# Work Package 4

*Development of an analytical framework to perform a comprehensive evaluation of integrated care programmes for multi-morbidity using Multi-Criteria Decision Analysis*

# Deliverable report 2

*MCDA framework*

**WP leader**
Institute of Health Policy & Management, Erasmus University Rotterdam, The Netherlands
**WP co-leader**
University of Bergen (UiB), Norway
**Authors**
Fenna Leijten, Apostolos Tsiachristas, Melinde Boland, Maaike Hoedemakers, Nick Verbeek, Kamrul Islam, Jan Erik Askildsen, Antoinette de Bont, Roland Bal, Maureen Rutten-van Mölken
**Date**
January 31, 2017

--------------------------------------------------------------------------------------------------------------------

# Table of Contents

# Executive summary

## Preface

This chapter constitutes the executive summary of the deliverable of Work Package (WP) 4 of the SELFIE project. In this WP we developed the methods to perform comprehensive evaluations of integrated care programmes for multi-morbidity using Multi-Criteria Decision Analysis (MCDA). The WP leader is the Institute of Health Policy & Management from Erasmus University Rotterdam in the Netherlands and the WP co-leader is the University of Bergen (UiB) in Norway.

The chapter is structured as follows. First, an overview of the SELFIE project is provided and an introduction to WP4 is given. Next, a summary of the two WP4 deliverable reports is given. In the first deliverable report we give an overview of the outcomes that will be included in the MCDA and we describe the methods used to identify and select them. The second deliverable report starts with a description of MCDA methods in general, followed by the MCDA methods that will be applied in SELFIE. At the end of this chapter we describe how this work feeds into the next WP in which the comprehensive evaluations will actually be carried out.

## Introduction to SELFIE and WP4

The EU Horizon2020-funded SELFIE ('Sustainable Integrated Care Models for Multi-Morbidity: Delivery, Financing and Performance') project aims to improve person-centred care for persons with multi-morbidity by providing evidence on the impact of promising integrated chronic care (ICC) programmes and supporting financing/payment schemes on health- and well-being, experience with care, and cost outcomes (i.e., the Triple Aim). This four year research project is divided into nine work packages (WP) conducted by eight European partners: The Netherlands (coordinator), Austria, Croatia, Germany, Hungary, Norway, Spain, and the UK. SELFIE distinguishes itself from other research projects on integrated care and/or multi-morbidity by aiming to not only identify and describe promising integrated care programmes for multi-morbidity, but to evaluate them using an innovative approach: Multi-Criteria Decision Analyses (MCDA). MCDA is an umbrella term for a set of methods that aid decision-making when this is based on more than one criterion, whereby the relative impact that all criteria have on the decision-making process is made explicit. MCDA thus aims to improve transparency, accountability, and acceptability of the decision-making process by explicitly defining aims, or criteria.

In WP1 we first developed a conceptual framework to guide the further descriptive and evaluative work on integrated care for multi-morbidity in SELFIE. This conceptual framework includes six key components centred around the holistic understanding of an individual with multi-morbidity: (1) Service delivery (2) Leadership and governance, (3) Workforce, (4) Financing, (5) Technologies and medical products (6) Information and research. Subsequently, in WP1 and WP2, 17 promising integrated care programmes for persons with multi-morbidity were identified in the countries of the SELFIE consortium (2-3 per county) and extensively described, using both document analyses and interviews. This resulted in 17 'thick descriptions' that are being compared across countries (see also the executive summaries of WP1 and WP2). Preparations for the empirical evaluations using an MCDA framework were made in WP4.

Integrated care programmes are complex interventions consisting of a mixed package of interacting patient-, provider-, and organisational-oriented interventions that are tailored to the context in which they are introduced and continuously improved as more experience is gained. They do not only aim to maximise health but also to improve well-being, experience of care and reduce costs. Therefore, traditional evaluation frameworks such as cost-utility analysis in which costs per Quality Adjusted Life Years (QALY) are calculated, were deemed insufficient. In contrast, using MCDA has the purpose of ensuring that these evaluations are broad enough to incorporate the wide range of different outcomes of these programmes.

The work done in WP4 is divided into two deliverables (i.e., two reports). In the first report we defined a set of outcomes for which we will **measure performance** of the promising integrated care programmes and for which we will **elicit weights**. In the second report, we created an MCDA evaluation framework by selecting an MCDA method, constructing study designs to measure the performance of the 17 promising programmes, and preparing for the weight-elicitation study whereby the **importance** of the various outcomes will be determined. With this MCDA framework we will, in WP5, conduct 17 comprehensive evaluations in which integrated care programmes are compared to usual care or a control group.

The results of the two deliverables will be summarised below.

## WP4 Deliverable Report 1: Selecting and defining outcomes for the evaluation

In the first deliverable report of WP4, entitled "*Outcomes and indicators in integrated care for persons with multi-morbidity",* we selected and defined a set of outcomes that are specifically relevant for the empirical evaluation of integrated care programmes for persons with multi-morbidity. Secondly, we selected instruments or indicators to measure these outcomes. A great number of outcomes and instruments exist to measure integrated care, and we witnessed a tremendous growth in new types of measures such as patient reported outcomes (PROMs) and patient reported experience measures (PREMs). Although multiple criteria, or outcomes, can be incorporated in an MCDA, feasibility in collecting performance information on these and the need to avoid cognitive overload for respondents in the weight-elicitation study forced us to select the most relevant ones. **Figure I** provides an overview of the steps in this selection process.

**Figure I:** *Steps to develop the list of outcomes measures and indicators included in the MCDA: a core set and a programme-type specific set*

| Creating a long-list of outcomes | | | |
| --- | --- | --- | --- |
| Literature | Stakeholder workshops | Focus groups | Selected programmes |

| Creating a core set of outcomes and programme-type specific outcomes |
| --- |
| Discussion with researchers |

| Selecting instruments/indicators |
| --- |
| Literature search and discussion with researchers |
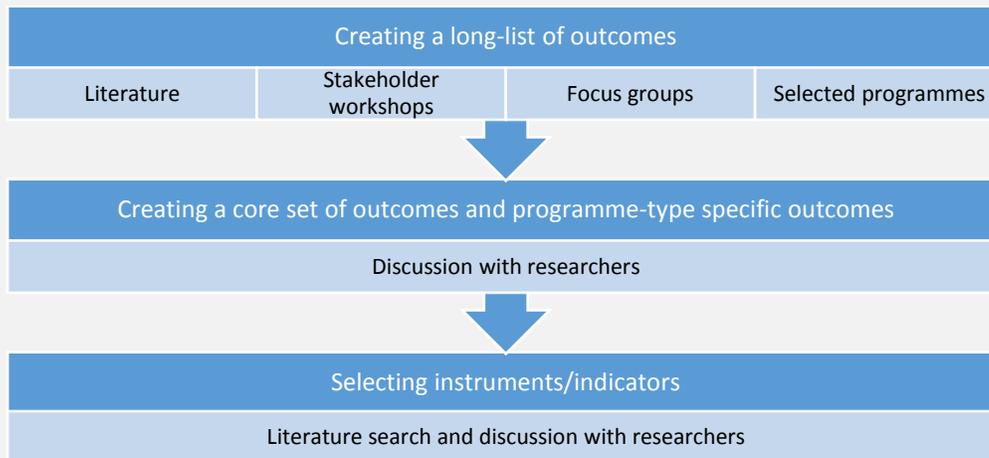
We used four main sources to create an initial 'long-list' of outcome measures: (1) literature review, (2) stakeholder workshops, (3) focus groups in individuals with multi-morbidity, and (3) a review of outcomes currently used in the 17 programmes selected for evaluation in SELFIE.

Key (grey) **literature** was used to explore innovative definitions and measures of health and integrated care. We focused in particular on outcomes to measure integrated care according to the Triple Aim (health- and well-being, experience of care, and costs). In scientific literature, we identified outcomes that were measured in integrated care programmes specifically targeting individuals with multi-morbidity.

Alongside the literature review, each country in the SELFIE consortium organised a **workshop with national stakeholders** that represented five stakeholder groups, the 5Ps: Patients, Partners (i.e., informal caregivers), Professionals, Payers, and Policy makers. During these workshops the stakeholders were asked to name and define what would make them reimburse, participate in, offer, or implement an integrated care programme for multi-morbidity. The outcomes that resulted from these workshops were added to those found in the literature.

In order to zoom into and learn more about the importance of person-centred integrated care for individuals with multi-morbidity, each country in the SELFIE consortium organised a **focus group**. In total 58 individuals with multi-morbidity attended these focus groups, in which they discussed what it means to them to be in 'good health' and how they define a good care process. Outcomes mentioned in the focus groups were again added to the list of candidate outcomes.

Lastly, we reviewed the 17 **promising integrated care programmes** included as case studies in SELFIE as a source of information on possible outcomes: what are the goals of these programmes, what are they already measuring, and what do they find important to measure?

