Evaluating chronic disease management

Recommendations for funders and users

Ellen Nolte, Annalijn Conklin, John L. Adams,
Matthias Brunn, Benjamin Cadier, Karine Chevreul,
Isabelle Durand-Zaleski, Arianne Elissen,
Antje Erler, Maria Flamm, Anne Frølich,
Birgit Fullerton, Ramune Jacobsen, Cécile Knai,
Robert Krohn, Boris Pöhlmann,
Zuleika Saz Parkinson, Antonio Sarria Santamera,
Andreas Sönnichsen, Hubertus Vrijhoef

RAND EUROPE

The reseach described in this document was prepared for the European Commission on behalf of the DISMEVAL Consortium.

RAND Europe is an independent, not-for-profit research organisation whose mission is to improve policy and decision making for the public good. RAND's publications do not necessarily reflect the opinions of its research clients and sponsors.

RAND® is a registered trademark.

Preface

This report presents the sixth formal deliverable (D6) to the European Commission of the DISMEVAL project (grant agreement 223277). DISMEVAL is a European collaborative project with the overarching aim to contribute to developing new research methods and to generating the evidence base to inform decisionmaking in the field of chronic disease management evaluation.

We here report on the overall findings of the project, bringing together the findings of work undertaken in DISMEVAL to derive evidence-based recommendations for such evaluation approaches that are relevant to planned and ongoing policies at the EU and wider European level and internationally.

The report was prepared by the project Consortium, with coordinating support by RAND Europe.

RAND Europe is an independent not-for-profit policy research organisation that aims to improve policy and decisionmaking in the public interest, through research and analysis. RAND Europe's clients include European governments, institutions, NGOs and firms with a need for rigorous, independent, multidisciplinary analysis. This report has been peer-reviewed in accordance with RAND's quality assurance standards.

The corresponding author for this report is Dr Ellen Nolte; for further information please contact:

Ellen Nolte
RAND Europe
Westbrook Centre
Milton Road
Cambridge CB4 1YG
United Kingdom
Tel. +44 (1223) 353 329
enolte@rand.org

Contents

List of Figures

List of Tables

List of Boxes

DISMEVAL Consortium

Beneficiary Number	Beneficiary name	Beneficiary short name	Country
1(coordinator)	RAND Europe Cambridge Ltd	RAND	UK
2	London School of Hygiene and Tropical Medicine	LSHTM	UK
3	Paracelsus Medizinische Privatuniversitat Salzburg	PMU	Austria
4	Københavns Universitet	UCPH	Denmark
5	Johann Wolfgang Goethe Universitaet Frankfurt am Main	GUF	Germany
6	Universite Paris XII – Val de Marne	UPVM	France
7	Universiteit Maastricht	UM	The Netherlands
8	Instituto de Salud Carlos III	ISCIII	Spain
9	Centre Anticancereux Léon Bérard	CLB	France
10	AQUA-Institut für angewandte Qualitätsförderung und Forschung im Gesundheitswesen GmbH	AQUA	Germany

Acknowledgements

Introduction

Responding to the burden of chronic disease presents challenges for all health systems. As populations age and advances in healthcare allow those with once-fatal conditions to survive, the prevalence of chronic conditions is rising in many countries.[1] In the European Union, in 2006, between 20 and over 40 per cent of the population aged 15 years and over reported a long-standing health problem and one in four currently receives medical long-term treatment.[2]

The implications for health systems and society as a whole are considerable. Chronic diseases pose a sizeable burden for national economies, with some studies estimating the associated costs at up to seven per cent of a country's gross domestic product.[3] Societal costs arise partly as a result of direct healthcare costs, including those associated with healthcare utilisation, medication and potentially costly interventions, alongside indirect costs caused, for example, by decreased work productivity.[4] These challenges require effective measures to prevent disease through reducing the major chronic disease risk factors and addressing influences that drive exposure, while also providing services to meet the requirements caused by chronic health problems, so ensuring that people with established illnesses can participate in society.

Structured disease management has been proposed as a means to improve the quality and reduce the cost of healthcare, and to improve health outcomes for those with chronic conditions. However, while intuitively appealing, the evidence on the ability of structured approaches to managing chronic disease to actually do so remains uncertain. What we know about the impact of chronic disease management is mainly based on small studies of high-risk patients, often undertaken in academic settings.[5] One important reason for the limited evidence is the lack of widely accepted methods for the evaluation of disease management, and indeed complex health interventions more generally[6-7], that allow for the attribution of observed effects to a given intervention and that are scientifically sound yet operationally feasible. This is, in part, because of the trade-offs between scientific rigour and practicability that frequently have to be balanced in routine operations. Thus, pilot programmes, in particular in academic settings, are frequently implemented and evaluated as randomised controlled trials. However, where such a programme is being rolled out there is typically little incentive, after completion of the pilot phase, for researchers to continue tracking a programme's impact with less rigorous observational research designs. Conversely, evaluation approaches that are used in routine practice tend to be limited methodologically, so reducing the credibility and usefulness of findings of programme effect. Overall, there is a need to better understand the effects of large, population-based

programmes using widely accepted evaluation methods that are scientifically sound and are also practicable in routine settings. Such evaluation methods are a precondition for the selection of effective and efficient interventions to address the growing burden of chronic disease.

1.1 The DISMEVAL project aimed to develop and validate methods and metrics for the evaluation of disease management

DISMEVAL (Developing and validating DISease Management EVALuation methods for European healthcare systems), a European collaborative project, aimed to support this process through contributing to developing new research methods and to generate the evidence base to inform decisionmaking in the field of chronic disease management evaluation. Specifically, the DISMEVAL project sought (a) to enhance our understanding of the use of various approaches to the evaluation of disease management in Europe and to identify examples of best practices and lessons learned; (b) to provide evidence-based recommendations to policymakers, programme operators and researchers on evaluation approaches that may be most useful in a given context; (c) to promote and support the networking and coordination of research and innovation activities relating to scientific knowledge and policy development in this area, building on existing work carried out in Member States and at the wider European level; and (d) to analyse scientific knowledge and developments as well as actions and policies within EU Member States, develop tools to assist policy analysis, and work in close collaboration with the Commission services, networks and experts in this area, and with stakeholder groups and various agencies to provide scientific input to support ongoing and planned actions and policies in the European Union.

This was to be achieved through a programme of work that can be divided into three phases: (1) review of current approaches to the implementation and evaluation of chronic disease management programmes; (2) testing and validation of methods and metrics for chronic disease management evaluation; and (3) development and recommendations for methods and metrics for chronic disease management evaluation. We here briefly outline each of the three phases and the work packages carried out therein.

Phase 1: Review of current approaches to disease management implementation and evaluation in Europe

Phase 1 comprised three work packages that sought (1) to review approaches to managing chronic conditions that have been developed and/or implemented in different countries in Europe, and to assess whether and how countries evaluate approaches to chronic disease management; (2) to provide an overview of the types of evaluation approaches that are being used in Europe to estimate the impact of structured approaches to disease management on the cost and quality of chronic illness care; and (3) to assess the overall policy framework for chronic disease management in selected European countries.

Phase 2: Testing and validation of methods and metrics for the evaluation of disease management

Phase 2 of the project included six work packages; the main aim of this phase was to utilise data from existing chronic disease management programmes, or their equivalent, in partner countries, in order to test and validate different evaluation options reviewed in Phase 1 of

the project. The countries included were Austria, Denmark, Germany, France, the Netherlands, and Spain.

Phase 3: Development and recommendations for methods and metrics for the evaluation of disease management

Phase 3 sought to summarise the findings and to present best practice and lessons learned from work undertaken in phase 2 of the programme. The present report presents the output of phase 3 of the DISMEVAL project; its specific aims are detailed further in the next section.

1.2 This report summarises our findings to inform those considering similar evaluations in the future

The overarching aim of this report is to bring together the findings of work undertaken in DISMEVAL to derive evidence-based recommendations for disease management evaluation approaches that are relevant to planned and ongoing policies at the EU and wider European level and internationally. Further details of work that has informed this volume is available from Conklin and Nolte (2010)[8] and the DISMEVAL Final Report.[9]

The work presented here is based on the premise that policymakers, programme operators and researchers need validated approaches for disease management evaluation to design effective, efficient and equitable interventions to improve the care for people with chronic conditions. Of particular relevance here are two components of evaluation: (1) the performance indicators that are used to capture the impact of a given programme and (2) the attribution strategy, which enables identifying any observed changes in selected indicators as programme effect rather than as a consequence of other factors, such as secular trends or changes in the treatment (ie establishing the counterfactual – could the change have been observed in the absence of the programme). These methods need to be scientifically sound but also operationally feasible to address common threats to validity in disease management evaluation.

The report aims to explain choices, options and trade-offs to policymakers, programme operators and researchers based on analyses undertaken within the DISEMEVAL project. At the outset it is important to note that many of the issues discussed in this report are not specific to disease management evaluation but can be seen to apply to any evaluation of complex interventions in healthcare. However, there are specific concerns around evaluation design and metrics that are of relevance to disease management evaluation and which will be given particular attention here. Thus, we begin, in Chapter 2, by setting out the context for evaluating disease management, exploring the reasons for undertaking evaluation in the first place and explaining some of the underlying principles for doing so. Chapter 3 examines methods and metrics of disease management evaluation, focusing specifically on themes that have emerged as being pertinent to work carried out within the DISMEVAL project and which are set against the background of the general literature on disease management evaluation. Chapter 4 explores practical considerations for disease management evaluation, based on experience of work undertaken in DISMEVAL, while Chapter 5 explores some of the broader challenges and lessons learned that may be relevant for policymakers, funders and practitioners interested in the use and usefulness of disease

management evaluation more generally. We close with Chapter 6, which draws together the evidence compiled in this report and identifies themes for future work.

The context for evaluating chronic disease management

Chronic conditions are defined by the World Health Organization (WHO) as requiring 'ongoing management over a period of years or decades' and cover a wide range of health problems.[10] The goals of chronic care are not to cure but to enhance functional status, minimise distressing symptoms, prolong life through secondary prevention and enhance quality of life.[11] These goals are unlikely to be accomplished through the traditional acute, episodic model of care that is characterised by fragmentation of service delivery across professions and institutions. Indeed, chronic conditions frequently go untreated or are poorly controlled until more serious and acute complications arise. Even when chronic conditions are recognised, there is often a large gap between evidence-based treatment guidelines and current practice.[12] Thus, an effective response to the rising burden of chronic disease will require healthcare delivery models that are characterised by collaboration and cooperation among health professionals over an extended period of time and that place patients at the centre as co-producers of care to optimise health outcomes.[13]

Against this background, health professionals, policymakers and institutions in many countries are initiating new models of service delivery designed to improve the quality and reduce the cost of healthcare, and ultimately improve health outcomes for the chronically ill.[14-16] We have previously shown how countries in Europe vary in their attempts to do this, with many implementing some form of (chronic) disease management – although the nature and scope of related approaches differ.[17] Germany and the Netherlands, for example, have introduced large-scale, population-based structured care or disease management programmes while others are experimenting with smaller-scale care approaches, although this is now changing.[15]

However, the available evidence on the value of different approaches remains uncertain, as does the evidence of what works in which contexts and for which populations.[18] This is in part because of the variety of terms and concepts that are used to describe efforts to improve chronic illness care and its components. There remains considerable need for further efforts to improve the scientific rigour of evaluating these approaches and the reporting of the results of such interventions,[14] which tend to be complex in nature and scope, with several interrelated components often acting at different levels of service delivery.[7]

In this chapter, we explore the rationale for undertaking rigorous evaluation of chronic disease management and related approaches and set out some of the principles of

evaluation. We conclude by briefly setting out some of the principle challenges related to evaluation design and implementation in Europe.

2.1 Disease management can be conceptualised in different ways

The DISMEVAL project focused on approaches that can be broadly subsumed under the heading of 'disease management' although it is important to recognise that conceptualisations of this notion vary widely.[18-19] Definitions range from "discrete programs directed at reducing costs and improving outcomes for patients with particular conditions"[20] to "a population-based systematic approach that identifies persons at risk, intervenes, measures the outcomes, and provides continuous quality improvement".[21] Ellrodt et al. (1997) defined disease management as "an approach to patient care that coordinates medical resources for patients across the entire delivery system"[22] while more recently, the Care Continuum Alliance (previously Disease Management Association of America, DMAA) defined disease management as "a system of coordinated health care interventions and communications for populations with conditions in which patient self-care efforts are significant".[23] The term 'disease management' is now frequently used interchangeably with other terms such as care management, case management and multidisciplinary care, among others, although these are conceptually different, at least in their origins. Disease management, by definition, traditionally targets patient groups with specific conditions such as diabetes, while case management, for example, is aimed more broadly at people with complex needs that arise from multiple chronic conditions, coupled with increasing frailty at old age.[18] However, as more recent definitions of disease management explicitly adopt a broader view towards a population-based approach that addresses multiple care needs, boundaries between concepts are becoming increasingly blurred.[24]

Within the DISMEVAL project we defined disease management as comprising the following components: (a) collaborative models of care among providers such as physicians, hospitals, laboratories and pharmacies; (b) patient education; and (c) monitoring/collection of patient outcome data for the early detection of potential complications.[19] According to this definition, disease management does not normally involve general coordination of care. It also does not normally include preventive services such as flu shots. However, as approaches that are being implemented and tested across Europe may not fully meet this definition, we also sought to capture the range of models that use a subset of disease management interventions or else are conceptualised in a different way while pursuing the same objective, ie to improve the care for those with chronic disease. We therefore considered a wider range of approaches that we termed 'chronic disease management' or chronic care. An overview of the range of approaches considered within the DISMEVAL project is provided in a separate publication reporting on the overall findings of the project.[25]

2.2 What is disease management evaluation for?

Although the introduction of new models of service delivery to better meet the needs of those with chronic health problems may be viewed as 'a good thing' in itself, against a

background of resource constraints in the healthcare sector, policymakers, healthcare managers, planners and funders or purchasers increasingly want to know whether implementing such approaches indeed provides a 'good investment' by improving the quality and reducing the cost of healthcare, and, ultimately, improving health outcomes for the chronically ill.

Knowledge about the effectiveness ('does it work?') and, in some settings, efficiency ('is it worth it?') through (economic) evaluation of healthcare programmes and interventions has become an important component of decisionmaking on the (public) funding of new health technologies and, more recently, wider public health interventions. Evaluation aims to help in understanding the contribution made by a given intervention to achieving particular outcomes.[26] Accordingly, evaluation of structured approaches to disease management may pursue a range of objectives, which will vary with the nature and scope of interventions that are being implemented. In some settings there may be an emphasis on assessing the (longer-term) economic impact of the intervention; often the interest is to determine whether costs were saved or whether the intervention yielded a positive return on investment. Other perspectives may emphasise the quality improvement achieved through structured disease management by assessing processes of care (eg adherence to clinical or practice guidelines, referral rates) and short-term outcomes such as disease control or satisfaction of programme participants.

Whatever the emphasis that is being pursued by a given evaluation, it must be able to provide information on whether or not observed effects (however defined) are indeed attributable to the intervention under consideration and whether these could have been observed in the absence of the intervention (the counterfactual). Existing research on structured chronic disease management provides examples of improvements in the quality of care and of benefits for patient outcomes.[27-29] The evidence for such programmes to reduce healthcare costs remains inconclusive, however,[5, 28] although observed improvements in care quality and/or patient outcomes without actual cost savings might be viewed as providing sufficient evidence of value for money.[30]

It is important to recognise, though, that where a given evaluation fails to identify an intervention effect this might mean that the intervention under consideration was not suitable to achieve the desired effect, or that the magnitude of the desired effect was too small for it to be measured reliably. This will depend on a range of factors, such as baseline levels of quality and/or health status of the population targeted by the intervention, or the design and execution of the intervention.[31] Alternatively, failure to demonstrate change in a given outcome might also suggest that the evaluation design was not suitable to adequately capture programme effect. Likewise, where an evaluation finds measurable change in a given outcome, this may not necessarily be attributable to the intervention under consideration if the evaluation failed to control for, for example, changes that are occurring simultaneously outside the actual intervention, for example a secular trend in improvements in the quality of care for people with chronic disease.[14]

As noted in the introductory section to this report, available evidence about the impact of interventions to manage chronic disease(s) tends to be based on small studies that frequently focus on high-risk patients, often undertaken in academic settings, and there is a need to better understand the impacts of large-scale, population-based programmes.[32]

However, even where such programmes are being implemented and evaluated, the resulting evidence may not be clear-cut, with different evaluation designs potentially yielding different findings (Box 2.1).

Box 2.1 Evidence of impact: Evaluating disease management programmes in Germany

In Germany, in 2002, the government introduced population-wide structured care or disease management programmes (DMPs), in an explicit effort to provide insurers and providers with incentives to encourage evidence-based chronic care.[33] Defined as 'an organisational approach to medical care that involves the coordinated treatment and care of patients with chronic disease across boundaries between individual providers and on the basis of scientific and up-to-date evidence', DMPs became the predominant approach to chronic illness care in Germany.[34] There are six types of DMPs for: breast cancer; type 1 and type 2 diabetes; coronary heart disease (CHD); asthma and chronic obstructive pulmonary disease (COPD). DMPs are principally offered by statutory health insurance (SHI) funds; in 2011 there were 10,340 DMPs in Germany as a whole, and almost 6 million people had enrolled in at least one programme (3.6 million for DMP diabetes type 2).[35] As the content and organisational structure of DMPs by disease is regulated at the national level, DMPs are very similar, however.

There is now increasing evidence from the range of evaluations of disease management in Germany. Evidence from the statutory evaluation of diabetes DMPs points to improved quality of care for patients with diabetes participating in such programmes[36], and limited evidence from (few) controlled studies also points to improved outcomes such as quality of life[37] and mortality[38, 39] as well as reduced costs.[39] However, the extent to which improved intermediate or definite outcomes, such as survival, can indeed be attributed to the diabetes DMP remains uncertain[38, 40-41], with other studies failing to provide evidence of improved medical outcomes.[42]

Furthermore, there is a need to better understand which (component) interventions, in what combination, achieve which effects under which conditions.[31] Much of the conceptual thinking and empirical evidence on disease management originates from the United States, which is characterised by a highly fragmented system of generalist and specialised care and where the baseline outcomes from common chronic diseases such as diabetes are worse than, for example, in Europe.[43] Given the differences in health systems, in particular as they relate to coverage and access, findings originating from the United States may not easily be transferable to healthcare systems that are characterised by (almost) universal access to healthcare such as those in Europe.[18]

Against this background there is a clear need for many more evaluations of the innovations being introduced in Europe in order to expand the existing evidence base; greater standardisation of evaluation approaches will go some way to facilitating this and so enable comparison of evidence on improvements to care quality and cost and support decisionmaking through learning across countries.

2.3 Evaluation of disease management can be organised around a common set of principles

As noted in the preceding section, the overarching aim of evaluating disease management initiatives and programmes is to understand the contribution made by a given activity to

achieving particular outcomes.[8] It involves forming a judgement about performance against a set of explicit criteria using a transparent and objective selection of measures.[26]

Fundamental to all evaluations is a clear definition of what constitutes effectiveness (or 'success'). This will be determined by the overarching aim(s) of the programme or intervention being evaluated and by the hypothesised mechanism(s) of expected effect(s). This requires good theoretical understanding of how the intervention causes change and of the links within the causal chain ('theory of change'[26, 44]) (for example, the intervention will cause behaviour change conducive to the control of high blood pressure as a main risk factor for stroke).[7] This understanding is important for establishing a causal relationship between the intervention and measured effects.

The aims and objectives of the intervention or programme to be evaluated should also inform the design of the evaluation, including the strategy for enrolling or recruiting participants, the timeframe for data collection, data sources and the type of control group for comparison. There is a range of research designs considered particularly appropriate for the evaluation of disease management interventions, which we have described in detail elsewhere[8] and will reflect on in Chapter 3. In brief, there are two principle design options that allow for deriving causal inferences about a programme or intervention–effect relationship: experimental (eg clinical and community trials) and non-experimental or observational (eg cohort, case-control, case-crossover, ecological, cross-sectional) studies. However, although experimental research designs are considered to provide the most robust level of evidence ('gold standard'), few evaluations of disease management interventions have employed such designs.[32, 45-46] Indeed, most published evaluations have followed a simple before-and-after study design, typically without a control group with which comparisons can be made.[45] Such designs are problematic, as they do not allow for conclusions about whether a given disease management intervention yields a 'true' effect although, as we will see below, there are various ways to construct a comparison strategy to allow for such conclusions to be drawn.

The choice of evaluation design will rest on a set of common principles[8]:

- characterising the intervention's scope, content, intensity and context (eg intervention setting, use of incentives, etc)

- characterising the planned target population of the intervention in terms of selection criteria (inclusion and exclusion), enrolment strategies and length of participation

- determining the length of observation period according to the expected effects to be measured

- planning the minimum number of evaluation participants needed to detect anticipated effects

- constructing an appropriate comparison strategy

- selecting the type, unit level and scale of performance measures that are sensitive to the design and goals of the intervention being evaluated

- analysing measured effects with robust methods, and validating findings by testing their sensitivity and assessing potential threats to their generalisability.

The choice of research design is, of course, closely linked to the goals and expectations of the evaluation itself. In particular, it is necessary to align evaluation length with the assessment objectives. There is a tendency for evaluations of chronic disease management to be conducted over 12 or even 24 months. Shorter timeframes are likely to be sufficient if, for example, the goal is to assess the process of implementation against a plan for learning. However, if the evaluation's goal is to determine 'success' in terms of medium- and long-term effects such as economic or health impacts, then a multiyear timeframe may be required before such effects can be expected to occur and be reliably measured. It has been suggested that three to five years are needed for a given intervention to be fully implemented and for any individual level effect to become evident.[45, 47] It is not known, however, what period is considered a sufficient evaluation timeframe to identify economic impact or the 'optimal' length of a given intervention to ensure sustainability of measured results.[5]

Although choosing the appropriate timeframe is only one of the many challenges that any successful evaluation will have to address, it presents an important consideration, particularly in relation to chronic disease management, for example:

- Disease management might increase costs in the short-term for some chronic conditions as a consequence of the initial investment required to implement the relevant intervention.[45, 47]

- Serious long-term complications can develop as a given disease progresses, which can influence whether disease management initiatives will have lasting effects.[48]

- There may be temporal influences leading to a plateauing of intervention effects, and which need to be distinguished from the effects of processes that bring about change through the intervention itself.[49]

It is worthwhile noting that much of the discussion about evaluation approaches rests on explicitly quantitative methods, and in this report we mainly focus on quantitative approaches. Cretin et al. (2004) have suggested, however, that because of the complexity and variability of disease management interventions there is a need to apply mixed-method research methods.[50] A number of evaluations of disease management interventions have applied qualitative methods[51-52], sometimes combining these with quantitative methods.[53-56] There is, however, relatively little research on methodological, analytical or conceptual aspects of the use of qualitative approaches in disease management evaluation. Recently, there has been a move towards emphasising 'realistic evaluation',[57] which uses pluralistic quasi-experimental methods for evaluating complex interventions with high contextual influence, such as disease management.[58-59] Realistic evaluation involves understanding what works for whom under what circumstances, and places equal emphasis on external validity, generalisability and cumulative learning. In section 3.3.4 we present an example of work undertaken in DISMEVAL that used a mixed methods design to advance our understanding of what works for whom in disease management.

