The Canadian Journal of Program Evaluation Vol. 16 No. 2 Pages 1–26
ISSN 0834-1516

USING THE RIGHT TOOLS TO ANSWER THE RIGHT QUESTIONS: THE IMPORTANCE OF EVALUATIVE RESEARCH TECHNIQUES FOR HEALTH SERVICES EVALUATION RESEARCH IN THE 21ST CENTURY

Evelyn Vingilis
University of Western Ontario
London, Ontario

Linda Pederson
Centers for Disease Control/National Center for Chronic Disease Prevention and Health Promotion
Atlanta, Georgia

Abstract: The marked changes in health care expenditures in recent years have led to a call for greater accountability in the areas of health education, policy, services, and reform. In recent years, evaluative research has been conducted in health arenas under the rubric of health services research. The research methods employed have evolved not from evaluation research methodology, but from frameworks that often do not lend themselves to deriving appropriate causal inferences in multi-causal environments. Given the complexity of the interrelated political, social, psychological, and economic factors that can affect health services, more complex evaluative techniques are needed. This article describes the epistemology of evaluative research and, through a series of examples from the health services literature, demonstrates the strengths of theory-driven approaches and statistical multivariate techniques compared to traditional black-box methods, as ways to increase validity for causal inference.

Résumé: À cause des modifications marquées des dépenses en soins de santé, la demande de reddition de comptes dans les secteurs de l'éducation, de la politique, des services et de la réforme dans le domaine de la santé s'est intensifiée. Ces dernières années, des évaluations ont eu lieu dans les secteurs de la santé dans le cadre de la recherche sur les services de santé. Les méthodes de recherche sur les services de santé ne s'inspirent pas de la méthodologie de recherche évaluative mais adoptent plutôt des mécanismes qui ne se prêtent pas toujours à formulation de déductions causales dans des milieux englobant plusieurs cau-

ses. Étant donné la complexité des rapports entre les facteurs politiques, sociaux, psychologiques et économiques qui influencent les services de santé, des techniques d'évaluation plus complexes s'imposent. Cet article décrit l'épistémologie de la recherche évaluative et, par une série d'exemples tirés de la documentation scientifique sur les services de santé, montrent les forces des approches fondées sur la théorie et des techniques statistiques à variables multiples par rapport aux méthodes traditionnelles de la boîte noire pour accroître la validité de la déduction causale.

Over the last 30 years, the industrialized world has witnessed major changes in health care expenditures (Nair & Karim, 1993). For example, between 1971 and 1991 health care expenditures as a percentage of GDP almost doubled in the U.S.A. (7.5% to 13.4%), increased by one third in Canada (7.4% to 10.0%), and increased between 25% and 33% in other industrialized countries except Sweden (Nair & Karim, 1993). This dramatic growth has led to a call for greater accountability in health care expenditures, and has spurred evaluative research in health education, policy, services, and reform.

Unfortunately, many evaluation studies in the health/medical field have been fraught with problems, arising from the use of simplistic and inadequate conceptualization, methodology, measures, and analyses (Vingilis & Burkell, 1996). If the results of these studies simply sat on the shelf, these problems would not be of concern. Unfortunately, as Terris (1999) recently observed, many of these problematic evaluations are the bases for decisions on health policies and the provision of services. Such decisions, made on the basis of weak, incomplete, or poor evidence, are at best no better than decisions made with no evidence, and are in many cases worse because they are likely made more confidently, in the misguided belief that they are based on facts.

Fundamental to all evaluative studies is the issue of causal inference. Cook and Campbell (1979) describe three necessary conditions for causal inference: (1) the cause must precede the effect in time, (2) the cause and effect must covary, and (3) there can be no plausible alternative explanation for the covariation. The third condition is the one most commonly unmet by researchers attempting to draw conclusions from the results of health services evaluative studies.

Research methods typically used to determine causation are those associated with control of various aspects of the condition/situation

under study (such as the random control trial), and are consistent with the simple cause-effect models of causation. These research methods were developed for problems that differ from those currently facing the health of a population and health care system, and are based on a conceptualization of health that differs from those currently in use (McMichael, 1999; Susser & Susser, 1996a, 1996b; Syme, 1996). Although simple cause-effect models may have been effective in determining the causes of ill health, such as the role of *E. coli* in gastrointestinal illness, they are often not appropriate for the study of changes in health services. As Terris (1999) succinctly points out, there are problems with using simple short-term studies to evaluate the effect of health care changes on a population's health status because such studies do not account for the fact that health care is only one of three major determinants of health status (the other two being living standards and the use of preventive measures, such as environmental protection of food, water, air, etc.). Long-term trends in these other two determinants may cause changes in health status that are mistakenly attributed to health care changes.

The seminal work of Campbell, Stanley, and Cook (Campbell & Stanley, 1966; Cook & Campbell, 1979) has been one of the foundations for the conceptualization of causal inference in evaluative research. The impetus for improved evaluation methods has also come from the need to evaluate the effectiveness of large-scale, government-funded social services implemented during the last half century. This need challenged social scientists to develop methodologically and statistically stronger designs for such evaluations (Rossi & Freeman, 1993; Rossi, Freeman, & Lipsey, 1999). Some of these methodological and statistical approaches, commonly described in social services evaluative texts (Bernstein & Sheldon, 1983; Cook & Campbell, 1979; Isaac & Michael, 1997; Mitchell, 1985; Posovac & Carey, 1997; Rossi & Freeman, 1993; Rossi et al., 1999) can provide relevant methods and approaches to health services researchers.

The purpose of this work is to present problems of causal inference associated with certain traditional approaches to evaluating health services, and to introduce approaches that better support causal inferences. Examples from published studies that illustrate the use and misuse of approaches, methods, and data analyses will be included. Finally, we will offer some suggestions about how an innovative combination of approaches can help researchers answer questions of critical importance as our health care systems head into the 21st century.