An abundance of outcomes was obtained from these four sources, many with large conceptual overlap. We clustered the outcomes into higher-level concepts and categorised them into the Triple Aim. There was considerable agreement between the outcomes mentioned by the various stakeholders. There was general consensus that we should focus on patient-reported outcome measures and patient-reported experience measured to extend the frequently used

structural indicators or indicators of the adherence to programme-components that are extracted from routine organisational and system-level databases. Furthermore, an overarching theme was that evaluations of integrated care for multi-morbidity should go beyond traditional clinical health outcomes, and should focus more broadly on well-being. Further, it was felt that in the selection process the outcomes that were frequently mentioned by the persons with multi-morbidity in the focus groups should be leading. The group of persons with multi-morbidity is in and of itself complex and varied; they deal with many different health- and social problems that may interact with one-another. For these reason, in SELFIE, we incorporate **'social relationships and participation'**, **'enjoyment of life'** and **'resilience'** as health- and well-being outcomes alongside the more traditional outcomes of **'physical functioning'** and **'psychological well-being'** as these apply to persons with all different types of disease- and problem-combinations. Persons with multi-morbidity often deal with care providers from different sectors, with a high risk of fragmentation and repetition. Thus we identified **'person-centeredness'** and **'continuity of care'** as the two key elements of experience of care that should be included in evaluations. Moreover, we also include **'total health- and social care costs'** in SELFIE, to capture the care utilisation in different sectors.

The above mentioned outcomes were also selected because they were found relevant and applicable across all 17 integrated care programmes. The discussion of these outcomes amongst the SELFIE researchers revealed the need to add important programme-specific outcomes to the MCDA. This resulted in the decision to construct a) a core set of outcomes to be included in all 17 programme evaluations, and b) programme-type specific outcomes. Defining outcomes that are relevant across multiple programmes is important because one of the SELFIE aims is to develop a reusable MCDA, where criteria-weights can be used again by others who want to monitor different integrated care programmes, facilitated by an online tool.

To keep the weight elicitation study feasible, the 17 integrated care programmes were categorised into four types of programmes: population health management programmes, programmes targeting frail elderly, palliative care and oncological programmes, and

programmes targeting persons with problems in multiple life domains. For each type of programme a list of outcomes was defined. The core set of outcomes includes the eight mentioned above; they are presented and defined in **Table I**. An overview of the core set and the programme-type specific outcomes are presented in **Table II**. Both the core set of outcomes and the programme-type specific outcomes will be included in the MCDA. This implies that we will elicit weights for both of them.

The outcomes in **Tables I** and **II** were defined at a conceptual level in order to allow the use of different instruments or indicators to measure a particular outcome-concept. The reason that we permit the use of different instruments and indicators to measure a particular concept is that some programmes have already been measuring certain outcomes for years, and this retrospective data is of great value. In the cases when data collection still needs to be set up, we make recommendations for instruments or indicators that best operationalise the outcomes in SELFIE.

**Table I**: *Core set of outcomes*

| Health / well-being |
|---|
| **Physical functioning** |
| *"Acceptable physical health and being able to do daily activities without needing assistance"* |
| **Psychological well-being** |
| *"Absence of stress, worrying, listlessness, anxiety, and feeling down"* |
| **Social relationships & participation** |
| *"Having meaningful connections with others as desired"* |
| **Enjoyment of life** |
| *"Having pleasure and happiness in life"* |
| **Resilience** |
| *"The ability to recover from or adjust to difficulties and to restore ones equilibrium"* |
| **Experience with care** |
| **Person-centeredness** |
| *"Care that care matches an individual's needs, capabilities and preferences and jointly making informed decisions"* |
| **Continuity of care** |
| *"Good collaboration, smooth transitions between caregivers, and no waste of time"* |
| **Costs** |
| **Total health and social care costs** |
| *"Total health and social care costs per participant"* |

**Table II:** *Overview of core set and programme-type specific outcomes in SELFIE*

| Outcomes for integrated care for persons with multi-morbidity | | | | | |
|---|---|---|---|---|---|
| | | **Programme-type specific outcomes** | | | |
| | **Core set outcomes** | *Population health management* | *Frail elderly* | *Palliative and oncology* | *Problems in multiple life domains* |
| **Health & well-being** | Physical functioning | Activation & engagement | Autonomy | Mortality | Self-sufficiency |
| | Psychological well-being | | | Pain and other symptoms | |
| | Social participation/ relationships | | | | |
| | Resilience | | | | |
| | Enjoyment of life | | | | |
| **Experience** | Person-centeredness | | Burden of medication | Compassionate care | |
| | Continuity of care | | Burden of informal caregiving | Timely access to care | |
| | | | | Preferred place of death | |
| | | | | Burden of informal caregiving | |
| **Costs** | Total health- and social care costs | Ambulatory care sensitive hospital admissions | Living at home | | Justice costs |
| | | Hospital re-admissions | Falls leading to hospital admissions | | |

## WP4 Deliverable Report 2: Creating an MCDA evaluation framework

In the **WP4 Deliverable Report 2**, entitled *"MCDA framework"*, the reason why MCDA was chosen as the evaluative framework in SELFIE is explained. As described above, this broad evaluation framework allows for multiple outcomes (in MCDA-terms 'criteria') to be included in the evaluation, and weighs these from specific perspectives. There are different MCDA methods, which can broadly be divided into the value-based, outranking, and goal- or reference point methods. Each is briefly described below.

In **value-based methods**, the aim is to assign values to alternatives and construct a value function. In SELFIE the alternatives are the integrated care programme and its comparator. In the commonly used value-based method Multi-Attribute Utility Theory (MAUT), a single overall value is created. The performance of each alternative, on all criteria, is determined. Separately, the importance of the criteria needs to be determined. For each alternative, the weighted performance on each criterion is aggregated into an overall value score. This overall value score is compared between the integrated care programme and its comparator.

In **outranking methods**, pairwise comparisons are made of the performance of all alternatives on all criteria. In the simplest case, if we compare the performance of alternatives on all criteria and one alternative scores better on all, then this is the preferred alternative. In less simple cases, patterns of dominance between alternatives are studied to reach a decision about the preferred alternative. For this method the performance of alternatives on criteria needs to be known, as well as the weights for these criteria.

In **goal** or **reference point methods**, alternative care programmes are compared by calculating the weighted deviations from a priori set goals. This method requires a specification of desirable levels of performance for each criterion.

Given that in SELFIE we aim to compare each integrated care programme to its comparator and not to rank all 17 programmes, and considering the theoretical foundations of all methods, we have opted for MAUT methods to be applied in the MCDA. The seven steps commonly undertaken in MCDAs, and MAUT specifically, are as follows:

1. Establish the decision-context
2. Identify and structure criteria
3. Determine performance on criteria
4. Weight-elicitation
5. Creating a global score
6. Sensitivity analyses
7. Examine results

The work done in earlier SELFIE WPs has helped us understand and establish the **decision-context** of integrated care for multi-morbidity (step 1). For most of the 17 case studies in SELFIE, the decision pertains to reimbursement, continuation, extension, and/or wider implementation of the integrated care programme. Step 2 was described in **WP4 Deliverable Report 1** "*Outcomes and indicators in integrated care for persons with multi-morbidity*".

To determine the performance on the criteria, step 3, empirical evaluations for all 17 promising programmes are being set up. In order to be able to attribute effects to the intervention, performance will be repeatedly assessed in both the integrated care group as well as a comparator group. Each SELFIE partner selected the most appropriate study design for their evaluation and started working on a study protocol to make this explicit.

There are different methods to elicit weights (step 4), including: direct ranking, visual analogue scales, point allocation, analytical hierarchy process, swing weighting, and **discrete choice experiments** (DCE). In a DCE, choice sets with scenarios are presented to stakeholders. The scenarios consist of various alternatives (e.g., care programmes) that systematically differ on performance criteria (i.e., outcomes, attributes). Stakeholders are asked which scenario they prefer. Hereafter, weights for each criterion can be statistically derived on the basis of the

likelihood that one scenario, with specific criteria performance, is preferred over another. The stakeholders are forced to make trade-offs in criteria and take the full range of potential performance into account. DCE was selected as the method to elicit weights, in all eight SELFIE partner countries, from the 5P perspectives. Due to the number of different outcomes/criteria, it is not possible to conduct a DCE for the core set and the programme-type specific outcomes. For this reason, a DCE will only be used to elicit weights for the core set of outcomes.

The perceived 'next-best' method, **swing weighting**, will be used to elicit weights for the programme-type specific outcomes. Swing weighting is also a trade-off weighting method, in which the relative importance is determined on the basis of moving from the worst to best score on a scale. Specifically, the SMARTER method will be used, whereby a stakeholder is asked to pretend there is an alternative (e.g., care programme) that has the lowest possible scores on all criteria. The stakeholder then ranks which criteria would be selected first to swing from the worst to the best level. This is subsequently done for the remaining criteria. These ranks are then turned into weights using, for example, the rank ordered centroid method.

