Skip all front matter: Jump to Page 1 ▼

TECHNICAL REPORT

Disease management evaluation

A comprehensive review of current state of the art

Annalijn Conklin, Ellen Nolte

December 2010

Prepared for the European Commission

RAND EUROPE

The research described in this report was prepared for the European Commission's Seventh Framework Programme.

Published 2011 by the RAND Corporation
1776 Main Street, P.O. Box 2138, Santa Monica, CA 90407-2138
1200 South Hayes Street, Arlington, VA 22202-5050
4570 Fifth Avenue, Suite 600, Pittsburgh, PA 15213-2665
Westbrook Centre, Milton Road, Cambridge CB4 1YG, United Kingdom
RAND URL: http://www.rand.org
RAND Europe URL: http://www.rand.org/randeurope
To order RAND documents or to obtain additional information, contact
Distribution Services: Telephone: (310) 451-7002;
Fax: (310) 451-6915; Email: order@rand.org

Preface

This report was prepared as part of the project "Developing and validating disease management evaluation methods for European health care systems" (DISMEVAL), funded by the European Commission's Seventh Framework Programme (grant agreement 223277). There are two overall aims of the DISMEVAL project: first, to review approaches to chronic care and disease management in Europe; and, second, to provide evidence for best practices and recommendations to policymakers, programme operators and researchers on which evaluation approach is most useful in a given context, through testing and validating possible evaluation methods.

This report forms part of the first phase of the project, which involves an assessment of disease management evaluation approaches in Europe. It presents a comprehensive review of the academic and grey literature on evaluation methods and metrics used in disease management, which was conducted between February and September 2009. The report, completed December 2010, aims to contribute a comprehensive inventory of existing evaluation methods and performance measures as a means to identify the key issues that need to be considered for the development and implementation of approaches to evaluate disease management programmes or their equivalent. It also aims to highlight the potential challenges and possible solutions to evaluating complex interventions such as disease management.

RAND Europe is an independent not-for-profit policy research organisation that aims to improve policy and decisionmaking in the public interest, through rigorous research and analysis. RAND Europe's clients include European governments, institutions and non-governmental organisations, and firms with a need for rigorous, independent, multidisciplinary analysis.

This report was peer-reviewed in accordance with RAND's quality assurance standards.

For more information on RAND Europe or this document, please contact:

Annalijn Conklin

RAND Europe
Westbrook Centre
Milton Road
Cambridge CB4 1YG
Tel. +44 (1223) 353 329
aconklin@rand.org

Contents

Table of Tables

Table of Boxes

Executive Summary

Chronic diseases account for a large share of healthcare costs while the care for people with such conditions remains suboptimal. Many countries in Europe are experimenting with new, structured approaches to better manage the care of patients with chronic illness and so improve its quality and ultimately patient health outcomes. While intuitively appealing, the evidence such approaches achieve these ends remains uncertain. This is in part because of the lack of widely accepted evaluation methods to measure and report programme performance at the population level in a scientifically sound fashion that is also practicable for routine operations. This report aims to help advance the methodological basis for chronic disease management evaluation by providing a comprehensive inventory of current evaluation methods and performance measures, and by highlighting the potential challenges to evaluating complex interventions such as disease management.

Challenges as identified here are conceptual, methodological, and analytical in nature. Conceptually, evaluation faces the challenges of a diversity of interventions subsumed under a common heading of "disease management" which are implemented in various ways, and a range of target populations for a given intervention. Clarifying the characteristics of a disease management intervention is important because it would permit an understanding of the effects expected and how the intervention might produce them and also allow for the replication of the evaluation and the implementation of the intervention in other settings and countries. Conceptual clarity on the intervention's target and reference populations is equally necessary for knowing whether an evaluation's comparator group represents the counterfactual (what would have happened in the absence of a given intervention). Other conceptual challenges relate to the selection of evaluation measures which often do not link indicators of effect within a coherent framework to the aims and elements (patient-related and provider-directed) of a disease management intervention and to the evaluation's objectives.

The establishment of the counterfactual is indeed a key methodological and analytical challenge for disease management evaluation. In biomedical sciences, randomised controlled trials are generally seen as the gold standard method to assess the effect of a given intervention because causality is clear when individuals are randomly allocated to an intervention or a control group. In the context of multi-component, multi-actor disease management initiatives, this design is frequently not applicable because randomisation is not possible (or desirable) for reasons such as cost, ethical considerations, generalisability, and practical difficulties of ensuring accurate experimental design. As a consequence, alternative comparison strategies need to be considered to ensure findings of intervention effect(s) are not explained by factors other than the intervention. Yet, as alternative strategies become less of a controlled experiment, there are more threats to the validity of findings from possible sources of bias and confounding (e.g. attrition, case-mix, regression to the mean, seasonal and secular trends, selection, therapeutic specificity and so on) which can undermine the counterfactual and reduce the utility of the evaluation.

As many design options have been suggested for disease management evaluation, choosing an appropriate study design can be an important methodological approach to selecting a suitable control group for an alternative comparison strategy that still achieves the goals of randomisation in disease management evaluation. Equally, there are various analytical approaches to such construction whereby controls are randomly matched to intervention participants can be created through predictive modelling or propensity scoring techniques, or they are created statistically by developing baseline trend estimates on the outcomes of interest. Whichever the approach taken to construct a sound comparison strategy, there will be a set of limitations and analytical challenges which must be carefully considered and may be addressed at the analysis stage. And, while some statistical techniques can be applied ex post to achieve the objectives of randomised controlled trials, such as regression discontinuity analysis, it is better to plan prospectively before a given intervention is implemented to obtain greater scientific returns on the evaluation effort.

Other methodological and analytical challenges in disease management evaluation also require thoughtful planning such as the statistical power of evaluation to detect a significant effect given the small numbers, non-normal distribution of outcomes and variation in dose, case-mix, and so on, typical of disease management initiatives. Several of these challenges can be addressed by analytical strategies to assure useful and reliable findings of disease management effects such as extending the measurement period from 12 month to 18 months and adjusting for case-mix to calculate sample size, for example. But, ultimately, what is required is a clear framework of the mechanisms of action and expected effects that draws on an understanding of the characteristics of disease management (scope, content, dose, context), those of the intervention and target populations (disease type, severity, case-mix), an adequate length of observation to measure effects and the logical link between performance measures and the intervention's aims, elements and underlying theory driving the anticipated behaviour change.

Glossary of Terms and Definitions

Bias. Extent to which there is any systematic error, observed or unobserved, in collecting or interpreting data (e.g. selection bias, recall bias). Only biases that can be observed can be controlled for in statistical analyses.

Confounding. Extent to which there is any additional variable that may influence the results outside the variables being observed. Confounding factors can cause or prevent the outcome of interest; they are not intermediate variables; and are not associated with the factor(s) under investigation. Confounding by indication is an example in this context where a provider might use a disease management intervention selectively for patients based on disease severity, which is also independently associated with the outcomes being evaluated.

Comparability. Extent to which two groups are probabilistically equal on all characteristics or risk factors except for an exposure or an intervention of interest; or, the extent to which a given measure allows for reliable comparison with external benchmarks or to other datasets collected in similar circumstances.

Reliability. Extent to which a given measure or diagnosis is reproducible across repeated tests, testers and time periods. *Inter-rater reliability* refers to different measurement taken by different persons with the same method or instruments; *test-retest (intra-rater) reliability* concerns measurement variation from a single person or instrument on the same item and under the same conditions; *inter-method reliability* is the variation in measurements of the same target when the same person uses different methods or instruments; and *internal consistency reliability* looks at the consistency of results across items within a test.

Replicability. Extent to which a given measure can be replicated in other contexts or by other researchers. This is a different concept from generalisability (below).

Validity. Extent to which tests, formulae, statements and arguments accurately represent a concept or phenomenon. *Measurement or construct validity* is the extent to which variables truly measure the intended concept(s), and *face validity* concerns the extent to which results of an instrument for one kind of person, setting, treatment or outcome can be generalised to elements that appear similar in important characteristics (e.g. a nurse disease management (DM) intervention is expected to elicit a similar response from all individuals with the same disease). Three types of validity relate to causal statements: *internal validity* means the extent to which causal attributions can be correctly made for an exposure-disease, or intervention-outcome relationship; *external validity*

(generalisability) is the extent to which a causal effect identified in a given study can be generalised across characteristics of persons, places and time periods (beyond the specific research setting); and ***ecological validity*** relates to the extent to which findings from population level (aggregate) data are applicable to relationships of interest at the level of an individual's everyday, natural social settings when individual level data are unavailable.

Acknowledgements

The authors wish to thank the independent peer-reviewers, Drs Jack Needleman and John Adams, for their comments and suggestions to help improve the document. Thanks are also due to those colleagues on the DISMEVAL project who provided constructive feedback and support as we developed this report.

We are also very grateful to the European Commission for its support of the DISMEVAL project. For further information about the project please see www.dismeval.eu.

The views expressed in this report are those of the authors alone and do not necessarily represent those of the European Commission. The authors are fully responsible for any errors.

CHAPTER 1 Introduction

1.1 Background

Chronic diseases place a substantial burden on individuals, their carers and society as a whole. They account for a large share of healthcare costs, while the care for people with such conditions remains suboptimal. Structured approaches to manage patients with chronic illness have been proposed as a means to improve the quality of care, to reduce the cost of healthcare, and ultimately to improve health outcomes for the chronically ill.

While intuitively appealing, the evidence such approaches achieve these ends remains uncertain. Existing research points to examples of benefits for patient outcomes (Powell Davies et al., 2008), while evidence of reduced direct healthcare costs is at best inconclusive (Mattke, Seid and Ma, 2007). What we know about the impact of interventions to manage chronic disease(s) tends to be based, mainly, on small studies that frequently focus on high risk patients, and are often undertaken in academic settings (Bodenheimer, Wagner and Grumbach, 2002). The effects of large, population-based programmes and initiatives are less well understood (Mattke, Seid and Ma, 2007). This is in part because of the lack of widely accepted evaluation methods to measure and report programme performance at the population level in a scientifically sound fashion and that are also practicable for routine operations. Indeed, although there is now considerable guidance particularly in the United States on methods to evaluate disease management programmes and initiatives, little effort has yet been made to standardise such approaches (Wilson et al., 2008). In addition, the different recommended approaches each face methodological, conceptual and analytical challenges whose impact on measured effects still need to be addressed and understood. Substantive methodological challenges in evaluating population-based programmes and initiatives include, for example, the critical need in evaluation for controlling for programme and patient choices about who receives a given intervention, since most programmes lack randomisation, and the need to overcome limitations in data quality.

Another challenge is the diversity of interventions that have broadly been subsumed under the heading "disease management" (Nolte and McKee, 2008) and the lack of standardisation of initiatives. Disease management interventions might include a provider network, a care pathway, case management, coordinated care, the chronic care model, Germany's Disease Management Programme, integrated care, managed discharge, a nurse-led clinic, or multidisciplinary team(s) or care. These tend to vary widely in design, scope, scale, operational detail and providers involved (Lemmens et al., 2008; Mattke, Seid and Ma, 2007; Pearson et al., 2007; Steuten et al., 2006; MacStravic, 2005; Weingarten et al., 2002). Within a particular initiative, the different components (both patient- and provider-focused) might be implemented differently across locations and levels of implementation (Lemmens et al., 2008; Cretin, Shortell and Keeler, 2004). Variation also exists in the definition of the scope, focus, purpose and range of component interventions of disease management itself (Norris et al., 2002). Definitions range from “discrete programs directed at reducing costs and improving outcomes for patients with particular conditions” (Rothman and Wagner, 2003) to “an approach to patient care that coordinates medical resources for patients across the entire delivery system” (Ellrodt et al., 1997). The challenge of diversity is further added to by related concepts such as “integrated care”, “coordinated care”, “collaborative care”, “managed care”, “case management”, “patient-centred care”, “continuity of care”, “seamless care” and others. Box 1 gives a brief history of the evolution of the concept of disease management in the United States. Detail on the development of approaches to manage chronic conditions in European countries is given elsewhere (Nolte and McKee, 2008; Nolte, Knai, McKee, 2008).

Thus, despite accumulation of evidence for the effects of approaches to chronic illness, methodological and analytical work is still needed to develop widely accepted evaluation methods that are scientifically sound and also practicable in routine settings. Given all the diversity and variability of disease management described above, a key issue for this work concerns the difficulties in establishing a useful “comparator” in settings where it is not practical or possible to execute an evaluation as a randomised controlled trial (RCT). This is indeed an important task because evaluation methods are a precondition to select efficient and effective programmes, or components within a programme, that can address the growing burden of chronic conditions. The DISMEVAL project aims to support this process by identifying and validating evaluation methods and performance measures for disease management programmes or equivalent approaches and to make recommendations to policymakers, programme officials and researchers on best practices that are both scientifically sound and operationally feasible.

Box 1 Brief history of the evolution of disease management initiatives

"**Disease management**" is a concept that was first introduced in the United States in the 1980s and has traditionally targeted a single (chronic) disease or condition. Disease management was initially used by pharmaceutical companies to promote medication adherence and behaviour change among people with diabetes, asthma and coronary heart disease through educational programmes offered to employers and managed care organisations. In the mid-1990s the US healthcare industry began to adopt disease management strategies, with nearly 200 companies offering disease management programmes by 1999. As growing evidence showed that treating people with chronic conditions could save costs, disease management was adopted more widely.

There was certainly a financial gain for disease management organisations, which had an estimated growth in revenues of nearly 40 percent between 1997 and 2005. The payers of disease management initiatives included the US federal government, individual states and large employers (with 200 or more employees). But although the concept of disease management was widely embraced by the early 2000s, the nature and scope of programmes under this heading varied from small initiatives focused on a limited set of people in a single payer group to widespread programmes for almost all chronically ill people across multiple payer groups.

There are two important trends in the evolution of this profitable disease management "industry". First, there were two basic types of initiative among the variety of programmes labelled disease management: "on-site" programmes and "off-site" or "carved-out" programmes. The on-site type of initiative was directed by the primary provider and delivered within a primary care setting, while the off-site type had a focus on specific care processes or clinical outcomes and was usually offered by commercial for-profit vendors. This second "carved-out" type of initiative usually involved information systems for patient education and self-management and was marketed as a cost-containment strategy to employers and health insurers. Notably, disease management can have a negative connotation for some people because the "carved-out" type initiatives lacked a focus on patient outcomes or tended to concentrate on short-term outcomes only.

The second trend in the evolution of disease management was a movement towards a broader, population-based approach to managing chronic conditions. Early programmes in this trend focused on single conditions or diseases, with commercial vendors targeting patient education and medication adherence. These programmes also evolved to shift the focus to addressing multiple needs of people with co-morbidities and multiple conditions. This second generation of population-based approaches attempted to provide a more integrated approach to care and, similar to "off-site" disease management programmes, led to the development of information systems aimed at identifying at risk patients.

Thus, the evolution of disease management gave rise to a situation where the initiatives varied in scope, focus, purpose and range of components, and the concept itself had a range of definitions and meanings.

SOURCE: Nolte and McKee (2008).

1.2 Research Approach

The research undertaken for this report is a literature review aimed at identifying existing methods and metrics to evaluate disease management programmes, approaches or interventions. We identified papers for possible inclusion by combining searches of electronic databases, hand searches of reference lists of papers, contact with experts in the field and a purposive electronic search of grey literature on key websites. Due to the differences between the specific focus, intended audience and type of relevant documents, we synthesised the included literature through narrative review of each document without any formal meta- or content analysis using a structured template or otherwise. Below, we describe the literature search approach in more detail.

1.2.1 Literature Review

We carried out a two-stage review. At the first stage, we performed a targeted search of (a) existing reviews of methodology papers in peer-reviewed journals and grey literature; (b) subject-specific journals (e.g. *Evaluation in the Health Professions*, *Disease Management and Health Outcomes*, *Disease Management*, *Managed Care Interface*, *The American Journal of Managed Care*, *Journal of Evaluation in Clinical Practice*); (c) papers and references provided by leading authors in the field of chronic disease management initiatives; and (d) the world wide web using common search engines and websites of organisations and companies that offer disease management programmes in the United States (American Healthways, Core Solutions, Health Dialog and United Healthcare) to identify additional documents on evaluation methods for disease management. We followed up references in articles and reports we identified. This search found 72 papers for inclusion in the first stage of the review.

This first stage of the review informed the identification of search term combinations for a systematic search of the literature using electronic databases including Medline/PubMed, Web of Science and CINAHL. The systematic literature search in the second stage focused on three concepts: (1) evaluation research; (2) disease management; and (3) indicator of effect. In order to capture the related literature we applied broad search terms, using combinations of PubMed "MeSH" terms ("/" indicating "or"): "evaluation/method/design/assessment" and "disease management/self care/case management" and "quality indicators/outcome/measure/effect". The terms "integrated care" and "care management" were not identified as alternative search terms for "disease management" because they also include concepts related to wider issues of how healthcare is organised and managed more generally rather than how chronic diseases are managed specifically. Since the PubMed search tended to result in papers with a strong focus on methods to evaluate financial outcomes and we wanted to capture non-financial indicators of effect in a European context, we supplemented our systematic search of PubMed by a systematic search of CINAHL and Web of Science using terms which reflected more qualitative indicators of effect (e.g. satisfaction and quality of life). The detailed search strategy is described in Appendix 1.

We limited the search in stage two of the review to papers published between 1 January 1990 and 30 September 2009. We screened titles and abstracts for eligibility for inclusion. We retrieved studies we considered eligible where possible and scrutinised them further for inclusion. Then we followed up references in studies considered eligible where appropriate. We excluded evaluation studies, evaluation protocols, reviews of effectiveness evidence or reviews of disease management approaches/interventions. We further excluded editorials and commentaries unless they explicitly commented on evaluation methodology, and articles in languages other than English, German, French and Spanish. We placed no limits on text options in Medline/PubMed in case we missed articles without an abstract.

The PubMed search identified 4,053 titles, of which we extracted 332 for abstract review and duplicate removal. This resulted in 34 new titles for inclusion in the review. The Web of Science and CINAHL searches identified 17 additional titles; we considered five eligible and retrieved them. We scrutinised articles considered eligible and retrieved them further for inclusion or exclusion in the review (15 could not be retrieved for review).

In total, our two-staged literature search yielded 111 papers which we considered eligible and included in this review. By publication type, this number comprised peer-reviewed articles (N=89), book chapters (N=6), working papers/reports (N=10) and other grey literature such as industry magazines or fact sheets (N=7).

1.3 Structure of the Report

This report has three main chapters. Chapter 2 outlines the principal objectives and design options available to evaluation research, along with a description of study design options available to evaluation in general and for disease management evaluation in practice. Chapter 3 discusses the threats to validity of the common research designs of various evaluation approaches and expands on a number of challenges, alternative strategies and consequences of method choices for specific consideration in disease management evaluation. The final chapter summarises the findings on the current state of the art to inform the design and conduct of evaluation approaches in Europe and elsewhere.

CHAPTER 2 Evaluating Disease Management: Objectives and Principles

The evaluation of disease management initiatives and programmes aims to understand the contribution made by a given programme or activity to achieving particular outcomes.[1] It involves forming a judgement about performance against a set of explicit criteria using a transparent and objective selection of measures (Ling, 2009). The set of indicators selected needs to be informed by a theory of change that makes a logical connection between the planned activities to the intended results and impacts (MacStravic, 2008). This chapter explores some overarching principles of disease management evaluation, including general objectives, principal design options and evaluation approaches applied to disease management in practice.