Two traditional health services evaluative approaches can be problematic for establishing causal inference (Cook & Campbell, 1979; Isaac & Michael, 1997; Mitchell, 1985). The first such approach is the "black-box paradigm." Black-box-oriented or method-oriented evaluation is atheoretical, and as such is characterized by a primary focus on the overall relationship between the inputs and outputs of an intervention (Chen, 1990). As Chen indicates, simple input/output or black-box evaluations may provide an overall assessment of whether or not a program works, but do not identify the underlying causal mechanisms that generate the intervention effects, thus failing to pinpoint the deficiencies of the intervention for future improvement or development. Chen continues: "A black-box evaluation is usually not sensitive to the political and organizational contexts of input and output, and it neglects issues such as the relationship between delivered treatment and planned treatment, between official goals and operative goals or between intended and unintended effects" (p. 18). Health services evaluation is often carried out using the black-box approach; that is, little consideration is given to *why* and *how* interventions work. Numerous authors have called for alterations in strategies used in health-related evaluations (Chen & Rossi, 1989, 1993; Lipsey & Pollard, 1989; National Institutes of Health, 2000; Palumbo & Oliverio, 1989; Petrosino, 2000; Rossi et al., 1999; Vingilis & Burkell, 1996).

A second source of issues related to causal inference concerns sources of random error variance and appropriate use of statistics and statistical tests (Mitchell, 1985). The statistical methods used in various health services evaluative studies may be inappropriate because of possible violations of assumptions, biased estimators, and ecological fallacies. Often, simple pre-post designs and statistical tools are used to evaluate complex changes to health services. Research designs and statistical methods (Cook & Campbell, 1979; Isaac & Michael, 1997; Posovac & Carey, 1997; Rossi & Freeman, 1993; Rossi et al., 1999) designed to address large-scale or complex interventions have unfortunately been absent in most health services studies. Time series, hierarchical regression, and structural equation modelling are examples of some statistical methods that, if *appropriately used,* can increase the accuracy of predictions and effects estimates, enhance interpretation of health services data, and lead to more valid causal inferences (Posovac & Carey, 1997; Rossi & Freeman, 1993; Rossi et al., 1999). These designs have generally not been addressed or used in health services studies and texts (Atkinson, Hargreaves, Horowitz, & Sorensen, 1978; Fink, 1993; Fink & Kosecoff, 1978; Smith, 1990; World Health Organization, 1981)

(although a recent epidemiology text by Rothman and Greenland [1998] describes the techniques and identifies the importance of using these more advanced designs for complex causal relationships).

SIMPLE CAUSE-EFFECT MODELS VERSUS MIXED-METHOD, MULTI-MEASURES, CAUSAL MODELLING APPROACHES: EXAMPLES

Evaluative research should begin with some "cause-and-effect link(s)" between intervention and outcome, based on explanatory (*why* the intervention should work) and change (*how* the intervention should work) theories (Chen, 1990; National Institutes of Health, 2000; Petrosino, 2000). Lipsey (1988) writes that most cause-effect links between interventions and outcomes fall along a continuum of causal complexity, with at one end "relatively molar, monolithic, undifferentiated causes applied at distinct points in time with effects that are produced directly, immediately and overtly" (p. 8). The simple cause-and-effect link is exemplified by immunization or other medical or pharmacological interventions that are delivered in measured doses and work primarily through physiological mechanisms (Lipsey, 1988; Petrosino, 2000).

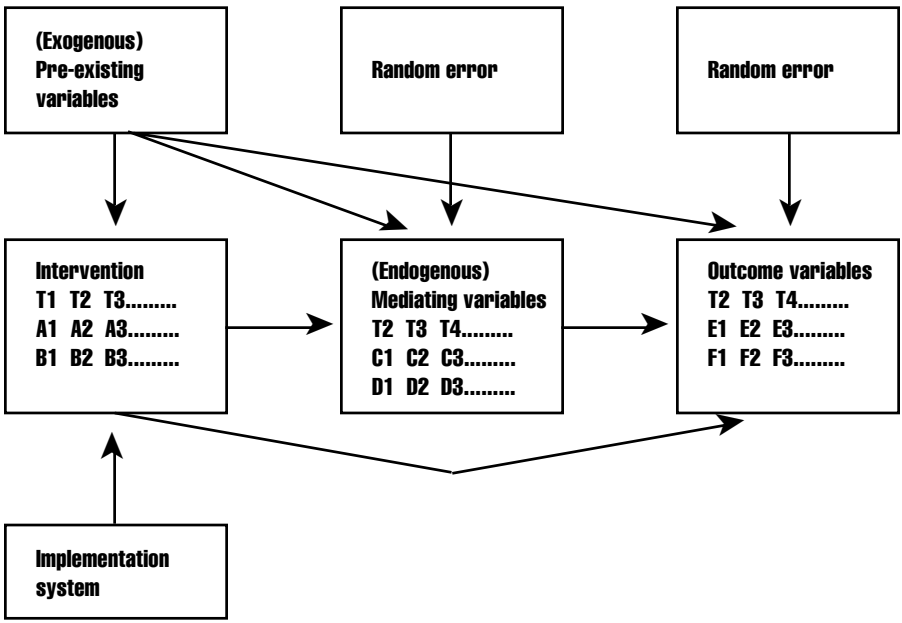
Lipsey continues:

> On the other end of the continuum of causal complexity we have processes in which the cause is applied over an extended period of time and is both multidimensional and variable from occasion to occasion. Such causal processes may have many components, may involve long chains of cascading direct and indirect effects, may be influenced by various extraneous events (i.e. be context dependent) and may have results that are delayed, diffuse, and subtle. (1988, p. 8)

A schematic diagram, or model, of this end of the continuum of causal complexity is shown in Figure 1 (Chen, 1989, 1990; Chen & Rossi, 1980; Lipsey, 1988; Sidani & Braden, 1998). The model indicates that interventions with causal complexity have four sources of potential causal impact on final outcomes: exogenous variables, implementation system, endogenous variables, and random error. Exogenous variables are confounders, which already exist prior to intervention. The implementation system refers to the actual delivery of the intervention, comparing hypothesized interventions to

those that are actually delivered and comparing official goals with operative goals. Endogenous (also called intervening or mediating) variables refer to factors representing a wide range of direct and indirect influences within the causal links occurring during the course of the intervention that could affect the final outcomes. Finally, random error refers to measurement error and the need to take this error explicitly into account through appropriate statistical methods. The specification of a causal model, although not necessary for evaluation research, is the recommended approach to developing designs to enhance researchers' abilities to draw causal inferences (Chen, 1989, 1990; Chen & Rossi, 1980; Lipsey & Pol-

Figure 1
Representation of Generic Causal Model



Note. T = time, A...F = variables.