**Table III** illustrates the type of information that will be obtained in the MCDA. This includes the (standardised) performance scores of two alternative care programmes (e.g., integrated vs. usual) on a number of criteria, the weights of these criteria from the viewpoint of different stakeholder groups (S1 and S2), and the weighted aggregation. In the simplified example, in **Table III** the aggregated weight for resilience is calculated by multiplying the criteria weight of stakeholder 1 (0.30) or stakeholder 2 (0.15) with the standardised performance (0.74 for the integrated care programme and 0.67 for the usual care). When these weighted performance scores are summed across all criteria the overall value of a programme is obtained (step 5 of the MCDA).

**Table III**: *SELFIE example of information needed for an MCDA of integrated care programmes for multi-morbidity*

| | | Weight | | Care alternatives | | Weighted aggregation | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | *Integrated* | *Usual* | *Integrated* | | *Usual* | |
| Triple Aim | **Criteria** | S1 | S2 | **Standardised performance*** | | S1 | S2 | S1 | S2 |
| Improved health | **Resilience** | .30 | .20 | 0.74 | 0.67 | 0.22 | 0.15 | 0.20 | 0.13 |
| | **Physical functioning** | .20 | .15 | 0.68 | 0.73 | 0.14 | 0.10 | 0.15 | 0.11 |
| Improved experience | **Person-centeredness** | .15 | .05 | 0.80 | 0.60 | 0.12 | 0.04 | 0.09 | 0.03 |
| | **Continuity of care** | .25 | .05 | 0.77 | 0.63 | 0.19 | 0.04 | 0.16 | 0.03 |
| Reduced costs | **Health care costs** | .05 | .30 | 0.28 | 0.30 | 0.01 | 0.08 | 0.02 | 0.09 |
| | **Social care costs** | .05 | .25 | 0.24 | 0.34 | 0.01 | 0.06 | 0.02 | 0.09 |
| | | | | | **Sum** | 0.69 | 0.47 | 0.64 | 0.48 |

*Note: Aggregation on the basis of Multi-Attribute Utility Theory (MAUT). *Standardised performance based on relative standardisation. S1 = Stakeholder 1 (e.g., patient), S2 = Stakeholder 2 (e.g., payer).*

In step 6, sensitivity analyses will be done. This will include subgroup analyses, such as per gender, educational level or types of morbidities. Furthermore, we will conduct deterministic analyses, whereby certain criteria are excluded, as well as probabilistic analyses, in which uncertainty in weights and performance is modelled simultaneously. In step 7, the results will be examined. This will be done by the SELFIE researchers, but will also involve reflecting and interpreting the findings with representatives from the 5Ps in international and national stakeholder workshops.

In the **WP4 Deliverable Report 2** we extensively describe the background of MCDA methods, the seven steps undertaken in MAUT, MCDA and weight-elicitation choices in SELFIE, possible study designs, the weight-elicitation procedure, and the draft study designs to measure the performance of the 17 programmes.

## Next steps

In the next phase of SELFIE research, each SELFIE partner will start data collection from the five defined stakeholders (5Ps) for the weight-elicitation. Parallel to this, the study designs for the empirical evaluations will be made definite and data collection in all 17 promising integrated care programmes will begin. Subsequently, the performances of the promising programmes on the (core) set of outcomes and the weights from the various stakeholders will be brought together in the proposed SELFIE framework, resulting in MCDAs of 17 promising integrated programmes for persons with multi-morbidity

In order to allow findings from the SELFIE study to be shared with others, an online MCDA tool will be developed that will allow others to also apply the criteria weights from the 5Ps to their own programme performance. The tool will stay available after the SELFIE research project has ended.

# 1. This report in the context of SELFIE

SELFIE is a Horizon2020 EU project that will contribute to the current state of knowledge of integrated chronic care (ICC) for persons with multi-morbidity and provide applicable policy advice. We aim to generate evidence on the impact of promising ICC programmes and supporting financing/payment schemes on health and well-being outcomes, experience, and costs. Specific ICC programmes for multi-morbidity will be empirically evaluated using multi-criteria decision analyses (MCDA). The definitions of multi-morbidity and ICC in the SELFIE project can be found in Box 1.

**Box 1:** *Definitions of multi-morbidity and integrated chronic care in SELFIE*

> **Multi-morbidity** in the context of SELFIE refers to multiple (i.e., at least two) chronic conditions, physical or mental, occurring in one person at the same time, where one is not just a known complication of the other.
>
> **Integrated chronic care (ICC)** in the context of SELFIE refers to structured efforts to provide coordinated, pro-active, person-centred, multidisciplinary care by two or more communicating and collaborating care providers that may work at the same organisation or different organisations, either within the healthcare or across the health care, social care, or community care sector (including informal care).

In SELFIE three research strands are distinguished, with 9 work packages, 7 of which are content based (WP8: Communication and dissemination, WP9: Management):

**Figure 1**: *SELFIE strands of research and work package (WP) overview* (Click to see overview on the SELFIE website)
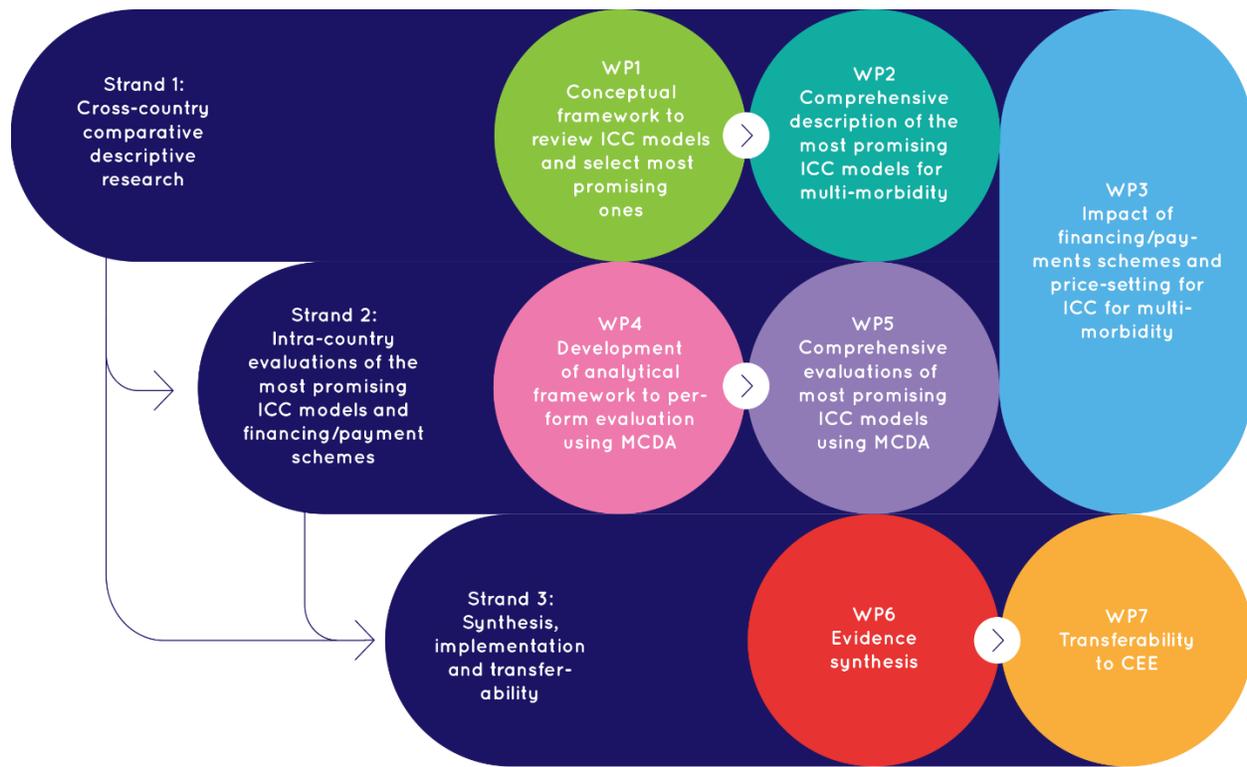


In WP1 a conceptual framework for integrated care for multi-morbidity was developed. Furthermore, promising integrated care programmes for multi-morbidity in each of the SELFIE partner countries were identified, and 17 were selected (2-3 per partner) (see **Appendix 1** for an overview of these programmes). In WP2 these 17 programmes were described on the basis of the framework developed in WP1, using both document analyses and interviews. This resulted in 17 'thick descriptions' that are being compared across countries. In WP3 the impact of different financing and payment schemes is being investigated. This WP makes use of the descriptive research on this topic in WP2 and the empirical evaluations in WP5, but it also investigates the impact of different funding and payment schemes, independent of the 17 programmes. Hence, WP3 overlaps strand 1 (descriptive, cross-country) and 2 (evaluative, intra-country).

The aim of WP4 is to develop an analytical framework to perform comprehensive evaluations of the 17 programmes using Multi-Criteria Decision Analysis (MCDA); the work in WP4 forms the theoretical basis and preparation of the actual empirical evaluation in WP5.