2.4 Evaluations of disease management will have to take account of the policy context

Ideally interventions are developed systematically, are based on the best available evidence and appropriate theory, and are tested using a phased approach to inform further development, alongside evaluation. However, as Craig et al. (2008) have pointed out, in practice interventions emerge from various sources, which may include theory, but may as well be based on weak evidence, depending on the drivers behind the intervention.[7] It is against this background that evaluation takes place, and the setting will impact on the choice of approaches to evaluation available vis-à-vis the nature of the intervention to be evaluated. Although, as noted before, a randomised controlled design is generally accepted as providing the most robust evidence, there may be powerful reasons for not applying this method, such as political or ethical considerations. Thus, the random assignment of participants to a disease management intervention or no intervention (control) group means that some participants may be excluded from the most appropriate care if the evaluation results demonstrate that the intervention does provide better care.[50] There may also be legal obligations on the side of the funder that prohibit such exclusion of individuals from receiving the most appropriate care.[49]

Overall, there is a need to consider the trade-offs between the importance of a given intervention and the value of the evidence that can be collected against the background of constraints, be they financial or otherwise.[7] While this will be important in any context, such trade-offs are likely to pose a particular challenge in those settings where the 'evaluation culture' is weak and/or of low priority.

Methods and metrics for evaluation: experiences from the DISMEVAL project

Experimental research designs, particularly randomised controlled trials, are generally considered to be the most rigorous way of determining the effectiveness of a given intervention.[8] In such designs individuals are randomly assigned to either the intervention group or a control group. Thereby, each person is given an equal chance to be chosen for the intervention so that any observed difference in outcome is not affected by systematic differences in factors, known and unknown, between those who receive a given intervention and those who do not.[60-61] However, use of an experimental design may not be possible (or desirable) for population-wide disease management interventions, which are frequently implemented in an operational setting and do not usually have a control group available that would allow for attributing observed change in a given outcome measure to the intervention. Furthermore, randomisation may be difficult in disease management 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 statutory health insurance systems such as those in Austria, Germany or France.[62]

To establish whether a given disease management intervention yields a 'true' effect ('attribution'), some form of comparison is required between individuals who received the intervention and those who did not.[32, 50, 58] If the two groups being compared are equal in every aspect other than the enrolment in the actual disease management intervention, then any observed differences in intermediate or definite outcomes, costs or other measures of interest can be reasonably attributed to the intervention. If randomisation is not possible, the next suitable evaluation approach for estimating intervention effect is the quasi-experimental, pre-post with a comparison group design, adjusting for known differences.[63] Figure 3.1 provides an illustration of the trade-offs between scientific rigour and practical implications of different methods of attribution. Further detail on methods of attribution is provided by Linden et al. (2004[45], 2006[61, 64-65], 2010[66]), Mattke et al. (2006)[67] and Conklin and Nolte (2010).[8]

no control group
- ➢ **Pre-post comparison**

- ➢ **Benchmarking**
  - Prior literature
  - Historic control

- ➢ **Statistical controls**
  - Risk adjustment
  - Trend adjustment

control group
- ➢ **Quasi-experimental approaches**
  - Predictive modelling
  - Propensity scoring
  - Regression discontinuity design
  - Difference-in-differences

- ➢ **Experimental approaches**

Practicability

Scientific rigour
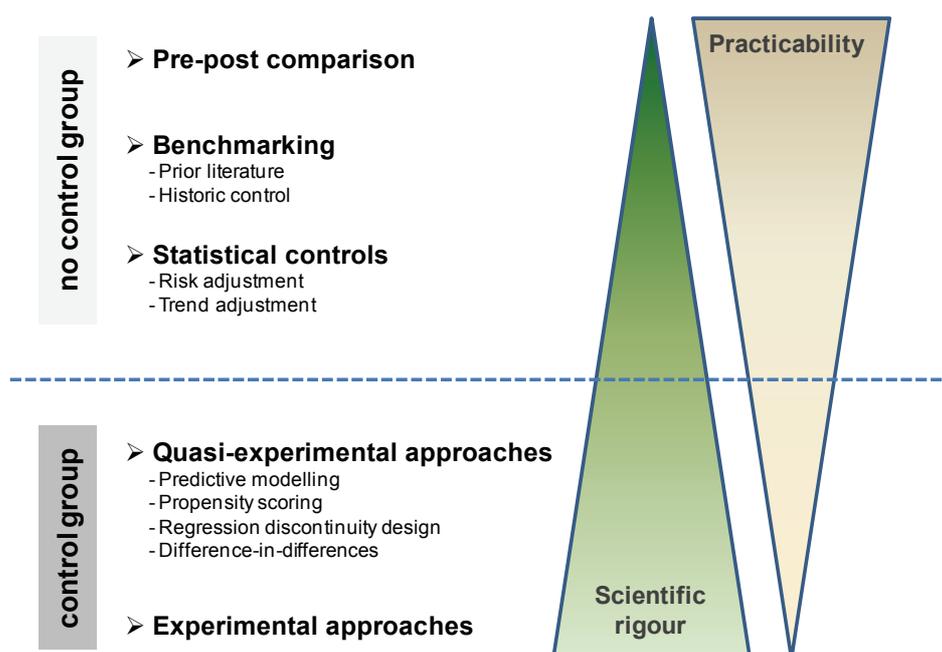
Figure 3.1 Methods of attributing observed effects to an intervention

SOURCE: adapted from Mattke et al. (2006)[32]

The DISMEVAL project comprised six work packages that sought to test different methods to evaluate chronic disease management approaches in six countries. Work carried out focused on *improving our understanding of the relative performance of different evaluation methods for measuring programme effect* in terms of processes and outcomes, typically in non-experimental settings, but also using experimental designs where feasible. A detailed overview of the findings of the work undertaken in DISMEVAL is provided in a separate report[9]; Table A1 in Appendix A provides a summary overview of the interventions underlying the evaluations carried out in six work packages (country case studies) and their main findings. In this chapter we present selected observations of the work undertaken, which we set against the background of the general literature on evaluation research design and methods. We focus on a select set of themes that have emerged as being pertinent to several work packages. We begin by examining various issues around testing and validating designs and methods, considering how selection bias in different experimental and non-experimental designs impacts intervention effect, exploring methods for constructing a comparison strategy in non-experimental settings and then we consider approaches to measuring the impact of different matching techniques on intervention effect. We then discuss various metrics ('performance indicators') that are being used to measure intervention or programme effect and close with a section that presents various approaches to advancing methodology in disease management evaluation.

3.1   Testing and validating designs and methods

Any design selected to evaluate the effectiveness of a given disease management intervention or programme is subject to exposure to factors outside the actual intervention that may influence the findings.[45] Confounding factors and sources of bias[a] pose a particular challenge for those designs that do not involve a comparison strategy, such as simple before–after (or pre-post) comparisons, although more rigorous designs are also vulnerable to what are frequently referred to as threats to the validity of findings. Threats to validity can occur at various levels, including the individual participant, the intervention itself, the provider of the intervention, the availability, accuracy and use of the data used to measure the intervention's effect, long-term secular trends, etc.[8] Not taking account of potential sources of bias and confounding and controlling for their impact may lead to results that look better or worse than was actually achieved by the intervention being evaluated.

The wide range of potential biases and confounding factors that have been identified as posing a threat to the validity of a disease management programme evaluation are reviewed in detail elsewhere.[8, 46] Here we focus on selection bias, which has emerged as pertinent to an evaluation of approaches to chronic disease management in European settings.

3.1.1   It is important to consider the impact of selection bias on intervention effect

Selection bias occurs when those who take part in a given intervention are different from those who do not. Thus, intervention participants are not representative of the population of all possible participants. For example, Beeuwkes Buntin et al. (2009) demonstrated, in an observational study of members of a large health insurer in the United States, that those who enrolled into the programme differed systematically from those who did not on demographic, cost, utilisation and quality parameters prior to enrolment.[68]

Motivation to change is an important factor, and volunteer enrollees who self-select to participate in a given intervention may be more likely to take a more active role in managing their own care and thus achieve desired behaviour and/or outcomes. Similarly, a higher proportion of 'sicker' individuals might enrol because of fear for worse outcomes, relative to the population potentially eligible for the intervention.[45] Conversely, an alternative model for enrolment in which those who do not wish to participate must actively 'opt out' is likely to include the group of unmotivated individuals otherwise missed by a volunteer-based model of enrolment.[8] Here, bias occurs when group assignment is non-random and the intervention group disproportionately includes more unmotivated individuals who are different from controls on this characteristic. In all these cases, the intervention effect is likely to be overestimated or underestimated if selection bias is not controlled for.[46]

Evidence of selection bias was observed in several analyses undertaken within the DISMEVAL project. One example is provided by the French case study, which examined

---

[a] *Confounding* is defined as the 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; are not intermediate variables; and are not associated with the factor(s) under investigation. *Bias* refers to the extent to which there is any systematic error, either observed or unobserved, in collecting or interpreting data (eg selection bias; recall bias; etc). Only biases that can be observed can be controlled for in statistical analyses.

patient characteristics of those enrolled with a diabetes provider network (Box 3.1). In the context of French provider networks, more than 90 per cent of the patients that enrol are encouraged to do so by their physician.[69] The quantitative analysis of diabetes provider networks undertaken within the DISMEVAL project demonstrated the degree to which patient selection takes place when patients enrol; these patients were of younger age and had a more recent diagnosis of diabetes but poorer glycaemic control (as measured by HBA1c levels) compared to those in the reference population of diabetic patients.

Box 3.1 Provider networks in France

In France, provider networks are considered the main approach to providing coordinated care for those with complex needs, with an estimated 1,000-plus networks operating across France.[70] They include disease-specific networks, with diabetes networks particularly well established, and provider networks for selected population groups, for example focusing on older people.

Patients can join a network through their physician (usually the general practitioner) or self-refer (five per cent). Participation is free of charge; in addition, patients may access services they would otherwise have to pay for. For example, in diabetes networks this might include access to educational sessions, dietary counselling, supervised weight loss and exercise programmes, in partnership with other networks. For patients with diabetes who cannot access health networks because these do not exist in their locality, another option is the *maisons du diabète* (diabetes home), located in 20 cities throughout France. These homes are non-profit institutions providing diabetic patients free access to nurses and dieticians for educational sessions, as well as information on diet and other lifestyle issues. Both networks and diabetes homes cover approximately five per cent of the diabetic patients in France.

In order to assess the impact of selection on intervention effect, the analysis used calibration to account for differences between the intervention and the control group. The method, often used in survey sampling, adjusts variables of a sample to the variables of the general population by generating 'calibration weights' (coefficients). Specifically, differences in patient characteristics at baseline were rebalanced by assigning a coefficient to each participant (eg a patient with very high values for a set of variables was assigned a low weight in order to take account of these 'extreme' values). Then pre-post evaluation was performed by (re-)weighting each patient with the weight determined initially.

By comparing standard uncontrolled pre-post evaluation to pre-post evaluation after calibration with a reference population at baseline, the analysis found that the pre-post only design overestimated improvements in glycaemic control (HbA1c level) and relative body weight (BMI) and underestimated deterioration in renal function (as measured by glomerular filtration rate (GFR)) in diabetic patients (Figure 3.2). Thus, calibrated evaluation gives an estimation of the impact that the intervention (here: the enrolment in the diabetes provider network for one year) would have on the diabetic reference population, as compared to the impact on patients enrolled with the provider network only.
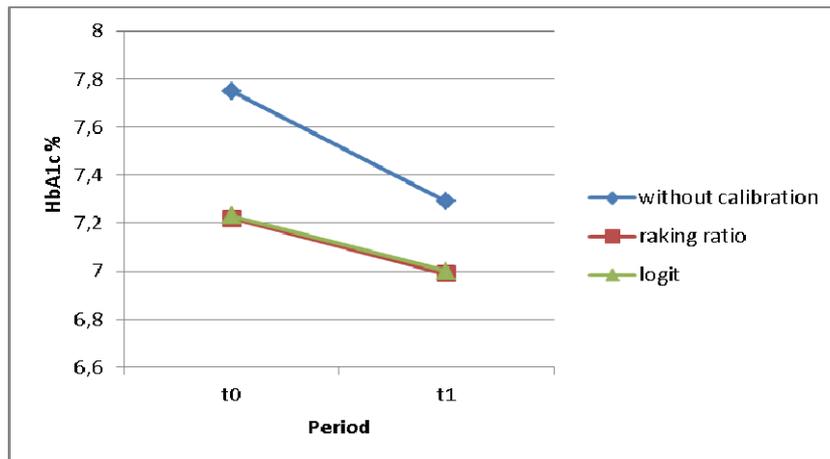
Figure 3.2: Differences in results of HbA1c after one year of enrolment in a diabetes provider network without calibration and with calibration using the raking ratio or logit method, France

Evidence of patient selection into the intervention was also observed in the German case study, which analysed a diabetes disease management programme (DMP) (see also Section 3.1.2) and observed a significant reduction of mortality in the intervention group.[71] Specifically, it observed a higher risk of death among those not participating in the programme compared to those participating in the programme who were similar in a number of characteristics and after adjustment for confounders. The hazard ratio (HR) for propensity-score weighted controls was 1.51 (95 per cent confidence interval 1.32, 1.75). This observation would suggest a truly beneficial effect of the intervention, with those not participating at a 50 per cent higher risk of dying than those participating in the programme. However, further analysis of the underlying data found the effect to be mostly due to a large mortality difference between the intervention and control group in the first year following DMP enrolment, and this difference fell considerably during the further follow-up period (Figure 3.3).
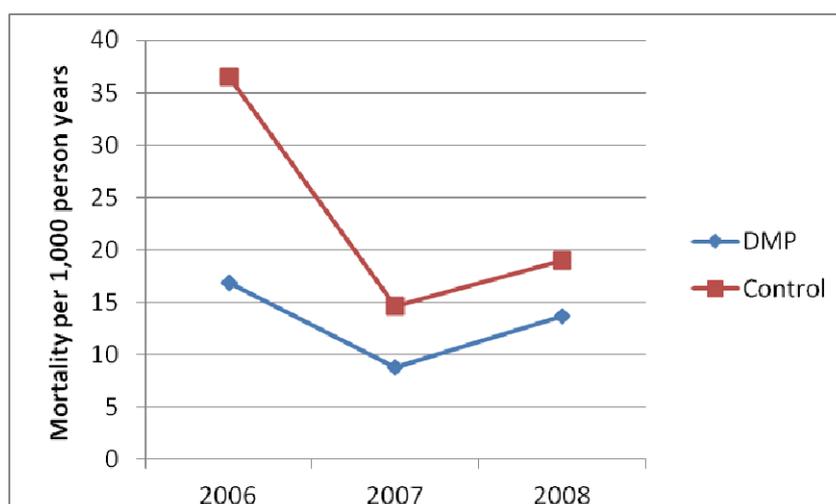
Figure 3.3 Mortality (per 1,000 person years) of DMP participants and non-participants (control), Germany, 2006–2008

If the outcome, that is, reduced mortality, had indeed been attributable to the intervention one would have expected the effect size to increase over time; instead, it declined. Therefore, rather than attributing the observed survival advantage of participants to the DMP, it is more likely that physicians who were responsible for recruiting patients into the programme may have systematically excluded those from joining who were more likely to die in the near future. Future work should include further adjustment for variables that predict short-term mortality risk to confirm the observed findings. Furthermore, a longer observation period would be required to assess whether the mortality difference diminishes over time.

It is important to note that selection bias can also occur in evaluations that use a randomised-controlled design, so compromising the generalisability of findings.[72] For example, the implementation of a diabetes disease management programme in Salzburg, Austria, was conceptualised as a pragmatic cluster-randomised controlled trial with an observation period of 12 months.[73-74] Randomisation was carried out at the district level of Salzburg province (see also Section 3.3.1). As is characteristic of pragmatic trials, blinding was not possible and the knowledge of being in the intervention or control group may have influenced the findings.

The overall evaluation found the disease management programme to reduce HbA1c levels in the intervention group by 0.13 per cent.[75] This finding was statistically significant in the unadjusted analysis but not after a mixed models adjustment for baseline values. There were statistically significant improvements in measures of process quality such as regular eye and foot examination and patient education.

Three types of selection bias were identified that could have influenced the findings of the evaluation. First, only one third of eligible physicians practising in Salzburg province participated in the study and the sample might have been skewed towards more motivated physicians who are early adopters. It could be hypothesised that the effect of the programme might have been larger if all physicians had participated, as less motivated

physicians might be associated with poorer levels of patient care and thus provide greater potential for improvement. Second, patient selection into the programme was undertaken at the physician level and there is a risk of differential recruitment through physicians selectively inviting those with a higher likelihood of adherence to the programme. This risk of differential patient recruitment was minimised, however, by inviting participating physicians to recruit patients consecutively and to enrol control group patients for later participation in the programme upon completion of the formal study. This approach to recruitment proved to be appropriate in addressing this potential bias; there were no significant differences between the intervention and the control group at baseline in relation to key characteristics of relevance to the intervention, with the exception of body mass index and cholesterol levels, which were elevated in the intervention group.[75]

However, although patients were recruited consecutively, selection bias could still have occurred as it used a volunteer-based enrolment strategy. As noted earlier, highly motivated patients seeking to maintain metabolic control and higher levels of adherence may be more likely to enrol in the programmes.[76] The potential for improvement in terms of measurable HbA1c reduction in these patients may be lower than in the general (eligible) population. It is conceivable that patients with poorer blood glucose levels may have been less motivated and less adherent despite their greater potential for improvement. Because the trial was conceptualised as a 'pragmatic' study, a disproportionate recruitment of 'healthy' patients is likely to reflect 'real' life, with unmotivated patients less likely to opt for such programmes.

Summary

In summary, therefore, in both experimental and observational designs, selection bias poses a threat to the validity of findings. We here provide examples of how the impact on programme effect of such bias can be identified and highlight how it will be important for any evaluation to describe in detail and understand the nature and sources of potential bias. It is important to note, however, that comparability of intervention and control can only be assessed on observed characteristics.[61] This poses a particular challenge for non-experimental settings. The application of sensitivity analysis has been proposed as a means to assess the potential impact of unobserved differences in the populations under study that are not accounted for through statistical means.[77] Sensitivity analysis provides an indication of the relative sensitivity of evaluation results to the size of the so-called 'hidden' bias, although it remains uncertain at what size 'hidden' bias would invalidate the measured effects of a given intervention.

The example of the diabetes disease management programme evaluation in Germany given above also illustrates the importance of the evaluation timeframe for assessing intervention effects and their validation, which we briefly reflected on in Section 2.3. Thus, there is a need to allow for sufficient length of follow-up to allow for robust conclusions on observed effects to be drawn.

3.1.2 The choice of approaches to constructing control groups in non-experimental designs impacts on observed intervention effect

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.[8] This involves selection of a suitable control group that most

closely mirrors what would have happened without the intervention (ie the counterfactual). For any evaluation design it will therefore be important to carefully document demographic and clinical details for intervention and control groups. If these differ on important observed baseline features, causal inferences about programme impact will be limited.[61]

A main challenge for disease management evaluation is the identification of a suitable comparator in practice. This is because the nature of the 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.[46] 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.

However, there is a range of techniques that can be employed in non-experimental settings to create control groups using readily available administrative data, for example predictive modelling and propensity scoring.[45] Within the DISMEVAL project, several analyses used propensity scoring to create control groups in non-experimental settings. These included the evaluation of a rehabilitation programme for people with chronic obstructive pulmonary disease (COPD) in Copenhagen, Denmark, and the assessment of the effects of a nurse-led intervention targeting a working-age population at risk of cardiovascular disease in Spain. The evaluation of the diabetes disease management programme (DMP) in Germany mentioned earlier used propensity scoring as one of several techniques to identify suitable treatment–control matches.

Using propensity scoring to match on disease severity: COPD rehabilitation in Denmark

The Danish case study in DISMEVAL aimed to evaluate the impact of a three-month rehabilitation programme for persons with chronic obstructive pulmonary disease (COPD) on healthcare utilisation using different designs.[78] The analysis was able to draw on data that are routinely collected including, for instance, vital statistics, disease registers and hospital admissions data. Availability of a unique identifier (social security number) makes it possible to link different types of data through Statistics Denmark.

The analyses of the impacts of the programme used three designs:

a)   pre-post comparison (without control) to assess changes in the intervention group over time

b)   post-period comparison (with control) to assess differences between intervention and control patients after the intervention

c)   difference-in-differences analysis (DID), which measures the differential change within intervention and control group (Figure 3.4).
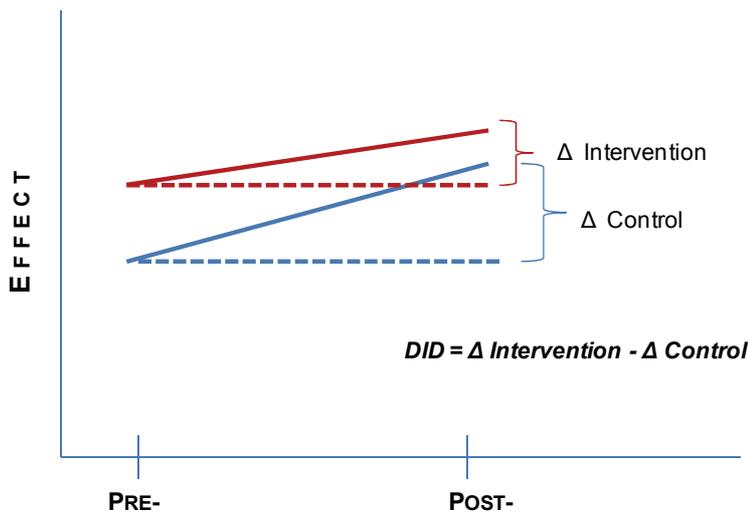
Figure 3.4 Difference-in-differences (DID) method

Control groups were created retrospectively, based on registry data from the population of individuals with COPD in the municipality of Copenhagen who had a documented hospital contact during the period 2005–2007 and who were aged over 35 years at the time of diagnosis, using three techniques: (1) random selection, (2) gender and age matching, and (3) propensity score matching. The rationale for using propensity scoring was to match on disease severity, which, in COPD patients, is related to healthcare utilisation.[79] Severity is defined by the Global Initiative for Chronic Obstructive Lung Disease (GOLD) standards based on spirometry test results,[80] but this data is not routinely documented in the Danish national registers. The analysis therefore used utilisation patterns as a proxy for disease severity. Propensity scores were calculated using socio-demographic characteristics, health service utilisation patterns and medication in the pre-intervention period, as well as disease duration (Table 3.1); matching was conducted by the nearest neighbour without replacement methodology. Pre- and post- intervention periods were defined as the two years preceding a rehabilitation course and two years after completion of the course.