Fox (1984) states in his statistics textbook: "the use of time-series data in regression or similar analyses generally cast doubt upon the assumption that errors from different observations are independent, since the observation period from one period is likely to share characteristics with observations from other periods in time. If, in fact, errors are autocorrelated in this manner, different estimation techniques than ordinary least-squares regression are called for." In addition, time series analyses can "de-trend" common trends, such as declining hospital bed occupancy rates, and seasonal trends, such as increased deaths due to MVIs in summer months, thereby reducing variation and possible estimator bias.

lard, 1989; National Institutes of Health, 2000; Palumbo & Oliverio, 1989; Petrosino, 2000; Posavac & Carey, 1997; Rossi & Freeman, 1993; Rossi et al., 1999; Sidani & Braden, 1998). Because causal inferences drawn from the results of black-box-oriented or method-oriented evaluation are purely empirical and not theory based — the causal links are identified and tested empirically — researchers cannot easily determine why an intervention does or does not appear to work. And when it does, the mechanisms are not known.

A central feature of most health services evaluations is the use of black-box evaluative studies. These approaches are appropriate in situations where the causal pathways of a particular outcome are simple and direct. However, black-box designs are problematic in studies of complex situations in which an outcome has multiple causal pathways. The use of theory and a causal model to explain the processes by which changes in outcomes are expected to occur is an important design tool to enhance the validity of research (Chen, 1989, 1990; Chen & Rossi, 1980; Petrosino, 2000; Sidani & Braden, 1998; Vingilis & Burkell, 1996).

An example of an intervention with causal complexity is a program providing public and professional education on reducing the delay in presentation and referral of stroke patients. Alberts, Perry, Dawson, and Bertels, as part of a larger experimental treatment study using tissue-type plasminogen activator, developed a "multifaceted program of public and professional education to encourage early hospital arrival upon signs and symptoms of a stroke" (1992, p. 352). Their educational efforts focussed on improving the recognition of stroke symptoms and emphasizing the need for rapid treatment of stroke victims. The intervention included six components: television and radio interviews, newspaper articles, lectures to local and regional primary care and emergency department physicians, mailings to several thousand local physicians, having neurologists on call 24 hours a day for referrals, and use of air ambulance helicopters. They used a simple pre-post design, where they compared the percentage of patients with cerebral and intracerebral infarction who presented or were referred to their emergency department within 24 hours of symptom onset for the 26-month period before their educational intervention with the percentage presenting within 24 hours for the 12-month period after their educational intervention had begun. A significant increase in presentation within 24 hours was found for patients with cerebral infarction (86%) during the intervention period; the percentage presenting within 24 hours

before the intervention was 37%. However, no significant change was observed between patients with intracerebral hemorrhage before and those afflicted during the intervention. No other data were collected. The authors concluded: "We found that a program of public and professional education was successful in raising awareness about stroke and significantly reducing the time delay seen in the presentation and referral of stroke patients to our facility" (p. 355).

The research and statistical methods used and conclusions drawn are inappropriate in this study because of the causal complexity of the intervention. Causal complexity exists for two reasons: the causal pathways for each of the components are complex, and the intervention has a number of distinct components, any of which could be causing the outcome. The causal mechanisms by which education leads to changes in behavioural outcome, based on health decision-making theory (Tones, 1985) or any other communication/education theories, require sensory detection of, attention to, and perception of the educational message by the audience, which generally requires high media saturation (high levels of message penetration). Then the audience must correctly interpret, understand, learn, and remember the message. Finally, the audience must try new behaviours and maintain them in order for an educational intervention to show actual impact on behaviour. Without measuring whether or not the audience actually heard, understood, and learnt the message, one cannot assume that any behavioural outcome was due to the message. Thus, the conclusion that "awareness" had increased in the target audiences is invalid, as we are provided with no measures of "awareness," and no direct causal link can be drawn to the actual educational components of the intervention.

Moreover, the intervention is multifaceted, with six different components, any one or combination of which could be causing the outcome results. Again, no data were collected to determine whether the observed changes were due to increased public or professional knowledge about stroke symptoms and early presentation. Perhaps the observed changes in reduced delay of presentation to the emergency department were due to the increased use of the air ambulance and not to education at all. Perhaps having neurologists on call 24 hours a day caused the changes in presentation patterns. Without detailed information and data measuring the implementation system, including what and how much intervention was provided, and measuring mediating variables, the outcome results cannot be readily interpreted. Thus, the cause-and-effect link can-

not be established with any certainty because this evaluation used a simple pre-post design within a black-box approach to test an intervention with causal complexity.

Another example of black-box evaluation whose output was not causally or conceptually linked to the intervention input was recently described in the newsletter of the Canadian Health Services Research Foundation (1999), a governmental granting agency for health services research. The headline of the article reads, "Research shows hospital cuts don't hurt health, access to care." The actual report, by the Health Services Utilization and Research Commission (HSURC) (1999) in Saskatchewan, concluded that closing small rural hospitals and cutting acute care hospital beds in major cities actually *reduced* death rates. In 1993, when 52 small hospitals were closed or converted to health centres because of spending cuts, the commission evaluated the impact of the closures.

First, the commission analyzed hospital use and mortality data in Saskatchewan before and after the cuts, from 1990 to 1996; this analysis, they stated, would "determine whether access to hospital care or the health of rural residents was affected" (p. 2). To determine whether or not these cuts affected the health of rural residents, they examined death rates as a proxy measure for health status. They compared hospital use and death rates among four groups: (1) communities that stopped receiving acute care funding to keep their hospitals open, (2) rural communities that never had hospitals, (3) rural communities that still maintained their rural hospitals, and (4) the rest of Saskatchewan. They tabulated heart attack, stroke, motor vehicle injury, premature (aged 0–74), and total (all causes) deaths per 100,000 population using a simple pre-post design (1990–92 versus 1993–96) for communities that never had small hospitals, for communities that kept small hospitals open, and for communities that closed small hospitals in 1993. They concluded that "health status as measured by death rates improved throughout the province during the study period. Communities that experienced the 1993 acute care funding cuts had the largest overall improvement in mortality rates; communities that still have small hospitals, the smallest. Communities that never had a hospital had the lowest mortality rates throughout the study period" (p. 3). Death rates per 100,000 population showed pre-post hospital closure declines for heart attacks, strokes, and all causes in all three communities (where hospitals closed, where hospitals remained open, and where no hospitals had existed) and for the rest of Saskatchewan.

