The Journal of Mental Health Policy and Economics
J. Mental Health Policy Econ. **3**, 97–109 (2000)

Using Randomized Controlled Trials to Evaluate Socially Complex Services: Problems, Challenges and Recommendations

Nancy Wolff

Department of Urban Studies and Community Health and Center for Research on the Organization and Financing of Care for the Severely Mentally Ill, Institute for Health, Health Care and Aging Research, Rutgers University, New Brunswick, NH, USA

Abstract

Background: Following the lead of evidence-based medicine, practice based on effectiveness research has become the new gold standard of contemporary public policy. Studies of this sort are increasingly demanded to evaluate services provided by mental health, social services and criminal justice systems.
Aims: The paper questions whether the simple randomized controlled trial (RCT) paradigm as applied in clinical trials can be used 'off the rack' to evaluate *socially complex service (SCS) interventions*. These are services that are characterized by complex, diverse and non-standardized staffing arrangements; ambiguous protocols; hard-to-define study samples and unevenly motivated subjects and dependence on broader social environments. The difficulty of ensuring precise protocols, equivalent groups (tied to a meaningful target population) and neutral and equivalent trial environments under real world conditions are explored, as are the implications of not achieving standardization and equivalence.
Methods: Limitations of effectiveness research as a research tool and information source are examined by comparing the assumptions underpinning the simple RCT to the characteristics of SCS interventions, as illustrated by programs targeted to mentally disordered offenders in Britain.
Results: SCSs violate the assumptions underpinning the simple RCT model in ways that draw into sharp question the validity, reliability and generalizability of inferences of SCS trials.
Discussion: The RCT is not a panacea. Effectiveness research of SCS interventions that is based on the RCT model is unlikely to yield valid, reliable and generalizable inferences without becoming more complex in design and more sensitive to issues of selection bias, unmeasured variables and endogeneity. Ten recommendations are offered for stylizing the RCT design to the characteristics of socially complex services.
Implications: It remains an empirical issue whether RCT-based services effectiveness research can inform mental health policy. Without major design innovations, it is more likely that the information generated by this research will have limited practical use, especially if the RCT model is unable to control for the effect of social complexity and the interaction between social complexity and dynamic systemic change. Scientific evaluations of services make clinical and economic sense so long as they are designed to meet the challenges of the services of which they promise greater knowledge.

Received 15 September 1999; accepted 7 April 2000

Introduction

There is a new norm in the delivery of publicly and privately financed services: evidence-based practice. For example, in Britain, practice based on research evidence is highlighted in most policy statements recently released by the Department of Health[1–5] and the Home Office.[6,7] Similarly, in the United States, payers and providers alike are looking to empirical evidence to inform their choices on what yields best value.[8] The policy drive for evidence-based practice in all service sectors has moved the techniques of effectiveness and cost-effectiveness analysis into center stage. It is expected that evidence from these evaluative studies will inform service delivery and funding policies and offer new opportunities to reshape the service system in ways that improve its overall performance.

This paper looks critically at these evaluation techniques and questions whether they are up to the task of informing practice and policy making. Over the past ten years, much has been written about the need for greater standardization in the framing of effectiveness research* in health and medicine,[9] in the methods used to measure costs[10–13] and in reporting practices.[9,14] Standardizing the frame and methods used to demonstrate and report best value makes sense but only if the design on which these studies are based is appropriate for the services being evaluated. It has been implicitly assumed with the diffusion of effectiveness

Correspondence to: Nancy Wolff, Ph.D., Institute for Health, Health Care and Aging Research, Rutgers University, 30 College Avenue, New Brunswick, NH 08901-1293, USA.
Email address: nwolff@rci.rutgers.edu

Source of Funding
Contract grant sponsor: British Government

* The term 'effectiveness research' is defined to include evaluations that compare two or more interventions in terms of their effects, where effects can be measured either narrowly—health and social outcomes—or broadly—health, social and economic outcomes.

research from medication and surgical interventions to socially complex service interventions that the design of the randomized controlled trial, the *sine qua non* of effectiveness research, is independent of the service intervention itself.

This paper explores the validity of this assumption. In the first section, the randomized controlled trial (RCT) model of effectiveness research is described. This model is based on three key assumptions; standardized interventions, equal groups and equal trial environments; all are necessary for making relative comparisons among interventions. The next section develops the concept of socially complex services (SCS) within the context of an intervention taxonomy. The third section examines whether the three assumptions of the RCT model can be satisfied in studies that evaluate SCS interventions. The last section summarizes the research and policy implications of the modified RCT model for socially complex service interventions.

Programs for mentally disordered offenders in Britain serve as an illustrative example of SCS in the third section. Service interventions for mentally disordered offenders (MDOs) in many ways represent the social complexity of interventions for persons with mental illness living in the community since they involve efforts to coordinate a variety of service sectors that are responsible for managing different behaviors,[15] but little research has been directed towards studying these programs in the United States and United Kingdom, although, with the increasing evidence of psychiatric morbidity within jails and prisons in both countries[16,17] and the rapid diffusion of liaison programs in the UK, efforts are advancing to study their effectiveness. Before this research movement gains too much steam, it is useful to look at the complexity of these programs as found in Britain to determine whether studies based on the standard RCT model will yield much useful information. Part of the evidence for this section is drawn from interviews and site visits undertaken in Britain by the author from September 1998 to April 1999.

Standard Clinical Model of Effectiveness Research

The effectiveness model is designed to answer a very simple question: which of two or three competing interventions achieves the best outcome? Here, best outcome can be defined in terms of treatment effect (therapeutic outcomes) or dollar per treatment effect (costs adjusted by therapeutic outcomes). In either case, the best intervention among those evaluated is determined by comparing the gains associated with doses of different interventions. Three key assumptions underpin the RCT paradigm.

Assumption 1: Standardized Intervention Protocols

One key assumption underpinning effectiveness research is that (dose) interventions—both experimental and control—can be defined precisely (i.e., standardized) and monitored specifically for adherence. Defining interventions involves describing *who* is doing *what* and *when* in ways that can be implemented uniformly and measured accurately and reliably. For purposes of validity, replicability and generalizability, protocols need to be defined in ways that capture the structure of the delivery mechanism and the process of the interactions among staff, as well as between the staff and patients.

Because the goal of effectiveness research is to identify the best relative program, all factors other than the dose intervention must be identical between the two interventions to rule out the possibility that some unknown and idiosyncratic factor correlated with one or the other program is contributing to the measured effects. There are two critical control factors: study samples and trial environment, which are assumed to be equivalent in a neutral therapeutic environment.

Assumption 2: Study Sample Equivalence

To ensure that the measured effects are the result of the dose interventions, competing interventions must be applied to a representative person. The groupings of individuals that receive the interventions are expected to be representative of a broader group of people that might gain from the diffusion of the intervention.

Population Definition

Defining the target population for intervention is complex. It requires first knowing the set of symptoms that the intervention is expected to impact and what disease/disorder label best fits these symptoms. Individuals with these disease or symptom clusters would be potential candidates for inclusion in the population. However, it is also necessary to know if there are internal and external factors that may mitigate or militate the symptom or intervention pathways and if these factors are representative of persons afflicted with the particular set of symptoms.