Throughout the SELFIE project five groups of stakeholders are repeatedly involved to reflect on findings and advice on how to move forward and interpret findings, the **5Ps**:

1. **P**atients; persons with multi-morbidity
2. **P**artners; informal caregivers
3. **P**rofessionals
4. **P**ayers
5. **P**olicy makers

The Dutch EUR team leads WP4, the Norwegian UiB team leads WP5 – they are also one-another's co-leaders in these WPs. The teams collaborate closely, as the work in WP4 feeds directly into WP5. WP4 started on February 1$^{st}$, 2016 and ends February 1$^{st}$, 2017. WP5 started on September 1$^{st}$, 2016 and ends September 1$^{st}$, 2018.

In WP4 we aim to develop an analytical framework and have two main deliverables:

  i. Outcomes report – due month 17 (Jan 2017 – submit before Feb 1$^{st}$ 2017)
  ii. MCDA framework report – due month 17 (Jan 2017 – submit before Feb 1$^{st}$ 2017)

This is the second deliverable report, 'MCDA Framework'. These reports correspond to the tasks set out in the grant agreement, the current report addresses tasks 1, 4 and 5 of WP4:

1. **To select the MCDA method to be applied** *[Ch. 2-5]*
2. To select, define, develop and prepare the collection of outcomes, including indicators specifically relevant to the evaluation of integrated care programme for persons with multi-morbidity *[Deliverable 1 report]*
3. To develop methods to collect healthcare utilisation and other cost categories that will be included in the evaluation *[Deliverable 1 report]*

4. **To determine the importance of the various outcomes and indicators from the perspective of the 5 stakeholder groups** *[Ch. 6]*
5. **To construct study designs and select statistical techniques to ensure causal inference between the integrated care programmes and outcomes** *[Ch. 7-8]*

Because WP4 feeds into WP5, we also present some basic information on WP5 below. WP5 aims to comprehensively evaluate the selected most promising integrated care programmes using MCDA.

There are three deliverables in WP5:

i. Comprehensive evaluation of 17 most promising integrated care programmes – due month 35 (Jul 2018 – submit before Aug 1st 2018).
ii. Performance-monitoring tool – due month 36 (Aug 2018 – submit before Sept 1st 2018).
iii. Submission of study questionnaires and ethical approvals – due month 24 (calendar month Sept 2017 – submit before Oct 1st 2017).

These deliverables will correspond to the five tasks set out in this WP:

1. Writing study protocols for the evaluation of each integrated care programme
2. Data collection
3. Data management
4. Perform MCDA
5. Development of a user friendly performance monitoring tool based on MCDA

In the remainder of this report a general introduction to the use of MCDA in SELFIE is provided (**Chapter 2**). Hereafter, the most commonly used MCDA methods, their assumptions, strengths, and weaknesses are described (**Chapter 3**) and the steps to be taken in conducting an MCDA are presented (**Chapter 4**). Subsequently, in **Chapter 5** the MCDA and weight-elicitation method selection in SELFIE is described. In **Chapter 6** instructions for the partners are provided on how to execute the weight-elicitation. Lastly, in **Chapter 7** an overview of possible study designs for the evaluation of the programmes' performance is provided and in **Chapter 8** per programme a brief overview of the selected design is described. The latter will be expanded upon in WP5.

## 2. General introduction to MCDA and how it will be applied in SELFIE

In order to evaluate the 17 selected integrated care programmes, multi-criteria decision analysis (MCDA) will be used. In SELFIE we use the following definition of MCDA:

> "A set of methods and approaches to aid decision-making – where decisions are based on more than one criterion – in which the impact that all criteria have on the decision and their relative importance is made explicit." (Devlin & Sussex, 2011)

MCDA aims to improve transparency, consistency, accountability, credibility, and acceptability of decision-making by assessing alternatives on the basis of explicit aims that are identified and deemed important for the decision problem at hand. According to the aims of the problem, several criteria are derived which are measurable and can be weighted. By assessing alternatives, MCDA allows for systematic trade-offs to be made between multiple, and sometimes conflicting, criteria. Because criteria weights can be elicited from different perspectives, i.e., from different groups of stakeholders, it makes differences in opinion about the relative importance of criteria more explicit. (Dolan, 2010) MCDA can be performed with different methods; the main types hereof will be described in the current report.

The application of MCDA is well-suited in the health technology assessment field as it can aid the difficult decision-making process surrounding complex interventions. Integrated care programmes for multi-morbidity can certainly be considered 'complex interventions': they consist of various interacting components, target individuals but also groups and organisations, have a variety of intended outcomes, are amendable to tailoring through adaptation and learning feedback loops, and their effectiveness is impacted by the behaviour of those delivering and receiving the intervention. (Craig et al., 2008) For such complex interventions applying a traditional cost-effectiveness evaluation is not sufficient. Namely, these complex intervention aim to effect a wide array of outcomes, reaching far beyond the traditional QALY. This is the reason that in the SELFIE evaluations comprehensive MCDA will be performed.

Currently, stakeholders involved in the decision-making in integrated care for multi-morbidity, often implicitly weigh criteria but incorporate them into decision-making in a deliberate way – in an MCDA this is all made explicit. In addition, performance measurements on criteria are combined with their relative importance weights from the perspective of a particular stakeholder group [in a specific country]. This will allow us, in SELFIE, to conduct comprehensive evaluations that compare integrated care programmes to usual care or a control programme, explicitly taking different perspectives into account. Our results can aid more transparent and nuanced decisions regarding the reimbursement, continuation, and wider implementation of integrated care programmes for multi-morbidity.

In the context of SELFIE, the following working definitions of MCDA concepts are used:

- **Aims** overall pertain to the Triple Aim (i.e., improved health and well-being, improved experience, and reduced costs) but can be divided into [programme-specific] sub-aims.
- **Alternatives**[a] pertain to the 17 integrated care programmes for multi-morbidity that were selected in WP1 and their comparators (together making up the case studies) (see **Appendix 1**). In SELFIE we will not be comparing the 17 programmes to one-another; instead each programme, if possible, will be compared to a control group (e.g., usual care).
- **Case study** thus pertains to the comparison between the two alternatives (i.e., the selected promising integrated care programme and the control).
- **Country specific criteria weights** refer to the breakdown of the weights from the perspective of the 5P stakeholders by country- or regional level, as differences may exist.
- **Criteria**[b] refer to specific, measurable, *concepts*. These concepts relate to [sub-]aims. Although our primary aim is not to compare the 17 integrated care programmes amongst one-another, we aim to use a core set of criteria in all 17 case studies in order

---

[a] Alternatives are also referred to as *choices* or *options* in the literature.
[b] Criteria are also referred to as *attributes* in the literature, and in the **SELFIE WP4 Deliverable Report 1** these are referred to as *outcomes* and *concepts.*

to enable knowledge transferability between the partner countries.[c] Programme-specific criteria will also be identified. Criteria will be operationalised with the use of specific indicators and will be measured on varying scales (e.g., continuous scores or ordinal scales). See **Figure 1** for an example of aims, criteria, and indicators.

- **Indicators** or instruments re means to operationalise or measure criteria (i.e., instruments).

- **Weights** are the relative value, preference, or importance that is given to each criterion.

- **Stakeholders** are persons directly or indirectly involved in the decision-making process in the field of integrated care for multi-morbidity who will have an explicit perspective in the MCDA. These, in SELFIE, are the 5P's: Patients, Partners, Professionals, Payers, and Policy makers. These stakeholders were involved in the identification of alternatives (WP1), and of relevant criteria (WP4-task 2 (focus groups with patients) & national stakeholder workshops). The importance of the criteria (i.e., weights) will be determined from their perspective (WP4-task 4).

---

[c] This will be explained in more detail in **Chapter 4.2.**

**Figure 1:** *Example of a value tree with aims, criteria, and indicators*
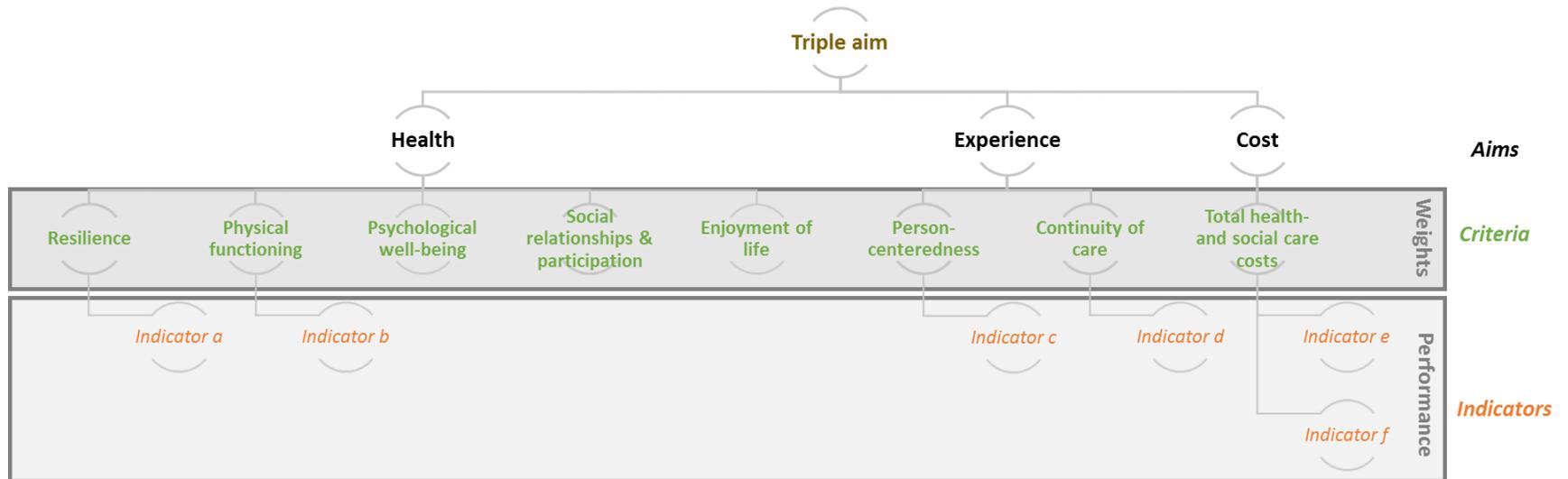


*Note: This example partly overlaps with the SELFIE core set presented in WP4 Deliverable Report 1, but is not identical to it.*

