  b. Global Analysis Tera Exploratory Statistics (GATES)
- Methodologies to disseminate research base
- Conceptualization & definition of new branch of nursing science called NURMETRICS incorporatng Computer scientists and computational mathematicians

High Performance Computing Direction (Output)

- Discover-actual and potential barriers
- Data Archival System:
  Hardware solutions
  a. 250 GB health data available-peruse entire dataset in 10 hours
  b. Data archival systems hardware solutions
- Transparent Software:
  a. Time domain & trend analysis
  b. VuSTAT-Graphic user interface visual
- DANI-Data Analysis Network Institute with the HI-DAN-Health Interactive Data Analysis Network
- Theory of Nursing Knowledge & Practice in Mathematical Form
- Interdisciplinary networks between nursing scientists, engineers, computer scientists and mathematicians

Nursing & Health Mega/Tera Data Research (Outcome)

- Establish foundational parameters of health across the lifespan
- New application to be discovered such as:
  a. Formula for predicting effects of nursing action and intervention on health trends
  b. Application of virtual reality techniques for educating both lay persons and professionals about health parameters
  c. Silicon Graphics Nursing Education Initiative

data bases [8–11]. The size of the data bases, of the order of 0.3 terabytes, far exceeds the capacities of present software packages and small computers unless the date is somehow reduced to fit the computer and the available software. As pointed out above, such reduction can bias the data, and may remove important factors affecting the meaning of the data. Further, the number of variables in these data bases far exceeds the capabilities of traditional software for providing meaningful analysis. Even traditional mainframe computers are not capable of such a task.

To perform the analysis of infant mortality data, it will be necessary to first develop the tools needed to cause a paradigm shift in procedures for analyzing the information contained in gigabyte/terabyte-size health care data bases. The foremost issues which must be met in developing these tools to make the information contained in the health care data bases easily accessible to health care researchers and practitioners are as follows:

(1) Development of statistical computer programs suited to analyzing data contained in massive data sets. This will involve a survey of existing statistical software for large data bases. It may involve modifications of current statistical software packages, so that they are suited to deal with gigabyte-size data sets, or it may require the writing of new computer codes. Use of the Cray Y-MP supercomputer at the University of Nevada, Las Vegas will remove the limitations of computer time and storage requirements imposed by smaller computing facilities. This portion of the research study will provide novel methodology and new computer software for rigorous monitoring over time of infant mortality.

(2) Development of interfaces so that data bases in varying formats can be analyzed by the same statistical and graphical packages. Present data bases are available in a wide variety of formats. Interfacing software is needed so that they can be analyzed by the same statistical software.

(3) Development of a visual graphical processing package, to allow rapid visual inspection of data base to confirm or modify stratification and significance of variables, or to establish new relationships between variables not previously discerned. While statistical numbers derived from the data bases provides important information concerning the data set, visual inspection of the raw data on a graphical workstation can suggest changes in stratification or analysis not apparent from calculated statistics [12,13].

However, the consequences of the research are even broader than infant mortality per se. It is generally agreed [14] among health care researchers that of the many statistical measurements in public health, infant mortality rates are an accepted indicator of the health of a population. The continued assessment of population-based pregnancy outcomes among minority ethnic groups must remain a public health priority if infant mortality rates in the United States are to be further reduced [15]. Infant mortality rates have long been of interest in the United State. The pace of research on its determinants has increased over the last decade upon recognition that

(1) the mortality rate is in decline,

(2) the mortality rates in the United States are high relative to other industrialized nations, and

(3) persistent and substantial infant mortality differentials exist between population groups [16].

While the principle objective of the IM research project is an analysis of the factors affecting infant mortality, major secondary benefits will occur. These include the following:

(1) Maximization of the utilization of gigabyte-size data bases by developing software to allow high performance computing and visual communication and interpretation of these data bases.

(2) Development of user-friendly software, graphic user interfaces, and computer model simulations to improve the analysis and understanding of the large data bases available in health research.

Exploratory analysis of the available data will be used as a means to gain a better understanding of the variables which are the major contributors to the infant death rate. Sensitivity analysis will be performed during the early stages of experimental design to gain an insight as to how a change in a variable affects the mortality rate. Extensive use will be made of computer graphics in order to visualize the variable mentioned and their effect on infant mortality. Other variables included in the available data sets will also be studied, resulting in not only a visualization of the change and behavior of each variable separately, but also the joint contribution of two or more variables to the infant death rate. During this exploratory analysis, identification will be completed regarding the measured variables in the available data base for inclusion in the statistical analysis, and which are important, and therefore, should be included in our mathematical model.