Random Assignment

The goal of randomly assigning individuals to interventions is to create equivalent groups. Random assignment ensures that if there are any systematic or unmeasured differences within the sample, the differences will be randomly distributed among interventions, but random assignment does not guarantee that sample groups will be balanced or equivalent. With small sample sizes, it is quite possible to have unequal assignment of cases such that one group has more high or low severity cases. Random assignment generates balanced groupings only when there are large enough numbers to average out any chance asymmetries.

Assumption 3: Trial Environment Equivalence and Neutrality

The trial environment is expected to be unaffected by factors such as financing, supportive assistance, inter-agency behavior and community dynamics, or, if it is affected, it

is assumed that the effect is equivalent between interventions. Keeping the environments 'clean' and balanced between the groups ensures that only the interventions, as specified in the protocols, are producing the relative differences between the outcomes.

A simple pairwise* RCT effectiveness model is shown in Figure 1. Effect1 and Effect2 can be uniquely attributed to Dose1 and Dose2, respectively, if Dose1 and Dose2 are uniquely defined and precisely and consistently implemented, Subject1 and Subject2 samples are identical, and the boxes representing the trial environment are equivalent. The relative results from a study based on this type of design can be generalized to other people equivalent to the sample characterized by Subject1 and Subject2 and to environments that are similar to the trial environment.

Real World of Services Effectiveness Research

The RCT effectiveness model has appeal because it provides robust evidence on which clinical practice yields the best outcomes. Yet the utility of the model for services research depends on whether it can be adapted to the services area where the interventions are more socially complex. The first part of this section develops the concept of socially complex service interventions and contrasts it with the ideal clinical trial intervention, represented by a particular drug or surgical protocol. In the second part of this section, issues of intervention standardization and population definition and randomization, as well as environmental neutrality and equivalency (the assumptions of the RCT model), are re-examined in the context of socially complex services, as illustrated by interventions for mentally disordered offenders.

Structural Taxonomy of Interventions

Interventions are produced using various types of inputs. It is, therefore, useful to begin by identifying the most commonly important inputs that help distinguish among types of intervention. They are: staffing arrangements, protocol specificity, subject involvement and environmental boundaries.

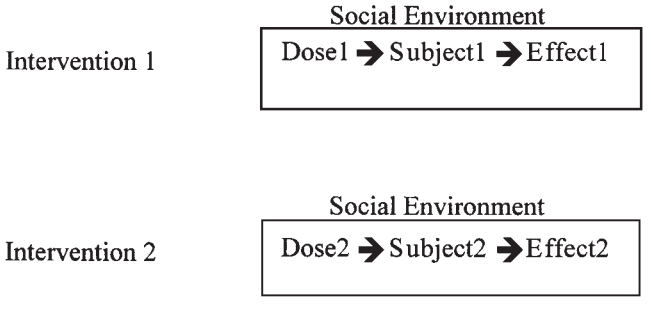

Figure 1. A diagram of a simple pairwise randomized controlled clinical trial (RCT) model based on hard boundaries between the intervention and the social environment

Staffing Arrangements

Interventions may be produced with varying numbers of staff, and staff of different skill types and motivation levels. The simplest staffing arrangement is one where there is a single staff member of a standardized skill level producing an intervention (e.g., surgical procedure or injection). This type of input is easy to standardize and replicate if it involves a professional whose skills are standardized by an accredited process (say, board-certified clinician, registered nurse or licensed social worker). Variation, in this simple case, relates to how motivated the staff member is to follow the intervention protocol. In contrast, more complex arrangements involve large numbers of differently skilled providers working together to produce an intervention. The concept of a 'team' is pervasive in socially complex services. However, like poetry, its interpretation is often highly idiosyncratic in terms of who is included on the team and how they work together. These interactional processes and the skills necessary to support teamwork are not easily standardized or professionalized through a certification or licensing process.

Protocol Specificity

The protocol defines what will be done to or for the study subject and when this will happen. Protocols vary in their degree of ambiguity. Some protocols are very concrete: they specify, say, the type of medication to be given to whom and under what conditions and for what length of time. There is little leeway for interpretation. In cases of complex services, however, the protocol involves a generalized approach or style of interaction that is, by its nature, subjective in its interpretation and application. For example, protocols involving case management are less clear in their definition and application.[18,19] Because case management is a process of interaction, how it is implemented will by stylized in part by the professional and personal characteristics of the staff.

Subject Involvement

Individuals choose to participate in research studies. Because there are often risks associated with new interventions, individuals must be willing to bear the costs associated with participation. Those who choose to bear the risks typically do so only because they believe they have a problem/illness and they see some prospect of being restored to greater function. Motivation to participate, therefore, is inextricably tied to subjects' belief in and insight into their illness, their understanding of the intervention's potential to alleviate aspects of illness, their willingness to bear risks and their desire to be healthy. Uncertainty in subjects' acceptance and understanding of their illness, and in their valuation of the intervention and its potential benefits increases complexity since those who may benefit most from the intervention may be least likely to participate. Evidence on resistance to treatment among persons with serious mental illness[20] and on treatment failure among

* Pairwise comparisons are the most common form of effectiveness or cost effectiveness test. While it is possible to use these analysis techniques to compare more than two interventions, there is usually insufficient statistical power in most studies to compare more than two or three possibilities.

persons with substance abuse problems[21] suggests that motivational issues are likely to be more salient among persons with serious mental illness and substance abuse problems than those with physical illnesses.

Environmental Boundaries

Environmental boundaries range along a continuum from hard to soft. Hard boundaries are those where the trial setting exists outside the broader social context and is itself unaffected by outside forces. Examples of hard boundary settings are structured therapeutic environments, such as hospitals or clinics, where confounding external effects can be controlled. In contrast, soft boundaries are those where the divide between the intervention setting and the social environment is permeable. This occurs when the intervention setting is within the larger social setting (say, the community) and each of the settings is directly or indirectly influenced by the other. Assertive outreach programs and jail/court diversion programs for mentally disordered offenders are examples of interventions with soft boundaries.

Table 1 shows the array of characteristics among the inputs that produce the ideal clinical intervention and the typical socially complex service intervention. The elegance of the *ideal clinical intervention* emanates from its simplicity, clearly defined and motivated populations, standardization, concreteness and independence, which when combined make this type of intervention easier to study and the findings more robust. *Socially complex service (SCS) interventions* are characterized by their complex and diverse staffing arrangements, ambiguous protocols, hard-to-define and unevenly motivated subjects and dependence on the broader social environment.

The structure of interventions, although presented here as a taxonomy, is best thought of as a continuum that varies between these two extremes: simple clinical and complex social services. The boundaries between these two types of intervention are not sharp ones. It is easy to imagine hybrid interventions (e.g., simple service interventions and complex clinical interventions) that have characteristics in common with both extreme types. Much of the value of the taxonomy, however, lies in its ability to differentiate among interventions along characteristics that are central to the key assumptions underpinning the RCT model. The task at hand is to determine whether these variations in characteristics threaten the robustness of the RCT model and, if so, whether new methodological approaches and tools are needed to assure the validity, reliability and generalizability of services effectiveness research.

Assumptions of RCT and the Characteristics of SCS: A Poor Match

Recognition that socially complex services are dissimilar from simple clinical services raises the issue of whether the differences between the (extreme) service types affect the utility of the RCT framework to test the effectiveness of SCS interventions.* If such differences do matter, it is important to know if there are any modifications that could be made to the traditional RCT design to enhance its utility for services effectiveness research. How particular characteristics of SCS interventions relate to the three key assumptions of the RCT model are discussed below. In cases where the assumptions are sufficiently challenged by SCS characteristics, remediable design recommendations are proposed.