2.1 Objectives of Evaluating Disease Management Interventions

The evaluation of a given disease management intervention may pursue a range of objectives. In some settings there may be an emphasis on assessing the (longer-term) economic impact of the intervention, most notably on determining whether costs were saved or whether the intervention yielded a positive return on investment (ROI). This has particularly been the case for so-called "off-site" or "carved-out" disease management programmes that focus on specific processes of care or clinical outcomes, and that are normally offered by commercial for-profit vendors, such as specialised disease management organisations, and are marketed to employers and health insurers primarily as a cost-containment strategy (Cavanaugh, White and Rothman, 2007; Bodenheimer, 2000). Other perspectives may emphasise the quality improvement achieved through structured disease management by assessing processes of care (e.g. adherence to clinical or practice

[1] The terms 'evaluation' and 'assessment' are often used interchangeably in the disease management literature, but this can lead to conceptual confusion over the many ways in which academic evaluation research is distinct from assessment as performance audit (Ling, 2009; Zajac, 2004). Assessment aims only to develop a narrative account of why something happened, whereas evaluation research seeks to make a judgement about programme performance which might be based on information provided by an assessment (Ling, 2009). If future literature on disease management evaluation includes more qualitative methods, then greater linguistic clarity of these two concepts may be warranted so the results of such future studies can be appropriately interpreted and understood against their aims.

guidelines, referral rates) and outcomes such as disease control or satisfaction of programme participants.[2]

Drawing on Avedis Donabedian's framework (1980) for evaluating the quality of healthcare and on standards for outcome measurement in disease management (Mulcahy et al., 2003), the evaluation of disease management may focus on single or multiple components of intervention structure, process, output and outcome or impact (Lemmens et al. 2008; Pearson et al., 2007; CBO, 2004):

- **Inputs** are the structural aspects of a given intervention, such as its financial inputs and human resources.
- **Processes** are actual activities such as how a disease management programme is delivered or implemented (how it worked) and to the fidelity of activities (the extent to which a programme was implemented *as intended* and/or implemented according to the evidence base).
- **Output** is defined as productivity or throughput – the immediate result of professional or institutional healthcare activities, usually expressed as units of service.
- **Outcomes** are the medium- and long-term effects of healthcare or a health intervention on the health status of individuals and populations. Outcomes can be further divided across a continuum of categories from immediate, intermediate, post-intermediate to definite or long-term outcomes of health status. A definite outcome might also be considered to be a health "impact".

Possible evaluation outcomes of interest are not only related to health status but can also include economic impact, social impact or even environmental impact.

Table 1 provides examples of measures for each of the dimensions of a logic model listed above. The selection of dimensions and actual measures will be determined by the specific design and goals of a given disease management intervention in order to ensure construct validity in the selection of evaluation measures (Pearson et al., 2007). These issues will be discussed in further detail below.

[2] The term participant or individual is used where possible, instead of subject or patient in recognition of (a) the goal of self-management in disease management programmes, and (b) ethics literature on the role of individual agency in medical care and research.

Table 1 Evaluation measures for disease management interventions

Evaluation measure	Variable	Example
Input measures		
Structure of DM programme		Staffing ratios; caseload size; staff qualifications; hours of training; experiential preparation; organisational supports;
Process measures		
Patient-related	Reach	Initial contact rate; enrolment rate; referral rate; targeted population
	Patient education	Education sessions; content covered
	Patient coaching	Contact frequency; call duration; call content; a written action plan; smoking cessation counselling
Organisational		Frequency of disease-specific diagnostic testing and/or follow-up (e.g. eye exam rate; foot exam rate; HbA1c tests; blood pressure tests); procedures performed (e.g. pulmonary lab procedure, perfusion imaging); adherence to standards of care as defined by relevant organisation; prescription rates
Output measures		
Utilisation	Utilisation	Hospital admissions; emergency room visits; physician or clinic visits; length of stay; in-patient bed days; urgent care visits; scheduled physician or clinic visits; readmission rate; number of insurance claims for medications; waiting times, discharge rates
Outcome measures		
Immediate/proximate	Knowledge	Participant knowledge (general and disease-specific); beliefs (general and disease-specific)
Intermediate	Self-care behaviour	Administration of oral/injectable medication, adherence to diet/exercise, glucose self-monitoring
	Self-efficacy	Self-efficacy; health locus of control; psychosocial adaptation/coping skills
Post-intermediate	Clinical	Physiological measures (e.g. HbA1c values; blood pressure; blood lipids); weight; self-reported severity of symptoms; shortness of breath; smoking rate; quantity and frequency of exercise; adherence to medication
	Satisfaction	Programme satisfaction; perceptions of migraine management; participant satisfaction with care
Definite/long term		Quality of life and well-being; health status; functional ability (emotional well-being, daily work, social and physical activities); self-reported health; fatigue; pain; disability; mortality
Impacts	Financial	Overall healthcare costs (direct and/or indirect); project cost savings; detailed financial performance measures; return on investment (ROI), cost-effectiveness; cost-benefit; cost-consequence; etc.
	Socio-economic	Absenteeism; presenteeism; return to work; productivity; (corporate) "image"

SOURCE: Adapted from Pearson et al. (2007); Dalton (2005); Aadalen (1998); and Donabedian (1980).

2.2 Design Options for Evaluating Disease Management Interventions

There are two principal design options available to evaluate disease management interventions or programmes that allow for deriving causal inferences about a programme-effect relationship: experimental and non-experimental (observational) studies (Webb, Bain and Pirozzo, 2005). Experimental studies can take three forms:

- **Clinical trials** assess two or more interventions (one can be "usual care") for individuals with a disease.
- **Field or preventive trials** intervene to reduce individuals' risk of developing disease.
- **Community trials** assess interventions targeted at (a) group(s) of individuals.

Non-experimental or observational studies can be analytical or descriptive. Analytical observational studies involve planned comparisons between people with and without disease or an exposure thought to cause disease; they include cohort studies, case-control studies, case-crossover studies and ecological studies (Table 2). By contrast, descriptive observational studies are generally used to explore patterns of disease and to measure the occurrence of disease or risk factors for disease (exposure) in a population, such as cross-sectional studies, prevalence surveys, routine data collection and case reports.

Table 2 Analytical observational study designs

Design	Description
Cohort study	A longitudinal study in which a group of individuals who share a common characteristic (e.g. being disease-free) within a defined period are selected and categorised based on their exposure to the factor under investigation and followed to determine a change in their status (e.g. disease onset). Cohort studies can be prospective or retrospective (historical cohort studies) and subgroups within the cohort may be compared with each other. The focus of a cohort study is the frequency of an outcome in the exposed compared with non-exposed populations, and one measure of association between exposure and disease is the *relative risk*. Variations of the cohort study include: natural experiments, nested case-control study*, case-cohort study^ and record linkage, or household panel survey.
Case-control study	A study of individuals that compares "cases" selected on the basis of a given disease (or other outcome) to individuals without the disease/outcome under study (controls). Exposure to the variable(s) of interest and other characteristics before the onset of the disease (or other outcome) are recorded through interview and by records or other sources for both cases and controls. The focus of a case-control study is the frequency and amount of exposure in individuals with a specific disease (cases) and people without a disease (controls); the measure of association between exposure and occurrence of disease/outcome in case-control studies is the *odds ratio*.
Case-crossover study	A design developed to study the effects of transient, short-term exposures on the risk of acute events (e.g. myocardial infarction), where the study population consists of individuals who have experienced an episode of the health outcome of interest. Control (comparison) information for each case is based on the individual's past exposure experience and a self-matched analysis is conducted. The design is most suitable for studying associations with the following elements: • The individual exposure varies within short time intervals. • The disease/outcome has abrupt onset and short latency for detection. • The induction period is short.
Ecological/correlation study	The unit of analysis is an aggregate of individuals where information is collected on this group and not on individual members. The design studies the association between a summary measure of disease and a summary measure of exposure. The design and analysis should assess or correct[†] for the ecological fallacy that what holds true for the group also holds true for the individual.

SOURCE: Webb, Bain and Pirozzo (2005).

NOTES: *A cohort can have a nested case-control study whereby people in the cohort who develop a disease of interest are identified as cases and the controls are selected from among those in the cohort who were disease-free at the time of case diagnosis. This design has the advantage of a cohort study but on a much smaller scale (Webb, Bain and Pirozzo, 2005).

^The comparison group is selected from the full underlying cohort defined at the beginning of the study. This design allows comparison with multiple different case groups; however, a drawback is that it requires more sophisticated data analysis than traditional cohort or nested case-control analyses (Webb, Bain and Pirozzo, 2005).

[†] See for example: King (1997) for potential solutions to the problem of ecological inference.

The Grades of Recommendation, Assessment, Development and Evaluation (GRADE) Working Group has developed a system for grading the quality of evidence (Schünemann et al., 2008). The GRADE system defines the quality of a body of evidence as the extent to which an estimate of effect or association reflects the quantity of interest, taking account of within-study risk of bias (methodological quality), directness of evidence, heterogeneity, precision of effect estimates and risk of publication bias. The approach defines four levels of quality (high, medium, low and very low), generally giving the highest quality rating for evidence from RCTs, while evidence from (sound) observational studies will normally be considered as of low quality. However, the rating will depend on the observed effect size, directness of evidence and probability of bias, among others. Thus when observational studies yield a large effect, show a dose-response gradient and take sufficient consideration of potential confounding of an observed finding, they will be rated to be of higher quality (moderate or high quality). Evidence obtained from case series or case reports is generally considered of very low quality.

Experimental research designs, particularly RCTs, are generally considered to be the most rigorous way of determining a cause–effect relationship between a given intervention and an outcome because they usually involve randomisation of individuals to an intervention or control group. This process ensures that the outcome is not affected by systematic differences in factors, known and unknown, between those who receive a given intervention or not (Susser et al., 2006; Sibbald and Roland, 1998). Since there are many ways to randomly allocate individuals to an intervention, it is important to describe the method of randomisation and whether efforts were made to conceal group allocation in the reporting of study results. Yet little attention is given in published studies to reporting these important details (Loveman et al., 2003) and we therefore provide a brief overview of different methods of randomisation in Appendix B. Random allocation, furthermore, does not have to be among individuals; it can occur at a group level where one or more clusters are randomised (Schünemann et al., 2008). As shown in Table 3, there are a number of study designs using group level randomisation and these can be useful design options for disease management evaluation where there are no individual controls available for comparison.

Table 3 Cochrane GRADE list designs for individual- and group-level allocation

	Individual level allocation	Group level allocation
Experimental designs	• Randomised controlled trial (RCT) • Quasi-RCT • Non-RCT • Controlled before and after study	• Cluster RCT • Cluster quasi-RCT • Cluster non-RCT • Controlled interrupted time series • Controlled cohort before and after study
Observational designs	• Prospective cohort study • Retrospective cohort study • Historically controlled trial • Nested case-control study • Case-control study • Cross-sectional study • Before-and-after comparison • Case report/case series	• Interrupted time series • Cohort before and after • Ecological cross-sectional study

SOURCE: Adapted from Schünemann et al. (2008) and Shadish, Cook and Campbell (2002).

NOTE: A cluster is an entity (e.g. an organisation), not necessarily to a group of participants; a group is one or more clusters.

Finally, it is worth noting that while randomisation is considered the gold standard for high quality evidence, the absence of randomisation in an evaluation study does not necessarily produce different results. Rather, it has been shown that estimates of treatment effect are not consistently larger with either a randomised or a non-randomised evaluation method (McKee et al., 1999). It has also been noted that the RCT design may not provide an appropriate standard of proof for intervention-effect relationships, which is discussed further in relation to disease management in the next section (Berwick, 2008; Norman, 2008; MacStravic, 2005; Woolf and Johnson, 2005).

2.3 Approaches Used in Disease Management Evaluations

While experimental research designs are considered to provide the most robust level of evidence, few evaluations of disease management interventions have employed such designs (Mattke, Seid and Ma, 2007; Linden, Adams and Roberts, 2003a; 2003c). Indeed, most published evaluations have followed a simple before-and-after (pre-test/post-test) study design with no control group for comparison; the design is often referred to as the "total population approach" (Tinkelman and Wilson, 2008; Linden, Adams and Roberts, 2003a; 2003c). As a standard methodology in the United States (Outcomes Consolidation Steering Committee, 2005), this method measures the experience of a chronic population in a base year and again in the intervention year, assuming equivalent populations between the base and measurement periods. When cost is an outcome of interest, this method then multiplies the base year cost (per member per month) by a trend to compare the resulting product with the measurement year costs (Farah et al., 2008). Several choices for an appropriate trend adjuster are given in section 3.5 on constructing a comparison strategy.

Reasons for not applying a randomised controlled design to disease management evaluations are pragmatic and conceptual in nature. They include the following (Cretin, Shortell and Keeler, 2004; Linden, Adams and Roberts, 2003b; 2004c):

- **Financial considerations**. RCTs require the commitment of substantial financial resources, which organisations that implement a disease management intervention may be unwilling to commit, in addition to the resources required to set up the intervention in the first place.
- **Design accuracy**. Experimental designs may be difficult to develop because of difficulties in identifying suitable participants in disease management interventions. This poses a particular challenge for settings in which potential participants (patients) are not registered with a given provider as is the case in many insurance-based health systems.
- **Diffusion effect**. Employing an experimental design in population-level disease management interventions requires that the control group is not exposed to the intervention through other sources and this is difficult to ensure in practice.
- **Generalisability**. RCTs are highly selective about the population under study, often excluding population groups typically representing the general population. This reduces the generalisability of findings to population groups, settings and contexts different from the study setting. Another issue related to generalisability is

that an RCT may assume high fidelity to a well-defined intervention, when the intervention may be evolving through "learning by doing", and fidelity across implementation sites may vary; ignoring these facts (not accounting for them explicitly or adequately), as many RCTs do, loses information and threatens the generalisability of findings.

- **Ethical considerations**. The random assignment of participants to either a disease management intervention or no intervention (control) group means that some participants may be excluded from the most appropriate care, which raises ethical issues of fairness and unnecessary harm if the evaluation results demonstrate that the disease management intervention does provide better care (Cretin, Shortell and Keeler, 2004). Thus, there may be an unwillingness to participate in a randomised study that imposes the additional cost of experimental evaluation while possibly withholding the same "value-added" benefit (improvement) from those randomised to a control group. Linden, Adams and Roberts (2003b) also note that there may be legal obligations on the side of the funder that prohibit such exclusion of individuals from receiving the most appropriate care.

From a practical perspective, in contrast to "traditional" RCTs, which randomly allocate individuals to a given intervention, randomisation is difficult in disease management interventions because individuals tend to join (or leave) such interventions voluntarily. Voluntary participation is typically the case for systems in place in the United States, but also in social health insurance systems such as Germany and France (Nolte, Knai and McKee, 2008). Moreover, the nature of the "placebo" in disease management interventions is non-participation. It will therefore be difficult to establish a clear attribution of any observed effect in the intervention group to that intervention alone as an effect may have simply occurred by virtue of giving the individual the opportunity to participate in the first place.

Conceptually, there is an argument that the RCT design does not provide an appropriate standard of proof for the evaluation of disease management interventions because it does not allow for consideration of fidelity and change in the intervention, or learning from this experience (Berwick, 2008; Norman, 2008; MacStravic, 2005; Woolf and Johnson, 2005). Developed for clinical trials of medicines, the RCT design assumes high fidelity to a well-defined intervention that it can control (fidelity being the consistency of implementation with plan). By contrast, disease management interventions tend to vary in practice, as care tends to be "customised" to individual participants and the actual intervention will therefore vary by patient even in the same intervention group. Similarly, providers are also likely to receive an intervention related to their behaviours as part of a disease management initiative and there may be variability within a provider-focused intervention group, further adding to the problems with using the RCT design in this context.

However, to establish whether a given disease management intervention yields a "true" effect, some form of comparison is required between participants who received an intervention and those participants who did not (Norman, 2008; Mattke, Seid and Ma, 2007; Cretin, Shortell and Keeler, 2004). If the two groups of participants being compared are equal in every aspect other than the enrolment in the actual disease management intervention, then any observed differences in health, costs or other measures of interest can be reasonably attributed to the intervention. If randomisation is not possible because of reasons stated above, the next best evaluation approach for disease management interventions is the quasi-experimental, pre-post with a comparison group design, adjusting for known differences (Motheral, 2008). In section 3.6 on robust research analysis we also discuss formal statistical methods to control for selection, such as instrumental variables.

Mattke, Seid and Ma (2007) have provided a list of research designs that are considered particularly appropriate for the evaluation of disease management interventions in healthcare and which are also considered by Campbell (1969) to be particularly appropriate for evaluating specific programmes of social amelioration. These include:

- **Interrupted time-series design**. This is applicable to settings that do not allow for a control group because the experimental intervention targets the entire population.

- **Control series design**. This is applicable to settings where a non-equivalent control group can be created without matching on pre-test scores in pre-test/post-test design (controlled before and after studies).

- **Regression discontinuity design**. This is applicable to settings where randomisation is not possible or acceptable. The design uses a "cut off" score on a pre-test measure to determine assignment of individuals to the intervention or control group.

- **Staged innovation or phased (multi-site) implementation design**. The introduction of a disease management programme would be deliberately spread out, and those units (or sites) selected to be first and last would be randomly assigned so that during the transition period the first recipients could be analysed as experimental units, the last recipients as controls. This design type is akin to the nested case-control study design.

Another novel approach to evaluate a disease management intervention is the **Regression point displacement design**, which is a simple pre-test/post-test quasi-experimental study design where one experimental group is compared with multiple control groups using aggregate-level data (Linden, Trochim and Adams, 2006).

It is worthwhile noting that the evaluation approaches outlined above rest on explicitly quantitative methods. Cretin, Shortell and Keeler (2004) have suggested, however, that because of the complexity and variability of disease management programmes there is a need to use multiple research methods. A number of evaluations of disease management interventions have applied qualitative methods, using for example thematic interviews, logic models, participant and non-participant observation and document review (Weaver et al., 2003; Stevenson et al., 2001; Lamb and Stempel, 1994; Newman, Lamb and Michaels, 1991), sometimes combining these with quantitative methods (Esposito, Fries and Gold, 2009; Schmidt-Posner and Jerrell, 1998). However, there is relatively little research on methodological, analytical or conceptual aspects of the use of qualitative approaches in disease management evaluation. Orkin and Aruffo (2006) have argued that the use of ethnography as observation in context has so far been overlooked as a potentially important approach in the design of quality improvement interventions such as disease management initiatives. Such approaches would for example allow better understanding of the functional requirements of disease management educational materials.

Recently, there has been a move towards emphasising "realistic evaluation", a term coined by Pawson and Tilley (1997), where pluralistic quasi-experimental methods are used for evaluating complex interventions with high contextual influence such as a disease management programme (Berwick, 2008; Norman, 2008). As a methodology commonly used by social programme evaluators, realistic evaluation approaches involve understanding what works for whom under what circumstances, and place equal emphasis on external validity, generalisability and cumulative learning.

2.4 Principal Challenges to Evaluating Disease Management Interventions

As noted earlier, a common approach to evaluating disease management initiatives is the total population approach, which is a simple before-and-after (pre-test/post-test) design with no control group. This design is however susceptible to the methodological problem of many potential sources of bias and confounding factors that may influence outcomes (Linden, Adams and Roberts, 2003a). Sources include the participants enrolled in a disease management programme, the programme itself, the providers of the intervention, the availability, accuracy and use of the data used to measure the intervention's effect, long-term secular trends and so on. Confounding poses a threat to the validity of evaluation findings: it gives rise to situations in which the effects of two processes are not separated, or the contribution of causal factors cannot be separated, or the measure of the effect of exposure or risk is distorted because of its association with other factors influencing the outcome of the evaluation.

Table 4 and Table 5 summarise key confounding factors that have been identified as posing a threat to the validity of a disease management programme evaluation (Susser et al., 2006; Linden, Adams and Roberts, 2004d; Linden, Adams and Roberts, 2003a; 2003c; Velasco-Garrido, Busse and Hisashige, 2003; Campbell, 1969). Several of these confounding factors which are given explicit attention in the literature are described in further detail below. We note here that the risks given in Table 4 are risks for some but not all non-RCT designs: pre-post without controls are very sensitive to history or maturation effects, while treatment-control designs may have these factors better controlled. The risk of confounders differs by type of design but detail on this goes beyond the scope of this section and can be found in the wider research methods literature.