HEALTH FACTORS TO BE CONSIDERED

Multiple variables impact on infant mortality, including improved socioeconomic status, housing, nutrition, immunization, pure water, pasteurized milk, antibiotics, improved prenatal care and delivery, and technological advances in infant care. Some of the variables affecting infant mortality, particularly the neonatal mortality rate, are heavily influenced by biological and genetic factors, quality of the delivery, low birth weight, and availability of the services. The postneonatal mortality rate is heavily affected by the health care system, but also by factors such as the availability of housing, sanitation, adequacy of nutrition, and other external and environmental factors which are not controlled by the health care system. However, at the present time a rigorous monitoring of the impact of these variables over time is not available.

Some of the elements/variables contained in existing National Center for Health data base which will be utilized in this study are shown in Table 3. They represent the variables which have received the most attention in previous studies [15,17–49].

The factors listed in Table 3 are clearly not independent. Interaction effects can be reasonably expected, with the effects of some of the independent variables varying depending on the magnitude of other independent variables [16]. While several strategies have been tried to empirically determine these interactions [50–53], we will test a generalization of a model proposed by Eberstein *et al.* [16], which identifies interactions already having empirical support and in combinations. This computation will be done, both using analysis of the data bases and also using the simulation model.

ANALYTICAL PROCEDURE

Nurses and health care researchers typically use classical linear analysis to formulate theories. That is, they first reduce the data base to a manageable size and from this reduced the set of data from generalizations for larger populations and broader theories. Frequently, the sample size is barely large enough to be statistically significant, and the effects of nonlinear interactions between the independent variables are ignored. In spite of this, the research results are added to the body of scientific health care knowledge, and health care practice is based on this theory without the benefit of validation [7]. Presently, a particular disadvantage accruing from these shortcomings is the lack of prediction methodologies for use by the health care provider or government planner.

The desired outcome of the IM project is to establish foundational parameters affecting the cause of infant mortality. In order to accomplish this outcome, gigabyte-size data bases must be accessible to analysis through transparent software using high performance computing and communication techniques. To date, the majority of statistical methods and software for analysis of large health data sets have not utilized the advantages of a high performance computing environment. Even at the National Center for Health Statistics, in Maryland, there has been an absence of workstations, and a high performance computing and communication environment is very limited.

Table 3. Data elements/variables contained in National Center for Health data bases.

1. GENERAL		
Match status	Year of birth	Year of death
2. OCCURRENCE		
Region State	Division County	Expanded state
3. RESIDENCE		
Region State	Division County	Expanded state City
4. INFANT		
Race Gestation Apgar score	Sex Birth weight	Age Plurality
5. MOTHER		
Origin or descent Education	Race Marital status	Age State of birth
6. FATHER		
Origin of descent Education	Race	Age
7. PREGNANCY ITEMS		
Interval since last live birth Month prenatal care began Total birth order		Outcome of last pregnancy Number of prenatal visits Live birth order
8. MEDICAL DAT		
Underlying cause		Multiple conditions
9. OTHER ITEMS		
Place of delivery Hospital and patient status Autopsy performed		Attendant at birth Place of accident

Analysis of the large data bases available to nursing and health research is necessary for establishing the scientific foundational parameters of health. High performance computing and communication have been validated as the most cost effective means for analysis of the gigabyte/terabyte-size data sets. However, statistical methodologies, transparent software, and availability of supercomputing resources greatly limit the access and analysis of data sets. This research effort confronts development of process and product for gigabyte/terabyte-size data analysis in nursing and health with application to associated disciplines.

Data analysis will be carried out in the following fashion. The data will initially be stratified according to the following factors:

(1) The mother's age group. The age groupings initially proposed include the following five:
   (a) less than 18 years of age;
   (b) between 18 and 23;
   (c) between 23 and 30;
   (d) between 30 and 40;
   (e) older than 40 years old.

The rationale for this age classification are the following.

(a) Mothers younger than 18 years of age constitute a special class of teenagers having special problems associated with them (accidental pregnancies, not having finished high school yet, having financial difficulties, often times single parents, not having marketable skills, etc.).