Socially Complex Interventions and Assumption 1: Intervention Standardization

SCS interventions have staffing arrangements and protocols that are hard to define and measure precisely. For example, SCS interventions for persons with severe mental illness typically include some form of case management. Yet case

* This is not meant to imply that the RCT model is 'ideal' or 'problem free' in its application to clinical services. As discussed by Feinstein[22] and others,[23] there are limitations associated with the application of the RCT model to evaluate drug and surgical interventions. The central point here, however, is that the characteristics of SCSs, as a rule, may inherently conflict with the underlying assumptions of the RCT design as applied in the clinical field, and as such it may be necessary to take additional steps to control for them within a modified design.

Table 1. A structural taxonomy of types of interventions ranging between two extremes: ideal clinical interventions to socially complex service interventions

Key inputs of interventions	Ideal clinical intervention	Socially complex service intervention
Staffing arrangements	Single provider Professional staff Standardized expertise Highly motivated staff	Many providers Mix of lay and professional staff Non-standardized expertise Differently motivated staff
Protocol specificity	Concrete and measurable	Ambiguous and hard to measure
Subject involvement	Illness/problem with low level of professional uncertainty High insight into illness High understanding of benefits and and risks Health is valued	Illness/problem with high level of professional uncertainty Variable insight into illness Variable understanding of benefits and risks Mental health has mixed value
Environment boundaries	Hard external boundaries	Soft external boundaries

management takes on different (i) philosophical principles, ranging from time-limited therapeutic management to long-term, holistic management and advocacy; (ii) organizational structures, varying from a one person, single agency approach to a multi-disciplinary, multi-agency team approach; (iii) processes of interaction and engagement, ranging from assertive to reactive—which also vary in their definitions; (iv) style of engagement, varying from impersonal and objective to personal and subjective, and (v) set of performance outcomes, ranging from outcomes that are quantitative—number of face-to-face contacts and referrals—to the qualitative—building trust and rapport with clients. Standardizing the form of case management included as part of an intervention is challenging in part because it involves processes that are hard to discern and quantify.

This lack of precision makes it difficult to model the casual pathways of interventions, which is central to the RCT model.[9,10] Because an intervention can induce a variety of behavioral changes that produce both internal and external effects and because causality is assumed between the intervention and these effects, definitional or measurement ambiguity adds 'noise' to the model, which is likely to compromise the integrity of the connection between intervention and effects. For example, if the complexity of the intervention cannot be defined and measured precisely, there is a possibility that unmeasured and idiosyncratic aspects of the intervention may have an impact on measured effects either directly or indirectly through an interaction with a measured aspect of the intervention. By itself, having unmeasured and idiosyncratic characteristics affecting outcomes may not matter if the SCS can be replicated completely—measured and unmeasured features together. The problem arises when the SCS intervention is *unbundled* and described by its measured features but inferences about its *bundled* performance are attributed to the measured features, because if the highly effective aspects of the intervention are those unmeasured aspects that are associated with highly stylized characteristics of the staff, say their interactional style or level of motivation, there is no certainty that if the intervention, as defined and measured, is exported, it will render consistent effects.

Court liaison (or diversion) schemes for mentally disordered offenders are good examples of SCS interventions. Currently, there are approximately 200 liaison schemes in England and Wales.[24] However, because of the natural variation among these schemes, it is difficult to standardize and categorize their characteristics into meaningful models of liaison. These schemes share only one characteristic: they seek to identity offenders with mental disorders. They differ in terms of their staffing arrangements—the number and type of providers, their array of professional and interpersonal expertise and motivation. Similarly, their protocols are unclear and subject to change depending on the allocation of resources and the willingness of agencies to work together. Moreover, their protocols are typically defined in terms of concepts that are known for their ambiguity, such as 'risk or needs assessment', 'case management', 'multidisciplinary team approach', 'liaison', 'diversion' and 'interagency collaboration'.

This type of intervention is hard to specify and model. It requires first deconstruction of the ambiguous concepts into their constituent parts and then development of ways to separately measure them, but it also requires modeling how the separate parts of the intervention work together and separately to create a process that is expected to elicit a set of effects. The level of specificity needed in the model depends on the complexity of the intervention.

> *Recommendation 1*. Services effectiveness studies need to define and measure the characteristics of each intervention with enough specificity and precision to assure that (i) causal connections can be drawn between the intervention and effects, (ii) it is accurately implemented and consistently operationalized and (iii) it can be replicated elsewhere.

Socially Complex Interventions and Assumption 2: Sample Equivalence

Though central to the RCT model, creating representative and equivalent samples for SCS interventions is problematic for three reasons. First, SCS interventions often focus on populations that have multiple, co-occurring problems, each of which is difficult to define uniquely and with precision, which invites professional discretion. Second, the populations targeted for SCS interventions are often resistance to treatment and difficult to engage, complicating the recruitment process. Third, programs vetting SCS interventions are oft-times resistant to random assignment. Characteristics of the population, combined with features of the intervention, have important implications for the purported benefits of randomization, as well as for how one would ideally implement randomization. This point is illustrated by considering one such population, mentally disordered offenders.

Population definition. A prerequisite for good sampling is a precise and unequivocal definition of the population from which the sample will be drawn, but this condition is difficult to satisfy when the characteristics of populations are ambiguous. For example, defining the population of mentally disordered offenders is complicated by the fact that both the mental disorder and offender labels include a range of symptoms and problem behaviors. The label of 'mental disorder' may include any set of behaviors that meet DSM-IV criteria[25] or ICD-10 categories of diagnoses.[26] That is, it could include in the population any person with organic mental disorders, schizophrenia, mood disorders, neurotic disorders or personality disorders. Britain's Mental Health Act of 1983 complicates the definition of mental disorder by identifying four sub-categories ('mental illness', 'severe mental impairment', 'psychopathic disorder' and 'mental impairment'), defining all except the sub-category of 'mental illness' and adding the requirement of treatability.[27]

Ambiguity in the definition of a mental illness opens the way for discretion in the definition of the population. For

example, according to the Mental Health Law of 1983, the population of mentally disordered offenders could include only those offenders who have mental disorders that are *treatable*. The treatability clause of the law limits the disorders to a particular interpretation of a therapeutic construct and invites professional discretion regarding which cases are treatable. However, defining a population of mentally disordered offenders is further complicated by the fact that different service systems in contact with offenders develop their own protocols for defining mental illness. For example, in Britain, the Prison Service defines mental illness according to the medical classifications of particular disorders as assessed by prison medical officers,[28] whereas the police tend to use the term 'mental disorder' as defined by the code of practice to the *Police and Criminal Evidence Act of 1984*, which considers whether persons 'cannot understand the significance of questions put to them or their replies'.

Definitional ambiguity can create tensions between researchers and service agencies, as the clinical definitions set by researchers, say using a structured clinical interview schedule, may not be consistent with mental health laws or the eligibility criteria of service agencies that guide the decisions of courts, prosecutors, and service agencies. Yet, setting definitional criteria by legal or administrative standards, which are themselves subject to interpretation, may contribute to the selection of an unrepresentative clinical population.