Table 4 Inventory of threats to internal validity for different evaluation research designs

Confounding factor(s)	Description related to disease management context	Applicability	
		RCT	Non-RCT
History (general)	External factors or events, occurring between pre-test and post-test, that may mask the impact of the disease management programme intervention thus providing alternate explanations of effects, e.g. secular trends.		✓
Maturation (age effects), natural disease progression	Processes intrinsic to each participant, i.e. biological development (e.g. ageing) and accumulation of exposure over time change disease risk/severity, etc. (e.g. participants get sicker with progressive disease).		✓
Instability (measurement reliability and validity)	Unreliability of measures, fluctuations in sampling persons or components, autonomous instability of repeated or "equivalent" measures. Can the same results be produced repeatedly? Does the outcome measure make sense? (This is the only threat to which statistical tests of significance are relevant.)		✓
Testing	The effect of taking a test on the scores of a second testing. The effect of publication of a health/social indicator on subsequent readings of that indicator.	✓	✓
Instrumentation	In which changes in the calibration of a measuring instrument (e.g. patient survey) or changes in the observers or scores (e.g. unit cost increases) used may produce changes in the obtained measurements.	✓	✓
	Costs will appear higher if unadjusted for changes in pricing of services. Also, reimbursement method and coding changes may alter unit cost with unknown effect on results.		
Regression artefacts	Pseudo-shifts occurring when persons or intervention units have been selected on the basis of their extreme scores.		✓
Selection	Biases resulting from differential recruitment of comparison groups, producing different mean levels of the measure of effects.		✓
Unequal attrition	A different proportion of treated and controls are lost to observation (through disenrolment, loss to follow-up or death) and the magnitude of this difference is unequal for the two groups. This occurs when the intervention and the outcome have a synergistic effect on attrition.	✓	✓
Selection–maturation interaction	Selection biases resulting in differential rates of "maturation" or autonomous change.		✓

SOURCE: Adapted from Susser et al. (2006); Linden, Adams and Roberts (2003a); and Campbell (1969).

Table 5 Inventory of threats to external validity for different evaluation research designs used in disease management (DM)

Confounding factor(s)	Description related to disease management context	Applicability	
		RCT	Non-RCT
Interaction effects of testing	The effect of a pre-test in increasing or decreasing a participant's sensitivity or responsiveness to the experimental variable, thus making the results obtained for a pre-tested population unrepresentative of the effects of the same variable for the un-pretested universe from which the participants were selected.	✓	✓
Interaction of selection and experimental intervention	Differences among individuals may lead to unrepresentative responsiveness of the "treated" population, i.e. participants in a disease management programme are, on an individual level, not typical of the population from which they are taken and thus disease management programme participants and persons in the general population may achieve different results with a similar intervention simply because the characteristics of the two groups are fundamentally different.	✓	✓
Health disparities/unequal access to services (interaction of causal relationship with settings)	Baseline disparities specific to the intervention site could ultimately affect the ability to extend inferences about outcomes to different settings or countries. Gaps in quality of or access to healthcare exist within countries and between countries, with people experiencing health disparity being more likely to have poor disease control, higher costs and a greater burden of disease or severity, thus making them inherently different from either the general population or other people with chronic conditions. This presents a problem for evaluation because underlying differences in the use of health services can have an unknown effect on possible evaluation results of a disease management intervention (Linden, Adams and Roberts, 2003a).	✓	✓
Reactive effects of experimental arrangements	Provider practice patterns may change as a result of being observed ("Hawthorne effects", which overestimates positive results), or because of risk and reimbursement models of the disease management programme.	✓	✓
Multiple-intervention interference	Where multiple interventions are jointly applied, effects may be atypical of the separate application of each intervention (synergy).	✓	✓
Responsiveness of measures	All measures are complex, and all include irrelevant components that may produce apparent effects, e.g. costs for the same DM intervention will differ within and across programmes and costs savings can be independent of programme impact.	✓	✓
Replicability of intervention	Disease management programmes are complex and interventions vary, and replications of them may fail to include those components actually responsible for the observed effects, or may not apply to all diseases or participants equally.	✓	✓
Context-dependent mediation	Mediating factors can change a disease management outcome independent of programme design (e.g. clinician referral to emergency department can nullify the intervention effect on reducing the rate of such visits), which is a threat to internal and external validity.	✓	✓

SOURCE: Adapted from Linden, Adams and Roberts(2004d) and Campbell (1969).

2.5 Threats to Valid Inference About Disease Management Effects

Attrition bias. Attrition concerns the loss to observation of members from a disease management treatment group or a control group. Loss to observation can occur because of participant disenrolment, loss to follow-up or death. In the context of disease management, many studies show fairly high levels of attrition between initial recruitment and reporting of results (Loveman et al., 2003). Attrition is of methodological concern as it can produce misleading results when it is unequal between groups. For example, if the most motivated participants in a disease management intervention remain while those who are less motivated drop out, then the estimate of effectiveness for an unselected group of chronically ill individuals would be overestimated. Similarly, if attrition is greater in the control group than the intervention or treatment group, the estimate of intervention effectiveness is reduced because the least motivated toward self-management and most ill are the most likely to leave the study. Differential attrition is a methodological challenge in disease management evaluation as statistical adjustment to test for differences in baseline characteristics will not adjust for effects such as motivational differences that are not captured in baseline evaluations (Loveman et al., 2003). Apart from inferential challenges, attrition may signal some substantive limitations or weaknesses of the intervention itself; high attrition rates may raise concern about appropriateness of a given disease management initiative and whether it is likely to attract large (enough) numbers of individuals with chronic conditions.

Case-mix. Case-mix is the heterogeneous mixture of characteristics among individuals enrolled in a disease management programme or intervention, or between enrolees and non-participants. It is important to account for variation, for example, in disease severity and age among individuals within and between a given intervention because there may be turnover in enrolment when systems allow individuals to change provider (e.g. general practitioner) or payer (e.g. health insurance) of a given intervention. Such change will alter the baseline characteristics of the reference population, which is often used to calculate baseline utilisation trends of chronically ill participants, and hence influence evaluation findings (Linden and Goldberg, 2007; Linden, Adams and Roberts, 2003a).

Measurement error. Measurement error is the misclassification (for dichotomous variables) of a person's exposure to or outcome from a given intervention because there was error in the measurement of either status (Susser et al., 2006). Differential misclassification can either strengthen or weaken the apparent association between an intervention and a measured outcome because misclassification of one status depends on the other. Random measurement error is a concern because it undermines the ability to obtain the same answer across different tests, testers and time periods (reliability). By contrast, systematic errors do not necessarily undermine the reliability of a measure but do limit its validity. For example, the identification of suitable disease management participants is often based on diagnostic measures retrieved from administrative databases that tend to be inaccurate, and this can lead to spurious evaluation results (Linden, Trochim and Adams, 2006; Linden and Roberts, 2005). Furthermore, data systems may change over time thereby threatening the validity of repeat measurement of observations (Linden, Trochim and Adams, 2006).

Regression to the mean. Regression to the mean is a statistical phenomenon that occurs when, for example, a given analysis compares the apparent reduction in the value of a repeated measurement after the initial measurement was taken from extreme non-random values. The more extreme the values used in the first measurement of a group's average, the more likely the average of subsequent measurements will get closer to the mean of the whole population (Linden, Adams and Roberts, 2003a; Bland and Altman, 1994). This phenomenon is particularly relevant in disease management interventions (or their equivalent) that aim to control disease by reducing high levels of a given variable (e.g. blood sugar), or that aim to reduce utilisation (e.g. emergency care visits) and where patients have been selected for the intervention based on their extreme value of such effect measures. For example, in a given disease management initiative where high cost patients are selected for the intervention, high cost participants in the first year will cost less in the second year, giving better results than baseline when re-measured, whereas low cost participants in the first year will cost more in the second year, showing worse results than the baseline (Linden, Adams and Roberts, 2003a). There are several other examples of regression to the mean occurring in evaluations of disease management interventions which are important to consider (Bland and Altman, 1994).

Factors that influence regression to the mean can be participant-related, provider-related and/or methodology-related (Tinkelman and Wilson, 2008; Klosterhalfen and Enck, 2006; Enck and Klosterhalfen, 2005). Methodological factors include: (1) size of sample and study – regression to the mean is more likely to occur in small clinical studies that involve fewer participants; (2) study design and duration; and (3) selection of endpoints. Participant-related factors include memory bias and disease-related factors (e.g. length of an average symptom spell or the nature of the disease course). Evidence suggests that regression to the mean is more likely to occur for disease states that have a cyclical course such as depression and irritable bowel syndrome when compared with more chronic and progressive disease. When participants are selected on an extreme value such as high utilisation rates where the chronic condition is progressive, it is unlikely that there will be significant reductions in the future in the absence of effective intervention. By contrast, a cyclical condition might incur high utilisation rates as a result of random effects of, for example, "a bad year" such that future utilisation would likely regress toward more typical mean values with or without the intervention. However, regression to the mean may not be as big a problem as usually seen. A recent study to test this issue found minimal evidence of regression to the mean over two consecutive years in Colorado Medicaid patients with moderate to severe chronic obstructive pulmonary disease (Tinkelman and Wilson, 2008).

Seasonality, or cyclical trends. Seasonality is a characteristic of a time series in which data exhibit regular and predictable changes (cyclical variation), such as peaks in hospital admissions during the winter season with declines during the summer season. If left unadjusted for in research design and analysis, seasonal trends can influence the measured outcomes of a given intervention. For example, if the outcome of interest is hospital utilisation and the end of the intervention coincides with the winter flu season, hospital admissions will likely be higher thereby leading to the "true" effect of the intervention being underestimated when compared with the baseline.

Secular trend. Secular trend or secular drift is the long-term non-periodic variation of a given time series produced by general developments or innovations that may be outside the healthcare sector, such as new technology(ies) or drug(s), new care protocols, human growth and longevity, and so on. Secular trends can confound observed effects of a given intervention if for example the rates of the particular illness are falling in the community as a whole. Controlling for secular trends requires the use of a comparison group to isolate the effect of the intervention on the outcome of interest from such external effects.

Sensitivity and specificity of existing data. Sensitivity of data is the precision of disease identification data (e.g. algorithms or medical insurance claims) used to select suitable participants for a disease management intervention. The specificity of data is the extent to which data sources are available or complete. When the sensitivity of data is poor and disease identification data are imprecise, the selection of suitable participants from a population eligible for a given intervention may miss people relevant to an evaluation. The effect of poor data sensitivity on evaluation results is unknown.

Selection bias. Selection bias is systematic error in the method of choosing individuals to take part in an evaluation and in assigning them to the disease management group. It can occur when those who consent to take part differ from those who do not. Motivation to change is key and thus volunteer enrollees who self-select to participate may be more likely to take a more active role in managing their own care. Volunteer-based interventions are likely to overestimate the effect of the given intervention. Similarly, the alternative "engagement" model for enrolment where those who do not wish to participate must actively opt out will include the group of unmotivated individuals otherwise missed by a volunteer-based model of enrolment. Here, bias occurs when group assignment is non-random and the intervention group disproportionately contains more unmotivated individuals who are different from controls on this characteristic. In this case, the results of a given intervention will appear worse than baseline measurements (Linden, Adams and Roberts, 2003a). Beeuwkes Buntin and colleagues (2009) recently demonstrated, in an observational study of members of a large health insurer, that those who enrolled into the programme differed systematically from those who did not on demographic, cost, utilisation and quality parameters before enrolment. Finally, enrolling only individuals at high risk or with high disease severity will create an intervention group that is very dissimilar from the general population (Linden, Adams and Roberts, 2005b), which in turn can bias the effect of an intervention when illness severity is unlikely to be "intervenable" (Tinkelman and Wilson, 2008).

Therapeutic specificity. Therapeutic specificity is the relation between the intervention and the outcome, when the observed effects are demonstrably attributable to the entire content of the intervention (MacStravic, 2008). The key issue here is that the evaluation takes into account the dynamics of human interactions because in practice those providing the intervention are likely to go beyond the defined intervention scope by addressing more than the signs and symptoms specific to one disease when they interact with disease management participants. Yet therapeutic specificity in evaluations has tended to be interpreted in a more limited manner, referring only to the specificity of disease management to a disease or its severity, but this misses the wider interactive contents of a given intervention (MacStravic, 2008).

CHAPTER 3 Implementing Evaluation of Disease Management Interventions

Fundamental to all evaluations of programme effectiveness is a clear definition of what constitutes effectiveness (or "success"). This is determined by the overarching aim(s) of the programme being evaluated and by the hypothesised mechanism(s) of expected effect(s). For example, behaviour changes induced by a given intervention should be shown to be connected to specific health metrics that are known to be risk factors for the condition(s) targeted for management (e.g. high blood pressure as a risk factor for stroke). This demonstration is important for preserving therapeutic specificity of the given intervention and establishing a causal relationship between a disease management intervention and measured effects. It is therefore suggested that an evaluation should include a tracking mechanism so there is certainty that the intervention produces specific changes in participants' behaviour that were intended and expected (MacStravic, 2008).

In the absence of standardised evaluations for disease management interventions, it has been recommended that an evaluation should as a minimum clearly identify and define the following: (1) the actual intervention, (2) the intervention population, (3) the reference population, (4) outcomes metrics, and (5) confounding variables (Wilson et al. 2008). Further, it is important to consider the potential effect if the target population had not been exposed to or received an intervention (the counterfactual) (Susser et al., 2006). Linden, Adams and Roberts (2004d) further recommend that the external validity of an evaluation's results is best optimised when the evaluator maps as many domains as possible before the intervention is implemented. Additional dimensions could include the disease state targeted by the intervention; the method of participant selection (e.g. whether it is opt-out enrolment or volunteer-based); regional and cultural domains (e.g. medical practice patterns, patient demographics, belief systems and perceived need); the behavioural change model of the intervention; intervention site or setting (e.g. hospital, workplace or community); and the process of intervention implementation and its consistency with plan (e.g. fidelity).

We here present key methodological, analytical and conceptual issues related to evaluation methods and metrics in disease management evaluation. Literature is limited on how to address the myriad of problems across each of these three issues together; however, a reliable health technology assessment disease management instrument with 15 items has been developed to address several of the methodological problems described below (Steuten et al., 2004). The methodological, analytical and conceptual issues discussed in this chapter include: scope, content, dose and wider context of a given interventions;

characteristics of the intervention population in relation to the reference population; length of evaluation period and evaluation planning; construction of a comparison strategy; approaches to robust analysis; evaluation measures; and validation of evaluation results.

3.1 Characterising Disease Management Interventions

One of the key challenges to developing standardised approaches to evaluation of disease management initiatives is the wide variation of interventions broadly considered under the heading "disease management" (Lemmens et al., 2008; Mattke, Seid and Ma, 2007; Pearson et al. 2007; Steuten et al., 2006; MacStravic, 2005; Weingarten et al., 2002).

While heterogeneity in itself may not be a problem, the challenge is to ascertain sufficient level of detail about the intervention that would allow for: (1) sound evaluation and understanding of causal effects, particularly the identification of subcomponents that influence participant outcomes; (2) replication of the evaluation study; and (3) implementation of the intervention in other settings and countries.

Further, heterogeneity relates not only to variation in the range of actual interventions used, but also to variation in the target population. In addition, individual components of a given intervention may change over time and across settings. Some authors have therefore proposed a set of key characteristics of disease management interventions that a given evaluation should consider (Wilson et al., 2008; Mattke, Seid and Ma 2007; Linden and Roberts, 2004). These characteristics concern the scope, content, dose and wider context of such interventions, which we will briefly discuss in turn.

3.1.1 Intervention Scope – Disease Type and Severity

One key aspect of a given disease management intervention relates to the actual target group – whether the intervention is designed to address an entire patient group with a given (chronic) condition, or concentrates on a subgroup (Mattke, Seid and Ma, 2007). Subgroups might be those with serious illness, those with illness that has not responded to unmanaged medical treatment, those with rare health problems and/or those with costly health problems. The target population must be clearly identified in order to allow for identification of an appropriate comparison strategy and selection of a control group to be drawn from the same underlying (reference) population as the intervention group. Conceptual clarity on the target and reference populations is necessary to determine whether the comparator group represents the counterfactual.

Second, the actual condition(s) or disease(s) that are targeted by the given intervention must be clearly defined so as to identify the scope of the intervention being evaluated, along with the size of population, clinical course and baseline utilisation patterns (Mattke, Seid and Ma, 2007). So far, evaluations of disease management initiatives have primarily focused on "traditional" chronic conditions including coronary artery disease (CAD), congestive heart failure (CHF), diabetes mellitus, asthma, chronic obstructive pulmonary disease and depression, but much less frequently on conditions such as cancer, dementia or rare diseases (e.g. haemophilia) (Mattke, Seid and Ma, 2007). This observation is however not surprising given that evaluations can only address existing interventions. Nonetheless, consideration of the disease type or severity as a determinant of the intervention's scope will be critical to designing the measures for subsequent analysis.

3.1.2 **Intervention Content and Dose**

It is also important to define the actual total content of a disease management intervention in order to allow for adjustment of possible variation of its individual components. For example, disease management programmes tend to customise individual interventions to individual participants so introducing an element of variation across the intervention group (MacStravic, 2005). Interventions may also vary from session to session, and between providers (e.g. physician or nurse). It will therefore be important to characterise elements of intervention content, including (Mattke, Seid and Ma, 2007; Pearson et al. 2007; MacStravic, 2005):

- staffing (and training)
- information support provided
- protocols used
- services provided (nature and scope)
- therapeutic specificity
- level of individual customisation.

In addition, interventions can be characterised not only by what they deliver (content) but also by how much they deliver (dose), although relatively little is known about how to measure the actual dose delivered to assess quality and manage outcomes (Huber et al., 2003). A first step is the identification of the components of dosage: intensity and frequency. The intensity of disease management can stretch from "low" (e.g. postal instructions for programme participants) and "medium" (e.g. telephone follow-up) to "high" (e.g. home visits) (Mattke, Seid and Ma, 2007). Frequency is the rate at which a given intervention is delivered (e.g. biweekly, monthly or quarterly). Both frequency and intensity of a particular disease management initiative, or its components, can also be graded according to levels of severity, with the highest level corresponding to those participants who are frequently the sickest and in need of most support (Villagra and Ahmed, 2004). In addition to frequency and intensity, other basic elements of the amount of intervention delivered for consideration include duration and breadth (Huber, Hall and Vaughn, 2001).

Thus, there are several aspects of the content and dosage of a given intervention that need to be clarified conceptually and taken into account methodologically and analytically.

3.1.3 **The Wider Context**

Interventions targeting people with chronic disease further vary by attributes related to the wider context in which they are delivered and which require careful consideration when designing an evaluation of a given disease management initiative (Mattke, Seid and Ma, 2007; Pearson et al., 2007):

- setting of the disease management intervention
    - provider: primary care centre, general practice, hospital, community
    - funder: health insurance (e.g. sickness fund or private insurance), government, health authority or out-of-pocket payment

- level of system integration between primary care practice and the disease management programme or between levels of care
- use of patient or provider incentives.

Characteristics of the wider context are also important for considering the extent to which evaluation results about “successful” disease management interventions delivered in one country can be transferred to another. In a systematic review of patient education models for diabetes, Loveman and colleagues (2003) noted how not only the healthcare context (public or private provision) but also the beliefs and attitudes of patients, reflecting on general aspects of culture and traditions, need to be considered for impact on measured outcomes and generalisability of evaluation findings.

3.2 Characterising the Disease Management Intervention Population

Disease management evaluations need to clearly define the planned target population for a given intervention as this process is linked to subsequent data analysis and the potential for causal inferences about the intervention-outcome relationship under evaluation. Principally, the characterisation of the disease management intervention population involves identifying: (1) inclusion and exclusion criteria for selecting intervention participants; (2) strategies to enrol participants in the intervention; and (3) the length of participation in the intervention. For participant selection, different approaches have been suggested:

- Considering all patients eligible for a disease management programme as the treatment group to be evaluated (the target disease management population), regardless of whether or not they actually enrol in the intervention. The analysis would then use the non-eligible patients in an insurance plan or health system as the comparison group for the evaluation (DMAA, 2007).

- Using predictive modelling to identify the “most suitable” patients from a given population for prospective enrolment in a disease management intervention, which can be done using different analytic methods such as (1) a two-by-two table with a threshold; (2) Receiver Operating Characteristic (ROC) analysis; or, (3) R2 analysis (Linden, 2004; Cousins, Shickle and Bander, 2002).