**Table 2** shows the seven steps often taken when performing an MCDA and how they are related to the SELFIE WP's and tasks. (Department for Communities and Local Government: London, 2009; Marsh et al., 2016) These steps are mostly consistent with the value-based MCDA approach (see **Chapter 3**), which we adopt for the SELFIE MCDA model. Not all these steps are always necessary in the development of an MCDA model. In **Chapter 4** of this report these seven steps and their implementation within the SELFIE context is extensively described.

**Table 2:** *MCDA steps – SELFIE*

| MCDA steps | SELFIE WP - Task |
|---|---|
| 1. Establishing the decision-context and identifying alternatives | • WP1, conceptual framework development and selection of integrated care programmes<br>• WP2, macro level description (decision-context)<br>• WP4, task 5 (study designs) |
| 2. Identify and structure criteria | • WP4, tasks 2 & 3 (outcomes) |
| 3. Measuring performance of criteria | • WP5, task 2 (data collection) |
| 4. Weight-elicitation | • WP4, task 4 (weight-elicitation) |
| 5. Combine weights and performance into global score | • WP5, task 4 (perform MCDA) |
| 6. Sensitivity analyses | • WP5, task 4 (perform MCDA) |
| 7. Examine results | • WP5, task 4 (perform MCDA) |

In order to provide the reader with a clearer idea of what an MCDA is, in general and within SELFIE, we provide an example of the type of information that is needed in **Table 3**. In the example, six criteria and specific indicators from **Figure 1** are used. An MCDA is presented of a case study in which two alternatives are compared: an integrated care programme and usual care. The overall aims of the two alternatives are to achieve the Triple Aim. For illustration purposes, we have chosen two criteria and indicators per aim. In the grey columns, the weights per criteria are presented from two perspectives, stakeholder 1 and 2, these can for example be Patients and Payers. In the yellow columns, the standardised performance of both alternatives (i.e., the integrated care programme and usual care) on each criterion is presented with a range from 0 to 100. In this example, performance and weight-elicitation is done separately, and

these are aggregated using Multi-Attribute Utility Theory (MAUT).[d] Throughout the report we refer back to the example provided in **Table 3** and expand on this.

---

[d] Determining weights and performance separately is a methodological choice in MCDA that is addressed in **Chapters 3** and **4** of this report.

**Table 3**: *SELFIE example of information needed for an MCDA of integrated care programmes for multi-morbidity*

| Triple Aim | Criteria | Indicator (hypothetical range) | Weight | | Care alternatives | | Weighted aggregation | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | Integrated | Usual | Integrated | | Usual | |
| | Criteria | Indicator (hypothetical range) | S1 | S2 | Standardised performance* | | S1 | S2 | S1 | S2 |
| Improved health | Resilience | a (0-100) | .30 | .20 | 0.74 | 0.67 | 0.22 | 0.15 | 0.20 | 0.13 |
| | Physical functioning | b (0-100) | .20 | .15 | 0.68 | 0.73 | 0.14 | 0.10 | 0.15 | 0.11 |
| Improved experience | Person-centeredness | c (1-5) | .15 | .05 | 0.80 | 0.60 | 0.12 | 0.04 | 0.09 | 0.03 |
| | Continuity of care | d (1-7) | .25 | .05 | 0.77 | 0.63 | 0.19 | 0.04 | 0.16 | 0.03 |
| Reduced costs | Health care costs | e (health care costs – € p/y) | .05 | .30 | 0.28 | 0.30 | 0.01 | 0.08 | 0.02 | 0.09 |
| | Social care costs | f (social care costs – € p/month) | .05 | .25 | 0.24 | 0.34 | 0.01 | 0.06 | 0.02 | 0.09 |
| | | | | | | Sum | 0.69 | 0.47 | 0.64 | 0.48 |

Note: Aggregation on the basis of Multi-Attribute Utility Theory (MAUT). *Standardised performance based on relative standardisation. S1 = Stakeholder 1 (e.g., patient), S2 = Stakeholder 2 (e.g., payer). P/Y = per capita per year, P/M = per capita per month.

# 3. MCDA methods

To help understand which decisions have been made regarding the MCDA method for SELFIE, in this chapter we give a brief and general overview of the literature on MCDA methodology.

MCDA is traditionally regarded as a sub-discipline of operations research and management science (Belton & Stewart, 2002). However, any method within a certain discipline that considers multiple criteria in an evaluation of alternatives may be regarded as an MCDA method. In line with this, in their recent review on MCDA in healthcare, Thokala and colleagues (2014) emphasise that MCDA calls for a very general definition. To reflect stakeholder viewpoints in an analysis, relative stakeholder preferences between criteria are often incorporated into the MCDA model. To elicit preferences for this purpose, there is a large body of preference elicitation literature. This largely originates from econometric theory (e.g., conjoint analysis and discrete choice experiment). Hence, MCDA offers an opportunity to combine methods from different disciplines.

MCDA methods can traditionally be categorised into 'value-based', 'outranking', and 'goal' or 'reference' point methods. In value-based methods, the aim is to assign values to alternatives that reflect preferences regarding the performance of these alternatives and construct a value function. In outranking methods, the interest is in the relative performance ranking of alternatives. In goal or reference point methods, the aim is to optimise a multiple objective programming formulation to attain pre-specified desired performance levels. Within value-based methods, the most common approaches are Multi-Attribute Utility Theory (MAUT) and Analytical Hierarchy Process (AHP) and within outranking a commonly used method is ELECTRE. (Marsh et al., 2014; Thokala et al., 2016) For reference level methods, goal programming is a well-known application. These are very distinct methods however, that have little in common with one-another aside from combing multiple criteria. As such, they each serve a distinct purpose but can also be used in a complimentary manner to address a decision problem.

In **Table 4** an overview of MCDA methods is presented. This table is not exhaustive; the most commonly used methods are presented. Overall MCDA methods underlined and blue and the weight-elicitation methods *italicized in orange* will be addressed in the current report. These are the methods that are deemed possibly suitable and applicable in SELFIE.

**Table 4:** Categorising MCDA methods

| Value-based | • Multi-Attribute Utility Theory (MAUT) (i.e., additive aggregate model, weighted sum approach)<br>• Weight-elicitation:<br>    ○ *Direct ranking*<br>    ○ *Visual Analogue Scales*<br>    ○ *Point Allocation*<br>    ○ *AHP*<br>    ○ *Swing weighting*<br>    ○ *Discrete Choice Experiment (DCE)*<br>• Analytical Hierarchy Process (AHP)<br>• Measuring Attractiveness by a Categorical Based Evaluation Technique (MACBETH)<br>• Potentially All Pairwise Rankings of all Possible Alternatives (PAPRIKA) |
|---|---|
| Outranking | • ELimination Et Choix Traduisant la REalité (ELECTRE)<br>• Preference Ranking Organization METHod for Enrichment of Evaluations (PROMETHEE) |

Different categorisations of MCDA methods exist. Another way to differentiate methods is by the means of preference elicitation: either weights for criteria are determined separately from performance or preferences for whole alternatives are determined and thus performance and weights are simultaneously derived. Some of the same methods can be used to elicit preferences for separate criteria *and* for whole alternatives, such as AHP. Thus AHP is both a weight-elicitation method and a method to determine preferences for overall alternatives. Both applications will be described in this report. Below a general description of the main MCDA categories is provided. In later chapters of the report this will be done more extensively and examples will be provided. Hereafter, in **Chapter 5** we will describe the exact methods that will be applied in SELFIE.