There is similar ambiguity in the definition of offender status. To acquire the offender label, an individual must show evidence of deviance, as measured by an encounter with law enforcement agencies. Again, the definition of deviance is fungible; it may include social deviance, such as vagrancy, disturbance of the peace and panhandling, as well as criminal deviance. In some circumstances, it may, however, be limited to offenses that involve violence or only those without violence. Alternatively, the population may be limited to those with mental disorder and offender labels that reside in particular locations, such as the community, jail or prison.

At least theoretically, the population of mentally disordered offenders is defined by the overlap area between two populations: mentally disordered and offender-level deviance. Yet because the boundary of each population is affected by methodological choices, both the size and characteristics of the conjoint population will change depending on the choice of definitional metrics (which may change by service system, locality and country).

Service evaluation studies of programs for mentally disordered offenders typically divide the full population by the following characteristics: *mental disorder* (e.g., severe mental illness, acute mental illness, personality disorder), *type of deviance* (e.g., non-violent or violent), *level of dangerousness* (e.g., low, medium, high) and *place of domicile* (e.g., community, jail or prison). For example, studies of court liaison programs focus on persons with mental disorders who are at risk of being criminally processed for their deviance. In contrast, police station liaison programs may limit their population to persons with any mental disorder who have been charged with particular types of non-violent offense, whereas evaluations of prison programs for personality disordered offenders target inmates with personality disorders. Each of these programs is defining a different sub-population of mentally disordered offenders for study.

> *Recommendation 2.* The target population should be defined in terms of characteristics that clearly define the boundaries of the group. In the case of mentally disordered offenders, the boundaries would include the definition of mental disorder, type of deviance, level of dangerousness and place of domicile. The size of the target population should then be estimated in absolute numbers and expressed as a proportion of the broader population.

Selection of subjects. There are two parts to the selection of study subjects. The first is the definition of inclusion and exclusion criteria. In theory, inclusion and exclusion screens shape the characteristics of the study sample to the characteristics of the target population. Yet, if there is a high level of diagnostic or problem uncertainty within the SCS population, it may be difficult to develop screens specific enough to distinguish a true case from a false one. This relates particularly to the definition of mentally disordered offender. Because the inclusion criteria for determining a positive case of mental disorder and offender may be based on different metrics, information is needed on the validity and reliability of the metric used to define caseness for inclusion.

> *Recommendation 3.* Any difference in the definition or measure of caseness between the study criteria and that used to estimate national prevalence figures (recommendation 2) needs to be explained and justified.

Exclusion screens are used to exclude persons who have characteristics that are not representative of the target population and that may confound the therapeutic pathway. This type of screen, while attempting to screen for the most representative sample for testing the intervention, could distort the sample in ways that make it less representative of the target population. For example, it is not uncommon to exclude mentally disordered offenders with co-occurring substance abuse problems from specialized programs. In some cases, the exclusion condition is written to exclude those individuals whose primary problem (which involves professional discretion) is substance misuse, whereas in other cases the substance abuse problem is expected to be treated prior to admission.

This is problematic for three reasons. First, evidence based on samples drawn from therapeutic,[29] community[30] and prison/jail environments[16,17] shows significant co-morbidity between mental illness and substance misuse. Consequently, a large portion of the target population could be excluded. Second, the excluded group may be the more difficult to

engage and treat. It has been found that relative to persons with single disorders, persons with dual diagnoses are sicker,[31] less functional,[32,33] heavier service users,[34] less compliant with medication and treatment interventions[35,36] and have poorer treatment outcomes.[33,37,38] Third, the excluded group may be precisely the one that generates the greatest societal costs in terms of violent offenses. Studies suggest that dual diagnosis of substance abuse and mental health is uniquely associated with more prevalent violent behavior.[39–41]

A wide assortment of selection criteria is used in studies of mentally disordered offenders. For example, studies of court liaison programs connected to regional secure units tend to focus on persons with psychotic disorders who (i) meet the criteria of involuntary commitment as defined by the Mental Health Act of 1983, (ii) have been charged with a crime and (iii) can be appropriately managed in a medium-secure unit. Excluded from these studies are persons with personality disorder, persons who are not sectionable under the Mental Health Act and persons with mental disorders who have higher or lower security needs. Similarly, evaluations of the Grendon Prison therapeutic community (TC) program for personality disorders excludes those who are not psychologically motivated[28] and have substance misuse problems. The evaluation of the Revolving Doors link worker scheme includes 'all mentally vulnerable adults in contact with the police' who (i) have unmet needs, (ii) are not currently connected with statutory services and (iii) are not considered dangerous.[42] Since each of these programs is targeting a different sub-group of mentally disordered offenders, it becomes impossible to compare across evaluations to determine the relative effectiveness of different interventions, comprising the very goals that motivated the effectiveness studies.

> *Recommendation 4.* Exclusion criteria should include cases that match the definition of the target population. That is, the sample should be equivalent to the target population on all characteristics that are significantly related to illness severity, level of impairment and service needs. If exclusion screens serve to change the sample in ways that alter the size and character of the target population, then the target population should be re-defined and comparative statistics (estimated in recommendation 2) re-estimated.

A second aspect of subject selection concerns voluntary participation: who wants to participate in the study? Because there is a distribution of cases around characteristics such as illness severity, functional impairment and other related problems, the ideal sample would replicate the distributional properties of the target population. The distribution of actual cases may be distorted by self-selection if only certain types of individual are willing and able to participate in the trial. Some biases may be generated because individuals who are identified by professionals as a true case may not agree with the professionals' assessment. In our core example, some individuals who are clinically assessed to be in the target population may not see themselves as mentally disordered and as such they may be unwilling to participate in a trial if offered the opportunity. This is in contrast to typical clinical trial where individuals have insight into their health problem and are willing to consider the risks and benefits of being restored by the intervention.

Selection bias is likely to be more problematic for studies evaluating services for persons with severe mental illness. Individuals who do not see themselves as mentally unwell are not likely to perceive or value the benefits of the therapeutic intervention. For example, medication compliance is a major source of preventable morbidity in the community-based treatment of schizophrenia.[43] Yet over three-quarters of persons with psychosis are non-complaint with anti-psychotic medications.[44,45] Compliance with medication regimes is associated with attitudes toward treatment, insight into illness, presence of psychosis and substance use.[46,47]

Getting a random sample of mentally disordered offenders to participate in a study is even more complex than for studies of other individuals with mental illness. Again, offenders may not agree to participate because they do not define themselves as being either psychologically unfit or mentally disordered. Moreover, because of the stigma associated with being a 'nutter' within the prison, offenders may be motivated to hide their mental health problems and resist any effort to reveal them. In addition, voluntary participation may take on a different meaning in services trials with mentally disordered offenders subject to correctional supervision. Participation in service trials may be tied to other valued benefits such as early release or dismissal of charges. These ancillary benefits may differently motivate offenders, with those with more serious offenses being more inclined to view the benefits favorably. For these reasons, those who do agree to participate in services trials may be systematically different from the target population.