- Basing the selection of intervention participants on the mean of several measurements of the variable of interest to reduce the effect of regression to the mean (Tinkelman and Wilson, 2008).

- Selecting intervention participants based on one measurement, followed by a second measure to establish baseline values for the evaluation study (Tinkelman and Wilson, 2008).

- Setting eligibility criteria to exclude extreme values (e.g. high-cost conditions) or limiting inclusion to a certain threshold figure for a measure since disease management interventions more typically include high cost or high severity patients, which can skew evaluations. With this approach, however, it would be important to conduct analyses both with and without these extreme values in order to assess any selection bias from such exclusion at selection (Serxner et al., 2008).

- Randomly sampling all eligible patients in a target population, coupled with further randomisation to either an intervention or a control group because this will evenly distribute explained and unexplained sources of variation between participants and non-participants (Linden, Adams and Roberts, 2004d). While this approach to selecting the intervention population is most rigorous for ensuring the generalisability of intervention effects, it assumes that the evaluation is designed concurrently (prospectively) with the intervention.

Once an evaluation identifies the method used to select the intervention population, the type of enrolment strategy used by the programme needs to be considered. Different types of enrolment strategies, for example volunteer-based or "opt out" (presumptive strategy; engagement model in the United States), will have different implications for the type of analysis and evaluation measures chosen to evaluate the intervention effects. For example, the opt-out strategy will determine a higher likelihood for patients to be enrolled in a disease management intervention than the volunteer-based strategy, and this difference in likelihood needs to be considered if the propensity scoring technique is used to create a control group for comparison of intervention effects (Linden and Roberts, 2004). It will be equally important to examine whether there are any unusual patterns in (dis)enrolment of intervention participants and their potential impact on external validity (Linden, Adams and Roberts, 2004d). This can be addressed through, for example, regular review of cross-sectional, stratified measurements during implementation of the intervention.

Understanding the enrolment strategy is important not only in relation to the potential bias in evaluation results introduced by different strategies but also in order to assess to what extent the enrolled population is indeed appropriate in relation to the group targeted by the intervention. Evidence suggests that the "success" of interventions targeting patients with chronic disease requires the active involvement of the individual patient (Nolte and McKee, 2008). The design of such interventions should therefore be informed by a clear theory of behaviour change as it requires behavioural changes in both providers and participants (Linden and Roberts, 2004). If however the population targeted by the intervention in order to assess intended behaviour change is different from the one actually enrolled, it will be difficult to draw causal inferences between the intervention and the effects of the intervention. The evaluation therefore needs to ascertain the theory underpinning the disease management intervention, or its component parts, to enable an understanding of intervention characteristics that are likely to be effective, and for which groups of individuals, and to select appropriate evaluation measures.

Participants' period of enrolment in a given intervention or programme ("tenure") needs to be identified. Tenure is a factor that is typically not controlled for statistically in disease management evaluations (Linden, Adams and Roberts, 2004a). The benefits of a disease management intervention tend to take more than 12 months to manifest themselves, yet this is the typical time period of published evaluations, which undermines the ability of an evaluation to determine programme impact. Tenure can also make evaluation results problematic in certain settings where participants in a given programme only stay for a shorter period of time than 12 months (MacStravic, 2008). Thus, it is important to adjust for participant tenure through: (a) assessing the outcome of interest (numerator) against the sum total number of months that all eligible individuals were enrolled in the programme ("member-months") (denominator); or (b) considering only those participants who were continuously enrolled long enough to be influenced by the programme (Linden, Adams and Roberts, 2004a).

Other alternatives to adjust for tenure include either using regression analysis that also adjusts for seasonality (Linden, Adams and Roberts, 2004a) or grouping participants based on the length of participation in a set of disease management programmes regardless of calendar year or year of joining (Healthways, 2007). However, this type of approach means that evaluation participants are clustered from multiple and geographically diverse programmes, which may in turn undermine validity of the findings. Finally, Fetterolf, Wennberg and Devries (2004) suggest separating new and existing cases among diseased individuals in the baseline and intervention populations to overcome the problem of differential tenure.

3.3 Length of Evaluation Period

The length of evaluations of disease management interventions is typically 12 months (DMAA, 2007; Healthways, 2006; 2007), yet it has been argued that this timeframe is likely to be too short to capture any long-term health benefits or cost savings accurately (Mattke, Seid and Ma, 2007; Velasco-Garrido, Busse and Hisashige, 2003). For example, it will take some time for a given intervention to be fully implemented and for any individual level effect to become evident (Linden, Adams and Roberts, 2003c), so multiyear evaluation timeframes may be required (Plocher, 2006). Serxner, Baker and Gold (2006) noted that it takes at least three to five years for health management initiatives to identify "true" programme effectiveness because of time lags in reaching full implementation. At the same time it is difficult to define an "optimal" length of a given intervention to allow for sustainable impact (Loveman et al., 2003). A recent evaluation of disease management programmes for diabetes and cardiac conditions in the United States found a three-year period to be a sufficient evaluation timeframe to identify a positive impact on clinical outcomes (Healthways, 2006).

An evaluation period of more than 12 months will be important for assessing medium- and long-term effects such as economic impact, since disease management may actually increase costs in the short term for some chronic diseases as a result of the investment required to implement such an intervention in the first place (Plocher, 2006; Serxner, Baker and Gold, 2006; Linden, Adams and Roberts, 2003c). Also, as disease progression can involve the development of serious long-term complications, it is critical to demonstrate that disease management initiatives can have lasting effects, which can only be evidenced by a longer period of evaluation (Loveman et al., 2003). Finally, sufficient length of evaluation follow-up is needed to account for any potential intermittent plateauing of intervention effects because the evaluation must be able to distinguish these temporal influences from the effects of processes that bring about change through the intervention itself (Linden, Adams and Roberts, 2003b), in order to ensure accurate and valid findings, and to inform strategies to assure sustainability of measured results (Mattke, Seid and Ma, 2007).

3.4 Evaluation Planning

It is perhaps a truism that evaluation studies must be well planned to increase statistical power of future results. Yet some reviews of the evidence have noted the absence of prior power calculations to determine an appropriate study size for assessing disease management effectiveness, because power calculations were either not performed or not reported (Loveman et al., 2003). Thus the disease management evaluation literature includes a focus on the set of inter-related parameters that need to be determined before an evaluation is implemented if the evaluation aims to establish causal inference and statistically robust findings (Farah et al., 2008; Linden, 2008). These are: (1) sample size, (2) effect size, and (3) level of statistical significance (Appendix C). There has been a particular focus on the challenge of small sample sizes, as interventions commonly centre on disease categories with small numbers of affected individuals, or observed differences between the intervention group and control group are small.

Analytical strategies to improve the statistical power of small sample sizes include increasing the baseline rate through for example extending the measurement period from 12 months to 18 months to calculate the sample size (Linden, 2008). This approach to enhance the power to detect a statistically significant change obviously depends on the actual rate not being lower than the baseline rate assumed for the power calculations (Berwick, 2008). An alternative approach is to reduce the level of significance required (Farah et al., 2008) or to focus on achievement of absolute numbers (e.g. number of individuals with diabetes who receive a blood sugar measurement) so as to address variation in expected results (Fetterolf and Tucker, 2008). Not without limitations, another suggestion to deal with small sample size is to use simple binary tests at the participant level (e.g. "did this diabetes patient have a glycohaemoglobin measured within the past six months?") to develop and trend a percentage "pass rate" so that statistically different results from the "average" result can be detected (Fetterolf and Tucker, 2008).

Small numbers, however, are not the only challenge to statistical significance in disease management evaluation. The wide variation in cost or resource consumption for most of the diseases selected for disease management interventions means that measured outcomes are not normally distributed and therefore evaluation data must be analysed robustly with methods appropriate to non-normal distributions (Fetterolf and Tucker, 2008). It is equally important that evaluation planning takes into account variability in the dose of a disease management intervention, particularly the behaviourally based activities, as too little or too much may not produce the desired effect on outcomes being evaluated (Huber et al., 2003). It may be that greater effect sizes are associated with dosage of a given intervention or its components, but if this variability is not accounted for at planning stage then results may be subject to different types of error (false negative or false positive). This will also help ensure data on the implementation and processes underlying a given intervention are not absent from the evaluation as this can hamper the interpretation of findings. Finally, variability of chronically ill cases in hospitalisation rates (case-mix) is another consideration for evaluation planning when sample size is determined using hospital admissions (Linden and Goldberg, 2007).

3.5 Constructing a Comparison Strategy

A comparison strategy is essential for any evaluation that aims to assess whether or not the intervention under study did indeed have an effect on the intervention group that would not have occurred otherwise. This involves selection of a suitable control group that most closely mirrors what would have happened without the intervention (the counterfactual).

The basic premise for establishing the value of a disease management intervention through causal inference about its effects is that the intervention group and the control group are equivalent (and the metrics used for both are comparable) (Wilson and MacDowell, 2003). That is, the groups must be equivalent in the probability of selection for treatment across those variables and characteristics that might influence outcomes or effectiveness. This means that selection and assignment of individuals to either group must be taken from the same reference population in order to ensure that the individuals share all the same characteristics except for exposure to the intervention.

Frequently, there is an implicit assumption that the control group is formed by those individuals who are receiving standard or usual care. However, what comprises "standard" or "usual" care and how it differs from the actual intervention (in scope, content, dose and context) often remains unclear. As a consequence, assessment of what subcomponents within the intervention may be effective and how the intervention differed from the comparator can be obscured, thereby compromising generalisability of the evaluation results (Loveman et al., 2003). Replicability of findings may be affected, moreover, because what constitutes usual care will differ within and across healthcare systems (Mattke, Seid and Ma, 2007).

A key remaining challenge is the identification of a suitable comparator in practice (MacStravic, 2005). The difficulty is that the nature of the disease management intervention and/or the participant population are likely to change over time, as does disease progression and the likely development of co-morbidity or multi-morbidity (Plocher, 2006; Linden, Adams and Roberts, 2003a). Also, where disease management interventions are implemented across entire populations, identification of a suitable comparison group that is not affected by the intervention, directly or indirectly, will be increasingly difficult.

Approaches to select a control group for a robust yet realistic comparison strategy include some of the following (Wilson et al., 2008):

- randomised controls
- pre-intervention period
- historical reference population
- concurrent reference population
- expert opinion
- statistical control groups.

The Disease Management Association of America (DMAA) has offered a list of design options (Outcomes Consolidation Steering Committee, 2005) that include a comparison strategy, which we summarise in Table 6.

Table 6 Design options that include a comparison strategy

Comparison design	Description
Randomised controlled trial	An experimental design involving the random allocation of participants to either an intervention group or a control group, which is the method for balancing confounding factors between groups.
Pre-post historical control design	This is the total population approach where a comparison is made between measurements taken after a chronic disease management intervention with the measures taken for the participants before the intervention; thus participants act as their own control.
Pre-post historical control design with concurrent control group	This design is similar to the pre-post with the participant as their own "historical control" for comparison, but it also includes a comparison with data from a concurrent control group.
Pre-post historical control with actuarial adjustment	This design can involve a trend-adjusted (per-member per-year, or pmpy) historical control (pre-post) population methodology used to estimate cost savings or ROI. A cost trend is taken from another population and multiplied by the baseline cost pmpy of the pre-test participants, which is then compared with the actual cost pmpy post-test. The change (ideally reduction) is then multiplied by actual member years in measurement period. Trend is an annualised rate of change and actuarial adjustments can also be benefit or risk related. (The design, like other variations of the total population approach, does not capture temporal effects of maturation in diseased population.)
Pre-post historical control with no actuarial adjustment	The same as the total population approach of using the participants as their own controls.
Pre-post multiple baseline (with randomised controls)	The design functions as an experiment in that it entails a priori specification of both baseline durations in a data series and random assignment of participants to baseline durations. An intervention occurs for one data series while the baseline continues for the remaining data series and comparisons within and across data series allow for an assessment that a marked change occurred post-intervention while the data pattern for the remaining data series was stable.
Benchmark study	This study involves evaluating various aspects of a chronic care management intervention, or its overall effect, compared with the best practice within the sector.
Pre-post time series	This design involves a comparison of participant data post-intervention with participant data pre-intervention.
Post-only control group	This design makes a comparison between a control group and an intervention group only after the intervention has taken place and thus cannot attribute a difference found to the intervention since pre-intervention data for the control group is not included in the comparison.

SOURCE: Outcomes Consolidation Steering Committee (2005).

An example of a pre-post design is provided by Healthways (2006) where intervention participants are used as their own "control". The 12-month baseline period before the intervention is extended another 12 months earlier in time to create a "pre-baseline" period, and these periods are compared on the outcome of interest (e.g. blood lipids) for the participating population. A limitation of this design is regression to the mean, as described earlier, when inclusion in the intervention was influenced explicitly or implicitly by high values of the outcome variables. Other issues in this design are of history and maturation – the post-intervention figures are expected to change because of the progression of the disease. One way of overcoming this problem is by using a control group that is matched on the intervention group on a number of key characteristics except for the intervention (Tinkelman and Wilson, 2004).

An alternative strategy is the historical control design (Table 6), which "controls" for regression to the mean as the phenomenon is present in both participant and non-participant groups and acts in the same direction, but the actual effect size cannot be known (Tinkelman and Wilson, 2008; Fetterolf, Wennberg and Devries, 2004; Yudkin and Stratton, 1996). However, the results of a historical control design can still be biased by secular trends such as changes in healthcare infrastructure (Fetterolf, Wennberg and Devries, 2004).

The controlled series design (controlled before and after) is useful because it can also reduce many, but not all, threats to internal validity of evaluation findings. For example, the design cannot control for selection biases resulting in differential rates of "maturation" or autonomous change (confounding due to selection–maturation interactions).

Other options recommended by the DMAA are post-only, cross-sectional and time-series (Table 6). These do not include a non-participant comparator, however, so establishing a (causal) intervention–outcome relationship will not be easily possible. Indeed, Wilson and MacDowell (2003) advise against the use of "post-only" designs. Similarly, cross-sectional designs are less suitable for the assessment of quality improvement strategies such as disease management interventions than a longitudinal design because of the difficulty of establishing causality (Weiner and Long, 2004).

Mattke, Seid and Ma (2007) identify three types of reference populations that have been used in practice to construct a comparison strategy:

- programme participants versus non-participants, controlled for socio-demographic and healthcare system relevant characteristics to establish population "equivalence"
- participants with condition versus individuals without the condition, matched by age and sex
- natural experiment created by phased implementation, plus pre-post comparison, adjusted for risk and demographic differences.

Villagra and Ahmed (2004) present an example of a natural experiment design that allows for more than one set of comparisons by first using a sequential pre-post comparison in which each intervention site served as its own historical control, and, secondly a parallel group comparison by matching five programme sites with a concurrent site in which the intervention was not yet in place.

3.5.1 Creating Controls

There is a range of techniques to introduce randomisation to an evaluation design, such as by randomly matching a control group to the intervention group of the same reference population. Randomly matched control groups can be created using existing data sources. For example, predictive modelling or propensity scoring are techniques to create controls using readily available administrative data (Linden, Adams and Roberts, 2005c; 2003c).

Predictive modelling assigns a risk score to disease management participants based on their likelihood of using health services in the near future. In the United States, the risk score has been used to match participants in a given disease management programme who are insured with a given health insurance (health plan) to those insured members not participating. In this instance, the evaluation comprises a comparison of outcomes between enrolled participant members and control members matched on those risk scores. Hu and Root (2005) discuss the basic concepts of prediction accuracy, the relevant parameters, their drawbacks and interpretation, and offer a new accuracy parameter, "cost concentration", to better indicate accuracy of a predictive model in the context of disease management.

Propensity scoring derives a score from logistic regression that is based on the likelihood of being enrolled in the intervention (Linden, Adams and Roberts, 2003c; Dehejia and Wahba, 2002), creating a single independent variable that represents the disease management participant's entire set of covariates. This enables matching programme participants with controls on that one variable (Linden, Adams and Roberts, 2004a). Application of propensity scoring requires knowledge of all relevant independent variables that predict eligibility (e.g. age, sex, utilisation, cost) as well as behavioural factors that influence enrolment or attrition, so as to enable statistical controlling of pre-intervention differences between likely participant and non-participant groups (Linden, Adams and Roberts, 2005c). Similar to predictive modelling, one approach to propensity scoring is to compare outcomes of intervention participants with controls who are matched one-to-one based on their propensity score (Linden, Adams and Roberts, 2003c). The higher the number of observed covariates that can be identified the higher the likelihood of "true" comparability between the two groups (Linden, Adams and Roberts, 2005b). Linden, Adams and Roberts (2005c) further demonstrate how propensity scoring can reduce selection bias by assigning a random order to all controls with an identical propensity score. Other approaches to using propensity scoring include stratification, which is not a one-to-one matching strategy, and weighted regression, which has a different conceptual frame.

Predictive modelling and propensity scoring depend on an adequate pool of chronically ill individuals who are eligible for a given intervention being evaluated but who do not receive the actual intervention or any of its components (Linden, Adams and Roberts, 2005c; 2003c). This is critical for settings that employ so-called presumptive enrolment strategies, in which individuals must actively opt-out of a disease management programme. Under opt-out enrolment, there will be higher enrolment penetration and thus a smaller pool of eligible individuals who do not participate in the intervention and from which to construct a control group. In this case, the matching process will need to take into account the level or intensity of the intervention as described earlier (Linden, Adams and Roberts, 2003c) so as to create a larger pool of potential controls. Limitations imposed by a small pool of non-participating eligible individuals who are ill can be overcome in three ways: (1) using only individuals actively participating in the intervention's activities as the "true" intervention group while those "enrolled" but not active are considered non-participants in a control group; (2) using a historical (rather than concurrent) control group; or (3) recruiting non-participants from population subsets who were not offered the intervention because of their geographic location, for example (Linden, Adams and Roberts, 2005c).

3.5.2 **Statistical Controls**

Statistical controls can be constructed by developing baseline trend estimates on the outcome of interest, which can be health-, cost- and/or utilisation-related. A number of baseline trends can be considered for creating statistical controls for comparison (Farah et al., 2008; Serxner et al., 2008):

- a total population trend (from a whole nation, or a whole health insurance plan)
- a group-specific "well" trend from a non-chronic reference population that might be adjusted for:
    - observable differences in risk (difference-in-difference approach)
    - relative historical differences (ideally using two or three years of data)
    - absolute historical differences
- client-specific, or industry-published trend (based on characteristics of service users using deterministic models).

The typical evaluation approach of using a concurrent "well" population trend as a proxy to estimate the trend that would have been experienced by the chronic group has several important limitations. First, the method assumes a relatively stable relationship between the chronic trend and the non-chronic trend for the same population in the absence of a chronic disease management initiative (Niu et al., 2009). While this assumption generally seems to hold, there is a need to validate trend relationships from a given population in order to derive a chronic trend from a non-chronic trend. There is equally a need for a valid trend adjustment to ensure equivalence between the programme year and the baseline year for comparison (Niu et al., 2009). A second limitation is the distortion or neutralisation of programme impact that could occur when methods to adjust risk profile mix are applied to a chronic population in a given intervention, where the bias in trend estimates may not be relevant for risk adjustment in a non-chronic population (Duncan et al., 2008). The authors suggest it is possible to overcome this challenge by re-weighting the baseline population outcome using risk strata subgroups based on three transition states (new, continuing, terminating) and co-morbidity subgroups.

A third limitation can be the impact of random fluctuations on the measured effect of a given intervention. This issue is relevant to the characteristics of the reference population for calculating trends and also the size of the intervention group. Although larger sample sizes help reduce the risk of random fluctuations concealing the "true" measured effect, a given intervention can still appear ineffective because of adverse fluctuations regardless of the size of the intervention group (Farah et al., 2008). Farah and colleagues (2008) proposed the use of truncation and/or utilisation-based measures to address this challenge. Random fluctuations can also occur when trends are calculated from non-chronic populations that are in employment and therefore other sources for trend calculation should be considered, such as independent research or health insurance organisations not based on employment (Farah et al., 2008).