However, after the 1993 hospital closures, deaths per 100,000 due to motor vehicle injuries and premature deaths increased for communities that still kept hospitals open, but decreased for the other two communities, with no and closed hospitals, and for the rest of Saskatchewan. In addition, they found hospital use declined throughout the province: communities with hospitals whose funding for acute care was cut had the sharpest decline in rates, and communities that maintained their rural hospitals continued to have the highest hospital use rates.

Second, the commission interviewed residents, through telephone surveys and focus groups, about their opinions of their hospital closures (HSURC, 1999). The telephone survey found that 18% were dissatisfied with health services prior to the 1993 funding cuts, and 54% were dissatisfied after the cuts. Concerns were lack of local emergency services, decreased availability of health services and doctors, and travel distance to access care. Yet the report also indicated that of those who were surveyed via telephone, 89% stated that the funding cuts had no direct effect on their personal health and 64% were in good health. The conclusions of the study were that

> [c]utting the acute care funding to 52 rural Saskatchewan hospitals has not adversely affected the health status of residents in these communities.... Community residents' perceptions were consistent with the mortality data. Despite widespread fears that health status would deteriorate, respondents overwhelmingly reported that the loss of acute care funding did not adversely affect their own health or their family's health. In addition, two out of every three people surveyed rated their current health status as either good or very good ... Although we have no data to confirm that removing acute care funding from communities improved residents' health status, some of the data are intriguing. (p. 9)

The implicit causal model here is that hospitals increase deaths.

Unfortunately, a simple pre-post design, within a black-box approach, was also not appropriate for conducting this evaluation. First, the three community groups had different population-based death rates even before the intervention, ranging from 680 to 880 per 100,000 population. The highest death rates pre- (and post-) intervention existed in communities with hospitals, and the lowest pre-

(and post-) death rates existed in communities that never had hospitals. If we consider the generic model in Figure 1, clearly there are pre-existing conditions (exogenous factors) that could affect the results, and that needed to be accounted for in the research design and statistical analysis. Moreover, no endogenous factors were discussed to explain the findings. That is, the causal mechanisms by which the changes, or lack thereof, occurred have not been measured, so that the how and why are not understood.

Of particular concern should be the findings that deaths due to motor vehicle injury (MVI) actually *decreased* in communities that closed hospitals and in which average driving distance to the nearest hospital had increased to more than 50 kilometres, yet *increased* in communities that kept hospitals open (HSURC, 1999). Empirical evidence indicates that most critically injured patients can be saved if definitive medical intervention is provided within one hour (the "Golden Hour" for emergency treatment) (Stewart, 1990). It is hard to believe that an injured patient's chance of survival would not be affected by having to endure an additional 30–60 minute drive. To explain the findings of this study and to eliminate other possible causal factors, one would need to assess the exogenous preconditions, the actual implementation of the closures, and endogenous mediating variables (see Figure 1). In other words, how and why did some death rates decrease in the experimental communities and increase in the communities that kept their hospitals open? The report does not attempt to explain the increase in MVI mortality in the still-open hospital communities. Rather, it states: "The data do not suggest that the affected communities have suffered ... as a result of the 1993 acute care funding cuts; if anything, death rates for heart attacks and motor vehicle trauma are no different and possibly better" (HSURC, 1999, p. 11). Had they conducted a theory-based, causal modelling evaluation, they would have questioned the use in a population health study of mortality rates (Vingilis & Burkell, 1996) to determine whether access to hospital care or the health of rural residents was affected (HSURC, 1999) by hospital closures over a short time period, and a priori presented the causal mechanisms for expected outcomes (Chen & Rossi, 1989, 1993; Lipsey, 1988; Lipsey & Pollard, 1989; Petrosino, 2000; Sidani & Braden, 1998). As it stands, this black-box evaluation raises more questions than it answers.

Regardless of results, the simple input/output or black-box type of evaluation often draws conclusions that are less than satisfactory.

On the one hand, black-box-based positive findings of intervention success may be difficult for policy-makers or practitioners to apply (Chen, 1990). For example, the Alberts et al. (1992) study, mentioned above, evaluated a multifaceted intervention of great causal complexity. In this study, even if the causal link that the researchers claim to have established was valid, one cannot determine which and how much of the above-mentioned interventions resulted in the desired positive outcome. Thus, this type of evaluation typically provides no information about which combination of components and intervention strength of components are necessary to obtain the desired outcome (Lipsey, 1988).

Findings of intervention failure that are based on a black-box evaluation may be even more misleading. Is the failure due to a poor choice of intervention or a poorly implemented intervention (Chen, 1990; Vingilis & Burkell, 1996)? Is the strength of intervention too low, or is the measurement not sensitive enough (Chen, 1990; Lipsey, 1988)? The true intervention effect may be hidden by a gross estimation of the intervention effect. This interpretation problem associated with black-box evaluations will be demonstrated in the cold self-care evaluations described later in this article.

Many argue for specification of prior theoretical assumptions within a causal framework in evaluative research in order to meet policy needs (Chen, 1990; Mitroff & Bonoma, 1978; Petrosino, 2000; Vingilis & Burkell, 1996). They disagree with the view that overcoming methodological difficulties alone can always render data collected in evaluations valid and scientifically precise. In the HSURC (1999) study, even given more rigorous methodology, researchers would still need to think through the causal model that closing small hospitals in communities reduced the motor vehicle death rate. An obvious policy implication of the HSURC study, for example, is that all small community hospitals should be closed so as to reduce the death rates further. Yet this obvious policy implication is plainly absurd.

The second problem besetting causal inference within traditional health services evaluative research concerns appropriate use of statistics. The challenges of evaluating complex, multi-causal outcomes have led to the development of mixed-method, multi-measures research and statistical designs (Bernstein & Sheldon, 1983; Cook & Campbell, 1979; Greene, Caracelli, & Graham, 1989; Posovac & Carey, 1997; Rossi & Freeman, 1993; Rossi et al., 1999). Detailed descriptions of the different designs and statistical methods have