Selection bias is typically measured by comparing the characteristics of the study sample to the sample invited to participate in the study, but this begs the question of what the salient characteristics are on which samples should be compared. Saliency here must be defined in terms of characteristics that relate to behaviors targeted for impact by the intervention. It may be that such comparisons require information about medication compliance, prior treatment, substance dependency, criminal offense and prior criminal history. However, comparing the samples on meaningful attributes such as these is problematic because individuals who refuse to participate may not be willing to reveal or give permission to access information necessary to make appropriate comparisons. In the absence of information on relevant attributes, comparisons are frequently made on observable or known attributes such as gender, race, age or diagnosis. Finding no detectable differences on these attributes may, however, be misleading. For example, in one study of an assertive community treatment program, selection bias was investigated in two ways. The first compared the diagnosis, age, gender, race and marital and employment status between the full and study samples. No statistically significant differences were found between participants and

non-participants. However, the second method revealed significant differences between participants and non-participants in terms of criminal justice activity. Those who refused to participate were more likely to be arrested.[48]

> *Recommendation 5.* Tests for selection bias should be based on attributes that are correlated with behaviors that are at the center of the services intervention.

Because some subjects may agree to participate but eventually leave the study for reasons that are unrelated or related to the intervention, it is vital that similar tests for bias be conducted on the final groups of clients who complete the full intervention. For example, the TC program at Grendon Prison for personality disordered offenders has an average attrition rate of 20 percent (personal communication with Director of Research at Grendon Prison, 24 March 1999). Attrition occurs because offenders choose to leave voluntarily or because they have violated a rule and are expelled from the program. The overall performance of the TC program may be biased upwards if, compared to those who leave the program, those who remain involved are significantly better suited for change.

Assignment of subjects. Random assignment is the gold standard of the RCT model. Effectiveness evaluations of programs for mentally disordered offenders rarely use random assignment to groups.[28,49] Rather, effectiveness is implied by improvement on key performance indicators (e.g., referral rates to services, engagement with social and health services, reduced rates of hospitalization or recidivism). Sometimes the key performance indicators are compared to matched or statistically constructed control groups of mentally disordered offenders that are not part of the intervention. Although such information is commonly reported, it is of questionable value. Descriptive statistics on lone interventions begs the question of whether the performance indicators would have been the same without the intervention—there is no way to prove added effectiveness without a comparison group. Yet, similarly, *ad hoc* comparisons with other groups of mentally disordered offenders draws into question the equivalence of the groups being compared—are the results different because the groups are different?

For example, there have been a number of evaluations of the therapeutic community at Grendon Prison. Inmates with personality disorder (as defined and certified by a prison medical officer) are admitted to Grendon if they are (i) serving a sentence of three years or more, (ii) in the later phase of their sentences, (iii) recommended by the prison medical officer, (iv) motivated to be involved in therapy, (v) psychologically minded, (vi) willing to accept responsibility for their offence, (vii) average or above average intellectual functioning, (viii) competent in English, (ix) drug free and (x) not on anti-psychotic medication. These criteria are met by roughly 400 of the estimated 30 000 sentenced inmates with personality disorders in prisons in England and Wales. Reconviction rates for inmates admitted to Grendon Prison have been found to be lower than those in a matched general prison sample.[50,51] The favorable results associated with the TC program at Grendon may be explained in part by the screens that only include inmates with above average scores on characteristics that predict treatment outcomes and in part by the comparison group which is drawn from the full prison population and matched by characteristics such as age, offense type and sentence length. But because the program screens for factors that may predict future outcomes, such as intelligence, motivation and willingness to accept responsibility, a randomized services trial would be appropriate to sort out the apples and oranges problem and to shed a reliable light on relative effectiveness.

There are two explanations frequently given for not using random assignment in evaluation studies of programs for mentally disordered offenders. The first focuses on issues of ethics. Randomization necessitates denying half of the sample access to a potentially superior intervention, giving them instead usual care. Because the usual care to which the other half would be assigned frequently amounts to no care, there is a non-trivial possibility that the behavior of individuals in the control group may deteriorate to a level that threatens their functioning and well-being, as well as that of society. This explanation, however, assumes that usual care is static and cannot be altered in ways that are consistent with appropriate and reasonable standards of care without replicating the characteristics of the experimental intervention. It is customary in clinical trials to make both groups better off by guaranteeing the control group a level of care that is consistent with good practice and the experimental group a possibility of better care but with risks.

The other explanation for not using random assignment is practical in nature. It has been claimed that there are too few cases for randomization, the staff is unwilling to withhold the experimental intervention from clients, agencies are unwilling to stylize an improved version of usual care, there is insufficient funding or there is lack of interest. These practical issues constrain the ability to appropriately test for effectiveness, and raise the broader and more relevant question of whether effectiveness evaluations should be diffused to the operations level—a practice encouraged by UK policies.

> *Recommendation 6.* Random assignment is a necessary condition for proving the effectiveness of services interventions. Non-randomized *in vivo* studies produce unreliable and potentially invalid results, unless all preexisting and expected differences that are likely to impact outcomes can be controlled between groups.

Developing a sensible randomization strategy, however, is not as straightforward as the standard RCT model would seem to suggest. Randomization to group must incorporate information about the distributional characteristics of the full population and how they relate to the characteristics of the therapeutic intervention. For example, the Netherlands has a special detention and treatment program (referred to

as 'TBS') for mentally disordered offenders who have committed serious violent crime.[52] The TBS program is comprised of six TBS clinics that manage roughly 1000 patients. Each clinic offers a unique therapeutic environment, each of which has been stylized to the needs of particular types of patient. TBS patients are not randomly assigned to the six clinics; rather, after a period of observation, patients are matched to the most appropriate clinic by the clinical staff of the Meijers Institute. Because the patients are judged to be clinically different in their behaviors and needs, it is not feasible to make comparisons across the six clinics. Similarly, it would be inappropriate to disregard the information about their differences and randomly assign them to the different programs on the same principle that it would be inefficient to randomly assign cars of different makes to different specialty repair centers.

In cases where there are therapeutically meaningful differences within the population, a segmented randomization strategy is appropriate. This involves first categorizing the patients by behavior and need and then randomly assigning them to a set of competing programs designed to manage particular behaviors. A segmented randomization strategy answers the question of which model is best for a particular type of patient, whereas the unsegmented randomization strategy answers the question of which model is best for all. A model that is best for all may be less effective than an array of programs that is best for particular types of patient if specialization of care increases the average therapeutic effect for each group or if the pooling of undifferentiated patients in the non-specialized model lowers the therapeutic effect for particular groups of patients.

> *Recommendation 7.* A segmented randomization strategy is appropriate if there is a preponderance of clinical or empirical evidence indicating that there are meaningful therapeutic subgroups within the target population and that these groups differ systematically in their complex of service needs and their responsiveness to treatment approaches.

Socially Complex Interventions and Assumption 3: Environment Equivalence and Neutrality

Compared to simple RCTs of drugs or surgical procedures, the design of 'clean and equivalent' environments for SCS trials is highly challenging. Three factors commonly represent the most substantial barriers to the assumption of equivalent environments.