A review of the evidence of approaches to constructing statistical controls based on trends suggested that more than one trend should be used because multiple trends enable the evaluator to obtain a range of estimates of effect rather than a single point estimate of intervention effect (Serxner et al., 2008). Also, multiple trends allow stakeholders to develop a deeper understanding of the "true" impact of a given disease management initiative. Serxner and colleagues (2008) also suggest using a baseline period of 24 months to construct statistical controls while recognising that longer baseline periods may be difficult to secure in the context of disease management evaluations.

3.6 Robust Research Analysis of Chronic Care Management Effects

As with sound evaluation research design, robust analysis of evaluation data of disease management interventions seeks to establish a significant (causal) association between measured effects and the intervention studied and ensure alternative explanations of the resulting relationship were taken into account. There is a wide range of methods and approaches for robust analysis, which can be categorised in different ways depending on how they are applied or used (Mattke, Seid and Ma, 2007; Susser et al., 2006; Linden, Adams and Roberts, 2005a; 2003c). This section will briefly discuss only a selection of such methods and approaches, which are grouped according to whether a control group exists and whether selection for treatment is based on observables only or on observables and unobservables.

3.6.1 Time Series Analysis at Population Level

Time series analysis is considered superior to the "total population approach" when a control group for comparison is lacking (Linden, Adams and Roberts 2003b; 2003c). Time series analysis involves analysis of repeat observations or measurement of a given variable over time (Linden, Adams and Roberts, 2003b; 2003c). It has been proposed as a suitable technique for disease management evaluation as it allows for control of a range of threats to validity (e.g. regression to the mean, selection bias, secular trend, seasonality) by monitoring the behaviour of observations over an extended period of time – at least 50 data points are recommended to establish a historical period (at least four years of past data up to one month before commencing a given intervention) (Linden, Adams and Roberts, 2003c). It provides a timeline for changes in outcome measures and allows assessment of population-level changes that disease management initiatives are intended to effect because outcome measures are aggregated (e.g. rate per specified population) (Linden, Adams and Roberts, 2004a; 2003c).

A key feature of time series models is serial dependence, or autocorrelation, in which any variable measured over time is potentially influenced by previous observations (Linden, Adams and Roberts, 2003b). It uses previous observations to predict future behaviour of an observed variable without attempting to measure independent relationships that influence it. It will therefore account for confounding variables that may influence the outcome of a given intervention (e.g. reduced hospitalisations) but which are not easily identified or accurately measured in evaluation. However, historical experience can act as a confounder and pose a threat to internal validity of evaluation findings, thus requiring thorough understanding of observed patterns (Linden, Adams and Roberts, 2003b). The power of time series analysis to ascertain causal relationships between the intervention and the measured outcome of interest can be strengthened by adding design features to the standard analysis, such as non-equivalent no-intervention control group time series, non-equivalent dependent variables, multiple replications or switching replications (Linden, Adams and Roberts, 2003b).

3.6.2 Survival Analysis

Survival analysis applies to bivariate variables (0 and 1) and is relevant to evaluating disease management effects on individuals and cohorts when quantification of impact is measured as "time to event" (e.g. from introduction and a change in a clinical or physiologic marker to the ultimate outcome of reduced morbidity and mortality). The appeal of survival analysis here is that it can account for loss to attrition through an unique "built-in" control, which comprises those participants who do not realise a well-defined "event" by the end of the study period and therefore "censored" (because the study ended before the event or the participant was lost to follow-up, etc.) (Linden, Adams and Roberts, 2003c). Participants join a given intervention at different points during the evaluation; those enrolling towards the end of an evaluation will be followed for a shorter time and will have a lower likelihood of experiencing a given outcome (event, such as utilisation measure, clinical indicator, etc.). Thus, survival analysis allows for the inclusion of each participant and comparison between cohorts as well as the inclusion of explanatory variables to ascertain which specific participant or programme characteristics are related to better outcomes (Linden, Adams and Roberts, 2003c). Since comparison is possible through this built-in control, survival analysis is considered more appropriate than the total population approach, which has no comparable controls (Linden, Adams and Roberts, 2004b; 2003c).

3.6.3 **Intention to Treat Analysis**

The intention to treat analysis is based on the initial intervention intent and not on the intervention as implemented. Intention to treat compares groups according to the original allocation to the intervention group (Susser et al., 2006). This method addresses the problems introduced when participants leave the intervention (attrition) or when there is cross-over of exposure to the intervention between the control and intervention group, which can rupture randomisation. High drop-out rates, for example, can reduce the comparability between treatment and control groups that result from randomisation and may overestimate the effects of an intervention. Likewise, high drop-out rates might signal that only highly motivated participants remain in the intervention, so skewing the impact of the intervention and limiting the results to individuals with sufficient motivation to complete the treatment regimen (Loveman et al., 2003).

Intention to treat analysis requires a complete set of outcome data for all participants (intervention group) and non-participants (control group) as originally allocated and this depends on good tracking. In a disease management evaluation, this analysis would include all eligible participants in the analysis, and not only those individuals deemed to be high risk, or with severe disease or high health costs, and so on. In doing so intention to treat analysis addresses the risk of overstating the effect of a given intervention due to selection bias and/or regression to the mean (Pearson et al., 2007). Diamond (1999) has argued that the inclusion of all individuals irrespective of actual participation in the intervention is the only way to obtain a "true" measure of disease management impact; intention to treat is more likely to reflect the efficacy of an intervention in clinical practice (Vlad and LaValley, 2008).

3.6.4 **Treatment Effects Model**

Where a randomised controlled design is not feasible to assess impact of a given intervention, the treatment effects model may be applied to obtain statistically unbiased estimates for programme effect (Wendel and Dumitras, 2005). Potential self-selection bias can be mitigated using pre- and post-intervention information about participants and non-participants. It first models the impacts of patient characteristics on decisions to participate in the programme. Probit estimation techniques are used to determine the probability that a patient with a given set of characteristics will choose to participate in a given intervention with a dichotomous dependent variable set. The estimation of the probit equation that models the decision to participate in a given intervention provides the information needed to compute a separate regression equation, known in statistics as the "inverse Mills ratio" (IMR) (to take account of possible selection bias). The likelihood of deciding to enrol in a given intervention is then included in the outcomes equation that models the impact of intervention participation on the pre- versus post-change in the outcome of interest (Wendel and Dumitras, 2005).

The treatment effects model has been considered superior to other techniques such as ordinary least squares (OLS) regression as it allows for examination of the individual participation decision and minimisation of selection bias on the estimate of disease management programme impact (Wendel and Dumitras, 2005). Through assessing the individual participation decision and consequent programme impact it further enables identification of whether the intervention under review successfully attracts those participants who are most likely to benefit from the intervention (a potentially relevant evaluation question about programme structure and process). However, as with most techniques, this one makes assumptions of normality underlying the specific function of self-selection, which may not hold in disease management.

3.6.5 Instrumental Variable Estimators

Instrumental variable estimators are used in cases when allocation to treatment is non-random and possibly associated with characteristics of the treatment group members which are not observed. The instrumental variable approach relies on randomisation as the perfect instrument and thus seeks to mimic this process through use of natural instruments (Jones and Rice, 2009). A natural instrument is a variable which is correlated with treatment but is not directly correlated with treatment outcome. Instrumental variable techniques therefore seek to find and use one or more of these variables to identify the treatment effect of interest. Its strength lies in the fact that it has an explicit selection rule that does not necessarily assume normality, which might be helpful in disease management evaluation that requires further investigation and validation.

Conceptually, the instrumental variable estimator could be illustrated within a two-stage framework. In the first stage (the selection equation), the treatment variable is regressed on the instrument(s) and all other predictors to be included in the second stage. A prediction of the treatment is obtained, which then replaces the actual treatment variable in the second stage (the outcome equation). The estimated treatment effect from this second stage is the instrumental variable treatment effect and standard software adjusts for the standard error to reflect this (Jones and Rice, 2009).

One of the most significant challenges to conducting instrumental variable analysis is that it is unlikely that an instrument can be found which perfectly predicts the treatment assignment of individuals. Further, often it may also be difficult to find appropriate instruments to use in evaluations of disease management interventions. Instrument appropriateness is generally established in relation to three particular conditions: (1) the instrument has to be valid, which implies that it has to be correlated with treatment but uncorrelated with outcome; (2) the instrument has to be relevant, which suggests that its relationship with the treatment has to be strong; and (3), the instrument should not be poorly correlated with the treatment (the instrument should not be a weak instrument). While the first condition is not empirically testable (unless the number of instruments is greater than the number of treatment effects estimated), the latter two conditions are testable with standard tests (Jones and Rice, 2009).

3.6.6 Regression Discontinuity Analysis

One way to obtain an unbiased estimate of the effect of a given intervention by reducing the major concerns of selection bias and regression to the mean is to remove them at the outset using the regression discontinuity design and analysis. This may be the closest method to the RCT for disease management evaluation because critical threats to the validity of evaluation findings can be removed through its strict adherence to a cut-off point for allocation of individuals to intervention or control groups. At the analysis phase, regression discontinuity analysis aims to identify discontinuities in the relationship between a given variable and outcome of interest at the point differentiating the intervention and control groups (Linden, Adams and Roberts, 2004c). This method identifies correctly the treatment effects when the underlying process generating the pre-test measure is not mean-reverting. Thus, in choosing this method, it is important to understand the epidemiology of the disease being managed so it is clear the measured values cannot return to the mean without intervention.

Unlike other methods, this methodological and analytical approach is useful because the pre-test measure does not need to be the same as the outcome measure, and this allows the intervention to use a maximum range of tools to identify those individuals in greatest need so as to assign them to the intervention (Linden, Adams and Roberts, 2004c). Similarly, the cut-off score used to assign individuals to the intervention on the pre-test measures can also be determined based on data sources different from those used to measure the intervention effects (e.g. clinical understanding of the disease process, empirically driven or resource-based) (Linden, Adams and Roberts, 2004c). Notably, multiple cut-off scores can be used when the intervention is tiered. The consequent maximisation of the range of data sources for intervention allocation and analysis of effects is a helpful advantage for evaluation in practice. Of course there are a number of requirements for using the regression discontinuity design that need to be borne in mind, not least of which is strict adherence to the cut-off point applied to all individuals with the disease targeted by the intervention (Linden, Adams and Roberts, 2004c).

3.6.7 Regression Point Displacement Design

While similar to the regression discontinuity analysis, the regression point displacement approach is worth noting because it uses a single treated-group score (aggregate data). This has the advantage of being more stable and precise than within-group data. This feature means that statistical power is not reduced by the lower samples sizes typically used in this approach, although there are methods to increase power (e.g. by raising the number of controls, or by demonstrating a large disease management programme effect) (Linden, Trochim and Adams, 2006). Another feature of the regression point displacement design is that pre-test equivalence between the treated and control groups is not required because the design assumes the post-test results of the treated group do not differ significantly from the regression line prediction (Linden, Trochim and Adams, 2006). However, this approach still faces the problem of selection bias and so randomisation in the design is valuable to mitigate this bias. Finally, this design should be avoided when the effect of a disease management programme is heterogeneous and a function of group characteristics such as the programme being more successful with strong support from the care provider (Linden, Trochim and Adams, 2006).

3.6.8 Matched Pairs by Propensity Score

In non-experimental evaluations where the intervention group may differ substantially from the pool of potential controls, propensity score-matching in the analysis stage may improve estimation of the intervention effect (Dehejia and Wahba, 2002). As noted above, propensity scoring can be used to construct comparable control groups by correcting for sample selection bias due to observable differences between the intervention and comparison groups (Linden, Adams and Roberts, 2005c; Dehejia and Wahba, 2002). However, the size of unobserved variation remains a potential threat to the validity of findings (Linden, Adams and Roberts, 2005c). Moreover, propensity scoring for matching intervention-control pairs requires sufficiently large samples, especially when using subgroup analysis, which can limit the application of this method to disease management evaluation (Linden, Adams and Roberts, 2005c).

Matched pairs using propensity scoring in the evaluation analysis allows for the reduction of a large potential comparison group to a smaller group thereby minimising the burden of outcome data collection. Programme participants can be matched to non-participants using propensity scoring through: (1) pair-wise matching, (2) matching with and without replacement, (3) matching using propensity score categories, (4) matching based on the Mahalanobis distance, (5) kernel-density matching, (6) stratification and (7) weighted regression (see references in Linden, Adams and Roberts, 2005c for details on each). The type of matching will be determined by the number of relevant comparison units required (Dehejia and Wahba, 2002). It is worth noting that control-intervention pairs can be either over-matched or under-matched using such statistical techniques, both of which have consequences for causal inference discussed in further detail elsewhere (Susser et al., 2006).

3.6.9 System Analysis Using Reliability Block Diagrams

As a complex intervention, disease management initiatives comprise a set of interconnecting components, or subsystems, that exist in a complex environment of healthcare delivery. It will therefore be important to account for interaction between the various components that determine the overall outcome of a particular initiative. To account for interaction, Sonnenberg, Inadomi and Bauerfeind (1999) propose using a so-called reliability block diagram as a means to estimate the contributions of many heterogeneous factors, medical and non-medical, to overall impact of an intervention. If applied to disease management evaluation, it might be possible to study how individual components, for example diagnostic and therapeutic procedures, are integrated within the larger healthcare system and how the interaction between medical and non-medical subsystems affects the overall outcome of a given intervention.

However, reliability block diagrams assume probabilistic independence of the component subsystems and the model focuses only on a dichotomous outcome, namely system function or failure; this approach is therefore most useful in demonstrating areas within a system where changes can be implemented, so identifying avenues to strengthen system performance (Sonnenberg, Inadomi and Bauerfeind, 1999).

3.7 Evaluation Measures

The selection of different types of evaluation measures should be driven by construct validity of the measure, which requires careful consideration of the measure's sensitivity to the specific design and goals of the disease management initiative, both the interim and long-term goals. At the outset, therefore, it will be important for any evaluation to identify clearly not only the evaluation's objectives but also the aim(s) of the actual intervention that will drive the identification of measures to be assessed (Zajac, 2004; Mateo, Matzke and Newton, 2002). Yet, as Pearson (2007) noted, existing evaluations have tended not to define clearly these relationships; rather, only a minority of published studies apply a coherent framework linking the aims of disease management to measures of structure, process and outcome (Steuten et al., 2006). Quality is a prime example of the need for conceptual clarity; whereas some of the assessment literature includes quality among possible "outcome" measures, because quality can be a key objective of disease management, the examples of quality indicators relate to processes (activities of the intervention) (Box 2).

Box 2 Measuring quality in disease management evaluation

Traditional **quality** measures have been based on processes of care as well as intermediate outcomes. Quality is commonly reported as an "outcome" measure in evaluations of disease management initiatives, which requires further advancement of new quality measures (Orkin and Aruffo, 2006; Sepucha, Fowler and Mulley, 2004; Kerr et al., 2001) and regarding the link to the goals of a given intervention.

Most of the measures used to evaluate quality are a type of "technical" quality (Kerr et al., 2001), such as annual retinal examination and annual screening for nephropathy for diabetic patients with disease management. Reviewing different sets of indicators of technical quality, Kerr and colleagues (2001) identified a number of key challenges to the use of technical quality outcome measures, including:

- measured processes of care lack strong links to outcomes
- actionable processes of care are not measured
- measures do not target those at highest risk
- measures do not allow for patient exceptions, including contraindications and patient refusals
- intermediate outcome measures are not severity adjusted.

The authors therefore argue there is a need to develop and test clinically meaningful technical quality measures that will yield the most appropriate quality improvement responses, and a need to develop and test additional quality indicators for various chronic diseases.

Others have conceptualised quality in disease management as the degree of patient-centredness of a clinical decision (Sepucha, Fowler and Mulley, 2004). This concept of "decision" quality requires direct input from patients to measure this outcome, an approach that is still fairly underdeveloped in practice and frequently based on routinely collected administrative data or medical records. Where decision quality is measured based on patient experience (using self-reports of satisfaction with decisionmaking, the nature of the interaction with the clinician, state of knowledge and decisions made), there are limitations to consider (Sepucha, Fowler and Mulley, 2004). Here too there is a need to develop and test measures of "decision" quality, which include a consideration of decision-specific knowledge, treatments chosen and quantitative values for salient aspects of decision quality.

Yet another way to define quality in disease management is its "fitness for use in the context of the end user" (Orkin and Aruffo, 2006, p.58), which emphasises the importance of assessing the design of the given intervention and the extent to which it functions in the context of use. This includes, for example, designing education materials on a given condition in a way that is understandable and interpretable by the target group (Orkin and Aruffo, 2006). Here the outcome of interest is the extent to which the intervention or its components meet the functional requirements of its users. This requires an understanding of functional requirements in the first place, either ex ante when designing a given intervention or ex post when evaluating its effectiveness, using such methods as ethnographic techniques. This approach to assessing quality of disease management in relation to "fitness for purpose" is still fairly new and requires development and validation of relevant measures if this borrowed concept of "product" quality of disease management initiatives is to be applied to their evaluation(s), as proposed by Orkin and Aruffo (2006).

Quality as an evaluation measure of disease management "outcome" has further been conceptualised in relation to patient empowerment in complex chronic disease. Gagnon and Grenier (2004) reviewed a wider range of quality indicators and identified 77 as outcome measures, broadly classified into three domains: (1) patient domain of experience of health or illness and participation in care; (2) interpersonal domain of patient–provider relationship; and (3) practice domain of procedures and approach to care. Gagnon and Grenier's (2004) method of validation of this new set of patient empowerment quality measures might offer a starting point for the development and validation of the other subtypes where further work is still needed.

The absence of a consistent analytical framework for evaluating outcomes (however defined) has made comparisons of reported results impossible and has rendered many reports unreliable (Villagra, 2004). There is therefore a need for developing a standard methodology in measures used to evaluate disease management, with Lemmens and colleagues (2008) recently presenting an analytical framework for determining evaluation measures that link to patient-related and provider-directed elements of disease management interventions.

While it is important to link the choice of evaluation measures with the aims of the intervention being studied, it will be equally important to clearly specify the hypotheses about the expected impact of the intervention on the effects of interest (Mulligan et al., 2005) so as to enable assessment of the conditions under which a disease management intervention is deemed to be successful (Loveman et al., 2003; Diamond, 1999). The testable hypothesis should further be informed by a theoretical foundation of behaviour change underpinning the design of the intervention and selection method (Linden and Roberts, 2004) (Section 3.2).

3.7.1 Measures Vary by Type, Unit Level and Scale

We have noted earlier that evaluation measures can be broadly distinguished according to whether they measure inputs, processes and outcomes (Table 1). We here discuss general attributes of measures of input and process, with the next section focusing on more detailed outcome measures. As noted in the introduction, one key challenge to drawing firm conclusions about the effects of disease management (however defined) on outcomes (however defined) is the inherent heterogeneity of disease management interventions. This means that relevance to evaluation of different measures of process will differ across interventions (Zajac, 2004). For example, interventions with elements targeting clinician behaviour will need process measures on adherence to practice guidelines whereas participant-centred elements will focus process measures on patient activities of disease control. Also, the method of delivery (e.g. mail, call centre, internet, face-to-face interventions) will differ within and between interventions and their multiple elements. Although comparison of process measures at this level of granularity may be difficult to interpret, it emphasises the importance of characterising all aspects of the disease management intervention in as much detail as possible (Section 3.1).

Another attribute that informs the selection of evaluation measures is the level or unit of analysis, which can be the individual, a group or a nation. Individual level performance indicators include clinical or physiological measures indicating control of disease (Linden, Adams and Roberts, 2003b). Measures at the aggregate or group level include utilisation variables, such as admission rates. The choice of unit of analysis will in part be determined by the audience of the evaluation. For example, clinicians may be more interested in individual participants and groups of participants whereas operators or purchasers of disease management interventions may be more concerned with populations and the intervention itself at the aggregate level (Zajac, 2004).

Other work has suggested measures of intervention components should also be included in disease management evaluations. Velasco-Garrido, Busse and Hisashige (2003) recommend measuring the incremental benefits of single components to determine the ideal mix of components. Such an approach requires the ability to single out and manipulate individual components of a given intervention, for example patient education, to enable drawing conclusions about the effects (Loveman et al., 2003). However, there is very little sound evidence of single component effects. Notably, the evidence on patient-related interventions is fairly well established, with measurement frameworks developed in particular for self- and family management interventions (Grey, Knafl and McCorkle, 2006; Jack et al., 2004). But guidance on choice of indicators to assess provider-targeted interventions remains limited.