(b) The group of mothers 18 years to 23 are likely mothers getting married immediately upon exiting high school, or before finishing college, or your mothers in families with blue collar types of occupation.

(c) The age group 23 to 30 represents the young group just graduated from college, or having a steady job. In general they can be expected to be more established than the previous group.

(d) The 30 to 40 year age group represents the more mature group, possibly having more education and better financial means that the previous group, or being a more mature group with other children in the family.

(e) The last group represents a special class of mothers that appears to be on the increase in our time. These mothers either wait to have children later in their lives or do not hesitate to have children while in their forties.

An estimate of the death rate based on each one of the age groups mentioned above will be calculated. Our analysis includes the effect of the mother's age group on the death rate of infants. Since there are more factors besides the age group of the mothers, the total variance connected to these estimates is a sum of the variance components of these additional factors. Furthermore, the variance components of a particular factor could be different within each on of the age groups mentioned above.

The stratification of the mothers into the age groups mentioned above is based on our initial model. After performing an exploratory analysis using large scale computer graphics, if the national "linked birth/infant death" data does not agree with our initial model the proposed stratification of the mothers to age groups will be changed to reflect the information obtained by the exploratory analysis of the national data.

(2) Birth weight factor is of importance since congenital malformations, delivery complications, and perinatal conditions seem to be most dependent on birth weight [16].

(3) The existence of prenatal care is known to have an important effect on the death rate of infants.

(4) The race/ethnicity of the mother and of the father, especially if the mother is married, is an important factor in infant mortality. An important question is whether the death rate of infants is race/ethnicity-dependent, or if the race/ethnicity of the mother affects the probability of the survival of the infant. The inclusion of the race/ethnicity of the mother in our model could provide information as to whether the attitudes of health care professionals towards various groups of women involved in our study, or the attitudes of the mother towards the newborn infant, might be race/ethnicity dependent.

(5) The question of health insurance for the mother is important, particularly as to whether the lack of health insurance for the mother and child was a contributing factor to the infant death rate.

(6) The experience of the mother on having and caring for children would be an important factor to the infant death rate. Mothers that have had experience in raising children might be able to cope and care for infants much better than mothers having their first child. Inexperienced mothers with no close support groups might impact on health problems related to infants.

(7) Marital status of the mother is of importance in that unwed mothers have the pressures of taking care of the infant on top of the normal pressures of providing income, housing, and food. Many times other emotional and social pressures are also associated with their status [54].

(8) The medical history of the mother, especially alcohol abuse, drug usage, or other substance abuse can be expected to be very significant in affecting infant mortality, since substance abuse of the mother is likely to have detrimental effects on the health of the infant and to be associated with underweight births. Also associated with substance abuse, might be a number of problems such as child neglect, child abuse, and a general lack of proper care for the infant.

(9) The frequency of visits to the pediatrician or other health care services can be expected to have an effect on the health of the infant.

(10) A geographic effect can play a role in infant mortality. States, counties or geographic areas in the country can reflect a favorable or non-favorable attitude on infant care, including the availability of health care facilities. The expected trend from the mentioned attitude may have a spatial correlation, and a zone of influence [22,55,56].

(11) A final factor to be analyzed will determine from the data bases the presence of a time dependency between months of a year and a trend over the years. A time series model [57] will be constructed, and time trend analysis methods will be used, to discover whether the data suggest a time trend or other time dependencies.

PROPOSED INFANT DEATH MODEL

The factors identified in the previous section as contributing to or affecting the infant death rate are not necessarily listed according to their importance, or effect, on the infant death rate. Furthermore, our exploratory analysis of the data might show us that some of these factors are not as important as we initially thought. It is expected then, that other factors, not considered initially, would be incorporated into later models.

The factors listed above are clearly not independent. Interaction effects can be reasonable expected, with the effects of some of the independent variables varying depending on the magnitude of other independent variables [16]. While several strategies have previously been tried to empirically determine these interactions [50–53], we will test a generalization of a model proposed by Eberstein *et al.* [16] which identifies interactions already having empirical support and then examines comparatively the contributions of these variables both individually and in combinations. This computation will be done both by an analysis of the data in the data bases and also by using the simulation model.