SCS interventions have soft boundaries. It is typically assumed in clinical trials that the trial environment is independent of the social environment (as shown in Figure 1) and that the characteristics of the trial environment are under the direct control of researchers. These assumptions rarely hold for services trials. (See the randomized services controlled trial (RSCT) model in Figure 2.) More likely than not, the dose intervention influences the social environment (labeled by arrow ①), and the social environment influences the dose intervention and the behavior of the subjects

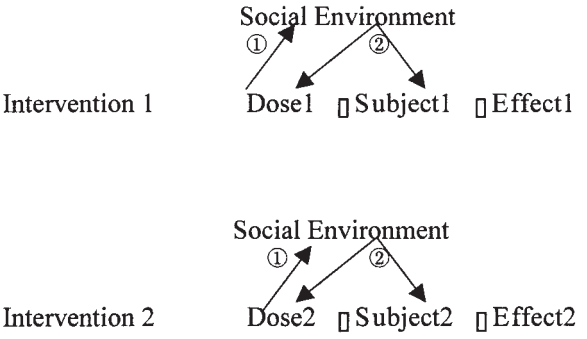

Figure 2. Diagram of the randomized services 'controlled' trial (RSCT) model that shows the interactions that arise when boundaries are permeable between the intervention and social environment

(labeled by arrows ②). If there is no hard boundary between trial and social environments, it is unclear to what extent (i) the magnitude of the absolute effects is due to the services intervention, the social environment or the interaction between the two and (ii) the difference in the relative effects between two programs is attributable to the differential impact of the social environment on the interventions.

Even with random selection and assignment of subjects, SCS trials may not be able to distinguish the most effective program (in isolation from the social environment) or tell us why an intervention was or was not effective. Ideally, one would need to randomly assign *programs* to environments. To do this, it would be necessary to identify the relevant features of the environment across which randomization will occur, in a manner parallel to that discussed earlier for individual subjects.

Services interventions have permeable boundaries in part because they are conducted in the community and are, therefore, part of it, in part because the intervention is attempting to affect the coordination of services delivered in the community to a particular group of users, and in part because the intervention draws on the resources in the community to 'dose' the user. For example, one of the primary goals of court liaison schemes is to connect mentally disordered offenders with an array of statutory services. Some court liaison schemes achieve this goal by way of multi-disciplinary teams comprised of representatives from various statutory agencies (e.g., health, probation and social services). Likewise, liaison programs based in prisons, like the Wessex project,[53] strive to engage released inmates with statutory services and to build collaboration among health, criminal justice and social agencies. These interventions cannot be effective without shaping and building the social environment within the community. In turn, interventions are shaped by the financial, social, and organizational characteristics and pressures that define the social environment, by the availability of resources in each of the local agencies and by the history of inter-agency relationships that influence their willingness to work together.

> *Recommendation 8.* Evaluation studies need to characterize the social environment in ways that facilitate inter-study comparisons and to measure how the

social environment interacts with the experimental and control interventions.

Implementation is influenced by local conditions. Although the design of an intervention may be theoretically driven, the way it is eventually implemented into practice depends heavily on local conditions. Unlike medical interventions, SCSs for mentally disordered offenders draw on resources from and the cooperation of agencies located within criminal justice, social services and health systems. For example, the success of liaison interventions depends in part on the level of social support and trust (i.e., macro-level social capital[54]) that exists among the key staff of agencies affected by the innovation. Innovative partnership interventions are more likely to be successful (dominate the usual uncoordinated approach) in those communities that want to work together. Indeed, these are the communities that typically apply (self-select) for experimental funding and that can produce letters of inter-agency support that are typically required before experimental funding is granted. This type of supportive local environment is biased against the control intervention and produces a non-neutral and non-equivalent therapeutic environment for the trial. That is, even if the experimental intervention is found to be more effective in this community, it may not be effective or cost-effective in another community because the environmental complex necessary to produce the results is absent. Moreover, this may suggest that the community itself may actually produce the effects that are attributed to the experimental intervention (which relates back to recommendation 1).

By contrast, interests of some local agencies may produce responses that undermine the performance of partnership building interventions. One of the goals of court liaison schemes is to identify people with mental disorders and to connect them to the appropriate mental health services in the community or hospital. These programs, in essence, create work for the health care sector. Local providers may resist the rise in demand for their services by frustrating the efforts of these schemes. This can be achieved by erecting administrative and statutory barriers that impede communication among liaison workers and local providers. For example, protocols for sharing medical or criminal history information with liaison workers may be written such that only particular information will be revealed to liaison workers with medical qualifications and only after they have submitted a request in writing and with appropriate authorization from the study subject. Slowing down and restricting the flow of information, through the strict enforcement of privacy laws, inhibits the referral process and retards the effectiveness of the intervention.

Even if a neutral and equivalent environment could be created at the beginning of an experiment, there is no guarantee that it would endure. Social environments are both complex and dynamic. This is particularly true in contemporary Britain, where New Labour has advanced a whirlwind of directives and new policies that are changing the way health, social and criminal justice services are organized and delivered, but rapid change is equally found in the US with the privatization of the public sector (e.g., jails and prisons) and the transformation of the health care system with the rise of managed care. Because the social environment is an integral part of the trial environment, changes within it may alter the relative performance of either intervention in ways that may have a differential impact on their effectiveness. Consequently, programs that were once effective (either relatively or absolutely) may no longer be effective because the local conditions have changed in ways that inhibit their performance or enhance the performance of usual care.

Recommendation 9. Evaluation studies need to define and measure local conditions, including levels of social capital, protocol arrangements, changes in service funding or organization and inter-agency staffing, that may interact with the experimental or control interventions. These factors need to be measured and monitored over the duration of the study.

Dose interventions are shaped by practical issues and personality factors. Experimental programs, by their very nature, are new and must be introduced to an existing social environment. Because new court liaison programs build on and feed off what already exists, how these programs are launched can affect their reception by the social environment and their eventual performance. For example, staff who are expected to make things work between systems are generally more effective when they have good relations with their colleagues on the other side. Having staff with reputations for being responsible, competent, trustworthy and pleasant and who are professionally well connected (i.e., micro-level social capital) may be the most vital part of an intervention. The development of micro-level social capital may be hindered if interventions are installed in ways that create physical or social distance among inter-agency staff.

The location and characteristics of accommodations for new programs can produce both physical and social distance. Opportunities for building rapport are reduced if the new program workers are placed in accommodations far removed from other professionals with whom they would need to work. Similarly, if the accommodations are superior for the new program or the staff is better resourced, resentments may form against the new program because it has advantages not extended to staff of collateral agencies. Such resentments are likely to be compounded if the new staff is seen as having more independence and less accountability. Although social and physical distance may be created in different ways, their effects are the same—to isolate the workers of the experimental program.

Recommendation 10. The effects of social and physical distance between the experimental intervention and collateral staff need to be examined to determine if local conditions shape the daily operations of control and experimental interventions in ways that are unique to setting and unequal to intervention.

Best Practice Guidelines for Services Effectiveness Research

Can effectiveness research based on the simple randomized controlled model yield valid, reliable and generalizable findings when it is applied to services that are socially complex? The best answer is: probably not without major design innovations. These services, by their nature, violate the assumptions underpinning the RCT design in ways that, even with random assignment, produce 'noise' between the dose and the effect, and the sample and the population, that threatens the validity, reliability and generalizability of findings. At a minimum, the simple RCT paradigm needs to be replaced with a more complex RSCT design that mirrors the complexity of services interventions. This new design would seek to minimize the distorting effects of ambiguous protocols and staffing arrangements, selection bias related to the population, sample and site and unmeasured variable problems associated with endogenous variables, which typically have been treated as exogenous. Yet, even with the recommendations proposed herein for restructuring the randomized controlled design, there are three important issues that will continue to challenge the utility of effectiveness research of socially complex services.