been previously published (Cook & Campbell, 1979; Posovac & Carey, 1997; Rossi & Freeman, 1993; Rossi et al., 1999) and are beyond the scope of this article. Suffice it to say that appropriate research designs and multivariate analytic tools, such as time series, hierarchical regression, and structural equation modelling techniques, can produce substantially more accurate predictions and effects estimates (Greenland, 1998). For example, when observational studies where data are collected over a time period use simple pre-post designs as described above, less rigorous statements about cause and effect can be made because simple pre-post statistical analyses may not be the appropriate statistical treatment for causal inference. A set of observational data, collected over time and comprised of equally spaced or at least consecutive observations, which may be denoted by $X_t\, t = 1,2, \ldots$ is called a time series. A time series design consists of numerous, repeated measurements of data points preceding and following an "intervention" (or some independent variable manipulation), taken on an aggregate unit, such as deaths due to MVIs (Ross, Campbell, & Glass, 1970; Vingilis, Blefgen, Lei, Sykora, & Mann, 1988), with the expectation of a change in the dependent variable (Rossi et al., 1999; Simonton, 1977). For such data standard statistical procedures such as *t*-tests, anova, and regression are likely to give misleading and spurious results due to lack of statistical independence inherent in the assumed error distribution.[1] Time series methods have been developed for such time-dependent data. The goal of intervention time series analysis is to model and describe statistically the effect of such external change on a time series. Time series analysis offers a sophisticated procedure for isolating the critical effects of an intervention and for ruling out a number of rival hypotheses, such as changes due to instability of data or to long-range trends already in motion prior to the intervention (Bernstein & Sheldon, 1983; Box & Jenkins, 1970; Cook & Campbell, 1979; Fox, 1984; Rossi & Freeman, 1993; Rossi et al., 1999; Simonton, 1977; Snortum, 1988).

Yet this design is rarely used in health services evaluative research, despite the fact that longitudinal, repeated measurements of administrative data commonly form the basis of many of these evaluations. Leaving aside the serious problem of potential violations to statistical assumptions when these administrative time series datasets are subjected to traditional statistical methods (Fox, 1984; Norusis, 1985), these longitudinal datasets of morbidity, mortality, procedure, hospitalization and readmission rates, and the like are ideally suited to the more methodologically and statistically valid

time series analyses. As Wagenaar (1998) summarizes in a recent article on research methods used to demonstrate intervention effectiveness:

> Some authors unfortunately lump together all non-randomized designs under the term "ecologic," implying that such studies contribute little to the literature. While a simple pre/post design without comparisons or with one comparison group is of limited value, time series intervention trials with hundreds of repeated measures over time, switching replications of one or more interventions, and multiple levels of comparison groups have very high levels of internal validity. Designs with repeated measures over time on a single unit also have the advantage of more statistical power than treatment/control independent group designs. I suggest we call such studies "controlled time-series trials" and include them with fully randomized trials in the high-quality group that provides strong evidence on which to base policy and programmatic decisions. (p. 10)

The previously mentioned Alberts et al. (1992) and HSURC (1999) hospital closing studies are examples of time series observational studies where appropriate analyses were not conducted. The HSURC report, for example, presents a graph of age- and sex-standardized mortality rates by community group for 1990 to 1996, which shows complex variation and a slightly decreasing trend over the six-year time period. A simple pre-post design cannot capture the variation or possible long-term trend. If daily or weekly data of this six-year period were subjected to time series analyses, the analyses could determine if statistically significant changes in mortality rates actually were related to hospital closings, once long-term trends and other variation were statistically removed.

Encouragingly, theory-driven evaluations within a causal framework and the use of more complex research and statistical methods such as time series analyses are beginning to appear in evaluative health services research (Alexander, Halpern, & Lee, 1996; Banaszak-Holl, Zinn, & Mor, 1996; Goes & Zhan, 1995). For example, Alexander et al. (1996) recently evaluated the short-term effects of mergers on hospital operations by using the causal framework of two merger motivation theories and employing a multiple time series design. The two merger motivation theories reflected the need for either

activities expansion or services consolidation. The authors reasoned that if the merger was motivated by the need to achieve access to capital and improve ability to attract other health care resources, little or no consolidation in staffing or service capacity and few improvements in operating efficiency would follow a merger of two or more hospitals. On the other hand, if consolidation was the main reason for mergers, they hypothesized that changes in operating practices would occur in those institutions involved in merger (namely a reduction in the overall scale of operation, improved operating efficiency, and reduced duplication in staffing in the period following merger). Although the use of only secondary data, as the authors state, prevented them "from assessing directly the issue of how and why changes in operative practices did and did not occur" (p. 845), their two theoretical models identified causal mechanisms. This design involved a six-year longitudinal assessment, using daily mean values, of change in hospital operating characteristics before and after merger, and a parallel analysis of change in a randomly selected group of non-merging hospitals. Changes in hospital operating characteristics before and after merger were assessed by six variables representing three areas of hospital operations: scale of operation, operating efficiency, and staffing practices. By using time series analyses they were able to control for secular trends other than the merger, and to identify changes in levels and slopes over and above general pre-existing trends in operating characteristics of merging partners. One of the time series analyses established that the decline in occupancy rates for merging hospitals between pre-post merger periods was most likely a continuation of the significant pre-merger trend towards declining occupancy in the merging hospitals. This finding could have been misinterpreted had a simpler pre-post case-control design been applied, as the lower post-merger rates of the merged hospitals, when compared to the rates of the non-merged hospitals, might well have been interpreted as due to the merger rather than to a significant downward trend found in the merging hospitals only, well prior to their mergers. Thus, the time series analyses provided a richer source of information by which to interpret the results of the evaluation, a source of information that may have prevented faulty conclusions being drawn about the impact of hospital mergers.

The consequence of using the black-box approach with simple pre-post research and statistical methods in studies evaluating health services interventions of causal complexity is that these studies cannot be sure that the intervention being evaluated is necessarily

linked to the outcome or that the conclusions and recommendations appropriately follow. In the remainder of this article we will present as examples two evaluations of the same health education intervention, one using a theory-driven approach within a causal framework and a mixed-method, multi-measures design and the other using the traditional black-box approach within a simple pre-post design.

In 1993, the Ontario Ministry of Health (MOH) determined that 12.6% of visits to physicians between January and March 1991 listed on the government's administrative, physician billing dataset, Ontario Health Insurance Plan (OHIP), were coded as "common cold." On the assumption that this health care utilization rate was too high, the MOH launched a public education campaign to increase patient self-care and thereby reduce the number of "inappropriate" visits to physicians' offices because of cold and flu symptoms. The campaign consisted of radio and newspaper ads and a patient information booklet to be delivered to every household in the city of London, Ontario. The MOH did not conduct a needs assessment of baseline data on health care practices to determine brochure content and approach, although some professional and consumer focus groups and consultations were set up for the purpose of testing ideas and materials. The materials focussed on informing the public that (1) the flu and colds account for the greatest number of unnecessary doctor visits, (2) unneeded cold visits cost taxpayers $200 million a year, and (3) doctors can't cure a cold. The materials also provided a number of home remedies for self-care (Vingilis et al., 1994).