The data will be analyzed in the following manner. The available samples of the mothers will be divided into n age groups. The age group will be denoted by $A_1 = 1, 2, \dots n$. The existence or nonexistence of prenatal health care will be denoted by A_2, where $A_2 = 0, 1$. Thus $A_2 = 0$ might denote no prenatal health care, while $A_2 = 1$ might denote the existence of prenatal health care. The variable A_3, taking on the values $1, 2, 3, 4, \dots,$ will denote the race/ethnicity of the mother where 1 denotes Caucasian, 2-African American, 3-Hispanic American, 4-Native American, etc. A_4 will be a similar variable describing the infant's father. A_5 will be a variable describing whether or not health insurance was available. The variable A_6 will be associated with the experience of the mother in raising children. Therefore, if the mother had other children prior to this infant, then $A_6 = 1$, else $A_6 = 0$. The factor A_6 will describe the marital status of the mother. Thus, if the mother is single $A_6 = 0$, and if the mother is married $A_6 = 1$. The variable A_7 would relate to possible substance abuse of the mother during pregnancy. Hence, if the mother did not use drug, alcohol, or cigarettes, $A_7 = 0$, if the mother smokes cigarettes $A_7 = 1$, if the mother uses alcohol $A_7 = 2$, if the mother uses drugs $A_7 = 3$, etc. The eighth factor would relate to visits of the infant to the pediatrician's office. Thus the factor A_8 would denote the geographic region of the residence of the infant, and A_9 would denote the time (day, month, year) the infant was born. A binary random variable X, taking on the values 0 or 1, would be used to denote whether or not the infant died in the first year after birth. A value 0 would denote that the infant died, and a 1 would denote the infant lived through the first year.

Exploratory analysis of the available data will be used as a means to gain a better understanding of the variables which are the major contributors to the infant death rate. Sensitivity analysis will be performed during the early stages of experimental design to get an insight as to how a change in a variable affects the mortality rate. Extensive use will be made of computer graphics in order to visualize the variables mentioned and their effect on infant mortality. Other variables included in the available data sets will also be studied so that we will have not only a visualization of the change and behavior of each variable separately, but also the joint contribution of two or more variables to the infant death rate. During this exploratory analysis, we hope to be able to identify which of the measured variables in the available data base should be included in the statistical analysis, and which are important, and therefore, should be included in our mathematical model.

The joint density function

$$Y = f(X, A_1, A_2, A_3, \dots, A_8), \qquad X, A_i, \quad i = 1, \dots, 8 \tag{3}$$

describes the probability of whether an infant will live or die when the conditions described by the A random variables are present. This joint density function will be estimated based on the given data. Also, the marginal density functions

$$f(X),\ f(X, A_1),\ f(X, A_2), \dots, f(X, A_1, \dots, A_7) \tag{4}$$

will be estimated. The above densities are discrete, and therefore, can be estimated for finite combinations of values of their parameters. Reliability measures of these estimates will also be found.

For the multivariate discrete distribution given by equation (1), the conditional distribution of X given the random variables $A_1 = a_1$, $A_2 = a_2, \dots, A_8 = a_8$, where the a_i's signigy specific values, is

$$f(X \mid A_1 = a_1, A_2 = a_2, \dots, A_8 = a_8), \tag{5}$$

where $a_1, a_2, \dots, a_8$ denote known values associated to the random variables A_i, $i = 1, \dots, 8$. These distributions can be estimated for $X = 0$ (infant death), and for $X = 1$ (infant survival). Thus the conditional distribution, or probability, or frequency, of infant mortality can be estimated for any given value of the other parameters.

Following the general theory of linear models, the expected value of X, given that the A_i's are known, is

$$E(A \mid A_1 = a_1,\ A_2 = a_2, \dots, A_8 = a_8) = G(a_1, \dots, a_8). \tag{6}$$

The above function involves the variance covariance matrix of multivariate distribution. A natural extension of the above formula is the equation

$$X = G(A_1, A_2, \dots, A_8) + \epsilon, \tag{7}$$

which can be used to predict X given the values for the A_i's. Unlike the regression model, the values for X can only be 0 or 1. Here ϵ is the error of estimation of X as a function of the A_i's, $i = 1, 2, \dots, 8$. This model includes, as a particular case, the model proposed by Eberstein *et al.* [16]. From the given set of data we can estimate the coefficients of the parameters included in the function F using a mean square error type of approach.