Table 3.1 Covariates in propensity score model

Socio-demographic	Healthcare utilisation	Other
Age	GP contacts	Medication
Sex	COPD-specific:	- delivered daily dose (DDD), COPD-specific
Ethnicity	- hospital contacts	
Marital status	- bed days	- delivered daily dose (DDD), all medication
Education	- outpatient visits	
Income	- emergency room visits	Disease duration
Assets		

Overall, findings pointed to disease progression as evidenced by increased utilisation over time (GP visits, hospitalisations, outpatient contacts, medication).[78] However, utilisation was more frequent among controls than among the intervention group, indicating that the

rehabilitation impacted on healthcare utilisation by decreasing COPD-specific hospital contacts, bed days, outpatient and emergency room visits. The magnitude of observed changes in the frequency of hospitalisation was smaller in the before–after design, while the difference-in-differences analysis found a larger effect size that was statistically significant. This points to a rehabilitation effect in the intervention group; that is, the intervention slowed disease progression.

Importantly, the magnitude of the predicted intervention effect as assessed by difference-in-differences analysis changed with the chosen control strategy. Thus, the effect size fell when moving from randomly selected to propensity score-matched controls for COPD-specific emergency room visits. This implies that matching by disease severity (or propensity score calculated based on healthcare utilisation pattern) is especially important in order to not overestimate the effect of pulmonary rehabilitation for selected utilisation measures.

Assessing the impact of different matching techniques on intervention effect: diabetes disease management in Germany

The impact of different matching techniques to reduce confounding due to baseline group differences between intervention and control groups was examined further within the German case study. It compared different techniques to adjust for baseline differences between the intervention (participants of a diabetes disease management programme, DMP) and control group (non-participants) using regression analysis with adjustment for confounding variables, regression analysis after matching or weighting by propensity scores and regression analysis after direct covariate matching on a selected number of variables.[71]

The analysis used routine data from a large nationwide German statutory health insurance fund covering three regions in Germany (North Rhine, Hesse and North Wurttemberg) for the years 2004 to 2008. The analysis was based on a cohort of insurance fund members with diabetes mellitus type 2 who entered a diabetes disease management programme (DMP) in 2005 and a control group, comprising insurance fund members who did not participate in the DMP until 2008. Routine data from 2004 were used to assess baseline differences between the two groups, who were further compared regarding mortality, micro- and macrovascular complications occurring during a three-year follow-up period (2006 to 2008) using Cox proportional hazard models.

The analysis compared a total of 11 different adjustment techniques. These were based on variations of propensity and direct matching methods regarding the selection of baseline variables used to determine the propensity scores. For each matching method, the baseline 'balance' between intervention and control group was calculated using the standardised differences for all baseline variables. The analyses found that propensity score weighting using a general boosted regression model for the calculation of the propensity scores provided the best balance. Overall, however, the choice of adjustment method appeared to play a minor role for the outcome analysis: all adjustment methods resulted in fairly similar hazard ratios for all three outcomes. The only exception was one matching method, the direct covariate matching method, which, in the case of macrovascular complications, showed a significantly increased risk for this outcome in the control (non-DMP) group (Figure 3.5).
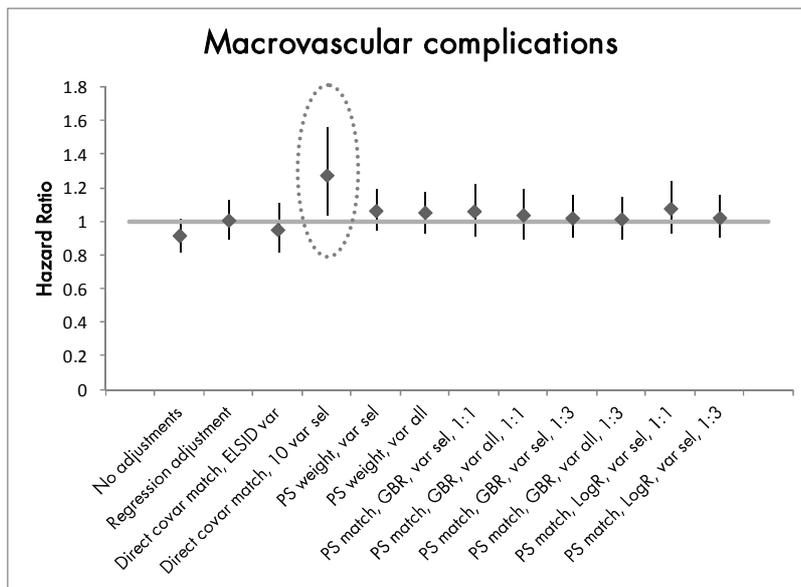
Figure 3.5 Risk of macrovascular complications of non-DMP patients compared to DMP patients with and without adjustment for potential confounders by multivariate regression and different matching methods (hazard ratios with 95 per cent confidence intervals), Germany, 2006–2008

However, this matching method did only allow for a minority of individuals to be matched, as 83 per cent of the study population had to be excluded from the analysis. The resulting small number of study participants that could be matched with this technique resulted in a large variance, causing the effect estimate to be imprecise and difficult to interpret. The exclusion of a large proportion of individuals is also likely to lead to another form of selection bias, as individuals with specific attributes are more likely to be excluded.

Overall, the analysis illustrated that the quality and completeness of the dataset had a greater influence on the validity of the results than the method of covariate adjustment (see also Chapter 4).

Summary

This section examined different approaches to constructing control groups in non-experimental settings and the impact of different techniques to reducing baseline differences between intervention and control groups. It finds that approaches using matching by propensity score are preferable over other approaches such as random selection or direct covariate matching. Based on the experience of the Danish case study, matching by propensity score as a technique for control group construction and difference-in-differences analyses is recommended as a method for the assessment of rehabilitation effect on healthcare utilisation among COPD patients in a non-experimental setting.

However the German case study, while also pointing to propensity scoring as the preferred method of adjustment for baseline differences between intervention and control group, highlighted that the choice of adjustment method may be less important in relation to determining intervention effect than the quality and completeness of the underlying dataset. The choice of matching method should, however, depend on the range of

confounding variables to be considered for adjustment. Thus, where a smaller number of confounding variables is expected to be relevant, simple methods such as direct covariate matching may sufficiently adjust for baseline differences between interventions and controls. In contrast, where a large number of confounding variables is being considered, methods including propensity score weighting may be more appropriate, as the findings of the Danish case study highlight.

3.2 Assessing measures of intervention effect

Drawing on Donabedian's framework (1980)[81] for evaluating the quality of healthcare as well as on standards for outcome measurement in disease management,[82] the evaluation of disease management may focus on single or multiple components of intervention structure, process, output and outcome/impact[83-84]:

- **Inputs** refer to structural aspects of a given intervention, such as its material and financial inputs and human resources.

- **Process** refers to: actual activities such as how the intervention is delivered or implemented (ie how it worked); also the extent to which the intervention or programme was implemented as intended and/or implemented according to the evidence base.

- **Output** is defined as productivity or throughput, ie the immediate result of professional or institutional healthcare activities, usually expressed as units of service.

- **Outcome** refers to the (medium- and long-term) effects of the intervention on the health status of individuals and populations. Outcomes are typically distinguished into intermediate (eg blood pressure, cholesterol levels) and definite, such as mortality, morbidity, quality of life or patient experience. Definite outcomes are sometimes also referred to as health 'impact'.[8]

Further evaluation outcomes of interest are not only related to health status but can also include economic impact, social impact or environmental impact. Examples of measures for the dimensions listed above are given in Appendix B.[8]

The selection of dimensions and actual measures to be assessed will be determined by the specific design and goals of a given disease management intervention and should be driven by construct validity of the measure. Existing evaluations have tended not to define these relationships explicitly[84]; indeed, only a minority of published studies apply a coherent framework linking the aims of disease management to measures of structure, process and outcome.[85] However, 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;[26] this will enable assessment of the conditions under which a disease management intervention is deemed to be successful, or indeed might have failed to achieve desired outcomes.[48]

Choice of measures of effect should be based on considerations of importance and practicality while ensuring the best use of data. It also needs to take account of timing of

(expected or hypothesised) change or rate of change and thus determine length of follow-up. Craig et al. (2008) further emphasised the need to consider sources of variation in outcomes and allow for potential subgroup analyses to be carried out.[7]

While outcomes are important to assess the effectiveness of a given intervention, it may be equally valuable to (also) carry out a process evaluation so as to help understanding why a given intervention works or indeed (unexpectedly) fails.[7] Process evaluation can also support effectiveness assessment by providing insights into the quality of implementation of the intervention and the contextual factors that may be associated with variation in outcomes.[6]

3.2.1 Measuring intervention effect: health outcomes

The goal of disease management interventions is most often to enhance the overall health status of the individual patient and, where applicable, the population at large. As an individual's health is influenced by many different factors that act over time, it has been suggested that the evaluation of related interventions should consider a range of outcomes to be assessed in the long term.[48]

As noted above, health outcomes are typically differentiated into intermediate, sometimes also referred to as clinical, and definite outcomes. Clinical outcomes, such as glycaemic control as assessed through measurement of HBA1c levels, blood cholesterol, blood pressure and body mass index (BMI), are the most frequently measured outcomes in evaluations of disease management interventions.[8] This is in part because of their ease of measurement with standard tests and procedures in place, although the relevance of some clinical measures for long-term health outcomes remains unclear.[86] Overall health status measures aim to quantify how a person's condition affects her/his life and can be distinguished into different domains such as (a) symptoms and their management; (b) functional limitations; and (c) quality of life.[87] Health status and health-related quality of life are generally captured using self-administered questionnaires, requiring the use of validated instruments so as to enable the reliable assessment of self-reported measures.

It is important to use valid and reliable outcome measures although changes in key outcome measures in the absence of a suitable control are difficult to interpret

Within the DISMEVAL project, the Danish case study focused on a rehabilitation programme for patients with chronic conditions, the Integrated Rehabilitation Programme for Chronic Conditions (SIKS) project.[78] The intervention was established as a single project in Copenhagen from April 2005 to September 2007 and was funded by the Ministry of Interior and Health. On entering and completing the three-month intervention, those joining (voluntarily) were assessed on a number of health indicators, including general measures (eg BMI, blood pressure), general- and disease-specific functioning (eg senior fitness tests, spirometric tests for COPD patients, glycated haemoglobin and lipid values for diabetes patients), and health-related quality of life (Table 3.2).

Importantly, the assessment only used well-known, valid, reliable and feasible instruments and tests to capture clinical measures and patient outcomes. These included, among others, a six-minute walk test, shown to be a feasible, reliable and valid measure of functional capacity targeted at older people and people with moderate-to-severe physical impairment;[88-89] a senior fitness standard and 2.45-meter test, shown to be valid and

reliable in the Danish context;[90] the 'endurance shuttle walking test' developed for measuring exercise tolerance, commonly used in pulmonary rehabilitation;[91-92] the Danish adaptation of Borg's dyspnoea scale, to assess the degree of breathlessness in relation to a certain physical activity;[93-94] the Medical Research Council dyspnoea scale, as a measure of self-reported disability in patients with COPD;[95-96] the Avlund scale, to assess ability to perform activities of daily living (ADL), which was shown to have high internal consistency reliability and construct validity;[97-98] the clinical COPD questionnaire developed by van der Molen et al. (2003), which measures symptom and functional state in daily clinical practice of patients with COPD;[99] and the Short-Form 36-Item Health Survey (SF-36), the most widely used generic health-related quality of life instrument,[100] shown to have high reliability and validity in multiple studies in many languages, including Danish.[101]

Table 3.2 Pre-post intervention measurements for individuals participating in the rehabilitation programme, Copenhagen, Denmark

	Persons with chronic obstructive pulmonary disease	Persons with type 2 diabetes
General health	- Weight - Body mass index - Waist circumference - Systolic blood pressure - Diastolic blood pressure	- Weight - Body mass index - Waist circumference - Systolic blood pressure - Diastolic blood pressure
General functioning	- Stand and Sit Standard Senior Fitness Test (SFT-st) 2.45 min Up and Go Senior Fitness Test (SFT 2.45) - Endurance Shuttle Walk test (End SWT) - Borg dyspnoea scale (together with End SWT) - Medical Research Council (MRC) dyspnoea scale	- Six-minute walk test - Stand and Sit Standard Senior Fitness Test (SFT-st) 2.45 min Up and Go Senior Fitness Test (SFT 2.45)
Disease-specific functioning (laboratory data)	- Forced expiratory volume at timed intervals of 1 second (as a percentage of the predicted value for people of similar characteristics (FEV1 % of predicted) - Forced expiratory volume as a part of forced vital capacity (FEV1/FVC)	- Glycosylated haemoglobin (HgbA1c) - Cholesterol - High density lipoproteins (HDL) - Low density lipoproteins (LDL) - Triglycerides
General and disease-specific quality of life	- Short form 36 (SF-36) - Clinical COPD questionnaire - Avlund scale	- Short form 36 (SF-36) - Avlund scale

In addition, the project employed several alternative instruments to assess, for example, exercise capacity and disease-specific health-related quality of life, and so assure the validity of taken measurements.[78]

The availability of this rich data set allowed predicting the impact of the rehabilitation on general functioning and quality of life outcomes by the means of multivariate linear

regression for patients with COPD and for those with diabetes.[78] However, in the absence of a suitable control group findings remain difficult to interpret, with evidence of what is generally known as 'regression to the mean' (Box 3.2); that is, the poorer selected health variables at baseline, such as (in the case of the Danish case study) poor glycaemic control among individuals with type 2 diabetes or self-rated quality of life, the larger the improvement of this outcome following the intervention.

Box 3.2 Regression to the mean

> Regression to the mean is also referred to as 'statistical regression'.[45] This phenomenon is particularly relevant in disease management interventions (or their equivalent) that aim to control disease through reducing high levels of a given variable (eg blood sugar), or that aim to reduce utilisation (eg 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.

Length of observation period is an important factor in interpreting observed changes in key outcome measures

As noted above, disease management evaluations often tend to focus on disease-specific intermediate health outcomes, in part because of ease of measurement. In contrast, measurement of 'hard' outcomes, such as mortality and disability, is frequently difficult because the duration of evaluation is of insufficient length to assess long-term (health) effects. However, it is important to recognise that intermediate measures are heavily influenced by an individual's health behaviour, improvements in which are frequently difficult to maintain over time. Therefore, it will be as important to consider an adequate length of follow-up to enable reliable measurement of intermediate outcomes.

To investigate in-treatment effects across different observation periods, the Dutch case study, which analysed data from 18 regional diabetes disease management programmes (see Section 3.3.2), conducted subgroup and meta-analyses incorporating the covariate 'length of follow-up'.[102] In line with previous research,[103] the findings suggest that evaluations of a shorter duration (< one year) tend to overestimate observed effects. Table 3.3 shows the outcomes of subgroup meta-analyses based on length of follow-up, with subgroups representing patients with a timeframe between clinical measurements of less than one versus one year or more.

Table 3.3 Subgroup analyses according to length of follow-up, the Netherlands

	Duration of care intervention		Heterogeneity (I^2)		Change in I^2
	≤1 year	>1 year	Total	Between group	
HbA1c			97%*	0%	−1%
N	57,069	18,058			
MD [95% CI]	0.02 [−0.77, 0.81]	0.53 [−0.22, 1.27]			
Total cholesterol			82%*	0%	−8%
N	45,912	15,466			
MD [95% CI]	−0.10* [−0.14, −0.05]	−0.11* [−0.14, −0.07]			
LDL cholesterol			88%*	0%	−5%
N	43,901	14,796			
MD [95% CI]	−0.09* [−0.13, −0.05]	−0.11* [−0.15, −0.06]			
HDL cholesterol			86%*	0%	−6%
N	40,620	13,836			
MD [95% CI]	0.02* [0.00, 0.03]	0.02* [0.00, 0.04]			
Triglycerides			54%*	0%	−21%
N	45,568	15,510			
MD [95% CI]	−0.05* [−0.08, −0.03]	−0.04* [−0.07, −0.01]			
Systolic blood pressure			61%*	94.3%*	+4%
N	55,686	17,751			
MD [95% CI]	−1.27* [−1.60, −0.95]	−0.04 [−0.52, 0.44]			
Diastolic blood pressure			30%*	78.3%*	−4%
N	55,456	17,659			
MD [95% CI]	−0.90* [−1.01, −0.78]	−0.57* [−0.84, −0.30]			

NOTES: * Statistically significant (p<0.05)

Figures shown in Table 3.3 demonstrate that for half of the endpoints, patients with a length of follow-up of less than one year achieved better results in terms of mean differences in health outcomes. The differences in clinical results between subgroups were, however, statistically significant only for systolic and diastolic blood pressure, with reductions achieved among those followed up for less than one year significantly larger compared with those who were followed up for a longer period. Meta-regression, which included length of follow-up as a continuous covariate (ie number of months), yielded significant negative linear relationships between this effect modifier and all clinical outcomes except for HDL. This suggests that as the duration of care increases, the positive effects of the DMPs on all outcomes except for HDL cholesterol are difficult to maintain. From an evaluation perspective, these observations highlight the importance of using longitudinal data for measuring intermediate health outcomes in disease management evaluation.

Where length of the evaluation period is limited, measurement of processes may provide useful insights into programme effect

Many evaluations of disease management programmes (DMPs) examine changes in both care processes and outcomes following the introduction of specific quality improvement strategies. For example, the Austrian case study of a diabetes disease management programme implemented in Salzburg province defined HbA1c levels as a primary outcome

measure.[73] While recognising the potential limitations of a surrogate measure such as HbA1c, it was considered the preferred choice for assessing metabolic control within a cluster randomised trial that was limited to an observation period of 12 months. Secondary outcome measures included systolic and diastolic blood pressure, blood lipids and body mass index. It further assessed self-rated health-related quality of life (HRQoL), using EQ-5D-3L, a standardised generic measure of health status designed for self-completion developed by the EuroQol Group.[104] In addition, the analysis considered measures of process quality including the frequency of HbA1c measurements, eye and foot examinations, as well as participation in patient education.

Overall, the evaluation demonstrated that the DMP did not significantly improve metabolic control for patients with type 2 diabetes mellitus, although measures of process quality improved significantly. Thus, a higher proportion of patients received patient education and regular screening examinations of the eyes, feet and HbA1c checks compared to those receiving usual care. However, within the timeframe of an evaluation period of 12 months it was not possible to validly predict the influence of improved process quality on clinical outcomes.

Summary

This section has examined the use of health outcome indicators to assess intervention effect. It notes that it will be important to consider a range of outcomes to be assessed long term as an individual's health is influenced by many different factors that act over time. The use of valid, reliable and feasible instruments to measure clinical and health outcomes will be important to enable systematic assessment of programme effect. Disease management evaluations tend to rely on measuring intermediate outcome measures such as metabolic control (HBA1c levels) or blood pressure, which are relatively simple to collect within a short observation period. However, the example of the Dutch case study has illustrated how a shorter duration observation period may overestimate observed intervention effects and it will therefore be crucial to allow for a sufficient length of follow-up to determine the degree to which a given intervention is likely to achieve sustained change in a given health outcome measure.

While it will be important to use valid and reliable measures, changes in key outcomes in the absence of a suitable control remain difficult to interpret, because of the regression to the mean phenomenon, suggesting an intervention effect that might still have occurred in the absence of the intervention.

3.2.2　Measuring intervention effect: costs

Disease management was first mentioned as a concept in the United States in the 1980s,[18] motivated initially by the prospects of such an intervention reducing hospital (re)admissions and hospital days, and, thus, controlling healthcare costs.[105] The potential of structured approaches to chronic disease care to reduce costs in the long term has remained a key aim for decisionmakers promoting the experimentation with innovative approaches to care,[18 62] although, as noted earlier, the evidence for such interventions to actually do this remains uncertain.[5 28]

Approaches to economic evaluation of healthcare interventions have been described in detail by, for example, Drummond et al. (2005),[106] and we have previously reviewed methodological considerations for capturing the economic impact of disease management

interventions in particular.[8] In brief, depending on the perspective taken (patient, provider, funder, society as a whole), comprehensive economic evaluation will have to take account of both the costs associated with the intervention (eg out-of-pocket expenses; set-up and operation costs) as well as the benefits accrued from improved chronic disease management, such as improved health experienced by the patient, long-term cost savings from complications avoided and reduced healthcare utilisation, and/or workplace productivity gains.[107] In practice, evaluations have tended to focus on expenditure incurred by the funder, for example the health insurer, frequently lacking a complete accounting of all relevant costs.[108]

In the context of evaluation, choice of design will be important for determining the measure of cost.[109] 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 (if any) were achieved. 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. 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 markedly reduce specialist utilisation.[8]

Linden et al. (2004) argued that economic impact is most appropriately measured indirectly, given that disease management interventions tend to use utilisation measures rather than cost as outcome measures.[45] Utilisation is price insensitive and can serve as a proxy for measuring financial return of investing in disease management.[110] Utilisation measures are considered to be less vulnerable to bias than cost over time. Recent work by Steuten et al. (2009) exemplified the use of utilisation measures to estimate short-term cost-effectiveness of chronic care programmes for people with COPD.[111]

Introducing measures of cost can provide additional insight into the impact of a given intervention although findings remain difficult to interpret

Evaluating a diabetes disease management programme (DMP) in Germany for the period 2004–2008, the DISMEVAL case study found evidence for improved process parameters in the group of diabetic patients participating in the programme. Thus, compared to a control group of patients not participating in the programme, a larger proportion of participants received an annual eye examination and regular measurement of HbA1c levels.[71] However, these improvements were accompanied by higher utilisation and costs as derived from standard fee schedules applicable to reimbursement within the German SHI system. Thus, DMP participants showed an increase in the number of outpatient visits, in prescription rates as well as overall costs, although inpatient days or hospital costs did not differ from the control group (Figure 3.6).
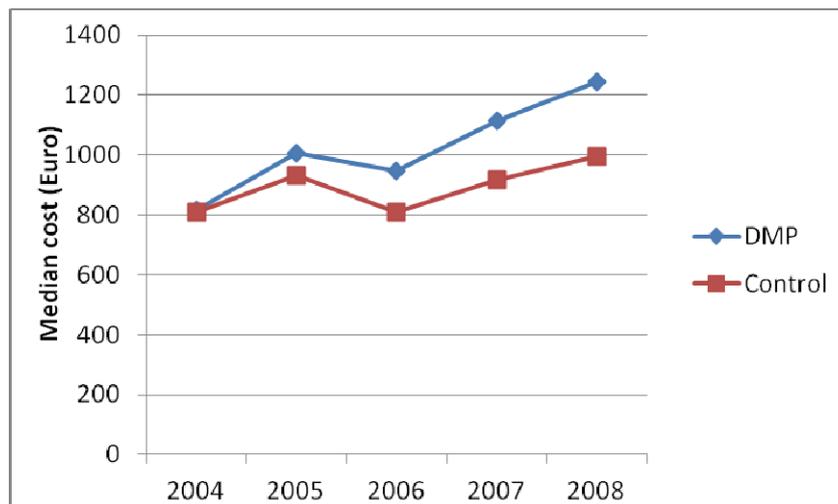
Figure 3.6 Median overall costs (Euro), per patient, in DMP and control groups, Germany, 2004–2008

The Dutch case study sought to understand whether observed differences in programme effect across different regional diabetes disease management programmes might be associated with the price of the intervention under investigation.[102] The intervention is delivered by so-called care groups – provider networks in primary care – who negotiate with health insurers a price for the package of services for diabetes they provide, on the basis of 'bundled payment contracts' (Box 3.3).