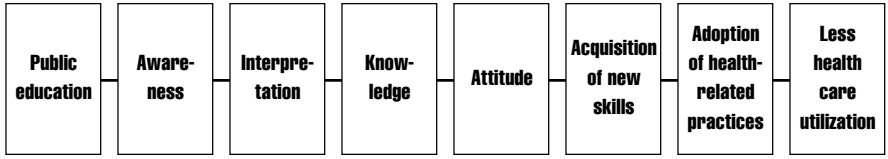
Vingilis et al. (1994, 1998; Vingilis, Brown, & Hennen, 1999; Vingilis, Brown, Sarkella, Stewart, & Hennen, 1999) conducted the initial evaluation of the MOH's cold/flu education campaign. As this campaign to change health practices had a high level of causal complexity, they used a mixed-method, multi-measures approach (Greene et al., 1989) informed by Tones' (1985) health decision-making theory, which hypothesized a series of causal links needed for successful health information campaigns (Figure 2). Vingilis et al. measured exogenous variables, the implementation process, and endogenous and outcome variables within the causal model. Exogenous variables that could affect outcome were the public's practices, knowledge, and awareness of cold/flu treatment prior to the intervention. This was assessed by using qualitative and quantitative (mixed) methods and multiple, independent measures. Vingilis et al. conducted a pre-intervention random-digit-dialling telephone survey in the experimental and comparison cities to measure the general public's

levels of awareness, knowledge, attitudes, skills, and health practices. They found that respondents were already quite knowledgeable about colds/flu and felt that physician visits were unnecessary unless the cold lasted longer than 7–10 days, showed signs of secondary infections, or affected people such as children or the elderly who were at high risk for complications. Only 4% self-reported that they would go to a physician with the "common cold"; self-care was the overwhelming treatment method of choice. An independent but convergent study of family physicians' patient caseload for one week during the campaign corroborated the self-report information from the public surveys. Results of the caseload study showed that patients reported experiencing symptoms for an average of 10 days before visiting a physician, and that the most common reasons for physician visits were persisting/worsening symptoms and the presence or indication of secondary infection (Keast, Marshall, Stewart, & Orr, 1999).

Intervention implementation was assessed by examining message penetration. Results of the post-intervention survey showed that two thirds of London respondents neither knew of the campaign nor had read the booklet.

The public survey assessed the causal links (mediating variables) of awareness, interpretation, new knowledge, and attitudes in the experimental and comparison cities. The pre-post comparisons showed that London residents correctly answered only 2 of 10 knowledge questions significantly more often than did residents of the comparison city. No pre-post changes were found for attitudes or for knowledge about new self-care remedies. In addition, a survey of family physicians assessed their opinions and observations on the effects of the campaign (Vingilis et al., 1994, 1998). The surveyed physicians were ambivalent about the intervention, expressing concerns about self-diagnosis. Moreover, physicians questioned the validity of using billing codes as a marker for "inappropriate" visits,

Figure 2
Causal Model for Cold Self-Care Public Education to Reduce Health Care Utilization



as many stated that the "cold" code was used as a catch-all for billing purposes only, and remarked that after the campaign they were less likely to use the "cold" code when presented with multiple health problems (Vingilis et al., 1994, 1998).

The outcome component of the evaluation measured whether the intervention (information booklet and ads) was specifically linked to outcome (increased self-care and decreased health care utilization in the London area). Vingilis et al. (1994, 1998) found no evidence that use of self-care remedies advised by the booklet had increased in the London area following the campaign. Moreover, of the respondents who remembered their last cold (about 90% of respondents), about 1% in London and the comparison city reported visiting the doctor for their cold symptoms, in both the pre- and post-surveys.

The results of this study suggest that the MOH's implicit hypothesis that the public are uneducated about colds/flu and "appropriate" health care utilization for colds/flu may have been questionable because most survey participants were knowledgeable and reported making "appropriate" physician visits for cold/flu. Campaign implementation appeared to have limited impact in that message penetration was low. The mediating variables of knowledge and attitudes and the outcome variables measuring health practices were virtually unchanged over the campaign period among London residents exposed to the campaign and among members of the comparison city. Vingilis et al. (1994, 1998) thus concluded that the campaign had little effect because of the limited success in implementation and the lack of self-reported behaviour change.

However, as Vingilis et al. (1994, 1998) did not have access to direct objective measures of health practices over time, they could not conduct intervention time series analyses in order to determine if objective health practices data converged with the results presented above within the causal model of health decision-making (Tones, 1985). However, the Institute for Clinical Evaluative Studies, with access to OHIP data, conducted a subsequent evaluation of the intervention (Brown & Goel, 1996).

In evaluating the effectiveness of this campaign, Brown and Goel used one outcome measure within a simple pre-post design. They used OHIP physician billing codes for visits for cold and flu symptoms and total billings for all types of visits. They compared the two

months before and after the start of the campaign with the same two periods in the previous year for London and for the rest of Ontario. Although they conducted no statistical tests, they stated that London showed a 28% reduction in "cold" billings during the campaign; the rest of Ontario, which was not targeted by the campaign, showed a 21% decrease in "cold" billings during the campaign period. Furthermore, there were virtually no changes in total billings in either London or the rest of Ontario. They concluded that "the modest relative reduction in physician billings for visits because of cold and flu symptoms in London following the introduction of the public education campaign may have been due to the intervention as well as to other factors" (p. 835).

The Brown and Goel (1996) evaluation raises a number of questions that cannot be addressed by the data in their black-box study. First, if the campaign caused a true reduction in health care utilization for the cold/flu, why did the overall billings not decrease in the experimental area? That is, how and why did the reduction in cold/flu billings occur? The 28% reduction in cold/flu billings should lead to some reduction in overall billings. These seemingly contradictory results, however, can be explained by the implementation information from the Vingilis et al. (1994, 1998) family physician survey, which suggested that family physicians, during the intervention phase in this non-blinded study, may have been changing their billing code practices.

Second, the reductions found in the number of billings for cold/flu in both the experimental and comparison regions suggest that historical factors were clearly reducing the cold/flu billings in Ontario as a whole. Although Brown and Goel (1996) do note that the cold/flu season had actually passed by the time the campaign was implemented, they would have had to conduct intervention time series analyses of multiple months of the OHIP billing claims in order to determine whether the decrease in cold/flu billings was related to the implementation of the campaign or to long-term or seasonal trends (Vingilis et al., 1994).