Issue 1: Selection Bias and Generalizability

Biased selection of populations, samples and sites have a direct impact on the generalizability of findings from trial to the real world. The challenge for researchers is to prove that their research samples and environments are representative of real world situations. Whether representativeness can be achieved depends critically on the role of factors such as professional discretion and motivation on individual and site participation. These factors may exert distorting effects that result in the creation of research samples that are non-representative of real world situations. Random assignment does not correct for this type of selection bias. Studies based on non-representative samples and environments will produce valid inferences but inferences that do not generalize to anything that exists in the real world. Whether research based on non-representative samples should or could inform policies that seek to shape best practices is arguable on clinical and economic grounds.

Issue 2: Unmeasured Variables and Validity

An incorrectly or selectively specified intervention is likely to produce invalid inferences about effectiveness. That is, if critical aspects of the protocol, staffing arrangements or social environment are left unspecified and unmeasured and these aspects are vital ingredients of an intervention, inferences about effectiveness may be falsely attributed to the known structural factors that have been specified and measured. While the bundled intervention may be more effective, it may, in practice, prove to be ineffective or inefficient if diffused in ways that replicate the known structural design but lack that vital interpersonal process or contextual variable that produces the superior effects. One way to minimize the effects of unmeasured variables is to add a comprehensive qualitative research component to services effectiveness research.

Issue 3: Endogeneity and Reliability

The ability to 'control' the boundaries of an intervention is essential for a randomized controlled trial. Yet, for socially complex service interventions, 'control' is very difficult since the social environment is part of the trial intervention. Even if the effect of the social environment can be defined and measured, it draws into sharp question whether the individual patient focus of the randomized controlled study is correct. That is, if the environment matters, then it may be necessary to implement a two-level structural design for the trial, with environmental conditions being the first level and individual patient the second. But, before this type of design could be used it would be necessary to know what features of the social environment influence effectiveness and whether these features could be replicated in other communities. If community factors are highly stylized and unique to place, it may be impossible to identify representative sites that could generate reliable inferences regarding effectiveness. Moreover, if environment does matter and the critical factors within the environment change as a result of policy changes, the effectiveness performance becomes unpredictable. The more dynamic the social environment the less reliable will be the inferences of randomized 'uncontrolled' studies, and the less useful will be results from randomized controlled studies.

Conclusion

This paper has focused on the problems and challenges associated with applying the simple RCT model to services that are socially complex. The recommendations suggested herein seek to strengthen the RCT design in ways that will improve the utility of SCS study findings. Such improvements, however, come at a price: an increased level of research effort and funding necessary to study the effectiveness of SCSs. At a minimum, the design of single site studies must broaden the scientific lens to monitor the possible effects of social environment, local conditions and social capital on outcomes. If these variables are found to be significant contributors to the production process, then a strong argument could be made for funding fewer, larger scale multi-site SCS trials, using a two-tier sampling design that incorporates these external factors in the site selection process. While there are a growing number of multi-site trials under way in the US, such as the Substance Abuse and Mental Health Services Administration funded effectiveness study of nine diversion programs located across the country, the underlying sampling designs are still single tier, reliant on sites' willingness to participate and insensitive to the possible effects of site on outcome. Researchers and

funders must accustom themselves to a more complicated research design of SCS evaluations if they are to achieve the desired outcome: meaningful information to guide best practice. However, it is important to note that even with the suggested modifications to the RCT model proposed here, there will still be uncertainty to the findings. The objective here is to enhance the performance of RCT design, not to eliminate error, which is unrealistic, as well as naïve.

Although the goal here is to raise the performance standards of the RSCT design, it is appropriate to question whether this outcome could be realized at a lower cost by using some alternative design. There are three possible options. First, efficacy studies could be substituted for effectiveness trials of SCSs. While this has theoretical appeal, it lacks practical utility since SCS interventions, by design, are part of the complex 'real' world, as they seek to change the community conditions in ways that will better serve the needs of persons residing there. Creating the efficacy condition of an 'ideal' noise-free environment would require, in the case of SCSs, artificially simulating the nature and complexity of the community. Whether this type of artificial modeling could be created is doubtful, but, even if it could be, the meaningfulness of these more internally valid results is more dubious as the community becomes less 'real'.

The next option is meta-analysis, which basically identifies the 'average of the average'[55] results from effectiveness trials. That is, it is expected that by averaging across the various trials the dominant finding will emerge once statistical methods have been introduced to control for inter-study variation. Implicitly it is assumed that the 'noise' within and among studies will wash out, but whether this happens depends again on the design of the individual studies. Meta-analysis is effective only if the individual studies measure the characteristics of confounding factors that interact with the intervention and information on these factors is reported.[23] The selection bias, unmeasured variables and endogeneity problems noted above limit the ability of meta-analysis to identify and measure the sampling and community effects on average findings. However meta-analysis would become more useful if unmeasured variables were measured as proposed in recommendations 2, 8, 9, and 10.

The quasi-experimental (Q-E) design, the last option, faces the same challenges as the RCT design, although these challenges are magnified by the 'naturalistic' attribute of the Q-E design. The primary challenge of the Q-E design is to statistically 'control' for all the confounding factors within the naturalistic settings and samples so that average differences can be attributed to the intervention. Here, again, the issues of unmeasured variables and selection bias become central to the statistical analysis. Even the most sophisticated Q-E studies are greeted with skepticism because of the difficulty of proving causation when there are so many ways in which the comparison groups may differ, as well as be affected by unmeasured factors, which cannot be adequately controlled for by statistical methods.

In conclusion, it remains an empirical question whether services effectiveness research can rely on the randomized controlled trial paradigm for valid, reliable and generalizable results. What is certain, however, is that the traditional design used to test the relative effectiveness of simple clinical trials is inappropriate for socially complex services, and that alternative options to the RCT are unlikely to perform any better. Our best hope still rests with the RCT design but with stylized modifications that will make each study more time-consuming and expensive. Without these modifications, we may, through our best but biased research practices, discover that the best effectiveness evidence yields ineffective or inefficient practice guidelines.

Acknowledgements

This research was supported by an Atlantic Fellowship in Public Policy, which is funded by the British Government.