Box 3.3 Bundled payment contracts for diabetes care in the Netherlands

In the Netherlands, the 2006 health insurance reform facilitated the development of new forms of service delivery and payment for more integrated care. This involved the initially pilot-based and diabetes-focused establishment of GP-formed 'care groups' who contract with health insurers on the basis of a 'bundled payment' for a defined package of (diabetes) care.[112] This approach was strengthened by the 2008 'Programmatic approach to chronic illness care' and proposals to generally fund chronic care through bundled payment schemes, accompanied by regulatory measures to strengthen the role of nurses in the care for chronically ill.

The care group (*zorggroepen*) is a legal entity that brings together providers in primary care (mostly general practitioners and affiliated personnel) and which enters into a contract with a health insurer to provide a package of care for a given condition according to a nationally developed care standard ('bundled payment contract'). The price for the care package is negotiated between the provider care group and the insurance fund on the basis of the performance of the care group and the expected case-mix of patients. Conceived as an 'experiment' in 2006, the government subsequently decided to roll out this strategy nationally for the delivery of care for patients with diabetes, COPD or vascular risk.[113]

The case study collected the price information from nine care groups and incorporated these into a univariable meta-regression model (for further details see Section 3.3.2). It found that, across the nine groups, prices ranged from approximately €299 to €458 per

patient per year, with a median of €367.48. However, despite this variation, the analysis identified a significant linear relationship between price and effectiveness of disease management only for metabolic control (HbA1c level) (Table 3.4). The negative regression coefficient for HbA1c suggests that patients treated in more costly DMPs achieve greater reductions in this clinical measure than their respective counterparts. Similar trends were observed for all other outcomes, although these did not achieve statistical significance. The only exception was HDL, for which the negative regression coefficient identified in the meta-regression suggests that patients in lower-priced care bundles achieve better results. Importantly, the heterogeneity in intervention effect between groups reduced after adjusting for differences in care bundle price for most of outcome variables.

Table 3.4 Univariable meta-regression results for the covariate 'care bundle price', the Netherlands

	Direction of the regression coefficient	Change in between-group variance (%)
HbA1c	–*	–50.6
Total cholesterol	–	–7.5
LDL	–	–5.8
HDL	–	–3.7
Triglycerides	–	–2.2
Systolic blood pressure	–	–17.0
Diastolic blood pressure	–	1.2

NOTES: * Statistically significant (p<0.05)

Introducing measures of cost can usefully highlight how lack of a controlled design can misrepresent an observed intervention effect

The German DISMEVAL case study demonstrated how failure to adjust for baseline differences between the intervention (DMP) group and the control group prior to enrolment into the programme can lead to observations that may wrongly be interpreted as an intervention effect. Thus, using the example of prescription costs, it found unadjusted data to point to a marked difference between costs for patients participating in the DMP and those for the control group. For 2005, this difference was calculated at €183.39 per DMP participant compared to the control group; this difference might be interpreted as an effect caused by the intervention. However, analysing baseline data before enrolment in 2004 found that those who were to enrol with the programme already had higher prescription costs than those who chose not to participate (difference of €108.10). Adjusting for this baseline difference yielded, for 2005, a cost difference between the two groups that was about less than half that calculated in the unadjusted model (€70.94) (Figure 3.7).
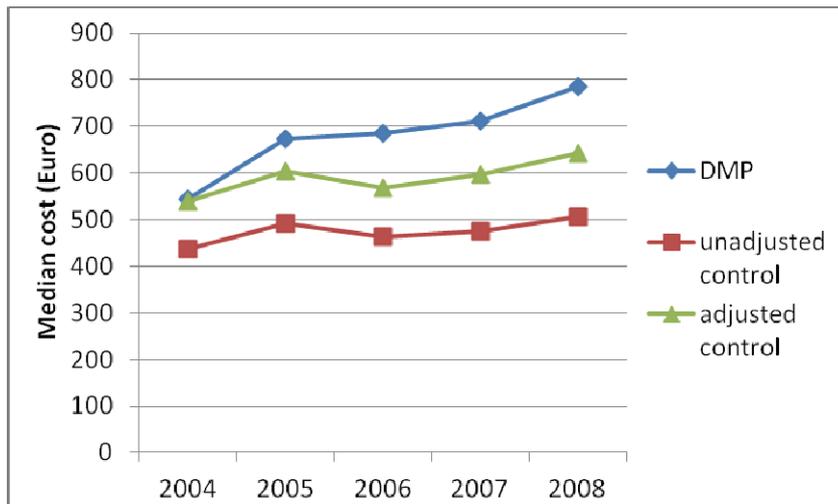
Figure 3.7 Prescription costs with and without adjustment for baseline differences between DMP and control group, Germany, 2004–2008

This point is further illustrated by the Austrian case study of the DISMEVAL project.[73, 114] It demonstrated, for a diabetes disease management programme ('Therapie Aktiv') established in Salzburg province in Austria, how an uncontrolled evaluation design led to an overestimation of programme net-effect for risk reduction of clinically relevant endpoints such as myocardial infarction and diabetes-related complications within ten years. It combined published data on the risk for patients with diabetes to develop long-term complications, as identified by the UK Prospective Diabetes Study (1998)[115] and a meta-analysis performed by Selvin et al. (2004),[116] with the observed decrease in HbA1c levels in the intervention and control groups in the Austrian DMP over a period of 12 months (at respectively 0.41 per cent and 0.28 per cent). In doing so, it was able to assess the impact of the control group on the interpretation of the number needed to treat (NNT) and the economic impact of the Austrian DMP.

Table 3.5 Effect of the DMP on relative risk reduction, absolute risk reduction and number needed to treat, Austria

	Uncontrolled pretest-posttest comparison	Randomised controlled comparison
HbA1c reduction (%)	0.41	0.13
Relative risk reduction cardiovascular disease (%)	4.6	1.4
Absolute risk for myocardial infarction / 10 years (%)	17.4	17.4
Absolute risk reduction for myocardial infarction (%)	0.80	0.24
Number needed to treat to avoid one case of myocardial infarction / 10 years	*125*	*417*
Relative risk reduction for any diabetes complication (%)	5.1	1.6
Absolute risk for any diabetes-related complication/ 10 years (%)	46.0	46.0
Absolute risk reduction for any diabetes-related complication (%)	2.35	0.74
Number needed to treat to avoid one case of diabetes complication/ 10 years	*43*	*135*

Figures in Table 3.5 shown how, using an uncontrolled before–after design, the estimated number needed to treat (NNT) to avoid one case of myocardial infarction or one case of any diabetes-related complication within 10 years was overestimated by a factor of three. This effect was then monetarised by applying reimbursement rates for physicians participating in the DMP, who receive a one-off payment for the initial examination for each patient joining the DMP plus additional payment for follow-up examinations every three months. Using current reimbursement rates (2010) and not accounting for inflation, the estimated reimbursement for each DMP patient translated into €1,066 in ten years. Combined with the number needed to treat for each design and endpoint as shown in Table 3.5, the estimated costs for the DMP to avoid one case of myocardial infarction within one year were then estimated to differ by more than €300,000 between the uncontrolled and controlled design; the difference was just under €100,000 for avoiding any diabetes-related complication within ten years (Figure 3.8).
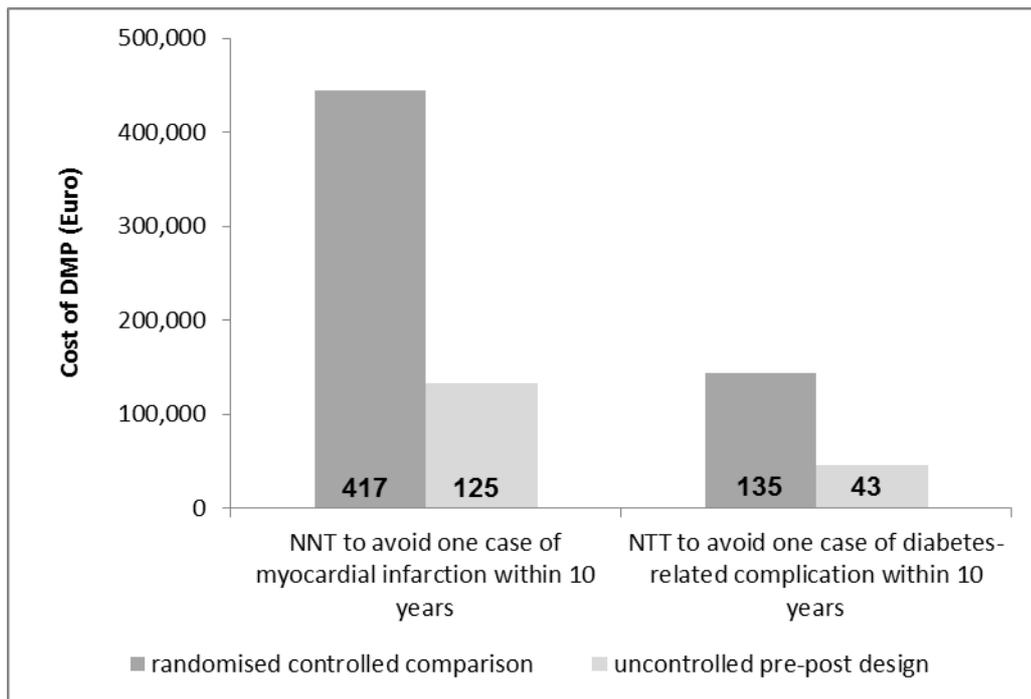
Figure 3.8 Estimated costs based on uncontrolled pretest-posttest vs randomised controlled comparison

NOTE: adapted from Flamm et al. (2011)[114]

Calculations only considered actual expenditure as operationalised by physician reimbursement; other costs such as those associated with programme development and evaluation costs, as well as costs induced by DMP-related diagnostic measures and medication, were not considered. Findings as presented in Figure 3.8 should therefore not be interpreted as absolute cost figures but rather as illustrating the potential for misinterpretation when relying on an uncontrolled evaluation design. Detailed assessment of return on investment and cost-effectiveness of a DMP would require further analysis and comparison of DMP-related costs to determine the savings achieved by the avoidance of myocardial infarction and other diabetes-related complications.

Summary

In summary, this section has provided examples of approaches to introducing cost estimates into disease management evaluation. The Dutch case study has usefully illustrated how estimates of cost, here conceptualised as price of a given intervention, can be used as a means to understand a possible association with intervention effect. The German case study found evidence for a disease management intervention to increase costs because of increased utilisation of GP visits and prescription, although utilisation of inpatient services did not change. Both the German and Austrian case studies provided important insights on how failure to adjust for baseline differences in intervention and control group, or failure to use a controlled design overall, may lead to substantial misrepresentation of intervention effect, with the Austrian example further illustrating how indirect estimates such as number needed to treat can be used to estimate long-term intervention effect.

3.3   Advancing methodology

Although the DISMEVAL project was primarily concerned with testing and validating different methods to evaluating disease management approaches, project work also offered opportunity to *advance further the existing portfolio of evaluation methods* and this section will illustrate examples from selected DISMEVAL case studies to inform the design of evaluation approaches elsewhere. These include:

- assessment of cluster randomisation as a 'pragmatic' experimental approach in DMP evaluation where experimental evaluation is feasible (Austria)

- testing of different evaluation designs (meta-analysis and meta-regression) to inform the advancement of evaluation approaches (the Netherlands)

- assessment of the impact of intervention intensity on estimate of effect (delivered dose analysis) (Spain)

- assessment of the intervention effects of different intervention components (the Netherlands).

3.3.1   Cluster randomisation can provide a 'pragmatic' experimental approach where experimental evaluation is feasible

Cluster randomisation is widely considered a feasible 'pragmatic' approach for the evaluation of complex interventions.[7, 117-118] The cluster design has been proposed as one solution to overcome the risk of contamination of the control group in population-level interventions.[7]

The implementation of the diabetes disease management programme 'Therapie Aktiv' in Salzburg, Austria, was conceptualised as a pragmatic cluster-randomised controlled trial with an observation period of 12 months. Randomisation was carried out at the district level of Salzburg province, resulting in a three-level cluster design in which the individual GP practice was nested within the district, and patients were nested within the GP practices.[75] Randomisation at district level was chosen to minimise contamination bias, however individuals were asked to give informed consent as the intervention is provided at the individual level. Individual randomisation was considered to be inappropriate because it was deemed not feasible for a single GP to treat some patients according to usual care and others according to the intervention protocol. Randomisation at the GP level would have also led to contamination bias because of overlapping patient groups, especially in rural areas. Blinding of physicians or patients was not possible due to the complexity of the intervention. This might have led to potential selection bias and had to be taken into account, as noted earlier (Section 3.1.1).

One important finding of the evaluation was that the Austrian DMP 'Therapie Aktiv' did not significantly improve metabolic control as measured by HbA1c after one year. However, the programme significantly improved indicators of process quality. Importantly, from an evaluation perspective, further analysis within the Austrian DISMEVAL case study that compared the findings of the randomised trial with an uncontrolled before–after design demonstrated a considerable overestimation of the programme net-effect, finding a more than threefold overestimation of reduction in HbA1c levels (see also Section 3.2.2).[73] Furthermore, risk reduction for clinically relevant

outcomes such as cardiovascular events was overestimated by more than 200 per cent within the uncontrolled setting.

In conclusion, the experience of the evaluation of the disease management programme in Salzburg province in Austria demonstrates that using experimental design in operational practice is feasible. Therefore, as DMPs are resource-intensive public health interventions, randomised controlled evaluation prior to programme roll-out should be indicated where possible; however, as we shall see further below, such designs also require considerable commitment of resources.

3.3.2 Advanced evaluation designs can help understanding 'what works for whom'

The primary aim of many disease management evaluation methods is to obtain a single effect size across many patients. Such a measure may, however, be less relevant for clinical practice where individual patients are treated, and it gives little guidance on which patient groups will benefit most from which intervention component.[119] In order to allow for a more granular assessment of the impact of components of the intervention on different patient groups, the Dutch case study used the techniques underlying meta-analysis of individual patient data[120-121] to assess the differential effects of 18 regional disease management programmes (DMPs) for type 2 diabetes on an array of intermediate clinical outcomes.[102] The intervention under consideration is delivered by so-called care groups that provide the DMPs for diabetes (Box 3.3, page 31).

The analysis used a two-step approach in which the individual patient data were first analysed independently per care group by means of paired-sampled t-tests (two-sided; $\alpha=0.05$). In the second step, the group-specific mean differences and associated standard deviations were synthesised into pooled treatment effects and 95 per cent confidence intervals by means of a random-effects meta-analysis model. Subsequent analyses involved examining the extent and potential causes of the heterogeneity in effects of the Dutch DMPs for diabetes (Box 3.4).[119]

Box 3.4 Quantifying heterogeneity in intervention effects in Dutch diabetes care groups

To quantify heterogeneity in effects, the I^2 statistic was calculated on the basis of the chi-square test. This statistic describes the percentage of total variation across care groups that is due to heterogeneity rather than chance and can be calculated from the basic results of a meta-analysis as:

$$I^2 = 100\% * (Q - df)/Q$$

where Q is Cochran's heterogeneity statistic and df the degrees of freedom. The principle advantage of I^2, which lies between 0 and 100 per cent, is that it can be calculated and compared across groups irrespective of differences in size and type of outcome data.[122] For outcomes showing moderate ($I^2>50\%$) to high ($I^2>75\%$) variance in effects, subgroup analyses were conducted to estimate how relevant characteristics, such as care group size, length of follow-up, frequency of clinical measurements, age and baseline health status, modify treatment effect. In all subgroup analyses, the study compared the level of heterogeneity before and after correcting for differences in these characteristics to determine whether the inconsistency in treatment effects (measured by I^2) was reduced once variation in these features was accounted for.

Meta-regression analyses were used to further examine heterogeneity in effects. Contrary to subgroup meta-analysis, meta-regression can examine multiple individual and group level factors together, the results of which can facilitate more stratified care approaches (see also Section 3.3.4). Analyses can even be adjusted for baseline (prognostic) factors, which can increase the power of the analysis to detect a true treatment effect.[123] Moreover, in contrast to meta-analysis, which requires categorisation into subgroups, meta-regressions can include continuous variables and identify potential linear relationships between, for instance, a patient's age and treatment effects.[124]

The analyses found that using a simple, observational study design would lead to the conclusion that the Dutch DMPs for diabetes do not achieve their intended goals. Over a median period of 11 to 12 months, minor average improvements were attained in total, LDL and HDL cholesterol, systolic and diastolic blood pressure, and BMI, whereas a slight deterioration occurred in HbA1c levels. However, applying meta-analysis and meta-regression methods, the analyses found effects to be differential, demonstrating that patients with poor baseline values showed clinically relevant improvements in all included outcomes. Thus, patients with first-year HbA1c levels of >75 mmol/mol, for example, achieved a mean reduction in this clinical measure of 16.8 mmol/mol during follow-up.

As noted earlier (Section 3.2.1), the analyses further demonstrated that the positive effects of the DMPs on clinical outcomes tended to decrease with longer-term follow-up. In addition, the findings further suggested that a greater measurement frequency of clinical measures was associated with progressively greater improvements in these measures, especially in patients with poorly controlled diabetes. The poorer patients' values of a particular endpoint at baseline, the higher the benefit accrued by frequent measurement of that clinical outcome in terms of achieved improvements. A broader measurement range was associated with significantly greater improvements in HbA1c, HDL cholesterol, and triglycerides. For the majority of these outcomes, measuring a broader range of indicators during follow-up became more beneficial as patients' baseline values were poorer.

Also, heterogeneity in intervention effect was more prevalent within than between care groups, and adjusting for known patient characteristics such as age, disease duration, baseline health status, co-morbidity and smoking status, substantially reduced within-group variance.

The Dutch case study aimed to further our understanding of how various processes and outcomes of care interact, taking into account the characteristics of the target population that might influence the effectiveness of certain care services.[102] It used meta-regression techniques to investigate the existence of linear relationships between three care process characteristics – the frequency and range of clinical outcome measurements and length of follow-up – and clinical outcomes, as well as potential three-way interactions between these process characteristics, outcomes, and various features of the target population.[102]

In conclusion, by enabling investigations of 'what works for whom', the use of meta-analysis and meta-regression techniques advances current approaches to disease management evaluation. Compared to the single treatment effects that inform many current healthcare reforms, using the differentiated insights gained from analysing the consistency of effects across care settings, care processes and patients can better support healthcare professionals and policymakers in their efforts to redesign chronic care.

3.3.3 Consideration of intervention intensity can provide important insights into differential levels of impact

The Spanish DISEMEVAL case study assessed the effects of a nurse-led, telephone-based intervention targeting a working-age population identified to be risk of cardiovascular disease.[125] Specifically, it aimed to undertake a delivered dose analysis, in terms of natural progression of independent cardiovascular risk factors, according to three groups: control group, low intervention (partial intervention), and high intervention (complete intervention). Partial intervention refers to those receiving only one or two of the structured, nurse-led telephone interviews instead of the three structured telephone interviews that were delivered in the complete intervention.

The analysis carried out a propensity score matching method in order to compare the control group to the intervention groups. Propensity score matching using the nearest neighbour algorithm, applying logistic regression. Variables considered for matching were: sex, age, occupation, smoking habits, total cholesterol, height, weight, creatinine levels, systolic and diastolic blood pressure.

The analyses found that, after 12 months, there were statistically significant improvements in the high intervention group compared to the control group in the following variables: body mass index, systolic blood pressure, diastolic blood pressure and level of cardiovascular risk as assessed by the SCORE (European Coronary Risk Evaluation[126]) risk chart. In contrast, partial intervention was found to be less effective. Specifically, there were no improvements in diastolic blood pressure or SCORE, although body mass index and total cholesterol levels did improve.

In conclusion, the use of a delivered dose approach provides important insights into the differential impacts of interventions and how evaluation can contribute to better understanding of the (potential) effects of suboptimal intervention implementation.

3.3.4 Disease management interventions typically comprise of different components, and methods are available to consider these simultaneously

By focusing on a single treatment effect, most disease management evaluation methods do not allow estimation of the effects of the different elements of a programme. As a consequence, questions concerning how best to redesign care for chronically ill patients are difficult to answer. One method that addresses this shortcoming of current disease management evaluation methods is meta-regression analysis. Similar in essence to simple regressions, meta-regression entails predicting an outcome variable (effect size) according to the values of one or more explanatory variables.