The above model can be tested using jackknifing. This involves estimating the parameters using all of the data except one, then inserting into the formula the values of A_1s associated with the ignored variable to see if it predicts the correct X value. This method is repeated n times, where n is the total number of data available. The predicted value of X, given the A_i, $i = 1, \dots, 8$ variables, would not necessarily be integer, and of course would not necessarily be 0 to 1. If the value of X is within an interval containing 1, then we associate the value 1 to X, otherwise we associate the value 0 to X. The length of the interval will be estimated from the data based on formulas to be developed. This jackknifing will help to verify that our theoretical development of the above interval is supported by practice.

When the value of X predicted by the above equation is within a certain interval we will assign the value 1 to X, otherwise we will assign the value 0 to X. In order to assess how well our model can predict the value of X given the values of A_i, $i = 1, 2, \dots, 8$, we have to estimate the probabilities of misclassification, namely the probability to classify the value of X as having the value 0, where the actual value of X is 1. Jackknifing is a good method for estimating the probability of misclassification. As we mentioned above, all of the data except for one is used to estimate the coefficients of the model. Then we estimate the X of the point left out using the model just estimated. If the value of X estimated by the model is the same as the true X value, then a correct classification was made, otherwise the misclassification count is increased by one. The initial misclassification count was 0. We repeat the method for each one of the data, and we divide the final classification count over the number of data, this gives us an estimate of the probability of misclassification. A low probability of misclassification implies that the model is a good predictor.

The complexity of this multivariate model can be increased by introducinɑ the geographic variable A_9 to detect geographic dependencies and their associated zone of influence. Finally, time dependencies can also be investigated. The question to be answered concerning time dependency is whether there has been any time trend since these data bases first became available. Additionally, has there been a fluctuation within the months of the year of the database that would suggest the existence of a time dependency as expressed by a significant autocorrelation function.

COMPUTER SIMULATION AND SENSITIVITY ANALYSIS

The estimation of the joint, marginal, and conditional distributions mentioned above, together with their associate reliability measures, would give us an understanding about the actual distribution of the random variables involved in the model, as well as their interdependencies. Multivariate functions, except for the multivariate normal, are very difficult to deal with. Along with the analysis recommended in the previous section, we will carry out a computer simulation of the multivariate phenomenon described here, in order to study the sensitivity of the infant mortality, to the various factors involved. The model is deemed sensitive to a particular random variable if small changes in that particular variable, result in large changes in the infant mortality model. On the other hand, the model is robust to a particular factor, or variable, if large changes in a variable result into relatively small changes in the infant mortality model.

CONCLUSION

The research tools to be designed and developed as a result of this research and development project will provide the first means for rigorous monitoring of trends and predictability for infant mortality. Health care providers can utilize the computer simulation model to predict birth/infant death cause and outcomes. This information can then be applied to a decision-making model for allocation of scare resources before birth, at birth, and after birth.

REFERENCES

1. U.S. Dept. Health Human Serv., Report of the secretary's task force on black and minority health, DHHS Publ. No. 1985 0487-647 (QL 3), US GPO, Washington, DC, (1985).
2. U.S. Dept. Health and Human Serv., A statistical methodology for analyzing data from a complex survey: The first national, health, and nutrition examination survey, DHHS Publ. No. (PHE)82-1366, National Center for Health Statistics, Hyattsville, (1982).
3. A. Wald, *Statistical Decision Function*, Wiley, New York, (1950).
4. D.A. Webster and R.J. Smales, Large database management in clinical dental research, *Austral. Dental J.* **36**, 397–400 (1991).
5. M. Witten, The Frankenstein project, *Int. J. Supercomputing Appl.* **6** (2), 127–137 (1992).