Finally, Vingilis et al.'s theory-driven, causal model study shows that the actual mechanisms of change were not in place; that is, the campaign did not fully realize the steps of the health education model that described the process by which changes in health practices could occur. Ultimately, the conclusions of Brown and Goel (1996) have to be tempered by the fact that two thirds of the surveyed community

had not even heard of the campaign, let alone absorbed the message, and that there was no evidence of changes in self-reported health practices as a consequence. The difference in the conclusions reached by the two studies discussed here highlight the problems inherent in black-box evaluations. The estimated cost of the campaign for London, a community of 300,000, was half a million dollars (Mickleburgh, 1994). Had the MOH expanded the campaign to all of Ontario, with a population of 10 million, on the basis of results from the Brown and Goel study, it would have done so at great taxpayer expense in the belief that the campaign had been a success. In contrast, the theory-driven evaluation of Vingilis et al. (1994, 1998), with its mixed-method, multi-measures design, suggested a different educational outcome.

CONCLUSIONS AND RECOMMENDATIONS

In this article we suggest that the black-box study design is inadequate for much health services evaluative research. Instead we suggest that researchers evaluating complex multi-causal outcomes should consider using mixed-method, multi-measures research and statistical designs in conjunction with a priori knowledge and theory to build causal models of the intervention process and implementation system. This could enhance the validity of the evaluations and provide more information about how to achieve desired effects (Chen, 1989; Petrosino, 2000).

This call for a broader approach to health services evaluative research into complex evaluative questions has become a recurring theme (Koopman, 1996; Lipsey & Pollard, 1989; Posovac & Carey, 1997; Susser & Susser, 1996a, 1996b; Syme, 1996; Terris, 1999). Yet despite the advantages of a theory-driven causal model approach and more complex methodological and statistical designs, the health services evaluation field is characterized by a remarkably low level of explicit theorizing about process (Lipsey & Pollard, 1989), although some recent health services evaluations have begun to move beyond the black-box paradigm to use theory-driven approaches and more appropriate research and statistical methods (Alexander et al., 1996). The challenge, however, will be to convince health services researchers and funding agencies of the feasibility and efficacy of such theory-driven approaches. The various groups involved in and affected by evaluative research will need to communicate with each other and agree upon common goals before theory-driven approaches can become more generally accepted. This proposal for the use of more

complex evaluation methods requires patience and re-education of the research community, policy-makers, funding agencies, and the public. Although this re-education process may take some time, we believe that improved health services evaluations will be well worth the time and effort.

NOTE

1 Fox (1984) states in his statistics textbook: "The use of time-series data in regression or similar analyses generally cast doubt upon the assumption that errors from different observations are independent, since the observation period from one period is likely to share characteristics with observations from other periods close in time. If, in fact, errors are autocorrelated in this manner, different estimation techniques than ordinary least-squares regression are called for." In addition, time series analyses can "de-trend" common trends, such as declining hospital bed occupancy rates or unemployment rates, and seasonal trends, such as increased deaths due to MVIs in summer months, thereby reducing variation and possible estimator bias.

REFERENCES

Alberts, M.J., Perry, A., Dawson, D.V., & Bertels, C. (1992). Effects of public and professional education on reducing the delay in presentation and referral of stroke patients. *Stroke, 23*, 352–356.

Alexander, J.A., Halpern, M.T., & Lee, S.D. (1996). The short-term effects of merger on hospital operations. *Health Services Research, 30*, 827–847.

Atkinson, C.C., Hargreaves, W.A., Horowitz, M.J., & Sorensen, J.E. (1978). *Evaluation of human service programs*. New York: Academic Press.

Banaszak-Holl, J., Zinn, J.S., & Mor, V. (1996). The impact of market and organizational characteristics on nursing care facility service innovation: A resource dependency perspective. *Health Services Research, 31*, 97–117.

Bernstein, I.N., & Sheldon, E.B. (1983). Evaluative research. In R.B. Smith (Ed.), *Handbook of social science methods* (pp. 93–132). Cambridge, MA: Ballinger.

Box, G.E.P., & Jenkins, G.M. (1970). *Time series analysis: Forecasting and control*. San Francisco: Holden-Day.

Brown, E.M., & Goel, V. (1996). Reducing demand for physician visits through public education: A look at the pilot cold-and-flu campaign in London, Ontario. *Canadian Medical Association Journal, 154*, 835–840.

Campbell, D.T., & Stanley, J.L. (1966). *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally.

Canadian Health Services Research Foundation. (1999). *Quid Novi, 2,* 4 [Newsletter].

Chen, H. (1989). The conceptual framework of the theory-driven perspective. *Evaluation and Program Planning, 12*, 391–396.

Chen, H. (1990). *Theory-driven evaluations*. Newbury Park, CA: Sage.

Chen, H., & Rossi, P.H. (1980). The multi-goal, theory-driven approach to evaluation: A model for linking basic and applied social science. *Social Forces, 59*, 106–122.

Chen, H., & Rossi, P.H. (1989). Issues in the theory-driven perspective. *Evaluation and Program Planning, 12*, 299–306.

Chen, H., & Rossi, P.H. (1993). Evaluating with sense: The theory-driven approach. *Evaluation Review, 7*, 283–302.

Cook, T.D., & Campbell, D.T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Chicago: Rand McNally.

Fink, A. (1993). *Evaluation fundamentals: Guiding health programs, research and policy*. Newbury Park, CA: Sage.

Fink, A., & Kosecoff, J. (1978). *An evaluation primer*. Beverly Hills, CA: Sage.

Fox, J. (1984). *Linear statistical models and related methods: With applications for social research*. New York: John Wiley & Sons.

Goes, J.B., & Zhan, C.L. (1995). The effects of hospital-physician integration strategies on hospital financial performance. *Health Services Research, 30*, 507–530.

Greene, J.C., Caracelli, V.J., & Graham, W.F. (1989). Toward a conceptual framework for mixed-method evaluation designs. *Educational Evaluation and Policy Analysis, 11*, 255–274.

Greenland, S. (1998). Introduction to regression modeling. In J.J. Rothman & S. Greenland (Eds.), *Modern epidemiology* (pp. 401–432). Philadelphia: Lippincott-Raven.