The Dutch case study used meta-regression methods to gain a differentiated insight into the effectiveness of the four practice-level components of disease management according to the Chronic Care Model (CCM), that is, self-management support, delivery system design, decision support and clinical information systems.[127-128] It used the Dutch version of the Assessing Chronic Illness Care (ACIC) instrument,[129] which assesses the degree of implementation of the four CCM components alongside their level of integration. This involved three professionals – a director/manager, GP and physician assistant – from nine care groups, who were invited to independently complete the ACIC survey. Respondents rated the four to six items per survey element on a four-point scale, with higher scores indicating greater implementation. The quantitative data gained from the ACIC survey

was cross-validated through semi-structured interviews with all participating professionals (N=27) as well as a document study of care groups' diabetes care protocols and annual reports.[102]

The mean ACIC scores per care group were included as covariates into univariable random-effects meta-regression models described earlier to determine how the level of implementation of the different components of high-quality disease management modifies effects. However, possibly due to the small spread in the group-specific scores, the meta-regressions yielded few statistically significant linear relationships between the ACIC elements and the effects of the diabetes DMPs on clinical outcomes. One positive linear relationship was found with the ACIC element delivery system design, suggesting that care groups with a higher rating of this component achieve greater results on HDL cholesterol. Moreover, the effects on LDL cholesterol and systolic blood pressure were significantly more positive in groups with a higher rating of the integration of CCM elements.

In conclusion, despite inconclusive quantitative evidence concerning the effects of the different elements of the CCM on clinical outcomes, characterising DMPs according to this model provided valuable insights into the working mechanisms of disease management for diabetes. It demonstrates how it is possible to bring together quantitative and qualitative elements within an evaluation to better understand 'what works' in what circumstances. The findings of the analyses appear to point to the need to move beyond the current approach of diabetes care in the Netherlands towards a more tailored approach that takes account of the level of risk of participants. Similar to the Spanish case study, the use of approaches that, in the case of the Dutch example, involved meta-regression methods provides important insights into the differential impacts of a given intervention. It further illustrates how evaluation can contribute to the better understanding of the (potential) limitations of existing interventions in achieving desired outcomes.

Practical considerations for evaluation

Work undertaken within DISMEVAL has identified a range of challenges, which may be deemed necessary to consider for users, funders and researchers interested in the evaluation of structured approaches to chronic disease management. Some of the practical considerations concern the actual data to be used, such as their quality, completeness and sources, its accessibility and management, the availability of and familiarity with analytical tools and capacity, as well as broader issues around costs. This chapter will draw to a large extent on findings of work undertaken within DISMEVAL, providing illustrative examples from case studies where feasible and appropriate.

4.1 The need for and quality of data is an important consideration in any evaluation

4.1.1 Routine data are an important source for evaluation

Data availability constitutes a considerable challenge to evaluation research, even where evaluation design and capacity is of high quality. Where routinely collected data is used, this may be incomplete and may require systematic scrutiny to assess the implications of missing data for the analysis and interpretation of findings; this may require additional resources where missing data has to be imputed from elsewhere. Even where routine data is fairly complete, it may be inadequate for the purpose of evaluation as it is typically used for administrative purposes only and may not record outcomes of interest. Evaluations may therefore necessitate new data collection, with consequent resource implications.[130]

Routinely collected data provide a valuable source for disease management evaluation but their use requires good understanding of their limitations

The German case study within the DISMEVAL project aimed to compare different techniques to adjust for confounding in an evaluation of a diabetes disease management programme (DMP) (Section 3.1.2).[71] It was able to draw on routine data from a large statutory health insurance (SHI) fund from three regions in Germany for the years 2004 to 2008. Routine data collected by SHI funds contain socio-demographic variables of insured members, such as age, sex, insurance group (employed, retired, family member) and geographical region, as well as data collected for reimbursement purposes from hospitals, pharmacies, GPs, specialists, other healthcare providers, other social insurance agencies and employers. These include diagnosis, medication, procedures and outpatient reimbursement codes as well as costs (eg for hospitalisations, medication, home healthcare, aids and devices, sick days).

SHI funds offering DMPs (see Box 2.1, page 8) also have available additional data on their members who participate in the given DMP(s). Although the regulatory framework for DMPs mandates for detailed information on programme performance to be documented (Box 4.1), for reasons of data confidentiality, SHI funds only receive an abridged version of the data set that does not contain clinical parameters such as HbA1c levels, blood pressure levels, body mass index or smoking status. Thus, data from SHI funds that can be potentially used for purposes of evaluation is limited to routine data.

Box 4.1 Mandatory data collection and evaluation of DMPs in Germany

The regulatory framework for DMPs in Germany sets out standards and measures for quality assurance, including for the documentation of information on patients participating in a given DMP. Documentation requirements include administrative data, information on the patient's condition, test results, medication regime and others.[33] This data is collected by the DMP physician, usually a GP, and documented at every follow-up visit. The data is then transferred electronically to an independent agency in order to perform the statutory evaluation.

An overview of the principles of the statutory evaluation is presented by Siering (2008).[33] In brief, its principle aims are to verify that programme targets are reached, that criteria for assessment are adhered to and that the costs of care and patient quality of life within DMPs are assessed. The minimum requirements for statutory evaluation are set by the Federal Insurance Office (BVA); evaluation costs are borne by SHI funds. As the statutory evaluation follows an observational, non-experimental design, it does not permit comparison of the quality of care provided in DMPs with usual care. Also, although it involves a longitudinal design, as the evaluation period has to cover three years, the interpretation of observed temporal changes among participants is difficult; the evaluation is not based on individual-level data but on average data across patients without adjusting for attrition.

Although the German DISMEVAL case study was able, in part, to draw on routine data from a statutory health insurance (SHI) fund already available to the research team in the framework of an earlier research project, additional data had to be obtained on DMP documentation. This data had to be extracted separately and then matched to the existing database. Routine data are collected for reimbursement purposes and different reimbursable items such as prescription, diagnosis and utilisation data are stored in separate tables. The collation of data tables for research purposes can be difficult and data cleaning and validation is time-consuming. Items do not necessarily match up when transferred to a single database; there may be missing values or incorrect and implausible entries. The DMP data set initially requested was found to be deficient, with several data fields not completed or missing specific values, requiring several rounds of corrections for converting the data set into one suitable for analysis, with implications for the timeframe within which to undertake analyses and resources to be committed.

Further scrutiny of data uncovered a substantial flaw, which only emerged during actual analyses concerning the outcome 'mortality'. This showed that the routine data obtained were incomplete with regard to the year of death for some individuals (see also Section 4.1.3). The DISMEVAL project therefore had to request supplementary data. However, as the data concerned had already been partly archived by the SHI fund, the process was

time-consuming and had to be first approved by the board of the SHI fund. This resulted in an additional delay and data for 2004 are still incomplete to some degree.

This experience illustrates that while the use of routine data from SHI funds has the advantage of relatively easy access and availability of data in electronic format, providing data on a large patient sample, the amount of time needed for data cleaning and multiple rounds of validity checks should not be underestimated and should be accounted for in any analysis plan.[131-132] Furthermore, the range of outcome variables to be considered for evaluation should be carefully chosen, taking into account the limited validity of a data set which was not collected for research but for administrative purposes.

Newly established data documentation on disease management interventions requires supportive infrastructure for data to be usefully utilised for evaluation

In the Netherlands, the implementation of bundled payment contracts for diabetes disease management (see Box 3.3, page 31) was accompanied by record-keeping obligations for care groups. These data, covering a specified number of care processes and outcomes, must be delivered to health insurers on an annual basis to allow for monitoring of services provided and their quality. The Dutch DISMEVAL case study used data gathered by nine such care groups retrospectively from their clinical information systems, covering a period of 20 or 24 months between January 2008 and December 2010.[102] In addition, the case study used data from nine further groups that were part of the national pilot phase, implemented in 2007, and provided by the National Institute for Public Health and the Environment (RIVM). The evaluation of these nine pilots conducted by the RIVM is described in detail elsewhere.[133]

Of the total of 18 care groups considered in the Dutch case study, only 2 groups were able to provide data on all 19 requested patient characteristics, care processes and intermediate health outcomes. The maximum number of missing indicators per care group was seven. Within the data provided, a considerable number of patients lacked (valid) registrations of one or more of the included variables. For illustrative purposes, Figure 4.1 sets out a flowchart for patient data on the indicator 'HbA1c level'.
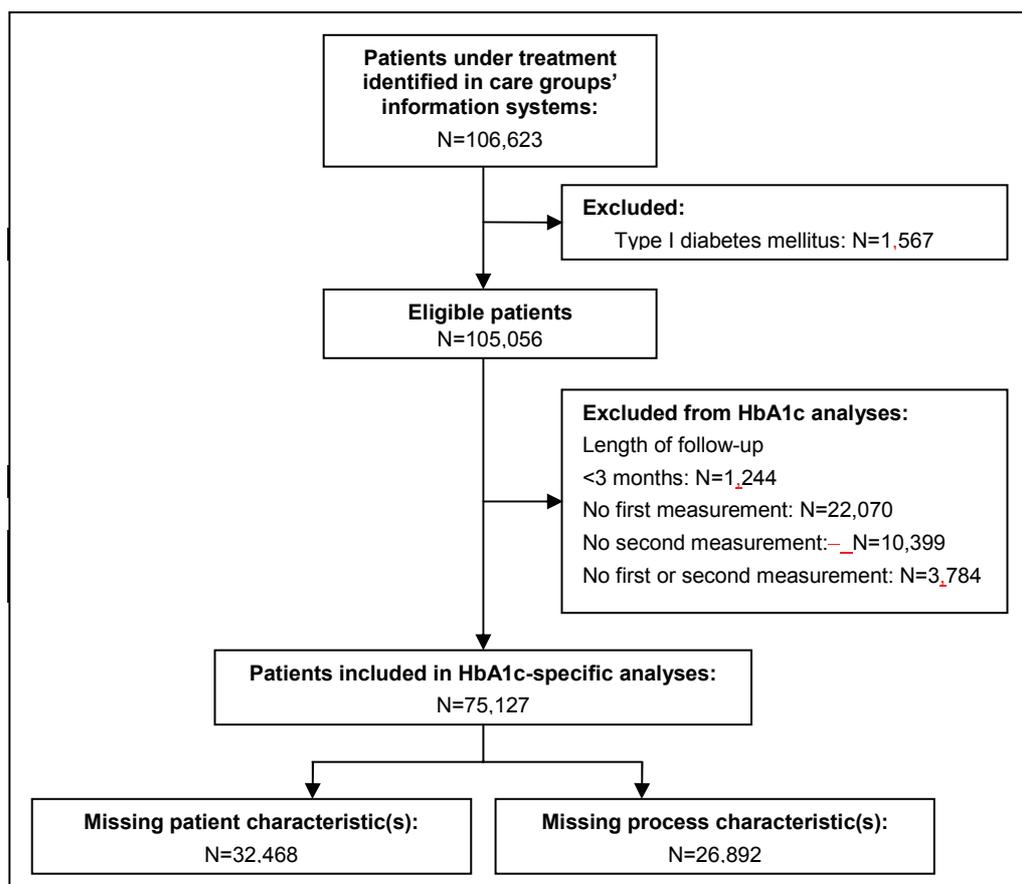
Figure 4.1 Flowchart for data availability on the indicator 'HbA1c level' in the Dutch case study

The main reason for lack of completeness of data available for the Dutch case study was the relative inexperience of care groups with the operation of structural data registration, alongside shortcomings in the quality of supporting clinical information systems.

As noted above, data were collected retrospectively, which is likely to have introduced bias as only a small set of analyses could be conducted on the basis of data from all groups and all patients. While it is difficult to assess the size of bias, the research population did not differ systematically from other diabetic populations studied in the Netherlands on a range of variables such as average age, disease duration or the percentage of smokers.[102] However, the prevalence of co-occurring conditions was lower than would have been expected based on the characteristics of the total population of Dutch diabetic patients; this is likely to be attributable to registration problems. Retrospective data collection further limited the choice of effect measures to the included set of intermediate clinical outcomes; definite outcome measures such as health-related quality of life, self-efficacy and patient satisfaction were, however, not available.

On the other hand, collecting data retrospectively meant that a large number of patients could be included, capturing information on daily practice of healthcare provision and registration. It also permitted assessment of relatively long-term effects within a limited time frame; furthermore, financial costs of data collection were low. These considerations

present important motivations for evaluations seeking to assess the context-specific effects of large-scale, population-based disease management programmes, potentially outweighing challenges associated with a considerable volume of missing data. Furthermore, prospective data collection on chronic care interventions in daily operation does not necessarily ensure more complete datasets. In the Dutch case study, there were no consistent differences in terms of number of missing values between the nine pilot care groups, for whom data was gathered prospectively by the RIVM, and the nine groups that provided data retrospectively from their clinical information systems.[133]

Similar challenges apply to analysts seeking to evaluate disease management approaches in Denmark. The Danish DISMEVAL study was able to draw on data from a specifically designed and funded (research) project and so had privileged access to a rich database that is not typically available to evaluators in routine settings (see Section 3.2.1).[78] However, with the roll-out of disease management programmes in Denmark (Box 4.2), some municipalities have implemented and are recording data sets that also include clinical data concerning rehabilitation, which is a part of DMPs. At present these data sets are mainly used for administrative purposes and a large proportion of data currently contained in these data sets are not suitable for DMP evaluation due to incompleteness. This is explained, in part, by clinical practice not yet being 'routinised' in terms of, for example, the selection of tests to be conducted, as well as attrition. Also, the quality of supporting IT systems has remained limited so far.

Box 4.2 Regional disease management programmes in Denmark

The five regions of Denmark are in the process of developing disease management programmes (DMP), working from an initial programme template developed by a working group at the National Board of Health in 2008.[134] Each region in Denmark is responsible for the development of its own DMP in the expectation that regional DMPs will operate with each other in addressing most common chronic conditions, so that all patients with a chronic condition are covered by a specific DMP and receive the health services and provisions described in the programme.

The actual content of regional DMPs is likely to differ, but the differences are not essential. Progress has varied with, for example, the Capital Region of Denmark having developed and approved DMPs for diabetes, COPD, dementia and cardiovascular diseases, with programmes for musculoskeletal disorders planned. DMPs for COPD and diabetes type 2 started to being implemented in 2010. The Central Denmark Region has developed and approved DMPs for diabetes, COPD and cardiovascular diseases while the Region of Southern Denmark is currently developing DMPs for diabetes, COPD, cardiovascular diseases and musculoskeletal disorders. The North Denmark Region is currently planning DMPs for dementia, cardiovascular diseases and COPD, whereas the Sealand Region has yet to develop related plans.

Heterogeneity of routine data collection systems poses challenges for their use in evaluation, highlighting the need for greater collaboration between those responsible for data collection

The French DISMEVAL case study used data from two separate diabetes provider networks.[69] In France, all health provider networks operate a routine database on administrative data and activity documentation such as the frequency and dates of therapeutic education workshops provided. However, operation of a clinical database is

mostly limited to provider networks that have a critical size and thus sufficient resources to set up such a database; those targeting conditions for which process and outcomes can be easily measured; and/or those with an inherent 'evaluation logic', as would be the case for provider networks that originate from an academic setting.

Diabetes provider networks are generally large, covering more than 1,000 patients, as diabetes is a relatively common condition and diabetes can be characterised by measures that are relatively easy to quantify (HbA1c, other laboratory parameters). Some diabetes provider networks explicitly target the improvement of measurable glycaemic control in their objectives and therefore frequently have established quantitative clinical databases that are relatively complete. Another example is provider networks for patients with multiple sclerosis, as this condition is often diagnosed and treated in academic settings, and clinicians who initiated the provider networks also have in interest in the scientific evaluation of their work and collect the necessary data.[135] This data is available for statutory evaluation use also.

Overall, there is considerable heterogeneity in data availability and completeness for French provider networks. While the relative lack of sufficiently detailed data hampers rigorous evaluation of provider networks in France at present, steadily growing network size and shared resources between networks (see also Section 4.1.3) are likely to improve this situation in the future.

Summary

This section has highlighted the opportunities and challenges arising from the use of routinely collected data for disease management evaluation. Routine data have the advantage of relatively easy access and data are available for a large patient sample, capturing information on daily practice of healthcare provision and permitting assessment of relatively long-term outcomes, such as mortality within a limited timeframe. Furthermore, financial costs of data collection tend to be low. These advantages potentially outweigh the challenges associated with the considerable volume of missing data that tends to be common, in particular for data collection systems that are newly introduced and/or for which validity checks are not routinely carried out. Therefore, any evaluation that makes use of routine data systems will have to allow sufficient time to enable validity checks and a good understanding of the context within which data collection takes place, in particular in relation to the range of outcome variables to be considered for evaluation where data is not collected for research but for administrative purposes.

4.1.2 Data access and confidentiality can act as a barrier to evaluation

Access to routine data for evaluation can present difficulties where their use requires approval from the relevant data holder (eg a health insurance fund), which may delay the use of the data; also, the data holder may impose certain restrictions on the use that may reduce the ability to employ the full potential of data for evaluation analysis (eg confidentiality issues may prevent disaggregation of data to smaller geographical units, which may be used to create controls). This is likely to be the case where data linkage is being considered as a means to analysis.

Also, where data is managed in a database external to the research team, this may have (additional) cost implications. This is particularly likely to be the case where not the entire data set is used but the analysis requires a subset to be extracted by the holder of the

database – or else the extraction is undertaken by the research team itself, so requiring relevant computing expertise and IT capacity.

Access to routinely collected data may require additional provisions and resources

Within the DISMEVAL project, the Spanish case study provides an illustrative example of restrictions to physical access to data. As noted previously, the case study aimed to assess the effects of a nurse-led intervention targeting a working-age population at risk of cardiovascular disease in Spain.[125] The intervention is offered by *Ibermutuamur,* a Mutual Fund for Workers' Injuries and Occupational Diseases and implemented through their network of healthcare professionals. Data are collected by Ibermutuamur and, although the research team was granted access, it was permitted to do so on Ibermutuamur's premises only.

Elsewhere, data extraction may require additional resources. For example in Denmark, as in other Scandinavian countries, government agencies operate a range of registries and routine administrative data collection systems including, for instance, vital statistics, disease registers and hospital admissions data.[136] Furthermore, as noted earlier, it is possible to link different sets of routinely collected data using a unique identifier, the social security number for instance, through Statistics Denmark.[137] Thus, to assess the effect of rehabilitation on healthcare utilisation in COPD patients within the Danish work package in DISMEVAL, it was possible to use (anonymised) linked data from the national registers. Given the richness of the data registers, it was possible to obtain data on a range of utilisation measures such as GP and specialist consultations, hospitalisations and bed days, emergency room visits and medication, by diagnosis. In addition, data was available on a wide range of socio-demographic variables beyond age and sex, such as ethnicity, marital status, education, employment status, income, assets and equivalent co-morbidity (as measured by summary dose of medications). This data was then used to construct a control group using propensity score matching.

Accessing these data is not a problem as such and data required for evaluation analyses can be extracted by the holder/manager of the database. However, this has cost implications. For example, one hour of statistician time at Statistics Denmark is costed at around DKK 1,500 (€200). With data extraction taking anything between 10 to 30 hours or more, depending on volume and specificity of data requested, this corresponds to between DKK 15,000 and 45,000 (€2,000–6,000) that would be have to set aside for the evaluation.

Data confidentiality can restrict the usefulness of a potentially rich routine data source for evaluation and needs to be accounted for in the evaluation plan

This richness of routinely collected data that are available, in principle, to researchers in Denmark is in contrast to that experienced by researchers in the German and Austrian case studies, where routine data held by SHI funds provide only limited information on socio-demographic data beyond age and sex and availability may be further restricted because of concerns about data confidentiality. For example, the German case study within DISMEVAL sought to access full postcodes of SHI fund members' home addresses. Due to data privacy issues, data were only provided with the first two out of five postcode digits. This, however, proved to be insufficient to even differentiate between urban and rural areas. As a consequence, analyses had to construct control groups based on large administration regions, approximately equivalent to federal states.[71] The lack of this

information impacted on the analyses in two ways: first, including postcodes as a baseline variable to be adjusted for in the outcome analysis could have provided a simple way of controlling for possible differences in socio-economic status between the intervention and control groups.[138] Second, if DMP participation rates vary by region, the availability of full postcodes would have provided a suitable variable to permit an instrumental variable analysis, a technique that allows controlling for observed and unobserved confounders.[139]

The Austrian case study experienced further challenges because of an inability to link different documentation systems that are used to record data on inpatient and outpatient care. Furthermore, analyses were unable to access hospital discharge codes of the patients concerned because of data confidentiality. It was therefore not possible to assess diabetes-related hospital costs for individuals included in the study, so estimates had to rely on hospital costs according to the length of the respective hospital stay as derived from the average costs per hospital day published by the Austrian ministry of health.[73]

Gaining permission to use routine data may be time-consuming, which needs to be accounted for in the analysis plan

The French case study faced different challenges related to data access.[69] In France, access to provider network data is not a problem for statutory evaluation as the provider network is held responsible by the financing body to grant such access and insight into the necessary documents. Generally, however, the use of (even anonymised) patient data for evaluation purposes requires permission from the National Data Safety Authority (CNIL). In the case of provider network data, the individual networks already have this permission, as this is required to collect and store data in the first place, therefore permitting networks to use patient data for evaluation purposes.

Access modalities for data other than those held by networks, for example for the purposes of constructing control groups for evaluation of impact of network interventions on patient outcomes, as undertaken in the French DISMEVAL case study,[69] depend on the body responsible for the relevant data. In France, potential sources for control group data include the National Institute for Sanitary Surveillance (INVS), which collects data on diabetes and asthma, for example; the French national health insurance fund (CNAMTS); and smaller bodies such as the Observatory for General Practice (OMG).

These institutions have distinct data access modalities that generally require a lengthy procedure to meet. The French case study used cross-sectional data from INVS. Here, access requirements changed over time: as part of the data request procedure, CNIL permission was initially requested even though only anonymised data, that is data not allowing identification such as name of the GP or patient address/postcode, was requested. Permission to access the data was granted only after a period of more than 12 months.

Long waiting times and difficult communication with CNIL have been reported by other researchers. This is explained, in part, by the growing workload of this rather young institution: founded in 2003, the number of decisions taken has increased by almost a factor of 100 during 2003 to 2009, while staffing has only increased by 74 per cent during the same period.[140] In order to circumvent lengthy waiting time to access data, evaluators might consider using databases for which CNIL permission has already been granted to those holding the relevant data. If this is not possible, however, waiting times are unpredictable and do represent a major obstacle for evaluation and research in France.

Similarly, the Dutch case study required considerable time commitment in order to access data requested for analysis, frequently requiring multiple meetings with the data holders, care group managers, and formal approval from the groups' boards of directors. Moreover, many care groups requested a contract to be drawn up, which restricted the use of the data and included stipulations concerning patient anonymity and data security.[102] Again, these experiences highlight the need for evaluation planning to provide a sufficient timeframe.