References

1. Department of Health. *The New Approach to Social Services Performance*. Department of Health: London, 1999.
2. Department of Health. *Modernising Health and Social Services*, No. 13929. Department of Health: London, 1998.
3. Department of Health. *Modernising Mental Health Services, Safe, Sound, and Supportive*. Department of Health: London, 1998.
4. Department of Health. *The New NHS: Modern, Dependable*. Department of Health: London, 1997.
5. Peckham M. *Research for Health*. Department of Health: London, 1993.
6. HM Inspectorate of Probation (HMIP). *Strategies for Effective Offender Supervision*, Report of the HMIP What Works Project. Home Office: London, 1998.
7. Home Office. *Home Office Circular 38/1998: Crime and Disorder Act of 1998*. Home Office: London, 1998.
8. The President's Advisory Commission on Consumer Protection and Quality in Health Care Industry. Fostering Evidence-Based Practice and Innovation. In *Quality First: Better Health for All Americans*, US Government Printing Office: Washington, DC, 1999; Chapter 11, 169–182.
9. Gold MR, Siegel JE, Russell LB, Weinstein MC. *Panel Report, Cost Effectiveness in Health and Medicine*. Oxford University Press: Oxford, 1996.
10. Wolff N. Measuring costs: what is counted and who is accountable? *Disease Mgt Clin Outcomes* 1998; **1** (4): 114–128.
11. Wolff N, Helminiak, TW, Tebes KJ. Getting the cost right in cost-effectiveness analyses. *Am J Psychiatry* 1997; **154**: 736–743.
12. Wolff N, Helminiak TW. Nonsampling measurement error in administrative data: implication for economic evaluations. *Health Econ* 1996; **5**: 501–512.
13. Drummond MF, Brandt A, Luce B, Rovira J. Standardizing methodologies for economic evaluation in health care. *Int J Technol Assessment Health Care* 1993; **9**: 26–36.
14. Mason J, Drummond M. Reporting guidelines for economic studies. *Health Econ* 1995; **4**: 85–94.
15. Wolff N. 1998. Interactions between mental health and law enforcement systems: Problems and prospects for co-operation. *J Health Politics Policy Law* 1998; **28**: 13–74.
16. Lamb HR, Weinberger LE. Persons with severe mental illness in jails and prisons: a review. *Psychiatric Services* 1998; **49**: 483–492.
17. Office of National Statistics. *Psychiatric Morbidity among Prisoners in England and Wales*. HM Stationery Office: London, 1998.
18. Kanter J. Clinical case management: definition, principles, components. *Hosp Community Psychiatry* 1989; **40**(4): 360–368.
19. McGrew JH, Bond GR. Critical ingredients of assertive community treatment: judgments of the experts. *J Mental Health Admin* 1995; **22**(2): 113–125.
20. Bachrach L. Young adult chronic patients: an analytical review of the literature. *Hosp Community Psychiatry* 1982; **33**(3): 189–197.
21. Office of National Drug Control Policy. *National Drug Control Strategy*. Government Printing Office: Washington, DC, 1989.

22. Feinstein AR, Horwitz RI. Problems in the 'evidence' of 'evidence-based medicine. *Am J Med* 1996; **103**: 529–535.
23. Krogh Johansen H, Gotzsche PC. Problems in the design and reporting of trials of antifungal agents encountered during meta-analysis. *JAMA* 1999; **282**(18): 1752–59.
24. Home Office. *Mentally Disordered Offenders: Survey of Inter-Agency Arrangements*. Home Office: London, 1997.
25. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders*, 4th edn. American Psychiatric Association: Washington, DC, 1994.
26. World Health Organisation. *The ICD-10 Classification of Mental and Behavioural Disorders: Diagnostic Criteria for Research*. World Health Organisation: Geneva, 1993.
27. Hoggett B. *Mental Health Law*. Sweet and Maxwell: London, 1996.
28. Cullen E, Jones L, Woodward R. *Therapeutic Communities for Offenders*. Wiley: New York 1997.
29. Ross HE, Glaser FB, Germanson T. The prevalence of psychiatric disorders in patients with alcohol and other drug problems. *Arch Gen Psychiatry* 1988; **45**: 1023–1031.
30. Kessler RC, Nelson CB, McGonagle KS, et al. The epidemiology of co-occurring addictive and mental disorders: implications for prevention and service utilization. *Am J Orthopsychiatry* 1996; **66**: 17–31.
31. Bartels SJ, Drake RE, McHugo GJ. Alcohol abuse, depression, and suicidal behavior in schizophrenia. *Am J Psychiatry* 1992; **149**: 394–395.
32. Newman DL, Moffitt TE, Caspi A *et al.* (1998). Comorbid mental disorders: Implications for treatment and sample selection. *J Abnormal Psychol* 1998; **107**, 305–311.
33. Osher FC, Drake RE. Reversing a history of unmet needs: approaches to care for persons with co-occurring, addictive and mental disorders. *Am J Orthopsychiatry* 1996; **66**(1): 4–11.
34. Bartels SJ, Tegue GB, Drake RE, *et al.* Substance abuse in schizophrenia: service utilization and costs. *J Nervous Mental Dis* 1993; **181**: 227–232.
35. Drake RE, Osher FC, Wallach MA. Alcohol use and abuse in schizophrenia: a prospective community study. *J Nervous Mental Dis* 1989; **177**: 408–414.
36. Osher FC, Drake RE, Noordsy DL, Teague GB. Correlates and outcomes of alcohol use disorder among rural outpatients with schizophrenia. *J Clin Psychiatry* 1994; **55**(3): 109–113.
37. Dakis CA, Gold MS. Psychiatric hospitals for treatment of dual diagnosis. In *Substance Abuse: A Comprehensive Textbook*, Lowinson JH, Ruiz P, Millman RM (eds). Williams and Wilkins: Baltimore, MD, 1992; 467–485.
38. McLellan AT. Psychiatric severity as a predictor of outcome from substance abuse treatments. In *Psychopathology and Addictive Disorders*, Meyer RE (ed.). Guilford: New York 1986.
39. Steadman HJ, Mulvey EP, Monahan J, Robbins PC, Appelbaum PS. Violence by people discharged from acute psychiatric inpatient facilities and by others in the same community. *Am J Gen Psychiatry* 1998; **55** (5): 393–401.
40. Link B, Andrews H, Cullen FT. The violent and illegal behavior of mental patients reconsidered. *Am Sociol Rev* 1992; **57**: 275–292.
41. Swanson J, Holzer DE, Ganju VK, Jono RT. Violence and psychiatric disorder in the community: evidence from the epidemiologic catchment area surveys. *Hosp Community Psychiatry* 1990; **41** (7): 761–770.
42. Revolving Doors. *Revolving Doors Agency, Link Worker Schemes*, Operational Policy Draft. Revolving Doors: London, 1997.
43. Kane JM. Problems of compliance in the outpatient treatment of schizophrenia. *J Clin Psychiatry* 1983; **44**: 3–6.
44. Weiden PJ, Dixon L, Frances A, *et al.* Neuroleptic noncompliance in schizophrenia. In *Advances in Neuropsychiatry and Psychopharmacology: Schizophrenia Research*, Vol. 1, Taminga CA, Schulz SC (eds). Raven: New York 1991; 285–296.
45. Corrigan PW, Liberman RP, Engel JD. For noncompliance to collaboration in the treatment of schizophrenia. *Hosp Community Psychiatry*. 1990; **41**: 1203–11.
46. Fenton WS, Blyler CR, Heinssen RK. Determinants of medication compliance in schizophrenia: empirical and clinical findings. *Schiz Bull* 1997; **23**: 637–51.
47. Weiden PJ, Rapkin B, Mott T, *et al.* Rating of medication influences (ROMI) scale in schizophrenia. *Schiz Bull* 1994; **20**: 297–310.
48. Wolff N, Helminiak TW, Diamond R. Estimated societal costs of assertive community mental health care. *Psychiatric Services* 1995; **46**: 898–906.
49. James A. *Life on the Edge: Diversion and the Mentally Disordered Offender*. The Mental Health Foundation: London, 1996.
50. Newton M, Thornton D. Grendon re-conviction study, Part 1 update. Unpublished report, 1995.
51. Cullen E. The Grendon reconviction study, Part 1. *Prison Service J* 1993; **90**: 35–37.
52. Ministry of Justice. *TBS, a Special Hospital Order of the Dutch Criminal Code*. Ministry of Justice: The Hague, 1994.
53. Lart R. *Crossing Boundaries: Assessing Community Mental Health Services for Prisoners on Release*. Policy Press: Bristol, 1997.
54. Putnam R. *Making Democracy Work*. Princeton, Princeton University Press: NJ; 1993.
55. Feinstein AR. Meta-analysis and meta-analytic monitoring of clinical trials. *Stats Med* 1996: **15**(12): 1273–80.