6. S. Meintz, Supercomputing for new branch of nursing science: Nurmetrics, *Research Poster Presentation*, WIN/WSNR, San Diego, CA, (1992).
7. S. Meintz, Supercomputer application to nursing and health data research, In *Supercomputer '91 International Conference*, Research poster presentation, Albuquerque, NM, (1991).
8. T.M. Goradia, K. Lange, P.L. Miller and P.M. Nadkarni, Fast computation of genetic likelihoods on human pedigree data, *Human Heredity* **42**, 42–62 (1992).
9. D. Lewin, Pour une large base de donne obstetricales, *Rev. Fr. Gynecol. Obstet.* **84**, 317–318 (1989).
10. B.M. Psaty, T.D. Koepsell, D. Siscovick. P. Wahl, J.P. Logerfo, T.S. Inui and E.H. Wagner, An approach to several problems in using large databases of population-based case-control studies of the therapeutic efficacy and safety of antihypertensive medicines, *Stat. Med.* **10**, 653–662 (1991).
11. M.G. Titler, D. Pettit, G.M. Bulechek, J.C. McCloskey, M.J. Craft, M.Z. Cohen, J.D. Crossley, J.A. Denehy, O.J. Glick and T.W. Kruckeberg *et al.*, Classification of nursing interventions for care of the integument, *Nurs. Diagn.* **2**, 45–56 (1990).
12. S.-K. Chang, Visual reasoning for information retrieval from very large database, Presented at the *IEEE Workshop on Visual Languages*, (October 1989).
13. S.-K. Chang and Y. Deng, Intelligent database retrieval by visual reasoning, In *Proceedings of the 14th Ann. Internat. Computer Software and Applications Conf.*, pp. 459–464, (1990).
14. W.S. Nersesian, Infant mortality in socially vulnerable population, *Ann. Rev. Public Health* **9**, 361–377 (1988).
15. R.L. Williams, N.J. Binkin and E.J. Clingman, Pregnancy outcomes among Spanish-surname women in California, *Am. J. Public Health* **76**, 387–391 (1986).
16. I.W. Eberstein, C.B. Nam and R.A. Hummer, Infant mortality by cause of death: Main and interaction effects, *Demography* **27** (3), 413–430 (1990).
17. N.I. Binkin, R.L. Williams, C.J.R. Hogue and P.M. Chen, Black neonatal mortality: How can it be reduced?, *J. Am. Med. Assoc.* **253**, 372–375 (1985).
18. D. Black, *Inequalities in Health: Report of a Research Working Group*, Pelican, Middlesex, U.K., (1982).
19. M.G. Boone, A socio-medical study of infant mortality among disadvantaged blacks, *Hum. Organiz.* **41**, 227–236 (1982).
20. H.C. Chase, Infant mortality and its concommitants, 1960–1972, *Med. Care* **15**, 662–674 (1977).
21. R.J. David and E. Siegel, Decline in neonatal mortality, 1968–1977: Better babies or better care?, *Pediatrics* **71**, 531–540 (1983).
22. G.T. Flatman and E.A. Yfantis, Geostatistical approaches to the design of sampling regimes, *Principles of Environmental Sampling*, pp. 73–84, Am. Chem. Soc., (1988).
23. S.L. Gortmaker, The effects of prenatal care upon the health of the newborn, *Am. J. Public Health* **69**, 653–660 (1979).
24. M.G. Kavar, Health status of U.S. children and use of medical care, *Public Health Rep.* **97**, 3–15 (1982).
25. K. Lee, N. Paneth, L.M. Gartner, M.A. Pearlman and L. Gruss, Neonatal mortality: An analysis of the recent improvement in the United States, *Am. J. Public Health* **70**, 15–21 (1980).
26. M. Learner and R.N. Stutz, Mortality by socioeconomic status, 1959–1961 and 1969–1971, *Maryland State Med. J.* **27**, 35–42 (1978).
27. *Children's Deaths in Maine*, Maine Dept. Human Serv., Augusta, (1983).
28. R.D. Mare, Socioeconomic effects on child mortality in the United States, *Am. J. Public Health* **72**, 539–547 (1982).
29. Massachusetts Dept. Health, Task force on prevention of low birthweight and infant mortality, *Closing the Gaps*, Boston, MA.
30. M.C. McCormick, The contribution of low birth weight to infant mortality and childhood morbidity, *N. Eng. J. Med.* **312**, 82–90 (1985).
31. M.C. McCormick, S. Shapiro and B. Starfield, High-risk young mothers: Infant mortality and morbidity in four areas of the United States. 1973–1978, *Am. J. Public Health* **74**, 18–23 (1984).
32. Natl. Cent. Health Statistics, Insurance coverage and ambulatory medical care of low-income children: United States, 1980, *Dept. Health and Human Serv.*, DHHS Publ. No. 85-20401, US GPO, Washington, DC, (1985).
33. W.S. Nersesian, M.R. Petit, R. Shaper, D. Lemieux and E. Naor, Childhood death and poverty: A study of all childhood deaths in Maine, 1976–1980, *Pediatrics* **75**, 41–50 (1985).
34. New York City Dept. Health, Sudden infant death syndrome, *City Health Information* **4** (21) (1985).
35. North Carolina Dept. Human Resources, SCHS studies, No. 36, Raleigh, NC, (1985).
36. N. Paneth, J.L. Kiely, S. Wallenstein, M. Marcus and J. Parker *et al.*, Newborn intensive care and neonatal mortality in low-birth-weight infants, *N. Eng. J. Med.* **307**, 149–155 (1982).
37. N. Paneth, S. Wallenstein, J.L. Kiely, C.P. Snook and M. Susser, Medical care and pre-term infants of normal birth weight, *Pediatrics*, 158–166 (1986).
38. N. Paneth, S. Wallenstein, J.L. Kiely and M. Susser, Social class indicators and mortality in low birth weight infants, *Am. J. Epidemiol* **116**, 364–375 (1982).
39. D.C. Shannon and D.H. Kelly, SIDS and near-SIDS, *N. Eng. J. Med.* **306**, 961 (1982).
40. B. Starfield, Family income, ill health, and medical care of U.S. children, *J. Public Health Policy* **3**, 24–259 (1982).
41. B. Starfield, Postnatal mortality, *Ann. Rev. Public Health* **6**, 21–40 (1985).