Summary

The experiences of DISMEVAL case studies in accessing data suggests that even where routine data can be accessed in principle, full exploitation of data available may be compromised. While maintaining data confidentiality is a serious concern in any research, there may be a need for decisionmakers to consider putting safeguards in place to allow for full access of routinely collected data sets while maintaining high standards of data privacy and confidentiality, such as is illustrated in the Danish case. An important lesson drawn from all case studies presented here is the necessity, in the evaluation plan, to allow for sufficient time to negotiate access to data and, where necessary, to set sufficient additional resources aside to allow for data extraction by the holder of the data.

4.1.3 Data quality and completeness determine the usefulness of routine data for evaluation

The introductory section to this chapter has already highlighted the notion of quality and completeness of data available for evaluation, in particular when relying on routinely collected data. Powell et al. (2003), in the context of assessing healthcare quality more generally, highlighted challenges around comparability over time and between providers or regions that are particularly pertinent to routine data.[130]

Within DISMEVAL, such challenges were confirmed in several case studies. Both the German and the Austrian case studies used data provided by statutory health insurance funds. Yet, there is a lack of oversight and routine monitoring by SHI funds of data that is collected using DMP documentation. As a consequence, the ability to check on data validity and reliability is limited, which poses challenges to evaluation in terms of its scientific soundness.

For example, as noted in the preceding section, the analyses of statutory health insurance data in the Germany case study found inconsistencies in the way death was (not) documented for some individuals in the database.[71] Until 2004, benefits covered by the German statutory health insurance system included a so-called 'death benefit' (or 'burial allowance') for SHI members and their dependants. With the abolition of this benefit with the 2004 health reform, SHI funds have experienced difficulties identifying whether a member, or their dependant(s), had died, if this is not specifically reported to the fund. Notifications of the death of an insured person by relatives occur with delay or not at all, and in some cases SHI funds will know about the death of a member only when post is returned or when they are informed by the pension fund. In addition, most SHI funds do not document deaths in a standardised way; there is no compulsory entry field for the date of death of an insured member in the software system used.

This implies that mortality data for the years 2007 and 2008 as used in the German case study is most likely to be incomplete. However, data incompleteness is not expected to have impacted on the findings of the case study as it relates to mortality. This is based on the assumption that reporting of deaths in the intervention group is not systematically

different from that in the control group. Thus, while the quality and soundness of the evaluation might not be directly affected by the specific challenges posed by this data set, it highlights the level of scrutiny and thoroughness required to understand the strengths and weaknesses of routine data to be used for evaluative purposes and, in particular, an understanding of the context within which these are collected to ensure accurate interpretation of observed phenomena.

This need is further highlighted by the Dutch case study, which worked with individual patient data that were obtained anonymously and, for most groups, without prior cleaning and validation.[102] To assess the plausibility of the data, range checks were conducted on the supplied variables. Extreme outliers or unusual values were removed based on cut-off points determined by expert Dutch professionals and researchers in the field of diabetes care. Table 4.1 displays these cut-off points as well as the number and percentage of removed clinical values from the overall dataset, incorporating data from a total of 105,056 individuals across 18 care groups. The highest percentage of removed values across clinical outcomes was 1.31 per cent for LDL cholesterol.

Table 4.1 Expert cut-off points and excluded patient outcomes

Indicator	Lower	Upper	Excluded (N)	Removed (%)
Glycated haemoglobin (HbA1c) (mmol/mol)	18	108	913	0.51
Total cholesterol (mmol/l)	–	14	51	0.03
LDL cholesterol (mmol/l)	1	7.3	2110	1.31
HDL cholesterol (mmol/l)	0.5	–	200	0.13
Triglycerides (mmol/l)	–	8	479	0.29
Systolic blood pressure (mmHg)	70	250	25	0.01
Diastolic blood pressure (mmHg)	40	120	121	0.07
BMI (kg/m^2)	16	70	123	0.08

In the Spanish case study, missing data posed the greatest challenge to undertaking rigorous analyses.[125] Thus, lack of data completeness greatly restricted the ability to apply propensity score matching using a wide range of variables, such as blood pressure measurements or cholesterol values. As a consequence of missing data, analyses had to exclude a comparatively large number of individuals, so reducing the sample size from a base of just over 1,700 individuals by between 20 per cent and two-thirds, depending on the technique used. Most of these limitations have to be related to the non-experimental nature of the intervention, which was carried out in the 'real world' clinical settings as part of routine clinical practice, using standard electronic medical records.

Summary

This section has further highlighted the particular challenges pertaining to the quality and completeness of existing (routine) data for the purposes of evaluation. As noted in the above sections, thus, if such data is being used for evaluation it will be important to have a good understanding of their limitations and their impact on evaluation findings. In practice, this will mean engaging in conversations with those collecting the data to assess the context for data collection, as well as allowing, in the analysis plan, for sufficient time to scrutinise data for consistency and completeness.

4.2 **Evaluation of disease management requires adequate financial and human resources**

Rigorous evaluation requires adequate resourcing, both in terms of financial resources and the identification of analytical capacity to address more sophisticated design issues as well as data collection and analysis. The level of necessary resources will be determined mainly by the aims of the evaluation, the complexity of the programme to be evaluated, the purpose of evaluation and intended use of findings, among others.

A fundamental challenge in several settings is the availability of funding to conduct rigorous evaluation studies. For example, the cluster randomised controlled trial drawn upon for the Austrian DISMEVAL case study was funded primarily by the evaluator itself, the Institute of General Practice, Family Medicine and Preventive Medicine of Paracelsus Medical University in Salzburg, Austria.[73] Additional funding was provided by the statutory public health insurance of the province of Salzburg, which provided additional reimbursement for those GPs acting as a control group equivalent to the DMP-reimbursement; it also financed the DMP in the intervention group (although the latter costs would have arisen in the absence of the study). The study was further supported by two unrestricted research grants from Salzburg Savings Bank and Roche Diagnostics Austria. Thus, although there is a general interest for valid assessment in Austria, provision of sufficient funds to implement a thorough programme evaluation represents one of the most challenging problems. Importantly, to ensure objective assessment, it will be important to set aside independent financial resources.

In France, the scale of a provider network evaluation is likely to be smaller than that of a programme evaluation in other countries.[69] The average size of the 742 provider networks receiving specific funding in 2009 was 1,000–2,000 enrolled patients per network; the average cost of a triennial external evaluation performed in 2009 was €20,875.[141] These figures illustrate that resources available for evaluation are limited due to the high number and relatively small size of provider networks. In this context, advanced evaluation designs including randomisation are not an option for the vast majority of provider networks and evaluators. Importantly, as provider networks are set up locally and participation is voluntary both for patients and physicians, randomised designs will be not feasible empirically. Further to these limitations, the sample size places a limit on the feasibility of certain statistical methods that require a large number of variables (eg propensity score matching).

One way of optimising the use of financial and human resources for evaluation purposes is to harmonise and share technical resources between provider networks with similar requirements. For instance, networks using quantitative databases can collaborate in setting up technical infrastructure (hardware and software, data transmission) or jointly employ staff responsible for the respective databases. Such collaborations between provider networks are currently being developed, for instance between diabetes provider networks in the Paris metropolitan region.

Summary
The examples presented in this section illustrate how robust evaluation will require considerable expertise, starting from the ability to conceptualise the evaluation design through to very practical issues around computing and IT capacity to enable manipulation

of data suitable for evaluation. Thus, epidemiological and statistical expertise will be necessary to devise a strategy for the identification intervention and control groups, to select variables for adjustment of differences between groups as well as meaningful outcome variables and to carry out robust analyses of evaluation data of disease management interventions. Most analyses undertaken within DISMEVAL used SAS statistical software, requiring acquisition of the software where this is not routinely available. Availability of powerful statistical software is particularly relevant where analyses consider a large body of data. This will have to be set in a framework of adequate financial resources to permit overall independent assessment of programme effect.

Challenges and lessons learned

This chapter is aimed at reflecting on the broader lessons learned from work undertaken within the DISMEVAL project with regard to evaluation of structured approaches to disease management. We begin by briefly describing existing evaluation practice and needs as assessed in two case studies. We then elaborate specifically on the potential challenges associated with interpreting evaluation findings and how evaluation can be used to inform programme development and improvement.

5.1 There is a need for the better understanding and communication of evaluation practice and needs between stakeholders

Work carried out within the DISMEVAL project aimed to better understand the existing context for evaluation in European countries and the variation in evaluation needs and practices from the perspective of different actors within a given setting, including practitioners, funders and decisionmakers. We here describe findings from two case studies undertaken within DISMEVAL that sought to provide additional insights into the broader context for evaluation. Thus, the French case study examined the existing statutory evaluation requirements of provider networks in France to inform current evaluation practice, while the Dutch case study collected qualitative data on information needs and measures of success by practitioners involved in evaluations. This section describes the findings of this work to illustrate the challenges experienced in practice by those undertaking evaluation in routine operation.

Framework for evaluating provider networks in France: theory and practice

In France, the majority of provider networks are publicly funded.[70] In 2002, this funding was tied to the obligation to conduct a triennial external evaluation.[69] Until 2007, the modalities of this obligatory evaluation were not specified; the only existing guidance consisted of recommendations on evaluation issued by the National Authority for Health (*Haute Autorité de Santé*, HAS). These recommendations are structured around five themes: (1) the network objectives, (2) integration of users and professionals in the network, (3) functioning of the network (4) quality of care and (5) economic evaluation. In 2007, the regulatory evaluation modalities were eventually set out in a ministerial circular, which added to the evaluation domains used by the HAS the requirement of pre-post comparison or comparison with 'other initiatives', although these 'other initiatives' were not defined further.

The analyses undertaken in DISMEVAL sought to better understand how, within this broad regulatory environment, provider networks approach evaluation.[69] In particular it aimed to assess how well a given evaluation linked with the provider network's objectives; the appropriateness of the evaluation approach chosen; the extent to which the chosen approach was based on national recommendations as set out above; and what evaluation method can be recommended considering the characteristics of a given disease management intervention. The analysis was able to draw on data provided by 12 provider networks addressing a range of common conditions such as diabetes and obesity, but also less common conditions such as multiple sclerosis and motor neurone disease. The size of the networks reviewed here varied from 231 to 4,985 patients (2009). Analyses included non-public external evaluations of the most recent triennial period as well as the three most recent internal activity reports (2007–2009). In addition, 13 semi-structured interviews were conducted with provider network coordinators, evaluators, stakeholders and funding entities.[69]

Although the external evaluations of the provider networks sample were found to generally follow the stipulations set out by the HAS and the Ministry of Health, in practice the evaluation requirements were very rarely met in their entirety. For example, one requirement stipulates the verification of 'the presence of organisational guidelines [...] as well as their application and their impact on professional practice'. However, the ability of provider network evaluations to report on this indicator requires such guidelines to be available, which may not be the case for less common conditions. As noted earlier, national guidelines also require the evaluation to be comparative; however, only 4 out of the 12 provider networks reviewed did use a comparative design. Furthermore, existing evaluations rarely carried out sound economic assessment beyond budget descriptions.

Overall, the analysis found that available external evaluation reports: tended to be limited to describing the structure and organisation of the provider networks; listed their activities; reported on patient and provider satisfaction; and, in some cases, performed analysis on quantitative data. They generally provided a narrative conclusion and recommendations. These reported approaches were not suitable to assess whether a provider network's objectives were met, because they lacked clear indicators associated with the specific objectives.

These observations illustrate that, in France, national recommendations appear to allow for considerable room for interpretation and lack sufficiently detailed guidance for those interested in conducting sound evaluation. This apparent mismatch between the political ambition to standardise and the multitude of existing initiatives can be explained, in part, by the observation that provider networks tend to have emerged as small bottom-up initiatives, founded and operated by health professionals, and such networks are not equipped to collect the exhaustive data necessary for comparative evaluation. Additionally, external evaluators may not have the skills and financial, as well as human, resources needed for a task as complex as comparative evaluation. In future, thus, it would be important to develop objective indicators that would permit for assessment whether or not the network's objectives have been met. These indicators should not be defined at the national level but agreed on regionally so that individual characteristics of the respective provider network can be taken into account. There is a need to better match evaluation

methods and provider network characteristics by giving more specific indications by provider network subtypes, based on disease type.

Attitudes to disease management evaluation on the ground: a perspective of practitioners in the Netherlands

The Dutch case study aimed to assess professionals' attitudes towards disease management evaluation.[102] It carried out interviews with a selection of 27 healthcare professionals from 9 diabetes care groups (a manager, a GP, and a practice nurse per group) who were provided with various propositions concerning four broad topics: (1) determining and measuring the goals of diabetes care; (2) function and possibilities of evaluation; (3) type and range of quality indicators; and (4) indicator development.

The vast majority of respondents (N=25) agreed that before indicators are developed to evaluate the quality of care, the goals of a care programme must be clearly specified. Respondents highlighted that it would be important to know in advance which benchmarks to measure against in order to be able to work towards improving quality. While 20 respondents noted that goals set out for their care programme were matched with performance indicators, others highlighted that the focus of evaluation tended to overemphasise clinical outcomes; they were also concerned about the lack of attention to self-management support in both care provision and evaluation.

Providers expressed a concern that evaluation was perhaps too focused on the actual 'end product' of the care programme, which in their opinion would not necessarily reflect the quality of care provided. The focus on clinical outcomes was seen as problematic and processes and structures were evaluated less structurally and/or solely when outcomes are poor. Respondents emphasised that quality systems, including regular audits and benchmarks, must be put in place to ensure good care quality processes and structures.

Eighteen respondents considered more patient-centred outcomes, especially quality of life, to be important indicators of quality. Respondents also noted the difficulties associated with both measuring and interpreting such measures, which they considered as 'subjective'. There appeared to be general consensus (N=20) that parameters focusing on the knowledge and behaviour of patients should form a standard component of measuring the quality of diabetes care, to allow for more insight into patients' self-management capabilities. Overall, however, there appeared to be some uncertainty about the optimum 'mix' of indicators suited for quality measurement. Thus, while 15 respondents considered a limited number of measures, mostly those on outcomes and some processes, to be sufficient to measure quality of care, 12 respondents argued for a wider range of indicators.

There was some concern about the involvement of government, research organisations and, in particular, health insurers in indicator development. A particular concern centred on providers' workload becoming too heavy, which might result in unreliable data registration, and on health insurers' tendency to emphasise efficiency and cost reduction rather than quality. At the same time, the majority of respondents believed that a cooperative approach might be beneficial and could allow for a broader evaluation perspective beyond the traditional medical approach. Eighteen respondents reported having experienced difficulties in evaluating the quality of diabetes care provided by their care groups. These difficulties were often of a practical nature and related to shortcomings in IT infrastructure or to lack of time for data documentation. Providers also noted that

the trend of quality measurement in primary care is new and that they are still in a learning phase, trying to determine how to measure quality, what indicators to use, and how to analyse the large amount of data gathered.

Similar to what was observed for France, there appears to be a need in the Netherlands to provide better support to care groups 'on the ground', to enable data collection and analysis in order to support evaluation of intervention effect.

Summary

This section examined the existing context for evaluation in two European countries and the variation in evaluation needs and practices from the perspective of different actors within a given setting, including practitioners, funders and decisionmakers. It finds that there may be different expectations about the scope and purpose of evaluations between practitioners and those collecting the data vis-à-vis those funding such initiatives. Evidence from France and the Netherlands points to some uncertainties among practitioners, in particular about the range of indicators that may be relevant and feasible to collect, and there may be a risk of a mismatch between the perspectives of funders and decisionmakers about key indicators that ought to be collected and those perceived to be important (from the perspective of practitioners) to enable quality improvement. Overall, there appears to be a need to provide better support to practitioners and those engaged in local evaluation to enable formulation and definition of indicators for use in routine practice as well as for data collection and analysis in order to support evaluation of the intervention effect.

5.2 Interpretation of evaluation findings needs to be placed into the context of evaluation design and of the intervention being evaluated

This section reflects briefly on the interpretation of evaluation findings and the necessity for placing these in the broader context of programme implementation specifically and issues around evaluation more widely. For example, an evaluation might find improvements in process indicators (so suggesting improved quality of care) but not in outcomes. This might be because the length of evaluation was not sufficient to demonstrate health improvements. Likewise, an evaluation might find that a given intervention improved outcomes for only around half of participants; this might indicate that the intervention was suboptimal or, perhaps more likely, that expectations of what improvements could be achieved were unrealistic. Further, a given evaluation might not be able to establish statistical evidence of any health improvement simply because of small numbers. Also, intervention effect will differ by disease type.

As noted earlier, the Austrian case study within DISMEVAL was able to draw on data obtained from a cluster randomised trial of a diabetes disease management programme (DMP) implemented in Salzburg province.[73] The trial aimed to assess whether the DMP 'Therapie aktiv' improved metabolic control (HbA1c) and quality of care for adults with type 2 diabetes managed in primary care compared to a control group with usual diabetes care. It hypothesised that the DMP would lead to a significant reduction of HbA1c levels and an improvement in guideline adherent care. However, the findings of the evaluation demonstrated that the DMP did not significantly improve metabolic control for patients with type 2 diabetes mellitus, although measures of process quality improved significantly.

Thus, a higher proportion of patients received patient education and regular screening examinations of the eyes, feet and HbA1c checks compared to those receiving usual care.

What does this mean in practice? Although an evaluation may detect statistically valid associations, the interpretation of such findings, in terms of their transferability into healthcare decisionmaking, remains somewhat challenging. Thus, effects found to be statistically significant have to be set against their clinical relevance and, as we have seen earlier in the context of the Dutch case study (see Section 3.3.4), the length of the observation period will be crucial for interpretation of some outcomes. In the case of the Austrian DMP evaluated here, the measurement of HbA1c levels alone will not predict the influence of disease management on clinical outcomes, such as morbidity and mortality, based on an observation period of 12 months only. In order to translate the findings into more 'tangible' observations that can then inform decisionmaking, analyses were therefore extended to estimate the number needed to be treated to avoid future events such as myocardial infarction or diabetes-related complications, as illustrated in Section 3.2.2. While recognising the limitations inherent in an approach that extrapolates findings observed in other settings (here the UK Prospective Diabetes Study (1998)[115] and other work[116]), implying that the estimated figures on the number needed to be treated and the associated cost should not be interpreted as absolute figures, the analyses presented provide a useful illustration of how findings may be used further. In this case it was used to demonstrate how uncontrolled evaluation designs can lead to misleading results, typically, although not always, overestimating intervention effect.

The Danish DISMEVAL case study provided further insights into how different approaches to evaluation can help understanding the effect of a given disease management intervention, in this case, a three-month rehabilitation programme for people with chronic diseases.[78] One set of analyses aimed to assess the effect of the rehabilitation programme on functioning and quality of life among persons with chronic obstructive pulmonary disease or type 2 diabetes, using two statistical techniques within a before–after design without control. Keeping in mind the limitations of an uncontrolled design to attribute intervention effect, the analysis did not identify substantial differences in the performance of either technique (paired t-tests and mixed model linear regression for repeated measurements with random effects on person level) in relation to observed effects. The only exception was that the mixed models approach identified more associations that were statistically significant, although this was largely attributable to differences in sample sizes between the two models.[142] This 'non-finding' is important as it highlights the need to develop a standard or criteria to guide evaluators on which approach to choose in what contexts, in particular where statistical expertise is lacking or suboptimal.

A second set of analyses undertaken within the Danish case study sought to understand the impact of different evaluation designs on intervention effect, using different sets of control groups (created through random selection; gender and age matching; propensity score matching), which is described in more detail in Section 3.1.2. In brief, the three designs were: before–after without control; intervention-control in post-period; and difference-in-differences analysis (DID). The difference-in-differences analysis produced estimates of intervention impact that were in principle more plausible than those based on a single difference.[78] At the same time, the simple before–after analyses provided additional insights when interpreted alongside the findings of the DID analyses. Thus, as we have seen earlier,

the magnitude of observed changes in the frequency of hospitalisation was smaller in the before–after design while the DID analysis found a larger effect size that was statistically significant. This points to a rehabilitation effect in the intervention group, that is, the intervention slowed disease progression. Based on these analyses, and in order to gain a broader understanding of programme effect, it appears to be useful to combine different methods in outcome evaluation, such as before–after with DID analyses.

At the same time, while the use of different approaches to assess intervention effect can provide valuable additional insights, one of the main challenges remains the data being used to evaluate impact. For example, the German DISMEVAL case study highlighted that choice of matching technique to construct a control group in a non-experimental design may be less important than the actual quality and completeness of underlying data used to create controls.[71] Thus, analyses assessing the effect of a diabetes disease management programme in Germany found that all matching/weighting methods that were used to adjust for baseline variables resulted in a good balance between intervention and control group and fairly similar effect measures for all primary outcome variables analysed (see also Section 3.1.2). However, it also noted that in order to assess the 'true' intervention effect, any adjustment should consider confounding variables only and that concerted efforts should be made to obtain a data set as detailed and valid as possible, since unobserved confounders can result in intervention effects that may be misleading.

The preceding chapter has already highlighted the challenges associated with using routine data for conducting rigorous evaluation. Their interpretation equally poses challenges. For example, the German case study had to rely on inpatient and outpatient ICD-10 diagnosis codes for disease complications to assess the severity and development of the illness over time, as clinical parameters such as blood pressure or HbA1c were not available in routine data.[71] However, the validity of ICD-10 diagnosis codes may be limited, especially for outpatient settings, and without a medical history it is difficult to assess pre-existing illnesses for a given patient before joining the DMP.[143]

One other phenomenon pertinent to interpreting evaluation findings and highlighted by the German case study is the so-called 'immortality bias'. Immortality bias can be introduced into a retrospective analysis when patients meet the criterion to receive the treatment only if they survive for a certain period of time. Within DISMEVAL, inital analyses included survival analysis with the follow-up period starting in 2005, the year of DMP enrolment of study participants. This analysis found a very strong beneficial effect for the DMP group by means of higher survival rates compared to the control group. However, scrutinising underlying data revealed that this effect was largely attributable to the observation that the intervention group had almost no deaths recorded in 2005: in order to enter the intervention in 2005, the participant had to be alive at that point, while no such requirement applied to those assigned to the control group. Analyses therefore had to be repeated using the year following enrolment, 2006, as the commencement date for follow-up. Although repeat analyses found a reduced overall effect, the survival benefit for the intervention group remained for 2006 but decreased over the following years, as we have show in Section 3.1.1 of this report. If the effect was truly attributable to the intervention, the survivial benefit should have increased. It is therefore more likely, as noted earlier, that GPs selectively did not enrol patients who were likely to die in the near future. If this was the case, a hypothesis that would need to be confirmed, the analysis

could still suffer from some form of immortality bias, since the routine data set that was used for baseline adjustment might have missed important variables for the prediction of this short-term mortality risk. This highlights that caution needs to be applied when interpreting observed effects, requiring a good understanding of the context within which observations are being made[6] and, as noted earlier, of the data that are being used for evaluative purposes.