A third attribute to consider is the scale of measurement (e.g. nominal, ordinal, interval or ratio). A frequently used measure is the rate (time-related ratio) as it follows a Poisson distribution (Linden, 2008).

A key challenge for selecting any evaluation measure is potential variation in data sources (Pearson et al., 2007). For example, in measuring "reach" it will be important to ascertain whether those reached by the intervention are disproportionately those who already are more likely to be actively engaged in disease management through for example self-management and related health-seeking behaviour. However, most datasets make this indicator difficult to assess. The types and limitations of data sources that can be used for different evaluation measures are discussed in more detail by Pearson and colleagues (2007).

3.7.2 Outcome Measures for Evaluating Disease Management Effects

The goal of most disease management interventions is to improve the overall health status of the individual patient and, where applicable, the population at large. How this benefit ("outcome") is being conceptualised may vary, and will depend on the overarching aim and length of a given intervention. Given that a person's health status is influenced by many variables over time, it is proposed that high quality evaluation research of disease management interventions should focus on a range of outcomes evaluated after extended follow-up intervals (Loveman et al., 2003). This section describes the likely range of outcomes and any concerns about their use in evaluations.

Outcome measures can be distinguished by type and level along a "continuum of categories" that extends over time (Fitzner et al., 2004; Mulcahy et al., 2003), in particular:

- "Immediate" or proximate outcomes (alternatively called "outputs") are measured at the time of the intervention such as learning and improved knowledge.
- "Intermediate" and "post-intermediate" outcomes occur over time, requiring more than a single measurement point and are sensitive to change. Intermediate outcomes include for example participant behaviour change as a consequence of the intervention. Post-intermediate outcomes tend to focus on clinical outcome measures as confirmed by laboratory or procedural testing and occur as a consequence of one or more components of the intervention including participant self-management and clinical management.
- "Long-term" health outcomes are definite and result from multiple variables over time and include measures for health status such as improved quality of life, morbidity, and mortality.

At the long-term end of the spectrum, there are impacts other than health status effects, which are often of interest for evaluation measurement; these can include socio-economic benefits for people with chronic illness and society at large.

Along this continuum, the range of outcome types includes: clinical, humanistic, quality, utilisation, financial and socio-economic. Table 7 summarises some of the different types of outcome metrics used in evaluations of disease management initiatives.

Table 7 Types of outcome measures in assessing disease management

Outcome type	Examples of outcome type
Clinical	Percentage of diabetics with improved Hemoglobin A_{1c} (Hb1Ac); percentage of new (co)morbidity; number of performed foot and/or eye exams; amputation rate
Humanistic	Patient (health-related) satisfaction, psychosocial adaptation; coping skills; quality of life; mortality
Quality	**Technical quality**: proportion of patients with diabetes who have a specified treatment appropriate for hypertension and hypercholesterolemia; thrombolytic, β-adrenergic blocking agent and aspirin treatment after myocardial infarction; angiotensin-converting enzyme inhibitor use in heart failure **Decision quality**: no existing set of measures **Product quality**: intervention design fit for use in context – readability and functional use of education materials **Patient empowerment**: various measures
Utilisation	Emergency room attendance rates; inpatient admission rates; resource consumption, "utility"
Financial	Direct costs; indirect costs; return on investment
Socio-economic	Absenteeism, presenteeism; return to work; productivity; (corporate) image; civic or social participation

SOURCE: Adapted from Lewis (2009); Orkin and Aruffo (2006); Serxner et al. (2006); Fitzner et al. (2004); Gagnon and Grenier (2004); and Sepucha, Fowler and Mulley (2004).

NOTE: For some measures, the unit of analysis could be individual, aggregate or both (e.g. utilisation).

Clinical Outcome Measures

Clinical outcomes, oftentimes disease-specific, are the most frequently measured outcomes in evaluations of disease management interventions. This is in part because of ease of measurement, with standard tests and procedures in place, although there are no standardised approaches for a set of clinical outcomes (Fetterolf, Wennberg and Devries, 2004). However, while methodologically simple to assess, the relevance of some clinical outcomes for long-term health outcomes of disease management interventions remains unclear (Bailie et al., 2006).

Satisfaction and Health Status

Other outcomes may be more difficult to assess or their validity may be compromised. This may be the case in particular for self-reported measures such as patient satisfaction where there is a risk of participants trying to anticipate the desired effect or giving socially desirable answers (Susser et al., 2006). Use of validated instruments, such as the Patient Assessment of Care for Chronic Conditions (PACIC) instrument, to assess self-reported measures can reduce this risk (Loveman et al., 2003). The PACIC instrument is congruent with the Chronic Care Model, which is but one of many disease management approaches. Thus, while this is an important step in the direction of the methodological development of a prevalent and systematic approach to measuring patient satisfaction in disease management, there may be scope for field-testing its use in the evaluation of other disease management initiatives. There also remains a more fundamental conceptual question of what "patient satisfaction" actually means since surveys on this fail to measure what is important in relation to quality of care (Coulter and Cleary, 2001).

Similar methodological concerns apply to the assessment of provider satisfaction with a given intervention. There are two challenges here: a relative lack of reliable and valid instruments appropriate to measuring provider satisfaction (Aadelen, 1998), and a practical problem of tracking the care providers, especially when they change roles, work units and services within or between institutions during the intervention period under evaluation (Aadalen, 1998). Wendel and Dumitras (2005) examined the impact of physician referral on participation decision, identifying a need for further research into physician attitudes and referral patterns to identify and measure salient physician characteristics related to disease management participation and its impact on measured outcomes.

Health-related quality of life (HRQL) is another commonly used measure of outcome of (disease management) interventions (Wyrwich et al., 2005). The Short-Form 36-Item Health Survey (SF-36) is the most widely used generic HRQL instrument for individuals with different diseases; it yields scale scores for eight domains (Brazier, Harper and Jones, 1992). Related disease-specific instruments determine clinically important differences in disease-specific quality of life, so enabling assessment of levels of improvement as a consequence of the given intervention (Guyatt et al., 2002). The ability to establish clinically relevant differentiation is an important issue in evaluation of disease management initiatives because statistically significant differences in quality of life scores may not be meaningful at the individual level. The interpretation of evaluation results for disease-specific changes in HRQL over time can be aided using expert panel ratings of improvement thresholds, but it will be important to confirm panel-derived thresholds with individual patients and their physicians themselves (Wyrwich et al., 2005).

Overall health status measures aim to quantify how a person's disease affects their life in three domains identified by Spertus (2008): (1) symptoms and their management (e.g. fatigue, edema, dyspnea); (2) functional limitations (physical, emotional, mental and social); and (3) quality of life (living as desired). Overall health status measures might also include self-efficacy scales which are not covered by the SF-36 (Brazier, Harper and Jones, 1992).

Overall health status and HRQL are generally assessed using self-administered questionnaires and quantification requires a valid, reproducible and sensible means of obtaining an overall score. Not only is interpretation of such scores a challenge, but obtaining an overall score can be particularly difficult when the experiences of chronically ill populations are aggregated (Spertus, 2008). Several measures of health status exist for some chronic conditions (e.g. heart failure), which might have a role in implementing and evaluating disease management interventions; however, their use requires further validation and refinement and calls for research on national benchmarks (Spertus, 2008).

Utilisation

Utilisation is commonly measured by calculating emergency room attendance and inpatient admission rates associated with the disease(s) being managed (Lewis, 2009). As we discuss below, changes in these rates might then be translated into a measure of cost. A key challenge in outcomes measurement of utilisation in evaluations is the need to consider the interaction of the disease in question with (a likely set of) comorbidities, through for example the use of a "comorbidity multiplier" (Lewis, 2009; Linden et al., 2007), which has certain assumptions to be considered (Lewis, 2009).

Financial Outcomes (Impacts)

One of the key motivations for introducing disease management programmes in the United States was an expectation that structured interventions would, ultimately, save the costs of care by reducing utilisation of expensive services, in particular hospital use (Bodenheimer, 1999). Costs can be measured directly or indirectly. If cost is measured directly, it will be crucial for an evaluation of a given intervention that uses a pre-test/post-test design to use an appropriately matched *concurrent* control group (Linden, Adams and Roberts, 2003c). Use of historic control group is likely to bias findings as cost categories will likely have changed over time and it will not always be possible to identify them for statistical adjustment. Table 8 summarises approaches to estimating economic impact using time series or survival analysis. However, a time series analysis that measures cost directly can still be biased if any part of the cost variable was measured differently between time periods, or if no adjustment was made to account for inflation, changes in reimbursement and so on (Linden, Adams and Roberts, 2003c).

The choice of evaluation design will also be important for determining the measure of direct cost (and any reduction thereof) (Crosson and Madvig, 2004). Where the potential cost impact of disease management is defined by an absolute reduction in baseline year costs, the interpretation of findings will have to consider the wider context within which the cost reductions were achieved (if any). For example, where an intervention is implemented in a system context that is already characterised by relatively low baseline costs, any additional saving achieved by the intervention is likely to be small (Crosson and Madvig, 2004). In contrast, systems that are characterised by high utilisation rates of specialist providers, for example, are likely to accrue relatively higher savings if, indeed, the intervention is suited to reduce specialist utilisation markedly.

Table 8 Estimating cost using time series and survival analysis

	Direct measure of cost	Indirect measure of cost
Time series analysis	Can compare actual normalised aggregate costs (after each measurement period is completed), to forecasted or predicted values with the difference indicating whether a savings effect was achieved.	Can translate utilisation measure into estimated cost for each observation period and compare to the actual utilisation value and its estimated cost.
Survival analysis	Can tailor analysis to estimate the probability of hitting a given cost threshold as the outcome measure, such that "programme success is indicated by a longer period of time until the intervention group reached that threshold. The difference in those costs over that time period can be easily computed."	Can estimate a marginal cost effect for the probability of hospitalisation and compare across enrolled and control groups.

SOURCE: Linden, Adams and Roberts (2003c).

Linden, Adams and Roberts (2003c) argue that economic impact is most appropriately measured indirectly given that disease management interventions tend to use utilisation measures rather than cost as outcome measures. Utilisation is price-insensitive and can serve as a proxy for measuring the financial return of investing in disease management (Fetterolf, Wennberg and Devries, 2004). Utilisation measures are less susceptible to bias than cost over time given that healthcare expenditure is constituted of price and units, which can fluctuate over time. The indirect measurement of economic impact further allows for matching participants in a disease management initiative either to a concurrent control group or to a historical control group (Linden, Adams and Roberts, 2003c). Day costs would then be assigned to the utilisation rate for both groups to derive an estimate for financial effect, for example. Recent work by Steuten and colleagues (2009) exemplifies the use of utilisation measures to estimate short-term cost-effectiveness of chronic care programmes for people with chronic obstructive pulmonary disease.

There are several practical recommendations available for shifting evaluation focus to utilisation as an outcomes measure for calculating financial impact. These are summarised in Box 3; several are equally relevant to measuring non-financial effects of disease management.

Box 3 Recommendations for indirect measurement of financial impact

Serxner, Baker and Gold (2006) have provided several practical recommendations for shifting the evaluation focus from using direct costs to utilisation as an outcome measure for calculating the financial return on investment in disease management. Although the recommendations are aimed at evaluators in the United States where the focus has been predominantly on cost savings, many of these recommendations are equally relevant to other indicators of effect:

- Perform population-level and individual-level analyses of utilisation trends to provide a range of financial impact attributable to the programme as a whole and its component parts (utilisation analyses should be based on date of service and not date of finance claim submission, which has a lag between submission and processing).
- Establish financial impact targets across an appropriate timeframe given that approximately three to five years are required to recognise true programme effectiveness of disease management initiatives, particularly relating to costs.
- Enhance baseline methodology for retrospective and prospective cohorts to help mitigate regression to the mean and selection bias evaluation challenges (identify all eligible participants in the 12 months proximal to launch date of a chronic care intervention and average their 24 months of health insurance claims).
- Conduct analyses with and without selected factors that affect results, such as eligibility criteria, capping or transforming outliers, and excluding selected conditions; doing so will provide a range of return on investment and impact estimates based on the “with and without” analyses rather than one single point estimate.
- Use quasi-experimental design and conduct multivariate analysis using several different combinations of participation variables to predict the utilisation trend difference score while adjusting for baseline group differences, length of intervention exposure, trend factors, co-morbidities, demographics, health service design and so on. Individual-level participation data need to be linked to outcomes and demographics.
- Use unique trend figures for each of the major utilisation categories (inpatient admissions, outpatient visits, emergency department visits, professional service visits and prescription drugs) and for each disease category.

It is also recommended to apply a retrospective cohort pre-post baseline population-based savings analysis methodology to evaluate economic impact from chronic disease management to address regression to the mean and selection bias (Serxner, Baker and Gold, 2006). This methodology includes the following procedures: (1) diseased individuals are identified in the two years before the intervention start date and their healthcare utilisation is averaged for the two-year baseline period; (2) the projected utilisation for the diseased population is created by applying the trend of the non-diseased population (a 24-

month average cohort) to the cohort averaged baseline for the diseased population; (3) the projected utilisation is compared with actual programme-year diseased-participant utilisation to determine utilisation difference; (4) average unit costs are applied to overall utilisation to calculate a total and per-diseased member per-month cost difference; and (5) ROI is calculated by comparing annual cost difference against total programme expenses and all fees (e.g. communications), making appropriate adjustments for inflation, for example (Serxner, Baker and Gold, 2006). Notably, the comparison group is the non-diseased population, not the non-participant population, which is often too small to be used as a comparator.

Despite the wide range of evaluations that have attempted to measure the economic impact of chronic disease management, the available evidence of whether costs are saved has remained remarkably slim (Mattke, Seid and Ma, 2007; Goetzel et al., 2005; MacStravic, 2005; CBO, 2004). There is a lack of rigorous scientific studies that would provide the evidence for a business case that investment in disease management is a means to reduce overall costs and achieve a return on investment (Plocher, 2006). This is in part because of the absence of designs that use a sound comparison strategy to control for a number of threats to validity such as regression to the mean effects (Plocher, 2006; Linden, Adams and Roberts, 2003a; 2003c). Furthermore, existing evaluations tend to lack a complete accounting of all relevant costs (MacStravic, 2005). Evaluations tend to focus on actual expenditure incurred by the funder (e.g. health insurer), while neglecting total cost savings, which ought to be considered by also taking account of indirect cost impacts of disease management on productivity, for example.

Indirect costs include the following (MacStravic, 2005):

- total labour costs
    - employers' short- and long-term disability costs
    - absenteeism effects
    - presenteeism effects
- non-financial (human resource) benefits
    - improvements in recruitment and retention from healthier employees
    - improvements in customer satisfaction
    - other benefits (e.g. reduced informal care by spouses, relatives or friends)
- indirect health effects and improvements in quality of life
- impacts on the federal budget.

Financial savings might also be found in reduced costs for hospital-based physicians, or reduced emergency room utilisation (Farah et al., 2008).

A challenge to complete accounting for all costs is the availability of relevant data on costs, prices and resource use data, especially when evaluations are performed across institutions and systems. Where institutions use claims-based systems with specific formulae rather than cost-based fiscal systems, resource use data have to be collected through other means such as retrospective record review of other administrative records, the completeness of which may however vary (Aadalen, 1998).

In addition, there is a conceptual challenge related to determining what costs are relevant for economic evaluation and what data sources can provide information appropriate to such measures. The economic impact of a given intervention for one or more of the traditionally targeted disease types will be conceivably different from the greatest total negative impact to an employer because productivity will be impaired by other chronic conditions and unhealthy behaviours not typically addressed by disease management (e.g. musculoskeletal disorders, migraines and other headaches, chronic fatigue, emotional disorders, stress, and so on) (MacStravic, 2008). Thus, if a broader economic perspective is taken that includes productivity losses, then the scope of a given intervention is expanded because a new set of health conditions would be indirect or co-targets of disease management. Such an expansion of scope complicates the evaluation because relevant sources of data may become more limited where productivity changes, for example, would need to be measured not only by medical claims or workers' compensation but also more directly by employers or employees (MacStravic, 2008).

Despite a call to evaluate the economic (and social and ethical) implications of disease management interventions and to standardise methods for doing so (Velasco-Garrido, Busse and Hisashige, 2003), there are concerns that the current tendency to use cost as a principal measure of "success" of such initiatives may be misplaced (MacStravic, 2005; Linden, Adams and Roberts, 2004d). It is based on the assumption that because chronically ill people tend to be costly, the prevention of chronic illness and its management would help control rising costs. Yet, as Fireman, Bartlett and Selby (2004) argue, national spending on healthcare is driven up by a range of interlinked factors independent of a population's health. Furthermore, disease management initiatives will only reduce costs by improving quality if the interventions for the chronically ill are themselves cost-saving. Yet, in practice, the evidence for most interventions targeted at chronic conditions is of cost-effectiveness rather than of cost-saving, an analysis which ought to consider that benefits related to length and quality of life are unlikely to outweigh savings that accrue when exacerbations and complications are prevented (Fireman, Bartlett and Selby, 2004).

It has therefore been argued that if economic impact is to be included as an outcome for disease management evaluation, then financial measures should be part of a combination of factors focusing on measures of clinical and health outcomes (Plocher, 2006) and/or programme-specific utilisation measures (Fetterolf, Wennberg and Devries, 2004; Linden, Adams and Roberts, 2003b). Plocher (2006) proposes adopting a "holistic" approach to (economic) evaluation in disease management including the following combination of factors:

- feedback from participants
- feedback from physicians
- clinical improvements in participants (using a variety of health improvement metrics)
- financial return on investment.

3.8 Outcome Validation

3.8.1 Sensitivity Analysis

Sensitivity analysis allows for examination of the robustness of observed data patterns to errors related to bias or chance (e.g. to gauge the magnitude of bias resulting from differential attrition) and thereby ensures that any differences found in outcome measures are in fact attributable to the actual intervention being studied and not to baseline differences in study group (Susser et al., 2006). Greater confidence that the observed effects are a consequence of the intervention and not introduced by chance is one reason some authors have advocated that sensitivity analysis should be considered common practice in any robust research designed to measure disease management effects (Linden, Adams and Roberts, 2005b).

Another reason is that it also enables an assessment of the potential impact of unobserved differences in the populations under study, which are not accounted for through statistical means (Linden, Adams and Roberts, 2005b). As an example, sensitivity analysis of case-control design in which participants are matched on propensity scores enables estimation of the odds of participants being assigned to the intervention based on unobserved characteristics (Linden, Adams and Roberts, 2005c). Sensitivity analysis therefore provides an indication of the relative sensitivity of evaluation results to small, moderate and high levels of "hidden" bias. Although some guidance exists in disease management evaluation, there is no standard threshold for declaring that "hidden" bias is large enough to invalidate the measured effects of a given intervention (Linden, Adams and Roberts, 2005b). This issue may warrant further exploration.

It is worth reiterating the importance of the evaluation time period for robust analysis of intervention effects and their validation. For example, when effects of a given intervention are evaluated using a time series analysis, it is suggested that the validation set should include one year's worth of monthly observations (12 data points) following the historical set (Linden, Adams and Roberts, 2003b). This presumes of course that the evaluation is implemented over more than 12 months, which is traditionally the case.

Bootstrap method

The bootstrap technique is a nearly universal method to make inferences from data that are influenced by extreme values as well as to reduce the threat of multiple comparisons bias, but it is rarely used in disease management evaluations. Yet, in the disease management setting, data on outcomes of interest are often not normally distributed with the measured average of a disease management population being highly skewed by extreme values due to case mix (Linden, Adams and Roberts, 2005a). Also, sample sizes tend to be small. It should be considered more often because, similar to other non-parametric techniques, the bootstrap method does not make assumptions about the resulting data and thus allows statistical inferences to be extended more appropriately to the target population based on potentially skewed results achieved in the intervention group (Linden, Adams and Roberts, 2005a). Since the technique produces standard errors for almost any quantity of interest, it enables a disease management evaluation to make use of a broader array of quantities of interest in the analysis by validating findings. The application of this method to first-year outcomes of a congestive heart failure disease management programme demonstrates its utility for answering additional questions and making inferences to the underlying reference population (Linden, Adams and Roberts, 2005a).