Health Services Utilization and Research Commission (HSURC). (1999). *Assessing the impact of the 1993 acute care funding cuts to rural Saskatchewan hospitals*. Summary Report No. 13. <www.sdh.sk.ca/hsurc>.

Isaac, S., & Michael, W.B. (1997). *Handbook in research and evaluation* (3rd ed.). San Diego: EdITS/Educational and Industrial Testing Services.

Keast, D.H., Marshall, J.N., Stewart, M., & Orr, V. (1999). Why do patients seek family physicians' services for cold symptoms? *Canadian Family Physician, 44*, 335–340.

Koopman, J.S. (1996). Emerging objectives and methods in epidemiology. *American Journal of Public Health, 86*, 630–632.

Lipsey, M.W. (1988). Practice and malpractice in evaluation research. *Evaluation Practice, 9*, 5–24.

Lipsey, M.W., & Pollard, J.A. (1989). Driven toward theory in program evaluation: More models to choose from. *Evaluation and Program Planning, 12*, 317–328.

McMichael, A.J. (1999). Prisoners of the proximate: Loosening the constraints on epidemiology in an age of change. *American Journal of Epidemiology, 149*, 887–897.

Mickleburgh, R. (1994, January 19). Ontario touts chicken soup as flu remedy. *The Globe and Mail,* pp. A1, A2.

Mitchell, T.R. (1985). An evaluation of the validity of correlational research conducted in organizations. *Academy of Management Review, 10*, 192–205.

Mitroff, I., & Bonoma, T.V. (1978). Psychological assumptions, experimentations and real world problems. *Evaluation Quarterly, 2*, 235–259.

Nair, C., & Karim, R. (1993). An overview of health care systems: Canada and selected OECD countries. *Health Reports, 5*, 259–279. (Available from Statistics Canada, Ottawa.)

National Institutes of Health. National Cancer Institute. (2000). *Theory at a glance: A guide for health promotion practice*. <http://rex.nci.nih.gov/NCI_Pub_Interface/Theory_at_glance/HOME.html>.

Norusis, M.J. (1985). *Advanced statistics guide: SPSSX*. Chicago: SPSS Inc.

Palumbo, D.J., & Oliverio, A. (1989). Implementation theory and the theory-driven approach to validity. *Evaluation and Program Planning, 12*, 337–344.

Petrosino, A. (2000). Answering the why question in evaluation: The causal-model approach. *Canadian Journal of Program Evaluation, 15*(1), 1–24.

Posovac, E.J., & Carey, R.G. (1997). *Program evaluation: Methods and case studies* (5th ed.). Upper Saddle, NJ: Prentice-Hall.

Ross, H.L., Campbell, D.T., & Glass, G.V. (1970). Determining the social effects of a legal reform: The British breathalyzer crackdown of 1967. *American Behavioural Scientist, 13*, 494–509.

Rossi, P.H., & Freeman, H.E. (1993). *Evaluation: A systematic approach* (5th ed.). Newbury Park, CA: Sage.

Rossi, P.H., Freeman, H.E., & Lipsey, M.W. (1999). *Evaluation: A systematic approach* (6th ed.). Newbury Park, CA: Sage.

Rothman, K.J., & Greenland, S. (1998). *Modern epidemiology*. Philadelphia: Lippincott-Raven.

Sidani, S., & Braden, C.J. (1998). *Evaluating nursing interventions: A theory driven approach*. Thousand Oaks, CA: Sage.

Simonton, D.K. (1977). Cross-sectional time-series experiments: Some suggested statistical analyses. *Psychological Bulletin, 84*, 489–502.

Smith, M.J. (1990). *Program evaluation in the human services*. New York: Springer.

Snortum, J.R. (1988). On seeing the forest and the trees: The need for contextual analysis in evaluating drunk driving policies. *Evaluation and Program Planning, 11*, 279–294.

Stewart, R.D. (1990). Pre-hospital care of trauma. In R.Y. McMurtry & B.A. McLellan (Eds.), *Management of blunt trauma* (pp. 23–29). Baltimore, MD: Williams & Wilkins.

Susser, M., & Susser, E. (1996a). Choosing a future for epidemiology: 1. Eras and paradigms. *American Journal of Public Health, 86*, 668–673.

Susser, M., & Susser, E. (1996b). Choosing a future for epidemiology: 2. From black box to Chinese boxes and eco-epidemiology. *American Journal of Public Health, 86*, 674–676.

Syme, S.L. (1996). Rethinking disease: Where do we go from here? *Annals of Epidemiology, 6*, 463–468.

Terris, M. (1999). The neoliberal triad of anti-health reforms: Government budget cutting, deregulation, and privatization. *Journal of Public Health Policy, 20*, 149–167.

Tones, B.K. (1985). The use and abuse of mass media in health promotion. *Health Education Resources* [Pilot issue], 9–14.

Vingilis, E., Blefgen, H., Lei, H., Sykora, K., & Mann, R. (1988). An evaluation of the deterrent impact of Ontario's 12-hour licence suspension law. *Accident Analysis and Prevention, 20*, 9–17.

Vingilis, E.R., Brown, U., & Hennen, B.K. (1999). Common colds: Reported patterns of self-care and health care use. *Canadian Family Physician, 45*, 2644–2652.

Vingilis, E., Brown, U., Koeppen, R., Hennen, B., Bass, M., Payton, K., Downe, J., & Stewart, M. (1998). Evaluation of cold/flu self-care public education campaign. *Health Education Research, 13*, 33–46.

Vingilis, E., Brown, U., Koeppen, R., Hennen, B., Bass, M., Stewart, M., Payton, J., & Downe, J. (1994). *Evaluation of the Ministry of Health's*

cold self-care public education project [Working paper]. London, ON: Faculty of Medicine, University of Western Ontario.

Vingilis, E.R., Brown, U., Sarkella, J., Stewart, M., & Hennen, B.K. (1999). Cold/flu knowledge, attitudes and health care practices: Results of a two-city telephone survey. *Canadian Journal of Public Health, 90*, 205–208.

Vingilis, E., & Burkell, J. (1996). A critique of an evaluation of the impact of hospital bed closures in Winnipeg, Canada: Lessons to be learned from evaluation research methods. *Journal of Public Health Policy, 17*, 409–425.

Wagenaar, A.C. (1998). Importance of systematic reviews and meta-analyses for research and practice. *American Journal of Preventive Medicine, 16*, 9–11.

World Health Organization. (1981). *Health programme evaluation*. Geneva: Author.