In value-based methods the degree to which one alternative is preferred over another is determined by constructing and comparing a single overall value that reflects the importance/preference for each alternatives. This value is made up of an aggregation of values for individual criteria. (Thokala & Duenas, 2012)

In the value-based **Multi-Attribute Utility Theory (MAUT)** method, a single overall value is created by weighted aggregation of standardised performance scores over multiple criteria. MAUT is based on expected utility theory, and the axioms from the Von Neumann-Morgenstern utility theorem in the presence of uncertainty. Its extension to multiple criteria became popular due to the work by Keeney and Raiffa (1993). They emphasised the use of axiomatic foundations for multi-attribute decision problems where uncertainty, and therefore risk, plays a role. Methods often used in health economic evaluations, such as standard gamble, are based on MAUT. For MAUT to satisfy the axioms, several assumptions are necessary, and whether they are required depends on the chosen value function form (e.g., multiplicative, linear, multilinear). However, the explication of the risk involved in a decision, makes the weight elicitation task very complicated. For example, in SELFIE eliciting uncertainty of preferences would require determining respondents' inclinations to choose between having, for example, high 'enjoyment of life' with a certain probability and run the risk of a worse outcome, or otherwise settle for just a modest improvement with a higher probability. Accounting for decision uncertainty when eliciting weights is interesting, but quickly becomes unwieldly using common preference elicitation methods such as those presented in Table 4. Therefore, this is often left out, as a result of which the theoretical framework is reduced to Multi-Attribute Value Theory (MAVT), which only requires the preferential independence condition to satisfy its axioms. Preferential independence means that the weight of one criterion can be elicited independently from the performance score on another criterion. Most of the value based methods explained in this report and used in SELFIE are based on MAVT. Nevertheless, in this report we will use the term MAUT because it is a generalisation commonly used in the literature.

The performance of each alternative, on all criteria, needs to be determined for MAUT. These are then standardised (e.g., recalculated to scales from 0-1 or 0-100). Separate from this process, the importance of criteria needs to be determined in the form of weights. Lastly, the partial value functions (weighted performance on each criterion for each alternative) need to be aggregated. The most commonly used aggregation is a linear additive model. This is usually appropriate, as long as performance differences between similar intervals are judged to be equivalent for the entire range of possible response/performance values: for example, changes in preferences from 0-5 should be the same as from 50-55 and 90-95 (on a performance scale from 0-100). (Dolan, 2010) The use of a linear scale is acceptable as long as the ratio between any two equally large response intervals is less than or equal to 2:1. (Dolan, 2010)

**Outranking methods** are another value-based method that can be seen as methods of 'dominance'. For this method the performance of alternatives on criteria needs to be known, as well as the weights for these criteria. In outranking, pairwise comparisons are made of the performance of all alternatives on all criteria. In the simplest case, if we compare the performance of alternatives on all their criteria and one alternative scores better on all, then this is the preferred alternative. In other cases, pairwise comparisons are made in order to construct a concordance and discordance matrix. The concordance matrix shows the concordance index of every pairwise comparison between alternatives. The concordance index of alternative A versus B is based on the criteria on which A scores better than B, i.e., outranks B. The discordance index of A over B is based on the criteria on which A scores worse, or are outranked, by B. Sometimes an indifference threshold is determined, saying that performance can only be said to differ between alternatives A and B if the difference is larger than $x$ (this is taking imprecision into account). Patterns of dominance between alternatives are studied to reach a decision. (Department for Communities and Local Government: London, 2009; Thokala & Duenas, 2012)

**Analytical hierarchy process (AHP)** is a pairwise comparison method that can be used both to determine criteria weights as well as performance scores. The method consists of having

stakeholders make pairwise comparison usually on a 1-9 point scale. For criteria, weights are elicited by having stakeholders compare two criteria at a time and indicating the 'intensity of importance' of one criterion over another, in other words, the strength of their preference for one over the other. The 9-point scale in AHP ranges from 1 'Equally as important' to 9 'Overwhelmingly more important', whereby if criteria or alternative $x$ is overwhelmingly more important than $y$, $x$ gets a score of 9 as compared to $y$, and $y$ of $^1/_9$ to $x$. After all criteria have been compared to one-another, a matrix is made with the strength of preferences. To transform the values in this matrix into weights, computer software or more pragmatic approaches can be used. When using AHP to determine performance scores this is most often done post-hoc. Stakeholders are presented an overview of information on alternatives, e.g., information on both the integrated care programme and the usual care programme. The stakeholder is, like for the criteria weight-elicitation, asked to judge which alternative, given their performance on separate criteria, is preferred and to what extent (on the same 9-point scale). All alternatives are compared (in this example just 2) on each criterion separately. Again a matrix with these strengths of preferences is created, after which computer software or pragmatic methods can be used to calculate so-called 'local priorities', i.e., which alternative is preferred and to what extent, on each specific criterion. These local priorities are transferred to global priorities by combining them with the criteria weights (for example also elicited through a separate AHP weight-elicitation procedure). (Dolan, 2010; Thokala & Duenas, 2012) Both the use of AHP for weight-elicitation and performance scoring will be described more extensively below with numerical examples (Ch. 4.4 and Interlude, respectively).

Besides the value-based and outranking methods, another category of MCDA methods are the **reference level** or **goal programming** methods. In essence, these methods compare alternatives by calculating the weighted deviations from a priori set goals. This requires a specification of desirable levels of performance for each criterion. Because the integrated care programmes in SELFIE have not specified the desired level of performance in detail and because we anticipate that it is difficult to specific goals for all indicators that we are planning to use in SELFIE, this method will not be used and is not described further. (Thokala & Duenas, 2012)

It is important to note that MCDA methods do not offer a mathematical solution to a difficult decision-making process. MCDA is primarily about structuring the decision to be made and supporting the appraisal process of, in this case, integrated care versus usual care or a control situation. By making the importance of different criteria that play a role in decision-making more explicit, stakeholders become more aware of their preferences and may better structure their argumentation in a deliberate appraisal process. This also underlines the importance of the first two and the last two steps in an MCDA, described in the next chapter: establishing the decision-making process, identifying and structuring criteria, sensitivity analyses and examining results.

# 4. Steps to be taken in MCDA

Conducting an MCDA generally involves the seven steps described in this chapter (also see **Table 2**). (Department for Communities and Local Government: London, 2009; Marsh et al., 2016) Regardless of the specific MCDA method chosen, the first two steps in an MCDA are always the same: (1) establishing the decision-context and alternatives and (2) identifying criteria that are relevant to inform the decision. (Thokala & Duenas, 2012) Steps 3-5 taken hereafter depend on which MCDA method is selected. Step 3 and 4 are described mostly from the perspective of applying MAUT. For step 5, creating a global score, we specify how this is done per overall MCDA method. Lastly, steps 6 and 7, which can be uniformly explained across different MCDA methods, are described. Each step will be described and then discussed in the context of the SELFIE project.

## 4.1. Step 1: Establishing the decision-context and identifying alternatives

It is generally acknowledged that an MCDA should start with a scoping process in which the topics described in bullet points below are clarified.

I. It is important that there is a shared understanding of the decision context.

II. Aims need to be clear.

III. The goal of the MCDA needs to be clear: is the objective to value or rank alternatives, is this a 'one-off' MCDA, or does it need to be reusable?

IV. Stakeholders need to be identified.

V. Determine when, how, and which MCDA method will be used and what practical constraints are.

VI. Describe and clarify the context of appraisal (how, by who, when will this be done).

VII. List the set of alternatives to be considered.

**Application in SELFIE**

**I.** For most of the 17 case studies in SELFIE, the decision pertains to reimbursement, continuation, extension, and/or wider implementation of an integrated care programme. The macro level country descriptions prepared in WP2 provide information on the broader decision context. The thick descriptions prepared in WP2 provide more specific information on the programmes and their unique decision context.

**II.** The overall aims across integrated care for multi-morbidity can be coined as improving the Triple Aim, i.e., improving population health/well-being, experience with care, and reducing costs. For each aim multiple criteria can be defined (see **Figure 1**). In SELFIE we have chosen to define a core set of concepts, which in the realm of the evaluation will be called 'criteria'. The core set of criteria will be used in each case study to measure the performance of the programmes. However, different indicators may be used to assess a particular criterion. The reasons behind the selection of the core set are explained in **WP4 Deliverable Report 1**, but one of them is the relevance across all programmes in SELFIE. In addition to the core set, we

have specified programme-specific criteria. These programme-specific criteria were standardised across four types of programmes: population health management programmes, programmes targeting frail elderly, palliative and oncological programmes, and programmes targeting persons with problems in multiple life domains. The way criteria are identified and structured is explained in step 2 below. The criteria lists are described more extensively in **WP4 Deliverable Report 1** and are presented in **Appendix 2**.

**III.** The goal in SELFIE is to conduct broad evaluations of the 17 programmes using MCDA. This should support the deliberate decision-making process on reimbursement, continuation, extension, and/or wider implementation. The goal is to **value** the alternatives, i.e., integrated care programmes. There will be 17 case studies, in which each of the 17 programmes will be compared to usual care or a control situation. The primary aim is not to directly compare the 17 programmes to one-another or rank them, because their aims, target populations, interventions, and context differ from one-another.