3.8.2 Threats to External Validity: Stratification and Indirect Estimation

External validity of results of a given disease management intervention can be enhanced using fairly simple methods such as stratification or subgroup analysis, which can provide powerful insights into areas where intervention effects may differ (e.g. age, ethnicity, healthcare setting) – areas deemed important among the evaluation's sample population and the underlying target population. Indirect estimation can then be applied to generalise observed effects across broader groups using for example population-specific weights to adjust for differences in the composition of subgroups (Linden, Adams and Roberts, 2004d). In many cases, however, adjustment through indirect estimation may not be sufficient where the evaluation population does not have adequate representation of individuals and organisational features to permit generalisation to another population (e.g. different social groups or different service delivery models) (Linden, Adams and Roberts, 2004d).

3.9 Identifying Crucial Components of Disease Management

With the growing number of disease management evaluations, there may be more efforts to analyse several studies in a meta-regression as a way to identify which programme characteristics seem crucial to improving a particular outcome of interest, such as reducing hospital admissions. But careful consideration is needed of the methodological approach taken for the meta-regression so as to avoid the risk of spurious findings that exist in some of the disease management evaluation literature (Koller and Steyerberg, 2006). Referring to an example of a meta-regression of heart failure disease management programmes, Koller and Steyerberg (2006) commented that conclusions about exercise as the crucial component were misleading because the meta-regression was based on a dichotomisation of several studies ("effective" or "ineffective" based on statistical significance of reported results) in a fixed-effects approach that is inappropriate for complex disease management programmes. One reason is that small sample sizes can produce results that suggest the interventions studied were "ineffective", when perhaps the intervention is effective but accrued too few patients. Another is that heterogeneity assessment and weighting of studies requires information about variation within and between them, but this is removed through dichotomisation. Thus, it is argued that intervention characteristics should have been regressed against the effect size using a random-effects meta-regression approach in order to identify characteristics that may be crucial to achieving a specific outcome in disease management. This point serves as a reminder about the need to characterise clearly the actual intervention that is studied for this information to be useful for future comparisons across studies of potential "active ingredients" in disease management.

3.10 Predicting Future Effects

On a final note, there may be policy interest in long-term predictions of disease management outcomes and impacts. While it is possible to predict future effects using modelling techniques, such estimations depend on evaluation data that has been collected over more than the traditional 12-months. For example, the five-year cost-utility of a Dutch Disease Management Programme for asthma, compared with usual care, was recently estimated using a Markov model (developed based on health state transition probabilities from published studies) (Steuten et al., 2007). Importantly, the model was informed by data from a 15-month evaluation of the programme.

Thus, while futures-oriented estimations of disease management effect are not generally within the scope of an evaluation, it highlights an issue of importance to implementing disease management evaluation: the time period of an evaluation is critical for not only what effects of interest can be reasonably expected but also what analyses can be performed and considered to inform policy and decisionmaking.

CHAPTER 4 Conclusion

Our synthesis confirms that the current state of the art in disease management evaluation is still shy of having a widely accepted standardised approach to design and measurement. However, considerable technical guidance on different aspects of evaluation in this subject area can be found in the literature. Much of the technical guidance relates to the design choices available in the absence of a RCT – a design deemed to be ill suited to the science of improvement, which involves complex, multi-component, multi-actor interventions, such as disease management.

Our literature review shows that it is not necessary, nor ideal, to rely strictly on the use of an RCT design to establish a causal inference about the results achieved in a disease management initiative. Rather, it is more important to understand the goals of randomisation in an RCT and consider all the possible threats to valid inference about the effects of a given intervention. Ultimately, whatever the design chosen, a comparison strategy must be constructed using a control group careful selected from the same reference population as the group selected to receive disease management.

The review highlights the array of sophisticated statistical techniques that can be used not only during the analysis phase of an evaluation to enhance the validity of findings (e.g. by controlling ex post for confounders or enhancing sample size), but also during the selection phase of the intervention itself when individuals are assigned to either a "treatment" or a "control" group. Importantly, the power of statistics depends greatly on pre-evaluation planning and a careful understanding of: (1) who is to receive the intervention; (2) what the intervention is doing; and (3) when the intervention is received. To answer these perceptively simple questions, more information than what is currently reported in published evaluations is needed about the characteristics of intervention and its intended population(s). Reliable evaluation findings must be informed by not only the intervention but also the fidelity of implementation with intention. In addition, there is also a need to draw more explicitly on one of the many possible theories of behaviour change to better link the choice of performance measures to the goals of the intervention and to ensure the strategy for enrolling evaluation participants is appropriate to the population the intervention aims to target.

Finally, our synthesis also confirms that there is great scope for further development and validation of several different types of indicators of effect for disease management. In particular, further research is needed to improve reliability and validity of instruments and/or metrics such as patient and provider satisfaction, health-related quality of life and overall health status, and quality measures where greater conceptual clarity on the link to disease management goals is also needed. Care in the development of measures and collection of data are as important as statistical methods in assuring useful and reliable findings from disease management evaluation.

REFERENCES

Reference List

Aadalen S. (1998). Methodological challenges to prospective study of an innovation: interregional nursing care management of cardiovascular patients. *Journal of Evaluation in Clinical Practice*, 4: 197–223.

Bailie R, G Robinson, S Kondalsamy-Chennakesavan, S Halpin and Z Wang. (2006). Investigating the sustainability of outcomes in a chronic disease treatment programme. *Social Science and Medicine*, 63: 1,661–1,670.

Beeuwkes Buntin M, A Jain, S Mattke and N Lurie. (2009). Who gets disease management? *Journal of General Internal Medicine*, 24(5): 649–655.

Berwick D. (2008). The science of improvement. *Journal of the American Medical Association*, 299: 1,182–1,184.

Bland JM and DG Altman. (1994). Statistics notes: some examples of regression towards the mean. *British Medical Journal*, 309: 780.

Bodenheimer T. (2000). Disease management in the American market. *British Medical Journal* (26 February), 320: 563–566.

Bodenheimer T. (1999). Disease management: promises and pitfalls. *New England Journal of Medicine*, 340: 1,202–1,205.

Bodenheimer T, E Wagner and K Grumbach. (2002). Improving primary care for patients with chronic illness: the chronic care model, Part 2, *Journal of the American Medical Association,* 288: 1,909–1,914.

Brazier J, R Harper and N Jones. (1992). Validating the SF-36 Health Survey Questionnaire: new outcome measure for primary care. *British Medical Journal*, 305: 160–164.

Campbell DT. (1969). Reforms as experiments. *American Psychologist*, April: 409–429.

Cavanaugh K, RO White and R Rothman. (2007). Exploring disease management programs for diabetes mellitus: proposal of a novel hybrid model. *Disease Management and Health Outcome*s, 15(2): 73–81.

CBO. (2004). *An Analysis of the Literature on Disease Management Programs*. Washington, DC: Congressional Budget Office.

Coulter A and PD Cleary. (2001). Patients' experiences with hospital care in five countries. *Health Affairs*, 20(3): 244–252.

Cousins MS, LM Shickle and JA Bander. (2002). An introduction to predictive modelling for disease management risk stratification. *Disease Management*, 5(3): 157–167.

Cretin S, SM Shortell and EB Keeler. (2004). An evaluation of collaborative interventions to improve chronic illness care: framework and study design. *Evaluation Review*, 28(1): 28–51.

Crosson F and P Madvig. (2004). Does population management of chronic disease lead to lower costs of care? *Health Affairs*, 23(6): 76–78.

Dalton J. (2005). Identifying outcomes of care for the home care patient with diabetes. *Caring*, 24(6): 14–17.

Dehejia RH and S Wahba. (2002). Propensity score-matching methods for nonexperimental causal studies. *The Review of Economics and Statistics*, 84(1): 151–161.

Diamond F. (1999). DM's motivation factor can skew study results. *Managed Care Magazine*, June. Available at http://www.managedcaremag.com/archives/9906/9906.dmstudies.html (accessed on 03/11/09).

DMAA. (2007). *Outcomes Guidelines Report Volume II*. Washington, DC: Disease Management Association of America.

Donabedian A. (1980). *The Definition of Quality and Approaches to its Assessment*. Ann Arbor, MI: Health Administration Press.

Duncan I, M Lodh, G Berg and D Mattingly. (2008). Understanding patient risk and its impact on chronic and non-chronic member trends. *Population Health Management*, 11(5): 261–267.

Ellrodt G, DJ Cook, J Lee, M Cho, D Hunt and S Weingarten (1997). Evidence-based disease management. *Journal of the American Medical Association,* 278: 1,687–1,692.

Enck P and S Klosterhalfen. (2005). The placebo response in functional bowel disorders: perspectives and putative mechanisms. *Neurogastroenteroogy and Motility*, 17: 325–331.

Esposito D, E Fries and M Gold. (2009). Using qualitative and quantitative methods to evaluate small-scale disease management pilot programs. *Population Health Management*, 12(1): 3–15.

Farah JR, K Kamali, J Harner, IG Duncan and TC Messer. (2008). Random fluctuations and validity in measuring disease management effectiveness for small populations. *Population Health Management*, 11(6): 307–316.

Fetterolf, D and T Tucker. (2008). Assessment of medical management outcomes in small populations. *Population Health Management*, 11(5): 233–239.

Fetterolf D, D Wennberg and A Devries. (2004). Estimating return on investment in disease management programs using a pre-post analysis. *Disease Management*, 7(1): 5–23.

Fireman B, J Bartlett and J Selby. (2004). Can disease management reduce health care costs by improving quality? *Health Affairs*, 23(6): 63–75.

Fitzner K, J Sidorov, D Fetterolf, D Wennberg, E Eisenberg et al. (2004). Principles for assessing disease management outcomes. *Disease Management*, 7(3): 191–200.

Gagnon J and R Grenier. (2004). Élaboration et validation d'indicateurs de la qualité des soins relatifs à l'empowerment dans un contexte de maladie complexe à caractère chronique. *Recherche en soins infirmiers*, 76: 50–65.

Goetzel RZ, RJ Ozminkowski, VG Villagra and J. Duffy. (2005). Return on investment in disease management: a review. *Health Care Financing Review*, 26(4): 1–19.

Grey M, K Knafl, R McCorkle. (2006). A framework for the study of self- and family management of chronic conditions. *Nursing Outlook*, 54: 278–286.

Guyatt G, D Osoba, A Wu, K Wyrwich, G Norman and the Clinical Significance Consensus Meeting Group. (2002). Methods to explain the clinical significance of health status measures. *Mayo Clinic Proceedings,* 77: 371–383.

Healthways. (2007). *Measuring the Impact of the Disease Management on Clinical Testing in Health Disparity Zones.* Nashville, TN: Healthways Center for Health Research: 1–4.

Healthways. (2006). *Measuring the Impact of the Diabetes and Cardiac Disease Management Programs on LDL Laboratory Values.* Nashville, TN: Healthways Center for Health Research: 1–4.

Hu G and M Root. (2005). Accuracy of prediction models in the context of disease management. *Disease Management*, 8(1): 42–47.

Huber D, M Sarrazin, T Vaughn and J Hall. (2003). Evaluating the impact of case management dosage. *Nursing Research*, 52(5): 276–288.

Huber D, J Hall and T Vaughn. (2001). The dose of case management interventions. *Case Management*, 6(3): 119–126.

Jack L, L Liburd, T Spencer and C Airhihenbuwa. (2004). Understanding the environmental issues in diabetes self-management education research: a re-examination of 8 studies in community-based settings. *Annals of Internal Medicine*, 140: 964–971.

Jones AM and N Rice. (2009). *Econometric Evaluation of Health Policies*. HEDG Working Paper 09/09. York, UK: Health Econometrics and Data Group of the University of York.

Kerr E, S Krein, S Vijan, T Hofer and R Hayward. (2001). Avoiding pitfalls in chronic disease quality measurement: a case for the next generation of technical quality measures. *The American Journal of Managed Care*, 7: 1,033–1,043.

King, G. (1997). *A Solution to the Problem of Ecological Inference.* Princeton, NJ: Princeton University Press.

Klosterhalfen S and P Enck. (2006). Psychobiology of the placebo response. *Autonomic Neuroscience*, 125; 94–99.

Koller M and E Steyerberg. (2006). An unusual meta-regression approach with high potential of misleading conclusions. *European Heart Journal*, 16 February.

Lamb G and J Stempel. (1994). Nurse case management from the client's view: growing as insider-expert. *Nursing Outlook*, 42: 7–14.

Lemmens K, A Nieboer, C van Schayck, J Asin and R Huijsman. (2008). A model to evaluate quality and effectiveness of disease management. *Quality and Safety in Health Care,* 17: 447–453.

Lewis A. (2009). How to measure the outcomes of chronic disease management. Population Health Management, 12(1): 47–54

Linden A. (2008). Sample size in disease management program evaluation: the challenge of demonstrating a statistically significant reduction in admissions. *Disease Management*, 11(2): 95–101.

Linden A. (2004). Measuring diagnostic and predictive accuracy in disease management: an introduction to receiver operating characteristic (ROC) analysis. *Journal of Evaluation in Clinical Practice*, 12(2): 132–139.

Linden A, JL Adams and N Roberts. (2005a). Evaluating disease management program effectiveness: an introduction to the bootstrap technique. *Disease Management and Health Outcomes*, 13(3): 159–167.

Linden A, JL Adams and N Roberts. (2005b). Strengthening the case for management effectiveness: un-hiding the hidden bias. *Journal of Evaluation in Clinical Practice*, 12(2): 140–147.

Linden A, JL Adams and N Roberts. (2005c). Using propensity scores to construct comparable control groups for disease management program evaluation. *Disease Management Health Outcomes*, 13(2): 107–115.

Linden A, JL Adams and N Roberts. (2004a). Evaluating disease management programme effectiveness adjusting for enrolment (tenure) and seasonality. *Research in Healthcare Financial Management*, 9(1): 57–68.

Linden A, JL Adams and N Roberts. (2004b). Evaluating disease management programme effectiveness: an introduction to survival analysis. *Disease Management*, 7(3): 180–190.

Linden A, JL Adams and N Roberts. (2004c). Evaluating disease management programme effectiveness: an introduction to the regression discontinuity design. *Journal of Evaluation in Clinical Practice*, 12(2): 124–131.

Linden A, JL Adams and N Roberts. (2004d). Generalizing disease management program results: how to get from here to there. *Managed Care Interface*; pp 38–45.

Linden A, JL Adams and N Roberts. (2004e). Using an empirical method for establishing clinical outcome targets in disease management programs. *Disease Management*, 7(2): 93–101.

Linden A, J Adams and N Roberts. (2003a). An assessment of the total population approach for evaluating disease management program effectiveness. *Disease Management*, 6(2): 93–102.

Linden A, JL Adams and N Roberts. (2003b). Evaluating disease management program effectiveness: an introduction to time-series analysis. *Disease Management*, 6(4): 243–55.

Linden A, JL Adams and N Roberts. (2003c). Evaluation methods in disease management: determining program effectiveness, paper prepared for Disease Management Association of America, (October): 1–19.

Linden A, T Biuso, A Gopal, A Barker, J Cigarroa, S Praveen Haranath, D Rinkevich and K Stajduhar. (2007). Consensus development and application of ICD-9 codes for defining chronic illnesses and their complications. *Disease Management and Health Outcomes*, 15(5): 315–322.

Linden A and S Goldberg. (2007). The case-mix of chronic illness hospitalization rates in a managed care population: implications for health management programs. *Journal of Evaluation in Clinical Practice*, 13: 947–951.

Linden A and N Roberts. (2005). A user's guide to the disease management literature: recommendations for reporting and assessing program outcomes. *The American Journal of Managed Care*, 11(2): 113–120.

Linden A and N Roberts. (2004). Disease management interventions: what's in the black box? *Disease Management*, 7(4): 275–291.

Linden A, W Trochim and JL Adams. (2006). Evaluating program effectiveness using the regression point displacement design. *Evaluation and the Health Professions*, December, 29(4): 1–17.

Ling, T. (2009). A framework for understanding the contribution of public services to public benefit, in T Ling and L Villalba van Dijk (eds), *Performance Audit Handbook: Routes to Effective Evaluation.* Cambridge, UK: RAND Europe.

Loveman E, C Cave, C Green, P Royle, N Dunn and N Waugh. (2003). The clinical and cost-effectiveness of patient education models for diabetes: a systematic review and economic evaluation. *Health Technology Assessment*, 7(22): 1–190.

MacStravic S. (2008). Therapeutic specificity in disease management evaluation. *Disease Management*, 11(1): 7–11.

MacStravic S. (2005). *Rethinking the Question "Does Disease Management Work?"* Marblehead, MA: Health Leaders Information to Lead.

Mateo M, K Matzke and C Newton. (2002). Designing measurements to assess case management outcomes. *Case Management*, 7(6): 261–266.

Mattke S, M Seid and S Ma. (2007). Evidence for the effect of disease management: is $1 billion a year a good investment? *American Journal of Managed Care*, 13: 670–676.

Matts JP and JM Lachin. (1988). Properties of permuted-blocked randomisation in clinical trials. *Control Clinical Trials*, 9(4): 327–44.

McKee M, A Britton, N Black, K McPherson, C Sanderson and C Bain. (1999). Interpreting the evidence: choosing between randomised and non-randomised studies. *British Medical Journal*, 319: 312–315.

Motheral B. (2008). 2008: A tipping point for disease management? *Journal of Managed Care Pharmacy*, 14(7): 643–649.

Mulcahy K, M Maryniuk, M Peeples, M Peyrot, D Tomky, T Weaver and P Yarborough (2003). Standards for outcomes measurement of diabetes self-management education. Position statement of the American Association of Diabetes Educators (AADE). *The Diabetes Educator,* 29(5): 804–815.

Mulligan K, S Newman, E Taal, M Hazes, J Rasker and OMERACT 7 Special Interest Group. (2005). The design and evaluation of psychoeducational/self-management interventions. *Journal of Rheumatology*, 32(12): 2,470–2,474.

Newman M, G Lamb and C Michaels. (1991). Nursing case management: the coming together of theory and practice. *Nursing and Health Care*, 12: 404–408.

Niu K, L Chen, Y Liu and H Jenich. (2009). The relationship between chronic and non-chronic trends. *Population Health Management*, 12(1): 31–38.

Nolte E and M McKee (eds) (2008). *Caring for People With Chronic Conditions: A Health System Perspective*. Maidenhead: Open University Press/McGraw Hill Education.

Nolte E, C Knai and M McKee (eds) (2008). *Managing Chronic Conditions: Experience in Eight Countries.* Copenhagen: World Health Organization on behalf of the European Observatory on Health Systems and Policies.

Norman GK. (2008). Disease management outcomes: are we asking the right questions yet? *Population Health Management*, 11(4): 183–187.

Norris S, P Nichols, C Caspersen, R Glasgow, M Engelgau et al. (2002). The effectiveness of disease and case management for people with diabetes: a systematic review, *American Journal of Preventive Medicine,* 22: 15–38.

Orkin F and S Aruffo. (2006). Achieving breakthrough outcomes: measurable ROI. *The Case Manager*, September/October: 50–58.

Outcomes Consolidation Steering Committee. (2005). *Disease Management Program Guide: Including Principles for Assessing Disease Management Programs*. Washington, DC: Disease Management Association of America: 1–33.

Pawson R and N Tilley. (1997). *Realistic Evaluation*. London: SAGE Publications.

Pearson ML, S Mattke, R Shaw, MS Ridgely and SH Wiseman. (2007). *Patient Self-management Support Programs: An Evaluation*. Final Contract Report for Agency for Healthcare Research and Quality, No. 08-0011.

Plocher DW. (2006). Measurement evolution: better support for the doctor-patient relationship. *Minnesota Physician*, December, 20(9).

Powell Davies G, AM Williams, K Larsen, D Perkins, M Roland and M Harris. (2008). Coordinating primary health care: an analysis of the outcomes of a systematic review, *Medical Journal of Australia,* 188 (8 Suppl): S65–S68.