In this context it is also important to emphasise the challenges decisionmakers, funders and practitioners are faced with when interpreting findings from different evaluations of essentially the same intervention. We have briefly highlighted this issue in Section 2.3 in relation to DMPs in Germany (see Box 2.1, page 8). The German case study within DISMEVAL has highlighted how different evaluations of the diabetes DMP used different approaches to identify the intervention and control populations.[71] For example, as noted above, the DISMEVAL case study used inpatient and outpatient ICD-10 codes or prescription data to identify the populations to be included in the evaluation. Although the analysis was able to cross-validate a diagnosis of type 2 diabetes for DMP participants on the basis of the DMP documentation that was available from the SHI fund, such a verification of diagnosis was not possible for the control group. The control group therefore might have included a certain proportion of patients with type 1 diabetes or without diabetes. On the other hand, while prescription data are known to have a higher validity, diabetic patients identified on this basis will include only more severe cases taking medication, so introducing selection bias. Some of the evaluation studies performed in Germany to date excluded diabetics who did not take any diabetes medication or were younger than age 50 years,[38] while others included all diabetics over the age of 18 years.[39] Understanding these ostensibly small but important differences between study designs will be crucial for the interpretation of observed intervention effects, which are likely to differ as a consequence.

Finally, the Dutch case study provides an example of how a mixed methods approach can be used to advance our understanding of the differential impacts of approaches to chronic disease management on different population groups.[102] Thus, analyses carried out within DISMEVAL demonstrated how structured care for those with diabetes implemented in the Netherlands was most beneficial for patients with poorly controlled diabetes. This finding is supported by a recent meta-analysis of the international literature by Pimouguet et al. (2011),[144] which provides a plausible explanation for the overall small average effects of the intervention. Among those with controlled diabetes whose clinical values leave limited room for further improvement, the programme successfully maintained health, so preventing or delaying more serious complications associated with deteriorating glycaemic control.

Summary

This section has illustrated how the interpretation of evaluation findings will have to take account of the broader context within which evaluations were carried out. This includes consideration of factors such as length of observation period, which may explain absence of intervention effect if duration is short – although experience from the Dutch case study highlights that short duration could also lead to overestimation of effect. It also highlighted how the combination of different designs may be useful to fully understand observed effects, as demonstrated by the Danish case study. Importantly, however, interpretation of

observed intervention effect will be determined, to a considerable degree, by the quality of the data underlying the analysis. As demonstrated by the German case study, data quality may be more important with regard to the observed effect size than the choice of design to create a control group in non-experimental settings. This section has also highlighted how imaginative use of novel techniques and approaches may help better understanding of observed effects, such as the use of meta-regression approaches to identify differential impacts of approaches to chronic disease management.

5.3 Evaluation findings can help inform future programme development and improvement

In this section we discuss how evaluation findings can be used to identify potential shortcomings of existing interventions and guide further improvement. For example, lack of evidence of health improvement (however conceptualised) of a given intervention may simply reflect that programme components were ill-suited to lead to health improvement in the first place. Thus, a culture of passivity or lack of proaction to self-manage among patients may pose a considerable barrier towards implementing successful patient self-management interventions, although this also implies that 'off-the-shelf-interventions' might not be the appropriate means to activate this type of population.[145] Also, evaluation may help to identify where a given intervention is likely to lead to inequities in healthcare delivery, for example, through demonstrating that the enrolment procedure leads to only those with higher health literacy actually benefitting from the programme (selection bias).[7] Furthermore, evaluation findings might also highlight issues around programme implementation; for example, where a given intervention is characterised by high attrition rates this might indicate problems with programme set-up and/or recruitment process, while also suggesting that the intervention might be ill-suited to the needs of the target population.

There are several examples of how work undertaken in the DISMEVAL project can inform, and in some instances already has informed, the further advancement of existing interventions. Thus, the Austrian case study, evaluating a diabetes disease management programme that was implemented as a cluster-randomised controlled trial in Salzburg, found only small effects of the intervention as regards the primary outcome (metabolic control).[73] It therefore concluded that the current intervention approach may insufficiently take account of patient self-management support whereas the combination of traditional patient education with ongoing peer support may present a more promising approach in diabetes care in the Austrian context. Such an approach is currently being implemented as a cluster-randomised controlled trial: 'Self-Efficacy and Peer Support Enhance the Effectiveness of Disease Management in Diabetes Type 2 (SPEED)'.[146] It aims to evaluate peer support regarding management of diabetes, diet and physical activity as an additional component of a standard DMP previously implemented at a nationwide level.

The experience of the Austrian case study is echoed by the findings of the evaluation of diabetes disease management in the Netherlands, which observed a differential impact of current approaches to structured care on different patient populations.[102] This emphasises the need for decisionmakers, funders and practitioners to consider moving away from the standardised approach towards a more tailored approach to diabetes management, which

emphasises the patients' direct involvement in the processes of care, including self-management support.[102]

Although the Dutch care standard for type 2 diabetes, which currently guides the approach to diabetes care provided by care groups[112], stipulates that 'patients should play a central role in their care', the case study findings suggest that to date, the practice of diabetes care remains highly paternalistic. In its current form, the Dutch bundled payment system for diabetes care motivates care providers to deliver highly standardised care based on performance indicators as stipulated in the national diabetes care standard.[147] These indicators, which are monitored by health insurers, prescribe a defined intensity of service delivery, regardless of patients' health, demographic or social status. Although frequent monitoring was shown to be especially useful for improving clinical values in poorly controlled diabetic patients, patients in relatively good health might be managed equally effective in a less physician-guided way that emphasises self-management. As we have discussed earlier in this report, there is reasonably good evidence that self-management support can improve patients' health behaviours, clinical and social outcomes.[148-149] There is thus potential for a more tailored approach to disease management that proactively considers patient characteristics in determining care processes, including self-management support, benefiting a relatively healthy population of diabetic patients for whom intensive monitoring may be inappropriate.

Similar issues are illustrated by findings of the French DISMEVAL case study, which demonstrated the degree to which selection takes place when patients enrol into provider networks; these were of younger age and had a more recently diagnosed diabetes but worse glycaemic control (as measured by HBA1c levels) as compared with those in the reference population of diabetic patients.[69] As demonstrated by analyses undertaken, uncontrolled before–after evaluation of networks will only measure the effect of the programme on this particular group of patients (see Section 3.1.1). However, decisionmakers also wish to know the effect of the programme on the wider eligible population (eg all diabetic patients) in order to assess whether it should be extended (eg to other regions, or by increasing the size of a single programme).

The observation of a selection effect is not surprising given that more than 90 per cent of patients who do enrol in a network are encouraged to do so by their physician. However, based on data that were available for analysis, it was not possible to form a judgement as to whether the observed selection is appropriate or not. Yet other studies suggest that clinical decisionmaking is influenced not only by medical, but also by non-medical factors such as socio-economic status, ethnicity, language and gender, as well as patient motivation.[150] The implication for the development and improvement of provider networks is to address the issue of patient selection in two steps. A first step is to raise awareness of this phenomenon and of the characteristics on which the selection appears to be based (younger age, worse glycaemic control, etc). A second step is to develop future research using qualitative methods to better understand the decisionmaking process leading to selective recruitment into the provider network.

Summary
In this section we have briefly examined how the findings of work carried out in DISMEVAL uncovered potential limitations of existing approaches to chronic disease

management and how this can inform further development of interventions and programmes to optimise care for people with chronic conditions. In several cases, it was demonstrated how existing interventions might fail to address those who are likely to benefit most and so potentially waste resources. Examples also appear to point to an underutilisation of advanced patient self-management support approaches and the need to develop a more tailored approach. In the case of Austria, such advancement is already being tested in the framework of a cluster-randomised trial.

CHAPTER 6 **Going forward: Conclusions and further research**

This report set out to bring together the findings of work undertaken in DISMEVAL to inform evidence-based recommendations for the use of various approaches to the evaluation of disease management in Europe and to identify examples of best practices and lessons learned. It aimed to explain choices, options and trade-offs to policymakers, programme operators and researchers.

In this concluding chapter, we draw together the evidence compiled from previous chapters and identify those factors that have emerged from work within DISMEVAL that might support the implementation of structured approaches to chronic disease management as discussed earlier.

We acknowledge that randomised controlled trials are widely considered as the gold standard for appraising a causal relationship between a complex intervention and clinical outcomes.[6, 117, 151] However, randomised controlled studies are underused and often lack methodological rigour in evaluation of community health interventions.[152] Given the strength of the design, randomisation should always be considered as a preferred approach as it is the most robust way of determining the effectiveness of a given intervention, ensuring that any observed difference in outcome is not affected by systematic differences in factors, known and unknown, between those who receive a given intervention and those who do not.[7] The Austrian case study in DISMEVAL has illustrated that it is feasible to employ a randomised design in routine settings where the context allows for such a design to be applied.[73]

However, using a randomised controlled design will not be feasible in settings where the intervention is implemented at a population level, such as in Germany[71] or the Netherlands.[102] Thus, in the Dutch case study use of experimental comparisons was not possible due to the nationwide roll-out of structured care approaches for diabetes and the unsuitability of using historic controls.[41] Also, although randomised studies are generally considered to form the most rigorous means to assess intervention effect, the scientific rigour of required designs limits the generalisability of findings to larger and inherently more heterogeneous populations of, for example, chronically ill patients. Selection bias poses a threat to randomised designs just as it does for non-randomised designs, as we have highlighted.

Observational study designs are more suitable for 'real-world' disease management evaluations, keeping their methodological limitations in mind.[8, 45, 67] Given that disease

management is essentially a population-based care strategy, advancing observational study designs is crucial to arrive at strong conclusions regarding how best to treat subgroups of chronically ill patients in the daily practice of health care. The DISMEVAL project has identified and tested a wide range of methods that can be employed in situations where randomisation is not possible, emphasising that rigorous evaluation can still take place even where baseline or predefined control groups are not available.

We have shown how routine databases can provide a useful resource in the design of rigorous studies while noting their disadvantages and limitations. Such data can be used retrospectively to create a control group and provide baseline data. They contain the large numbers and opportunities for long-term follow-up required to investigate clinical endpoints. Results of such evaluations are sufficiently valid if certain data problems are taken into account and results are interpreted with caution in the light of these limitations.

Different (combinations of) care components and processes might be effective for managing chronic disease in patients with varying age, disease duration, health status, co-morbidity, education level, socio-economic status, and so on. Contrary to most disease management evaluation methods, which focus on assessing a single treatment effect, meta-analysis and meta-regression allow for investigations of which patient groups will benefit most from which treatment. Therefore future evaluation work drawing on such approaches can provide insight into what works for whom in the area of disease management, a question that randomised trials have thus far been unable to answer.[153] In addition, meta-regression analyses can be adjusted for baseline (prognostic) factors, which can increase the power of the analysis to detect a true treatment effect and allows adjustment for confounding factors, which is a particular advantage for analyses of observational data.

Given that the implementation of disease management is essentially a process of social change, it is important to combine quantitative data on effects with qualitative information concerning contexts. Use of mixed methods can ensure that disease management evaluation provides insight into how specific local conditions influence the outcomes of a given programme.

Work undertaken within DISMEVAL on evaluation metrics and methods was limited to disease-specific programmes, mirroring much of the existing research evidence that has focused on the management of a few specific diseases, such as diabetes. There has been less focus on individuals with coexisting conditions or multiple health problems,[154-155] even though it is this rapidly increasing population, with multiple disease processes and with diverse and sometimes contradictory needs, who pose the greatest challenge to health systems.[156] Furthermore, as we have shown, the impact of chronic disease management interventions will depend, to a considerable extent, on the specific features of the healthcare setting within which they are introduced, and this observation seems to hold both within and between care systems. However, this work has shown how it can be possible to learn from the experiences of others.

REFERENCES

Reference list

1. Yach D, Hawkes C, Gould C, Hofman K. The global burden of chronic diseases. *JAMA* 2004;291:2616–22.

2. TNS Opinion & Social. Health in the European Union. *Special Eurobarometer 272e*. Brussels: European Commission, 2007.

3. Suhrcke M, McKee M, Sauto Arce R, Tsolova S, Mortensen J. *The contribution of health to the economy in the European Union*. Brussels: European Commission, 2005.

4. Suhrcke M, Fahey D, McKee M. Economic aspects of chronic disease and chronic disease management. In: Nolte E, McKee M, eds., *Caring for people with chronic conditions: A health system perspective*. Maidenhead: Open University Press, 2008:43–63.

5. Mattke S, Seid M, Ma S. Evidence for the effect of disease management: Is $1 billion a year a good investment? *Am J Manag Care* 2007;13:670–76.

6. Campbell N, Murray E, Darbyshire J, Emery J, Farmer A, Griffiths F, et al. Designing and evaluating complex interventions to improve health care. *BMJ* 2007;334:455–59.

7. Craig P, Dieppe P, Macintyre S, Michie S, Nazareth I, Petticrew M. *Developing and evaluating complex interventions: new guidance.* London: Medical Research Council, 2008.

8. Conklin A, Nolte E. *Disease management evaluation. A comprehensive review of current state of the art*. Santa Monica, CA: RAND Corporation, 2010.

9. Nolte E, Hinrichs S, eds. *DISMEVAL. Developing and validating disease management evaluation methods for European healthcare systems. Final report*. Santa Monica, CA: RAND Corporation and DISMEVAL Consortium, 2012.

10. World Health Organization. *Innovative care for chronic conditions: building blocks for action*. Geneva: World Health Organization, 2003.

11. Grumbach K. Chronic illness, comorbidities, and the need for medical generalism. *Ann Fam Med* 2003;1:4–7.

12. McGlynn E, Asch S, Adams J, Keesey J, Hicks J, DeCristofaro A, et al. The quality of health care delivered to adults in the United States. *N Engl J Med* 2003;348:2635–45.

13. Nolte E, McKee M, eds. *Caring for people with chronic conditions: a health system perspective*. Maidenhead: Open University Press, 2008.

14. Coleman K, Mattke S, Perrault P, Wagner E. Untangling practice redesign from disease management: how do we best care for the chronically ill? *Ann Rev Public Health* 2009;30:385–408.

15. Nolte E, Knai C, McKee M, eds. *Managing chronic conditions: Experience in eight countries.* Copenhagen: World Health Organization on behalf of the European Observatory on Health Systems and Policies, 2008.

16. Ouwens M, Wollersheim H, Hermens R, Hulscher M, Grol R. Integrated care programmes for chronically ill patients: a review of systematic reviews. *Int J Qual Health Care* 2005;17:141–46.

17. McKee M, Nolte E. Responding to the challenge of chronic diseases: ideas from Europe. *Clin Med* 2004;4:336–42.

18. Nolte E, McKee M. Integration and chronic care: A review. In: Nolte E, McKee M, eds. *Caring for people with chronic conditions. A health system perspective.* Maidenhead: Open University Press, 2008:64–91.

19. Krumholz H, Currie P, Riegel B, Phillips C, Peterson E, Smith R, et al. A taxonomy for disease management: a scientific statement from the American Heart Association Disease Management Taxonomy writing group. *Circulation* 2006;114:1432–45.

20. Rothman A, Wagner E. Chronic illness management: what is the role of primary care? *Ann Intern Med* 2003;138:256–61.

21. Epstein R, Sherwood L. From outcomes research to disease management: a guide for the perplexed. *Ann Intern Med* 1996;124:832–37.

22. Ellrodt G, Cook D, Lee J, Cho M, Hunt D, Weingarten S. Evidence-based disease management. *JAMA* 1997;278:1687–92.

23. Care Continuum Alliance. *Care Continuum Alliance (CCA) definition of disease management.* As of 30 December 2011: http://www.carecontinuum.org/dm_definition.asp

24. Geyman J. Disease management: panacea, another false hope, or something in between? *Ann Fam Med* 2007;5:257–60.

25. Nolte E, Knai C, eds. *Chronic disease management in Europe. An overview of 13 countries.* In preparation.

26. Ling T. A framework for understanding the contribution of public services to public benefit. In: Ling T, Villalba van Dijk L, eds. *Performance audit handbook: Routes to effective evaluation.* Cambridge: RAND Europe, 2009:1–10.

27. Ofman J, Badamgarav E, Henning J, Knight K, Gano A, Levan R, et al. Does disease management improve clinical and economic outcomes in patients with chronic diseases? A systematic review. *Am J Med* 2004;117:182–92.

28. Powell Davies G, Williams A, Larsen K, Perkins D, Roland M, Harris M. Coordinating primary health care: an analysis of the outcomes of a systematic review. *Med J Aust* 2008;188(8 Suppl):S65–68.

29. Zwar N, Harris M, Griffiths R, Roland M, Dennis S, Powell Davies G, et al. *A systematic review of chronic disease management.* Sydney: Australian Primary Health Care Institute, 2006.

30. Norman G. All things considered, the answer is a resounding yes. *Am J Manag Care* 2008;14:e1–e4.

31. Mattke S. Is there a disease management backlash? *Am J Manag Care* 2008;14:349–50.

32. Mattke S, Bergamo S, Balakrishnan A, Martino S, Vakkur N. Measuring and reporting the performance of disease management programs. Santa Monica, CA: RAND Corporation, 2006.

33. Siering U. Germany. In: Nolte E, Knai C, McKee M, editors. *Managing chronic conditions – experience in eight countries.* Copenhagen: World Health Organizaton on behalf of the European Observatory on Health Systems and Policies, 2008:75–96.

34. Nolte E, Knai C, Hofmarcher M, Conklin A, Erler A, Elissen A, et al. Overcoming fragmentation in healthcare: chronic care in Austria, Germany and the Netherlands. *Health Econ Pol Law* 2012;7:125-146.

35. Bundesversicherungsamy. *Zulassung der Disease Management Programme (DMP) durch das Bundesversicherungsamt (BVA).* As of 30 December 2011: http://www.bundesversicherungsamt.de/cln_108/nn_1046154/DE/DMP/dmp__inhalt.html

36. Schäfer I, Küver C, Gedrose B, Hoffmann F, Russ-Thiel B, Brose H, et al. The disease management program for type 2 diabetes in Germany enhances process quality of diabetes care – a follow-up survey of patient's experiences. *BMC Health Serv Res* 2010;10:55.

37. Ose D, Wensing M, Szecsenyi J, Joos S, Hermann K, Miksch A. Impact of primary care-based disease management on the health-related quality of life in patients with type 2 diabetes and comorbidity. *Diab Care* 2009;32:1594.

38. Miksch A, Laux G, Ose D, Joos S, Campbell S, Riens Bea. Is there a survival benefit within a German primary care-based disease management program? *Am J Managed Care* 2010;16:49–54.

39. Stock S, Drabik A, Büscher G, Graf C, Ullrich W, Gerber A, et al. German diabetes management programs improve quality of care and curb costs. *Health Affairs* 2010;29:2197–205.

40. Schäfer I, Küver C, Gedrose B, von Leitner E, Treszl A, Wegscheider K, et al. Selection effects may account for better outcomes of the German Disease Management Program for type 2 diabetes. *BMC Health Serv Res* 2010;10:351.

41. Berthold H, Bestehorn K, Jannowitz C, Krone W, Gouni-Berthold I. Disease management programs in type 2 diabetes: quality of care. *Am J Manag Care* 2011;17:393–403.

42. Linder R, Ahrens S, Köppel D, Heilmann T, Verheyen F. The benefit and efficiency of the disease management program for type 2 diabetes. *Dtsch Arztebl Int* 2011;108:155–62.

43. Nolte E, Bain C, McKee M. Diabetes as a tracer condition in international benchmarking of health systems. *Diabetes Care* 2006;29(5):1007–11.

44. MacStravic S. Therapeutic specificity in disease management evaluation. *Dis Manage* 2008;11:7–11.

45. Linden A, Adams J, Roberts N. *Evaluation methods in disease management: determining program effectiveness*. As of 30 December 2011:

   http://www.lindenconsulting.org/Papers/White%20Paper%20-%20Final.pdf

46. Linden A, Adams J, Roberts N. An assessment of the total population approach for evaluating disease management program effectiveness. *Dis Manage* 2003;6:93–102.

47. Serxner S, Baker K, Gold D. Guidelines for analysis of economic return from health management programs. *Am J Health Prom* 2006;20(Suppl):1–17.

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

49. Linden A, Adams J, Roberts N. Evaluating disease management program effectiveness: an introduction to time-series analysis. *Dis Manage* 2003;6:243–55.

50. Cretin S, Shortell S, Keeler E. An evaluation of collaborative interventions to improve chronic illness care: Framework and study design. *Eval Rev* 2004;28:28–51.

51. Stevenson K, Baker R, Farooqi A, Sorrie R, Khunti K. Features of primary health care teams associated with successful quality improvement of diabetes care: a qualitative study. *Fam Pract* 2001;18:21–26.

52. Weaver T, Tyrer P, Ritchie J, Renton A. Assessing the value of assertive outreach: Qualitative study of process and outcome generation in the UK700 trial. *Br J Psych* 2003;183:437–45.

53. Esposito D, Fries E, Gold M. Using qualitative and quantitative methods to evaluate small-scale disease management pilot programs. *Popul Health Manage* 2009;12:3–15.

54. Boaden R, Dusheiko M, Gravelle H, Parker S, Pickard S, Roland M, et al. *Evercare evaluation: final report*. Manchester: National Primary Care Research and Development Centre, 2006.

55. Gravelle H, Dusheiko M, Sheaff R, Sargent P, Boaden R, Pickard S, et al. Impact of case management (Evercare) on frail elderly patients: controlled before and after analysis of quantitative outcome data. *BMJ* 2007;334:31–4.

56. Sargent P, Pickard S, Sheaff R, Boaden R. Patient and carer perceptions of case management for long-term conditions. *Health Soc Care Commun* 2007;15:511–19.

57. Pawson R, Tilley N. *Realistic Evaluation*. London: Sage Ltd, 1997.

58. Norman G. Disease management outcomes: Are we asking the right questions yet? *Popul Health Manage* 2008;11:183–87.

59. Berwick D. The science of improvement. *JAMA* 2008;299:1182–84.

60. Sibbald B, Roland M. Understanding controlled trials. Why are randomised controlled trials important? *BMJ* 1998;316:201.

61. Linden A, Roberts N. A user's guide to the disease management literature: recommendations for reporting and assessing program outcomes. *Am J Manag Care* 2006;11:113–20.