Furthermore, the aim is that the MCDA is **reusable**. We plan to be able to use the MCDA and the weights again if more programmes are added and to create an online tool (WP5-task 5) that others can use to monitor integrated care programmes. This reusability aim has driven us to develop a core set of criteria across the 17 evaluations.

**IV.** The stakeholders whose value judgements are considered relevant to the decision problem (e.g., reimbursement, continuation, wider implementation) are representatives of the 5P's: Patients, Partners, Professionals, Payers, and Policy makers. This will include stakeholders that are both directly and indirectly involved in the specific promising programmes we are working with in SELFIE. These stakeholders will be involved in the weight-elicitation process of the SELFIE MCDA. See **Chapter 6** for information as to how these stakeholders will be involved.

**V.** Commonly used MCDA methods were described above. Because one of the aims of SELFIE is to contribute to the development of methodology for the evaluation of integrated care for multi-morbidity, multiple methods will be applied and compared.

There are two main types of practical constraints that need to be considered in developing the SELFIE MCDA framework. First, practical constraints need to be considered related to the SELFIE overall project timeline and budget, as described in the grant agreement. Second, there are practical constraints that come forth from the collaboration with the programmes. Namely, the indicators that can be collected depend on what is already being collected in a programme and what possibilities there are to set up new data collection. Furthermore, issues related to study design also depend on the collaboration with the programme: how often can criteria be measured, and is it going to be possible to create a control group or compare the programme to usual care? These issues have been addressed throughout WP4 and a planned study design per case study (programme vs control) is described in **Chapter 8**.

**VI.** Each partner in SELFIE will conduct MCDAs for the case studies in their own country, for the most part using a common methodology. The final synthesis of information will be done at the end of the evaluation period when performance scores have been collected and criteria weights have been elicited. The SELFIE [Inter]national Stakeholder Advisory Boards with representatives from the 5Ps will advise us during this process and help us interpret the MCDAs.

**VII.** The list of alternatives has been created in WP1, and consists of at least 2 integrated care programmes per partner country (see **Appendix 1**). The comparator / control per integrated care programme will be addressed in WP4 (task 5) (see **Chapters 7-8** of this report) and WP5 (task 1).

## 4.2. Step 2: Identify and structure criteria

Decision criteria are the measures of performance of the programme that are considered relevant for the decision. (Marsh et al., 2016) When identifying, structuring, and defining these criteria the following points are important to take into account.

I.  Identify criteria; the measures of performance by which the alternatives will be judged. Ask yourself, 'what would distinguish a good alternatives from a bad one'? Use sources such as documents used in previous decision-making, study stakeholder's priorities, and use guidelines. Create a 'long list' of criteria. (Dolan, 2010)

II. It can be useful to group criteria, for example, using a value tree (see **Figure 1**). This way the aims can be decomposed into sub-aims and criteria. There are two options to consider:

    a.  Top-down (value-focused thinking): in this option the fundamental aims are the starting point, which are then decomposed into sub-aims by asking 'how do we achieve this'? A downside to this approach is that it may result in criteria that are quite general, which makes it difficult to relate these to each particular alternative.

    b.  Bottom-up: in this option, characteristics are identified that distinguish alternatives from one-another, these are then grouped to form higher level aims. A downside to this approach is that the criteria may be too specific to apply across different alternatives. (Marsh et al., 2016)

    Grouping can also be useful in large MCDA applications, as weights can first be determined for sets of criteria, thereafter within these sets for specific criteria.

III. Typically 3-20 criteria are used (mean circa 8). (Marsh et al., 2014) The upside of having a complete list of criteria is greater validity, however, a downside is a reduced validity in performance measuring and weighting due to practical infeasibility (e.g., missing data) and cognitive burden, respectively. It is good practice to have as few criteria as is necessary and essential for making a well-founded decision. (Marsh et al., 2016)

    It is thus important to move from a long-list to a short-list (e.g., core set), and ensure:

a. **Completeness**: have all important criteria been included?

b. **Non-overlap (i.e., no double counting)**: it is important to have criteria with as little overlap as possible, otherwise this can give too much weight to an overall dimension. However, this should not be confused with correlations – criteria can measure separate dimensions and still be correlated (there is nothing wrong with this).

c. **Preferential independence (i.e., mutual independence)**: when applying an additive model to calculate a weighted aggregate score (in MAUT), it is necessary that the weight on one criterion can be elicited independently from the performance score on another criterion. An example of dependence: health gain and disease severity – a stakeholder may not be able to weigh health gain without knowing the baseline health status (i.e., disease severity). Because in a DCE there is the possibility to model interactions, this assumption is less strict for this weight-elicitation method. However, sufficient statistical power is needed to do so.

d. **Operational**: can we in practice measure how well an alternative performs on the criteria?

    i. Fundamental outcomes are preferred over proxies, for example: including an outcome such as number of fractures due to falls is preferred over bone-mineral-density.

    ii. It is easier to operationalise criteria in absolute scales as opposed to change estimates, as preferences for change estimates always require having baseline information.

    iii. Using original, and widely accepted, indicators of criteria is preferred above self-constructed scales only for the purpose of the current MCDA.

MCDA methods are capable of combining different types of performance measures like quantitative scales, based on objective and subjective measurements (e.g., patient-reported outcomes) and qualitative scales.

**Application in SELFIE**

**I.** The identification of criteria forms the work in WP4 Deliverable Report 1. Four sources were used to create the long-list of criteria: literature review, national stakeholder workshops in the eight SELFIE countries, focus groups with persons with multi-morbidity in these countries, and lists of what is already being measured in the programmes.

**II.** Because in SELFIE we aim to develop a reusable MCDA model, we have started with a top-down approach – namely, in essence each programme's objective is to improve the Triple Aim. Subsequently, criteria were identified and defined. Per criteria, indicators were proposed and where indicators were already being measured these were mapped onto criteria (see **WP4 Deliverable Report 1** for more details).

**III.** The long-list was shortened into a core set; this is described in more detail in **WP4 Deliverable Report 1**. Ideally, all 17 case studies will use the same core set of criteria, but may apply different indicators to measure these. For each case study additional programme-type specific criteria will also be incorporated into the MCDA. Specifically, programmes have been divided into one of four types: population health management programmes, programmes targeting frail elderly, palliative or oncological care programmes, and programmes targeting persons with problems in multiple life domains. See **Appendix 2** for an overview of the criteria.

Although the primary aim in SELFIE is not to compare the 17 integrated care programmes amongst one-another, we have decided to work with a core set of criteria across the 17 case studies, because we aim to develop a reusable MCDA model. Weights will be elicited for the same set of criteria among the 5Ps in each SELFIE partner country. This will allow us to compare weights given to the core set not only across the 5P stakeholder groups, but also across countries. By standardising the MCDAs across the 17 case studies, weights are linked to higher level criteria and aims as opposed to specific indicators or even instruments. This will benefit the reusability of the SELFIE MCDA and the uptake of the MCDA tool (WP5-task 5): future programmes can map their measured indicators onto the SELFIE core set of criteria. Although there are advantages to working with such a core set, we need to be cautious when

mapping indicators to criteria, and thus to weights, to ensure that these are conceptually linked.

A SELFIE questionnaire was developed with indicators per criteria proposed (**WP4 Deliverable Report 1**). Per case study indicators are mapped onto the SELFIE core and programme-type specific sets, to determine what will be included in the MCDA. For performance indicators, absolute values and change-estimates will be looked at; for weight-elicitation also both options are possible. For the weight-elicitation, the literature points out that it is complex to weigh criteria that are expressed in terms of change estimates, we will thus elicit weights for criteria defined in absolute scales in SELFIE.

## 4.3. Step 3: Determining performance scores

Once the criteria are defined, the performance of the alternatives needs to be measured and made uniform. We describe this for when performance and weights are determined *separately* from one-another.

- Measuring performance per criteria can be based on: expert opinion, literature, empirical evaluation, modelling, or administrative data.
- Report performance, for example in a performance matrix (see **Table 3**). This pertains to 'natural' performance measures per criteria per alternative – these are reported in natural, and thus varying, scales.
- Criteria can be included and measured on different types of scales, e.g.:
  a. **Categorical** (i.e., nominal): Mutually exclusive but not ordered categories. E.g., gender.
  b. **Ordinal**: The order matters, but not the difference between values. E.g., a team climate score of 5 is better than that of a score of 4 and of 2, but the difference between the 5 and 4 may not be the same as between 3 and 2. Only order is expressed. This can also be expressed for example in 'high', 'medium', and 'low'.
  c. **Interval**: The difference between two values is meaningful. E.g., the difference between a temperature of 10 degrees Celsius and 20 degrees Celsius is the same as between 20 and 30 degrees Celsius.
  d. **Ratio**: has all the properties of an interval variable, but also has a clear definition of 0.0 = 'none' of that variable. E.g., height and weight.
- Once performance measures have been collected on their natural ranges and scales, these need to be transformed to standardised scores that are all on the same scale (e.g., 0-100, 0-1). The advantage of a 0-100 score is that it does not contain decimals, and thus may avoid confusion with probabilities.
- For standardising it is important to report what the end-points (e.g., 0 and 100) mean – this impacts interpretation and weighting (step 4).