Rothman A and E Wagner. (2003). Chronic illness management: what is the role of primary care? *Annals of Internal Medicine*, 138: 256–261.

Schmidt-Posner J and J Jerrell. (1998). Qualitative analysis of three case management programs. *Community Mental Health Journal,* 34(4): 381–392.

Schünemann HJ, Oxman AD, Vist GE, Higgins JPT, Deeks JJ, Glasziou P and Guyatt GH. (2008). Interpreting results and drawing conclusions, in JPT Higgins and S Green (eds), *Cochrane Handbook for Systematic Reviews of Interventions* Version 5.0.1. The Cochrane Collaboration, 2008. Available at: http://www.cochrane-handbook.org (accessed on 14/12/10).

Sepucha K, F Fowler and A Mulley. (2004). Policy support for patient-centered care: the need for measurable improvements in decision quality. *Health Affairs*. (October): 54–62.

Serxner S, K Baker and D Gold. (2006). Guidelines for analysis of economic return from health management programs. *American Journal of Health Promotion*, 20(6): Suppl 1–17.

Serxner S, S Mattke, S Zakowski and D Gold. (2008). Testing the DMAA's recommendations for disease management program evaluation. *Population Health Management*, 11(5): 241–245.

Shadish WR, TD Cook and DT Campbell. (2002). *Experimental and Quasi-experimental Designs for Generalised Causal Inference*. Boston: Houghton Miffin.

Sibbald B and M Roland. (1998). Understanding controlled trials: why are randomised controlled trials important? *British Medical Journal* (17 January), 316: 201.

Sonnenberg A, J Inadomi and P. Bauerfeind. (1999). Reliability block diagrams to model disease management. *Medical Decision Making*, 19: 180–185.

Spertus, J. (2008). Evolving applications for patient-centered health status measures. *Circulation*, 118: 2,103–2,110.

Steuten L, K Lemmens, A Nieboer and H Vrijhoef. (2009). Identifying potentially cost effective chronic care programs for people with COPD. *International Journal of COPD*, 4: 87–100.

Steuten L, S Palmer, B Vrijhoef, F van Merode, C Spreeuwenberg and H Severens. (2007). Cost-utility of a disease management for patients with asthma. *International Journal of Technology Assessment in Health Care*, 23(2): 184–191.

Steuten L, H Vrijhoef, H Severens, F van Merode and C Spreeuwenberg. (2006). Are we measuring what matters in health technology assessment of disease management? Systematic literature review. *International Journal of Technology Assessment in Health Care*, 22(1): 47–57.

Steuten L, H Vrijhoef, G van Merode, J Severens and C Spreeuwenberg. (2004). The health technology assessment-disease management instrument reliably measured methodologic quality of health technology assessments of disease management. *Journal of Clinical Epidemiology*, 57; 881–888.

Stevenson K, R Baker, A Farooqi, R Sorrie and K Khunti. (2001). Features of primary health care teams associated with successful quality improvement of diabetes care: a qualitative study. *Family Practice*, 18(1): 21–26.

Susser E, S Schwartz, A Morabia and EJ Bromet. (2006). *Psychiatric Epidemiology: Searching for the Causes of Mental Disorders*. Oxford: Oxford University Press.

Tinkelman D and S Wilson. (2008). Regression to the mean: a limited issue in disease management programs for chronic obstructive pulmonary disease. *Disease Management*, 11(2): 103–110.

Tinkelman D and S Wilson. (2004). Asthma disease management: regression to the mean or better? *The American Journal of Managed Care*, 10(12): 948–954.

Velasco-Garrido M, Busse R and Hisashige A (2003). *Are Disease Management Programmes (DMPs) Effective in Improving Quality of Care for People with Chronic Conditions?* Copenhagen: WHO Regional Office for Europe.

Villagra V. (2004). Strategies to control costs and quality: a focus on outcomes research for disease management. *Medical Care*, 42(4): 24–30.

Villagra V and T Ahmed. (2004). Effectiveness of a disease management program for patients with diabetes: testing the impact on health care quality, use and spending shows that disease management has many positive effects. *Disease Management*, 23(4): 255–266.

Vlad S and M LaValley. (2008). Intention-to-treat analysis may better represent the actual efficacy. *Archives of Internal Medicine*, 168(11): 1,228.

Weaver T, P Tyrer, J Ritchie and A Renton. (2003). Assessing the value of assertive outreach: qualitative study of process and outcome generation in the UK700 trial. *British Journal of Psychiatry*, 183: 437–445.

Webb P, C Bain and S Pirozzo. (2005). *Essential Epidemiology: An Introduction for Students and Health Professionals*. Cambridge: Cambridge University Press.

Wei LJ and JM Lachin. (1988). Properties of the urn randomisation in clinical trials. *Control Clinical Trials*, 9(4): 345–364.

Weiner M and J Long. (2004). Cross-sectional versus longitudinal performance assessments in the management of diabetes. *Medical Care*, 42(2): 34–39.

Weingarten SR, Henning JM, Badamgarav E, Knight K, Hasselblad V, Gano A Jr and Ofman JJ (2002). Interventions used in disease management programmes for patients with chronic illness – which ones work? Meta-analysis of published reports. *British Medical Journal*; 325 (7370): 925–928.

Wendel J and D Dumitras. (2005). Treatment effects model for assessing disease management: measuring outcomes and strengthening program management. *Disease Management*, 8(3): 155–168.

Wilson T and M MacDowell. (2003). Framework for assessing causality in disease management programs: principles. *Disease Management*, 6(3): 143–158.

Wilson T, M MacDowell, P Salber, G Montrose and C Ham. (2008). Evaluation methods in disease management studies 2004–07. *Disease Management Health Outcomes*, 16(5): 365–373.

Woolf SH and RE Johnson. (2005). The break-even point: when medical advances are less important than improving the fidelity with which they are delivered. *Annals of Family Medicine*, 3: 545–552.

Wyrwich K, W Tierney, A Babu, K Kroenke and F Wolinsky. (2005). A comparison of clinically important differences in health-related quality of life for patients with chronic lung disease, asthma, or heart disease. *Health Services Research*, 40(2): 577–591.

Yudkin PL and IM Stratton. (1996). How to deal with regression to the mean in intervention studies. *The Lancet*, 347 (January 27): 241–243.

Zajac, B. (2004). Performance measurement of disease management programs. *Disease Management Health Outcomes*, 12(4): 211–220.

APPENDICES

Appendix A: Summary of Literature Search Strategy

Table 9 summarises our search term combinations for the three electronic databases we searched systematically. We note that our search strategy for CINAHL and Web of Science databases did not use the same structured combination because the aim was to saturate our identification of disease management evaluation methodology papers concerned with non-financial indicators of effect.

Table 9 Summary of search terms

	Evaluation		Disease management		Indicator of effect	Limits
PubMed (MeSH terms only)	"Evaluation Studies as Topic/method" OR "Health Care Evaluation Mechanisms"	AND	"Disease Management" OR "Self Care" OR "Case Management"	AND	"Quality Indicators, Health Care" OR "Outcome Assessment (Health Care)" OR "Treatment Outcome" OR "Outcome and Process Assessment (Health Care)"	Publication date from 1990/01/01 to 2009/09/31; Humans; English, German, French, Spanish
CINAHL	1. SU program* evaluation and SU design or SU method* 2. SU program* evaluation 3. SU evaluation method* 4. SU assessment design 5. SU assessment method* 6. SU evaluation design	and	1. SU disease management 2. SU disease management program* 3. SU chronic disease management	and	1. SU quality of life 2. SU success 3. SU measure* 4. SU indicator* 5. SU effective* 6. SU outcome	Publication year from: 1990-2009; Peer reviewed; exclude MEDLINE records; publication type: abstract, journal article, review, statistics, systematic review; Language: English, French, German, Spanish
Web of Science	1."Evaluation" SAME "method*" 2. "evaluation" design OR method* 3. "evaluation" 4. "evaluation" methods 5. "assessment" OR "methods"	SAME	1. "chronic disease" SAME "management program*" 2. "disease management program*" 3. "disease management" 4. "disease management" OR "self care" OR "case management"	SAME	1. "acceptability" 2. "feasibility" 3. "acceptability" OR "feasibility" 4. "care coordination" 5. "clinic* measure" 6. "clinic* measure*" 7. "medical parameter*" 8. "cost*" 9. "effective*" 10. "outcome" 11. "quality" 12. "self-management" 13. "satisfaction" 14. "disease control" 15. "quality of life" 16. "success" OR "measure*" OR outcome indicator 17. "measure*" 18. "outcome" OR "quality of life" OR "disease control" OR clinical measure 19. "outcome*" OR "indicator*" OR "measure*" 20. "outcome*"	Timespan=1990 -2009; Databases=SCI-EXPANDED, SSCI, SPCI-S, SPCI-SSH

Appendix B: Methods of Randomisation and Types of Control Groups

Types of Randomisation

Randomisation is the statistically equal chance (50/50) of participants being assigned to different groups, either an intervention group or a control group. It involves two inter-related processes: a randomisation procedure and allocation concealment (Susser et al., 2006). Types or methods include the following:

- complete randomisation
- permuted block randomisation, blocked-randomisation[3]
- multi-site cluster randomisation
- urn randomisation[4]
- covariate-adaptive randomisation[5]
- outcome-adaptive randomisation.[6]

It is important to consider several statistical issues implicated in the randomisation procedure:

- Balance. As most statistical tests will have greater power when the groups being compared have equal sizes, the randomisation procedure should aim to generate similarly sized groups (imbalance is usually a concern for small samples sizes under 200).

[3] The blocked-randomisation method can be subject to selection bias when block sizes are not masked and the investigator can discern the next assignment (Matts and Lachin, 1988). The authors describe what tests should be employed when blocked-randomisation is used, such as the block-stratified Mantel-Haenszel chi-square test for binary data, the blocked analysis of variance F test, and the blocked non-parametric linear rank test.

[4] The urn design forces a small-sized study to be balanced but approaches complete randomisation as the size (n) increases and is not as vulnerable to experimental bias as are other restricted randomisation procedures. Also, in the urn design permutational distribution, post-stratified subgroup analyses can be performed. For more details, see Wei and Lachin (1988).

[5] When many variables influence outcomes, balance across these variables can be achieved with a separate list of randomisation blocks for each combination of values; the statistical method called minimisation can be used when the number of possible values is large.

[6] Randomisation probabilities are adjusted continuously throughout a study in response to the data when evidence may accumulate that one treatment is superior, placing an ethical demand on more patients being assigned to the superior treatment.

- **Selection bias**. The structure of the randomisation procedure can affect whether investigators or clinicians can infer the next group assignment by guessing which of the groups has been assigned the least at that point. The effect of breaking allocation concealment can lead to bias in the selection of participants for enrolment in the study.
- **Accidental bias**. Randomisation procedures can affect the magnitude of potential bias in estimates that result from statistical analysis in which important covariates related to the outcome are ignored.

It is understood that proper analyses of study results should employ statistical tests that incorporate the properties of the method used for randomisation.

Types of Control Groups

Equally important for subsequent data analysis and for establishing a causal finding in an association of interest (chronic care management intervention–outcomes relationships) is the type of control group. The type of control group chosen will be critical to simulate accurately the counterfactual of what would have happened to the participants had they not received the disease management programme; therefore controls must be taken from the same reference population as those individuals who are exposed to the intervention.

Ultimately, defining the boundaries of the reference population is a matter of choosing between either a "well control" or an "unwell control" (Susser et al., 2006). The choice of well or unwell controls has different implications for sampling controls from the same underlying cohort or population as the intervention group regarding potential introduction of bias that could threaten validity of findings (see Susser et al., 2006, on choosing controls).

Broadly, the chronic care management research literature offers the following possible types of control groups also used in other public health research areas:

- no treatment or unexposed concurrent control group
- dose-response concurrent control group
- active concurrent control group
- historical control
- population control
  - hospital control
  - neighbourhood control.

To give some examples for illustration, historical control groups have been used in a sequential pre-test and post-test comparison, which includes phased implementation of a chronic care management programme so as to create the conditions of a "natural experiment".

Population controls are an ideal reference group in public health research and evaluation and these can be selected from one of the following sources:

- population registers
- comprehensive electoral rolls
- residential telephone numbers randomly sampled (random-digit dialling)
- patient lists of GPs
- neighbourhood postal codes.

For practical reasons, studies often use hospital controls, which are individuals admitted to the same hospital but for conditions other than the one being studied. However, a key drawback of using hospital controls is that individuals in this reference group are also ill and therefore are different from the general population, which includes both healthy and ill people – a difference that threatens the external validity of study findings (Susser et al., 2006).

Appendix C: Determining Statistical Significance A Priori

The four inter-related parameters that must be determined at the evaluation planning stage are: alpha, power, sample size and effect size. Several papers in the disease management evaluation literature describe the different calculations for the parameters that will set the level of statistical significance for possible results to be credible. Importantly, these papers clearly demonstrate how modifying any one of these parameters can have drastic effects on possible study results, in particular estimates of cost savings (Farah et al., 2008; Linden, 2008).

Although the literature on determining statistical significance assumes, and implicitly argues for, the evaluation to be done on a prospectively identified panel of disease management participants and non-participants, there may also be ex post evaluations where data (e.g. administrative data) are being retrospectively analysed and there is little scope to influence the sample size and power. An ex post evaluation may not need to obtain results that are statistically significant if the point estimate of effect reaches operational or policy relevance in a particular context. Nonetheless, it would still be important to make an a priori assessment of the power and sample size required to reach the desired relevance level and to compare this with the scientific approach of levels of statistical significance.

Alpha and Power

The only statistical significance parameters under the control of an evaluator are alpha and power. For both parameters, it is critical to set the values for each at the outset of the evaluation before any effect is measured. The power parameter is often set at 80 percent, or even higher at 90 percent. The level of significance, or value of alpha, is generally set at 95 percent, but can be 90 percent or 99 percent, depending on how much error the evaluator may be willing to accept in detecting an effect when there is truly is none. Alternatively, some would advocate using confidence intervals instead of alpha (or, p-values) as these are often easier to understand and interpret than a p-value of 0.05 (Susser et al., 2006).

Minimum power is critical to determining sample size, but subgroup analysis is often conducted to assess whether program effects are comparable across relevant strata in a population. Thus, power calculations and related sample size will need to consider the likely implementation of stratified analyses. Some argue that estimates of adequate power for many studies are off target, since the data will always be disaggregated to conduct subgroup analyses to the maximum extent allowed by the data.

Sample Size

Determining the sample size of a disease management evaluation is an integral component of the needs analysis, which should be performed during evaluation planning (Linden, 2008). It is important to determine whether the given population of a given intervention is of sufficient size to allow the expected programme effect to achieve statistical significance. The larger the group size being evaluated, the less chance that random fluctuations in measures will overwhelm programme effects such as real savings (Farah et al., 2008). When a disease management programme is implemented and evaluated using an intervention and control group, the sample size calculation estimates the number of participants required for each group (Linden, 2008).

The author suggests that sample size calculations should be performed at the unit level of the individual (e.g. admissions per person per unit of time) for two reasons: (1) in population-based analyses this basis of measure is less influenced by population turnover and length of enrolment; and (2) this is the standard method of reporting outcome comparisons between two or more groups (cohorts). Moreover, sample size calculations are one of three analyses that should be conducted before evaluating a given intervention. If cost savings are the expected effect, then the other two analyses would be: (1) a time-series analysis of historic utilisation rates in the target chronically ill population to determine if the level and trending patterns suggest an opportunity for reduction; and (2) a number to decrease analysis to determine the number of hospital admissions that must be reduced by the programme to achieve a ROI (Linden, 2008).

Sample size is a function of four parameters:

- significance level: the probability of finding a significant difference when there truly is one, false positive
- power: the probability of not finding a significant difference when there truly is one, false negative
- effect size: the magnitude of change between two groups or within one group, pre- and post-programme
- standard deviation: the degree to which observations are spread out around a mean in a given data set.

As noted above, subgroup analysis will require a higher level of power and thus larger sample sizes, which is an important consideration for this parameter. In addition, sample size can also be influenced by the type of randomisation used in an evaluation design. For example, block randomisation methods can result in correlated errors among subjects that increase the standard errors relative to simple random sampling and samples need to be larger than in simple random sampling.

Another added complication to determining sample size is the fact that standard deviation is often assumed to follow a normal distribution. However, in the disease management context where the outcome measured is often a rare event such as hospitalisations or emergency department visits, the variable does not follow a normal distribution and therefore the use of statistical methods for normal distribution to evaluate such rare event outcomes brings the risk of increasing a type II error. As a result, it has been suggested that, to calculate standard deviation in disease management evaluations, outcomes should be expressed as a *rate* (a value expressed relative to a population size for a number of rare events occurring over a given period of time). The reason for expressing evaluation outcome measures as rates is that they follow a Poisson distribution which has the benefit that the variance is equal to the mean (Linden, 2008). The reader is referred to Table 2 in Linden (2008) for estimated sample sizes based on starting admission rate per person and predicted effect size.

Effect Size

It is important for a disease management evaluator to *anticipate* the expected difference of means between the intervention group and the control group in order to assess whether the measured effect size is meaningful or reasonable. This fourth parameter is effectively what the evaluator is measuring (Linden, Adams and Roberts, 2004e). Hence, Linden, Adams and Roberts (2004e) dedicate a paper to illustrate how setting different sizes of an expected effect will alter the other parameters – the smaller the expected effect to be detected, the larger the sample size must be; a larger effect size will result in greater power, and so on. In some cases, there may be a review body that pre-determines the level of effect (e.g. 10 percent reduction in performance gap between years one and two), or a disease management purchaser will want a threshold value of cost-saving to be reached, setting a given effect size for which it will be critical to calculate the right size sample beforehand (Linden, Adams and Roberts, 2004e).

During implementation of an evaluation, disease management evaluators might consider using principles of statistical quality control such as time series analyses, which can show statistical significance *before* routine statistical tests are passed. Fetterolf and Tucker (2008) give the example of cumulative sum plots that compute the summed difference of each data point in a series from the average expected result in order to detect graphically early on whether a shift in the mean is occurring. The authors note that the use of group sequential analysis techniques for continuous monitoring is growing in clinical trials because of their improved sensitivity over historical statistical methods.

Appendix D: Five Principles of Generalised Causal Inference

Generalisations of disease management evaluation results must be based on more than face validity and common sense and there are five guiding principles for extrapolating causal inferences to populations other than the one studied (Shadish, Cook and Campbell, 2002). In a paper dedicated to threats to external validity, Linden, Adams and Roberts (2004d) briefly remind disease management evaluators of these principles:

- **Surface similarity**. This principle is the general assumption that results for one kind of person, setting, intervention or outcome can be generalised to elements that appear similar in important characteristics (e.g. a disease management nurse is expected to elicit a similar response from all individuals with the same disease).

- **Ruling out irrelevancies**. This principle pertains to the identification of variables among persons, intervention, settings and outcomes that are irrelevant to the size or direction of the cause–effect association (e.g. rural–urban location may be deemed irrelevant to medication adherence among individuals with diabetes).

- **Making discriminations**. This principle involves discriminating between variations of a factor that are all considered to have a similar cause–effect relationship with a particular outcome when, in reality, one or more of these variations might change the direction and magnitude of the causal association (e.g. a computerised telephonic response system may not elicit the same results as a nurse-led telephonic dialogue as presumed although both types of interventions have initially been classified as "telephonic interventions").

- **Interpolation–extrapolation**. This principle allows for inferences to be drawn regarding values both within the range of observations (interpolation) and outside the range of observations (extrapolation). Accuracy of interpolation is improved when a given intervention is evaluated for its impact on a broad range of observations of the given factor. Extrapolation accuracy is stronger when calculated values are relatively close to the upper and lower boundaries of only the actual range of observations since it is more difficult to estimate the outcome of the next observation the further away one starts from the last known observation.

- **Causal explanation**. This principle ties all factors, including underlying mediating processes, together by explaining how the given intervention affects participants and achieves the desired impact. Given the innumerable sources of variation at every level of every factor, a proper investment in resources is imperative to analyse fully the causal relationships that exist within any one disease management programme dimension.