42. G. Stickle and P. Ma, Some social and medical correlates of pregnancy outcome, *Am. J. Obstet. Gynecol.* **127**, 162–166 (1977).
43. E.G. Stockwell and J.W. Wicks, Infant mortality and poverty in Ohio, *Ohio's Health*, Columbus, Ohio Dept. Health (1983).
44. K. Suries and G. Daughtry, SCHS studies, No. 29, Dept. Human Resources, Raleigh, NC, (1983).
45. Dept. of Urban and Environmental Policy, *Boston at Risk*, Tufts Univ., Boston, (1985).
46. M.E. Wegman, Annual summary of vital statistics–1984, *Pediatrics* **76**, 861–871 (1985).
47. K. Wicklund, S. Moss and F. Frost, Effects of maternal education, age and parity on fatal infant accidents, *Am. J. Public Health* **74**, 1150–1152 (1984).
48. R.L. Williams and P.M. Chen, Identifying the source of the recent decline in perinatal mortality rates in California, *N. Eng. J. Med.* **306**, 207–214 (1982).
49. P.H. Wise, M. Kotelchuck, M.L. Wilson and M. Mills, Racial and socioeconomic disparities in childhood mortality in Boston, *N. Eng. J. Med.* **313**, 360–366 (1985).
50. J. Cramer, Social factors and infant mortality: Identifying high-risk groups and proximate causes, *Demography* **24**, 299–322 (1989).
51. E. Powell-Griner, Infant mortality differentials by mother's marital status and race/ethnicity, In *Ann. Meeting of the Population Assoc. Amer.*, Washington, DC, (1988).
52. E. Powell-Griner, Differences in infant mortality among Texas Anglos, Hispanic, and Blacks, *Social Sci. Quart.* **69**, 452–467 (1988).
53. R. Rogers, Ethnics and birth weight differences in cause-specific infant mortality, *Demography* **26**, 335–343 (1989).
54. B. Berkov, Does being born out of wedlock still make a difference?, In *Ann. Meeting Population Assoc. Amer.*, Washington DC, (1981).
55. E.A. Yfantis, G.T. Flatman and J. Behar, Efficiency of kriging estimation for square, triangular, and hexagonal grids, *J. Math. Geology* **19** (3), 183–206 (1987).
56. E.A. Yfantis and G.T. Flatman, On sampling nonstationary spatial autocorrelated data, *Comp. and Geoscience* **14** (5), 667–686 (1988).
57. E.A. Yfantis and R. Bronson, Tests for validating time series models, *Trans. Soc. Comp. Simulation* **4** (1), 77–96 (1988).
58. S. Meintz, Supercomputers power discoveries in nursing, *Cray Channels* **14** (3), 20–21 (1992).
59. S. Meintz, New branch of nursing science, nurmetrics with the theory of nursing knowledge and practices, In *First Japanese International Research Conference,* Podium presentation, Tokyo, Japan, (1992).
60. S. Meintz, Supercomputer application to nursing data research, In *First Japanese International Research Conference,* Podium presentation, Tokyo, Japan, (1992).
61. S. Meintz, *Supercomputers Open Window of Opportunity for Nursing* (to appear).
62. M. Mori and K. Suzki, Very large database system to serve national welfare, In *Proc. Twelfth Internat. Conf. on Very Large Data Bases*, pp. 496–501, (1986).