62. Nolte E, McKee M, Knai C. Managing chronic conditions: An introduction to the experience in eight countries. In: Nolte E, Knai C, McKee M, eds. *Managing chronic conditions: Experience in eight countries*. Copenhagen: World Health Organization on behalf of the European Observatory on Health Systems and Policies, 2008:1–14.

63. Motheral B. 2008: A tipping point for disease management? *J Manage Care Pharm* 2008;14:643–49.

64. Linden A, Adams J, Roberts N. Evaluating disease management programme effectiveness: an introduction to the regression discontinuity design. *J Eval Clin Pract* 2006;12:124–31.

65. Linden A, Trochim W, Adams J. Evaluating program effectiveness using the regression point displacement design. *Eval Health Prof* 2006;29:407–23.

66. Linden A, Adams J. Using propensity score-based weighting in the evaluation of health management programme effectiveness. *J Eval Clin Pract* 2010;16:175–79.

67. Mattke S, Bergamo S, Balakrishnan A, Martino S, Vakkur N. *Measuring and reporting the performance of disease management programs*. Santa Monica, CA: RAND Corporation, 2006.

68. Beeuwkes Buntin M, Jain A, Mattke S, Lurie N. Who gets disease management? *J Gen Int Med* 2009;24:649–55.

69. Chevreul K, Brunn M, Cadier B, Durand-Zaleski I, Farsi F, Labrosse H. Evaluation of disease management in France. In: Nolte E, Hinrichs S, eds. *DISMEVAL. Developing and validating disease management evaluation methods for European healthcare systems. Final report*. Santa Monica, CA: RAND Corporation and DISMEVAL Consortium, 2012.

70. Durand-Zaleski I, Chevreul K, Farsi F, Brunn M, Knai C, Nolte E. France. In: Nolte E, Knai C, eds. *Chronic disease management in Europe. An overview of 13 countries*. In preparation.

71. Fullerton B, Erler A, Pöhlmann B, Krohn R. Evaluation of disease management in Germany. In: Nolte E, Hinrichs S, eds. *DISMEVAL. Developing and validating disease management evaluation methods for European healthcare systems. Final report*. Santa Monica, CA: RAND Corporation and DISMEVAL Consortium, 2012.

72. Linden A, Adams J, Roberts N. Evaluating disease management programme effectiveness adjusting for enrolment (tenure) and seasonality. *Res Healthcare Financial Manage* 2004;9:57–68.

73. Flamm M, Panisch S, Sönnichsen A. Evaluation of disease management in Austria. In: Nolte E, Hinrichs S, eds. *DISMEVAL. Developing and validating disease management*

evaluation methods for European healthcare systems. Final report. Santa Monica, CA: RAND Corporation and DISMEVAL Consortium, 2012.

74. Sönnichsen A, Rinnerberger A, Url M, Winkler H, Kowatsch P, Klima G, et al. Effectiveness of the Austrian disease-management-programme for type 2 diabetes: study protocol of a cluster-randomized controlled trial. *Trials* 2008;9:38.

75. Sönnichsen A, Winkler H, Flamm M, Panisch S, Kowatsch P, Klima G, et al. The effectiveness of the Austrian disease management programme for type 2 diabetes: a cluster-randomised controlled rrial. *BMC Fam Pract* 2010;11:86.

76. Dunkelberg S, Zingel D, Noack A, van den Bussche H, Kaduszkiewicz H. [Which patients are (not) included in the DMP diabetes programme?]. *Gesundheitswesen* 2006;68:289–93.

77. Linden A, Adams J, Roberts N. Strengthening the case for disease management effectiveness: un-hiding the hidden bias. *J Eval Clin Pract* 2006;12:140–47.

78. Frohlich A, Jacobsen R. Assessment of disease management in Denmark. In: Nolte E, Hinrichs S, eds. *DISMEVAL. Developing and validating disease management evaluation methods for European healthcare systems. Final report.* Santa Monica, CA: RAND Corporation and DISMEVAL Consortium, 2012.

79. de Boer A, Wijker W, de Haes H. Predictors of health care utilization in the chronically ill: a review of the literature. *Health Policy* 1997;42:101–15.

80. Global Initiative for Chronic Obstructive Lung Disease (GOLD). *Pocket guide to COPD diagnosis, management, and prevention.* As of 30 December 2011: http://www.goldcopd.org/guidelines-pocket-guide-to-copd-diagnosis.html

81. Donabedian A. *The definition of quality and approaches to its management.* Ann Arbor, MI: Health Administration Press, 1980.

82. American Association of Diabetes Educators. Standards for outcomes measurement of diabetes self-management education. *Diabetes Educ* 2003;29:808–16.

83. Lemmens K, Nieboer A, van Schayck C, Asin J, Huijsman R. A model to evaluate quality and effectiveness of disease management. *Qual Saf Health Care* 2008;17:447–53.

84. Pearson M, Mattke S, Shaw R, Ridgely M, Wiseman S. *Patient self-management support programs: an evaluation.* Rockville, MD: Agency for Healthcare Research and Quality, 2007.

85. Steuten L, Vrijhoef B, Severens H, van Merode F, Spreeuwenberg C. Are we measuring what matters in health technology assessment of disease management? Systematic literature review. *Int J Technol Assess Health Care* 2006;22:47–57.

86. Bailie R, Robinson G, Kondalsamy-Chennakesavan S, Halpin S, Wang Z. Investigating the sustainability of outcomes in a chronic disease treatment programme. *Soc Sci Med* 2006;63:1661–70.

87. Spertus J. Evolving applications for patient-centered health status measures. *Circulation* 2008;118:2103–10.

88. Enright P. The six-minute walk test. *Respir Care* 2003;48:783–85.

89. Enright P, Sherrill D. Reference equations for the six-minute walk in healthy adults. *Am J Respir Crit Care Med* 1998;158:1384–87.

90. Rikli R, Jones C. *Senior Fitness Test. Fysisk formåen hos ældre – manual og referenceværdier*. Copenhagen: FADL's Forlag, 2004.

91. Eaton T, Young P, Nicol K, Kolbe J. The endurance shuttle walking test: a responsive measure in pulmonary rehabilitation for COPD patients. *Chron Respir Dis* 2006;3:3–9.

92. Revill S, Morgan M, Singh S, Williams J, Hardman A. The endurance shuttle walk: a new field test for the assessment of endurance capacity in chronic obstructive pulmonary disease. *Thorax* 1999;54:213–22.

93. Borg G. Psychophysical bases of perceived exertion. *Med Sci Sports Exerc* 1982;14:377–81.

94. Borg G. *Borg's perceived exertion and pain scales*. Champaign, IL: Human Kinetics, 1998.

95. Bestall J, Paul E, Garrod R, Garnham R, Jones P, Wedzicha J. Usefulness of the Medical Research Council (MRC) dyspnoea scale as a measure of disability in patients with chronic obstructive pulmonary disease. *Thorax* 1999;54:581–86.

96. Fletcher C, Elmes P, Fairbairn A, Wood C. The significance of respiratory symptoms and the diagnosis of chronic bronchitis in a working population. *BMJ* 1952;2:257–66.

97. Avlund K, Schultz-Larsen K, Kreiner S. The measurement of instrumental ADL: content validity and construct validity. *Aging* 1993;5:371–83.

98. Schultz-Larsen K, Avlund K, Kreiner S. Functional ability of community dwelling elderly. Criterion-related validity of a new measure of functional ability. *J Clin Epidemiol* 1992;45:1315–26.

99. Van der Molen T, Willemse B, Schokker S, ten Hacken N, Postma D, Juniper E. Development, validity and responsiveness of the Clinical COPD Questionnaire. *Health Qual Life Outcomes* 2003;1:13.

100. Brazier J, Harper R, Jones N. Validating the SF-36 health survey questionnaire: new outcome measure for primary care. *BMJ* 1992;205:160–64.

101. Bjørner J, Damsgaard M, Watt T, et al. *Dansk manual til SF-36. Et spørgeskema om helbredsstatus*. Copenhagen: Lif Lægemiddelindustriforeningen, 1997.

102. Elissen A, Duimel-Peeters I, Spreeuwenberg C, Vrijhoef H. Evaluation of disease management in The Netherlands. In: Nolte E, Hinrichs S, eds. *DISMEVAL. Developing and validating disease management evaluation methods for European healthcare systems. Final report*. Santa Monica, CA: RAND Corporation and DISMEVAL Consortium, 2012.

103. Linden A, Adler-Milstein J. Medicare disease management in policy context. *Health Care Financ Rev* 2008;29:1–11.

104. The EuroQol Group. EuroQol – a new facility for the measurement of health-related quality of life. *Health Policy* 1990;16:199–208.

105. Bodenheimer T. Disease management – promises and pitfalls. *N Engl J Med* 1999;340:1202–05.

106. Drummond M, Sculpher M, Torrance G, O'Brian B, Stoddart G. *Methods for the economic evaluation of health care programmes,* 3rd ed. Oxford: Oxford University Press, 2005.

107. Suhrcke M, Fahay D, McKee M. The economic case for investing in the prevention and management of chronic disease. In: Nolte E, McKee M, eds. *Caring for people with chronic conditions. A health system perspective.* Maidenhead, UK: Open University Press, 2008:43–63.

108. MacStravic S. *Rethinking the question "does disease management work?"* As of 30 December 2011:

http://www.healthways.co.uk/newsroom/articles/rethinking_question_McStravick.pdf

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

110. Fetterolf D, Wennberg D, Devries A. Estimating return on investment in disease management programs using a pre-post analysis. *Dis Manage* 2004;7:5–23.

111. Steuten L, Lemmens K, Nieboer A, Vrijhoef H. Identifying potentially cost effective chronic care programs for people with COPD. *Int J COPD* 2009;4:87–100.

112. Elissen A, Duimel-Peeters I, Spreeuwenberg C, Vrijhoef H, Nolte E. The Netherlands. In: Nolte E, Knai C, eds. *Chronic disease management in Europe. An overview of 13 countries.* In preparation.

113. Vrijhoef HJM. Chronische zorg in Nederland anno 2010: leidt verandering tot verbetering? *Forum* 2010;2:59–62.

114. Flamm M, Panisch S, Winkler H, Sonnichsen A. Impact of a randomized control group on perceived effectiveness of a Disease Management Programme for diabetes type 2. *Eur J Public Health* 2011; in press.

115. UK Prospective Diabetes Study. Group-intensive blood-glucose control with sulphonylureas or insulin compared with conventional treatment and risk of complications in patients with type 2 diabetes (UKPDS 33). *Lancet* 1998;352:837–53.

116. Selvin E, Marinopoulos S, Berkenblit G, Rami T, Brancati F, Powe N, et al. Meta-analysis: glycosylated hemoglobin and cardiovascular disease in diabetes mellitus. *Ann Intern Med* 2004;141:421–31.

117. Campbell M, Fitzpatrick R, Haines A, Kinmonth A, Sandercock P, Spiegelhalter D, et al. Framework for design and evaluation of complex interventions to improve health. *BMJ* 2000;321:694–96.

118. Edwards S, Braunholtz D, Lilford R, Stevens A. Ethical issues in the design and conduct of cluster randomised controlled trials. *BMJ* 1999;318:1407–09.

119. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Control Clin Trials* 1986;7:177–88.

120. Stewart L, Tierney J, Clarke M. Reviews of individual patient data. In: Higgins J, Green S, eds. *Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0* (updated March 2011). As of 30 December 2011: www.cochrane-handbook.org

121. Thompson S, Higgins J. Can meta-analysis help target interventions at individuals most likely to benefit? *Lancet* 2005;365:341–46.

122. Higgins J, Thompson S, Deeks J, Altman D. Measuring inconsistency in meta-analyses. *BMJ* 2003;327:557–60.

123. Riley R, Lambert P, Abo-Zaid G. Meta-analysis of individual participant data: rationale, conduct, and reporting. *BMJ* 2010;340:c221.

124. Higgins J, Green S, eds. *Cochrane Handbook for Systematic Reviews of Interventions* Version 5.1.0 (updated March 2011). As of 30 December 2011: www.cochrane-handbook.org

125. Saz Parkinson Z, Sarria Santamera A. Evaluation of disease management in Spain. In: Nolte E, Hinrichs S, eds. *DISMEVAL. Developing and validating disease management evaluation methods for European healthcare systems. Final report.* Santa Monica, CA: RAND Corporation and DISMEVAL Consortium, 2012.

126. European Society of Cardiology. *SCORE Risk charts.* As of 30 December 2011: http://www.escardio.org/communities/EACPR/toolbox/health-professionals/Pages/SCORE-Risk-Charts.aspx

127. Wagner E. Chronic disease management: What will it take to improve care for chronic illness. *Effective Clinical Practice* 1998;1:2–4.

128. Wagner E, Austin B, Davis C, Hindmarsh M, Schaefer J, Bonomi A. Improving chronic illness care: translating evidence into action. *Health Affairs* 2001;20:64–78.

129. Bonomi A, Wagner E, Glasgow R, VonKorff M. Assessment of Chronic Illness Care (ACIC): a practical tool to measure quality improvement. *Health Serv Res* 2002;37:791–820.

130. Powell A, Davies H, Thomson R. Using routine comparative data to assess the quality of health care: understanding and avoiding common pitfalls. *Qual Saf Health Care* 2003;12:122–28.

131. Arbeitsgruppe Erhebung und Nutzung von Sekundärdaten (AGENS) der Deutschen Gesellschaft für Sozialmedizin und Prävention (DGSMP) und Arbeitsgruppe Epidemiologische Methoden, der Deutschen Gesellschaft für Epid e miologie (DGEpi), der Deutschen Gesellschaft für Medizinische Informatik Biometrie und Epidemiologie (GMDS), (DGSMP) dDGfSuPv. Good practice of secondary data analysis. First revision. *Gesundheitswesen* 2008;70:54–60.

132. Hoffmann F, Andersohn F, Giersiepen K, Scharnetzky E, Garbe E. Validation of secondary data. Strength and limitations. *Bundesgesundheitsblatt* 2008;51:1118–26.

133. Struijs J, van Til J, Baan C. Experimenting with a bundled payment system for diabetes care in the Netherlands. Bilthoven, The Netherlands: National Institute for Public Health and the Environment, 2010.

134. Jacobsen R, Frohlich A, Knai C. Denmark. In: Nolte E, Knai C, eds. *Chronic disease management in Europe. An overview of 13 countries*. In preparation.

135. Créange A, Debouverie M, Jaillon-Rivière V, Taithe F, Liban D, Moutereau A, et al. Home administration of intravenous methylprednisolone for multiple sclerosis relapses: the experience of French multiple sclerosis networks. *Mult Scler* 2009;15:1085–91.

136. Cacace M, Ettelt S, Pedersen J, Nolte E. *How health systems make available information on service providers – Experience in seven countries*. Santa Monica, CA: RAND Corporation, 2011.

137. Danmark Statistik. *Introduktion til Forskningsservice*. As of 30 December 2011: http://www.dst.dk/da/TilSalg/Forskningsservice/~/media/C28A3220AAB649E08A8D6B8F7D684744.pdf

138. Krieger N. Overcoming the absence of socioeconomic data in medical records: validation and application of a census-based methodology. *Am J Public Health* 1992;82:703–10.

139. Rassen J, Brookhart M, Glynn R, Mittleman M, Schneeweis S. Instrumental variables I: instrumental variables exploit natural variation in nonexperimental data to estimate causal relationships. *J Clin Epidemiol* 2009;62:1226–32.

140. Commission Nationale de l'Informatique et des Libertés. *Rapport d'activité 2009*. Paris: Commission Nationale de l'Informatique et des Libertés, 2010.

141. Fond d'Intervention pour la Qualité et la Coordination des Soins. *Rapport d'activité 2009*. Paris: CNAMTS – DGOS, 2010.

142. Altman D. *Practical statistics for medical research*. Boca Raton, FL: Chapman & Hall/CRC, 1991.

143. Erler A, Beyer M, Muth C, Gerlach F, Brennecke R. Garbage in – garbage out? Validity of coded diagnoses from GP claims records. *Gesundheitswesen* 2009;71:823–31.

144. Pimouguet C, Le Goff M, Thiébaut R, Dartigues J, Helmer C. Effectiveness of disease-management programs for improving diabetes care: a meta-analysis. *Canadian Medical Association Journal* 2011;183:E115–E27.

145. Rijken M, Jones M, Heijmans M, Dixon A. Supporting self-management. In: Nolte E, McKee M, eds. *Caring for people with chronic conditions. A health system perspective*. Maidenhead, UK: Open University Press, 2008:116–42.

146. Soennichsen A. *Self-efficacy and peer support enhance the effectiveness of disease management in diabetes type 2 (SPEED)*. As of 30 December 2011: http://apps.who.int/trialsearch/trial.aspx?TrialID=ISRCTN10291077

147. Nederlandse Diabetes Federatie. *NDF zorgstandaard. Transparantie en kwaliteit van diabeteszorg voor mensen met diabetes type 2*. Amersfoort, The Netherlands: Nederlandse Diabetes Federatie, 2007.

148. Kennedy A, Reeves D, Bower P, Lee V, Middleton E, Richardson G, et al. The effectiveness and cost effectiveness of a national lay-led self care support programme for

patients with long-term conditions: a pragmatic randomised controlled trial. *J Epidemiol Commun Health* 2007;61:254–61.

149. Lorig K, Sobel D, Ritter P, Laurent D, Hobbs M. Effect of a self-management program on patients with chronic disease. *Effect Clin Pract* 2001;4:256–62.

150. Haider A, Sexton J, Sriram N, Cooper L, Efron D, Swoboda S, et al. Association of unconscious race and social class bias with vignette-based clinical assessments by medical students. *JAMA* 2011;306:942–51.

151. Rychetnik L, Frommer M, Hawe P, Shiell A. Criteria for evaluating evidence on public health interventions. *J Epidemiol Comm Health* 2002;56:119–27.

152. Smith P, Moffatt M, Gelskey S, Hudson S, Kaita K. Are community health interventions evaluated appropriately? A review of six journals. *J Clin Epidemiol* 1997;50:137–46.

153. Shojania K, Ranji S, McDonald K, Grimshaw J, Sundaram V, Rushakoff R, et al. Effects of quality improvement strategies for type 2 diabetes: a meta-regression analysis. *JAMA* 2006;296:427–40.

154. Piette J, Kerr E. The impact of comorbid chronic conditions on diabetes care. *Diabetes Care* 2006;29:725–31.

155. Struijs J, Baan C, Schellevis F, Westert G, van den Bos G. Comorbidity in patients with diabetes mellitus: impact on medical health care utilization. *BMC Health Serv Res* 2006;6:84.

156. Nolte E, McKee M. *Caring for people with chronic conditions. A health system perspective*. Copenhagen: World Health Organization on behalf of the European Observatory on Health Systems and Policies, 2008.

APPENDICES

Appendix A: Summary overview of six DISMEVAL work packages on disease management evaluation methods and metrics

Work package	Country	Intervention that formed the basis for evaluation
WP5	Austria	Diabetes disease management programme 'Therapie aktiv' *Target*: type 2 diabetes *Key components*: Patient management through coordinating physician following care pathways developed by the Austrian Society of Diabetes (ÖDG); Patient education through group instruction; involvement in goals setting and timelines, with agreed targets signed jointly; regular follow-up Standardised documentation of clinical and diagnostic measures and treatment *About 17,000 patients are enrolled in DMP across Austria (~4.3 percent of all people with diabetes type 2)*
WP6	Denmark	Integrated Rehabilitation Programme for Chronic Conditions (SIKS) project *Target*: chronic obstructive pulmonary disease; type 2 diabetes; cardiovascular disease, balance problems among elderly *Key components*: Multidisciplinary team supports the delivery of rehabilitation; regular patient follow-up; regular inter-organisational meetings Patient education and regular documentation of self-management needs and activities; involvement in developing individualised treatment plans and goal setting; access to physical exercise intervention Monitoring of practice team performance; systematic collection of clinical and other data *During 2005–2007 about 80,000 patients were covered by the SIKS project*
WP7	Germany	Diabetes disease management programme Target: type 2 diabetes *Key components:* Coordination of three care levels by GP on the basis of evidence-based guidelines developed by the German Institute for Evidence-based Medicine and Institute for Quality and Efficiency in Health Care Patient education in group sessions; involvement in agreeing treatment goals; regular follow-up , with patient reminders for missed sessions Standardised electronic documentation of treatment, patient's condition and test results, medication regime, and agreed treatment goals; central data analysis to produce quality reports. and provider feedback on performance

Work package	Country	Intervention that formed the basis for evaluation
		and for benchmarking
		By end of 2010, ~3.4 million individuals were enrolled in a diabetes type 2 DMP (70–85 percent of diagnosed diabetics in the statutory health insurance system)
WP8	France	Diabetes provider networks
		Target: type 2 diabetes
		Typical components:
		Multidisciplinary healthcare team; development of individualised care plan by core team; discussion forum and quality circles; regular follow-up
		Patient involvement in developing treatment plan towards a 'formal' agreement between patient and network
		Shared information system involving a database collecting routine clinical indicators and used for evaluation and quality control
		In 2007, around 500,000 people with diabetes were enrolled in diabetes networks (~20 percent of people with diabetes in France)
WP9	Netherlands	Bundled payment contracts for diabetes care ('diabetes care group')
		Target: type 2 diabetes
		Typical components:
		Stratification of patients according to disease severity; GP oversees referral to secondary care and ensures follow-up according to Nationally defined standards for diabetes care and multidisciplinary care protocol
		Patient education on self-management by practice nurses / specialised diabetes nurses, depending on the level of need
		Disease-specific electronic patient record with check-up and referrals data within care programme which allows for information sharing and automatisation of care protocols
		An estimated 750,000 people with diabetes are covered by a bundled payment contract
WP10	Spain	Nurse-led intervention for the prevention of cardiovascular disease
		Target: cardiovascular risk
		Key components:
		Structured telephone interview after 1, 4 and 8 months from initial medical check-up conducted by a trained nurse to assess knowledge about cardiovascular risk; adherence to recommendations (eg quitting smoking); awareness of clinical symptoms
		The programme is provided by a mutual fund (Ibermutuamur); between May 2004 and May 2007 just under 1 million medical checkups were carried out, of which around 630,000 were first medical checkups; of these 5,200 persons were identified to be at moderate to high risk to develop cardiovascular disease and offered the intervention

Appendix B: Evaluation measures for disease management (DM) interventions

Table B1 provides examples of measures for components of intervention structure, process, output and outcome. The table is taken from Conklin and Nolte (2010).[8]

Table B1 Evaluation measures for disease management (DM) 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 (eg eye exam rate; foot exam rate; HbA1c tests; blood pressure tests); Procedures performed (eg 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; Inpatient 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 (eg 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
Other effects/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: Conklin and Nolte (2010)