- Using criteria for multiple decisions, and having a reusable MCDA, requires that the range of a scale covers best and worst performance that could realistically occur (i.e., global, fixed scale), as opposed to when making one decision (i.e., local, relative scale).
- There are two commonly used methods of standardising performance scores, 1) relative standardisation, and 2) ranging standardisation.

### Application in SELFIE

Performance on criteria will be assessed, by means of an empirical evaluation (WP5) in SELFIE. The study design for the empirical evaluation will be determined per programme, by the partner country in close collaboration with the EUR, UiB, and the programme itself. Partners received information on possible study designs, presented in **Chapter 7** of the current report. Where possible, the performance of a programme on the criteria will be repeatedly (≥2) assessed and compared to a control group. Throughout WP4 discussions have been held with the partners to determine which type of design best suits each programme. A summary of the proposed design per case study is presented in **Chapter 8** of the current report. This will be elaborated on in WP5.

In SELFIE, mostly we will be using categorical, ordinal, and interval scales. Specific methods of standardisation are described in **Appendix 3**. In the remainder of this report we use the example presented in **Table 3**, standardised using relative standardisation methods, as presented in **Table 5** below. The following formula describes the standardisation:

$$S_{aj} = \frac{x_{aj}}{\left(x_{aj}^2 + x_{bj}^2\right)^{1/2}}$$

$x$ = performance score (on the natural range/scale)

$a$ = alternative a

$b$ = alternative b

$j$ = criteria j

**Table 5**: *SELFIE example of relative standardising performance scores*

| | Alternatives | |
| --- | --- | --- |
| | *Integrated care* | *Usual care* |
| Criterion – *(Indicator a-f range)* | **Performance** : **Standardised** | |
| Resilience – *(0-100)* | 70 : **0.74** | 63 : **0.67** |
| Physical functioning – *(0-100)* | 60.50 : **0.68** | 65.00 : **0.73** |
| Person-centeredness – *(1-5)* | 4 : **0.80** | 3 : **0.60** |
| Continuity of care – *(1-7)* | 5.5 : **0.77** | 4.5 : **0.63** |
| Costs – (*health care costs – € per capita per year*)* | 9200 : 0.72 = **0.28** | 9000 : 0.70 = **0.30** |
| Costs – (*social care costs – € per capita per month*)* | 1500 : 0.76 = **0.24** | 1300 : 0.66 = **0.34** |

*Note: *Because for costs higher values are worse, 1 minus the standardised performance score is taken.*

## 4.4. Step 4: Weight-elicitation

Below different forms of weight-elicitation are described and examples are given. Again, this pertains to the approach whereby performance and weights are determined *separately* from one-another. Important with weight-elicitation is that the descriptions used to elicit weights conceptually match the way in which performance scoring of criteria is done.

### 4.4.1. Direct ranking

This is a very simple weighting method in which criteria are ranked (in our example, from 1-6). These rank orders are then converted into weights [and numerical scales] using rank order centroids. A rank order centroid is a value that estimates the distance between adjacent ranks on a normalised scale from 0 to 1. Rank order centroids can be calculated as follows: (Dolan, 2010; Byeong, 2011)

$$W_i = (1/M)\sum_{n=i}^{M}(1/n)$$

     $W_i$ = weight for the $i$th ranked criterion
     $M$ = total number of criteria
     $n$ = rank order of the criterion

Weights on all criteria taken together will add up to 1. A benefit to using ranking is that discrete weights are assigned to criteria, the only input needed is a rank ordering which is easy to understand and accomplish, and not time consuming. A downside, however, is that even though it is decided which criteria is more important than another, the magnitude of the difference is not taken into account (e.g., a step from rank 1 to 2 may be different than from rank 2 to 4). (Dolan, 2010)

*SELFIE example*: Stakeholders would be asked to rank the 6 criteria in order of importance (from 1, most important, to 6, least important). Below this has been done, and using the rank order centroids formula weights are obtained.

1: **Resilience** = (1 + 1/2 + 1/3 + 1/4 + 1/5 + 1/6) / 6 = 0.41

2: **Continuity of care** = (1/2 + 1/3 + 1/4 + 1/5 + 1/6) / 6 = 0.24

3: **Physical functioning** = (1/3 + 1/4 + 1/5 + 1/6) / 6 = 0.16

4: **Person-centeredness** = (1/4 + 1/5 + 1/6) / 6 = 0.10

5: **Health care costs** = (1/5 + 1/6) / 6 = 0.06

6: **Social care costs** = (1/6) / 6 = 0.03

## 4.4.2. Visual Analogue Scales (VAS)

Using VAS is a direct weighting procedure in which the importance of each criterion is considered separately and directly on a scale.

*SELFIE example*: For each criterion the following scenario would be presented, e.g.: How important is it for you that in a care programme **resilience** is high? *(30 points allocated)*



**0**
Not at all
important

**100**
Extremely
important
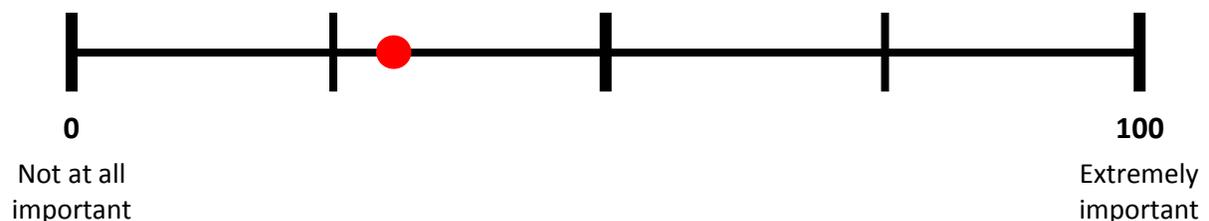
Initially, such weights across all criteria will not add up to 1 (or 100), thus they need to be rescaled afterwards, for example by taking the specific weight and dividing it by the sum of all weights given. A downside to this method is that persons might mark all criteria as quite important, which will result in little variation.

### 4.4.3. Point allocation

In this direct weighting method, points are allocated between criteria in proportion to their relative importance (e.g., 100 points to divide over *M* criteria).

*SELFIE example*: How important are the following criteria for you in a care programme? Divide 100 'importance' points among these:

- **Resilience** = 30
- **Physical functioning** = 20
- **Person-centeredness** = 15
- **Continuity of costs** = 25
- **Health care costs** = 5
- **Social care costs** = 5

As compared to the VAS method, here the points will already add up to 100. Because this is not an easy task for stakeholders, it is possible that stakeholders divide the points relatively equally over the criteria. This will again result in similar weights across all criteria.

### 4.4.4. Analytical Hierarchy Process

AHP is a pairwise comparison method that can be used both to determine criteria weights as well as performance scores (methods for the latter are described on page 32-35). The method consists of having stakeholders make pairwise comparison of the 'intensity of importance' of one criterion over another criterion, usually on a 1-9 point scale. (Department for Communities and Local Government: London, 2009)

*SELFIE example*: How important is **physical functioning** relative to **resilience**?

1 = Equally as important

3 = Moderately more important

5 = Strongly more important

7 = Very strongly more important

9 = Overwhelmingly more important

Whereby 2, 4, 6, and 8 represent 'shades' of the above judgements. This can also be shown in a figure:
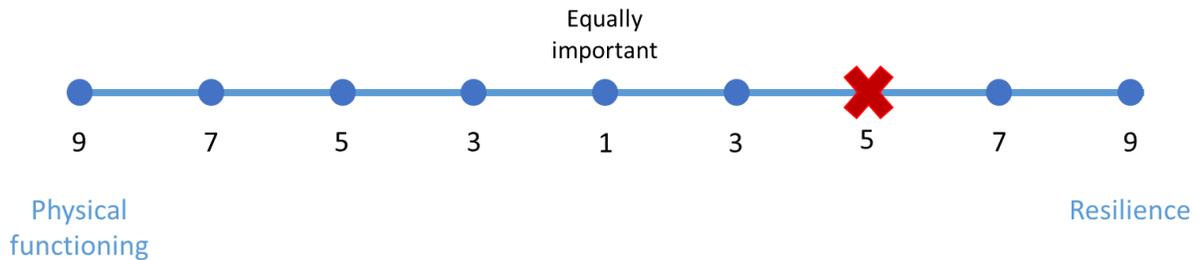


Hereby marking a 5 on this scale on the right side would mean that resilience is five times more important, and gets a score of 5, and physical functioning relative to resilience of $^{1}/_{5}$.

One of the main assumptions of AHP is that the stakeholder is consistent in making these importance judgements, and thus that only $^{1}/_{2}n(n-1)$ comparisons need to be made.

For the SELFIE example this would result in the following matrix, based on 15 pairwise comparisons of criteria: $^{1}/_{2}n(n-1) = {^{1}/_{2}}6(6-1) = 15$

**Table 6**: *SELFIE example of an AHP pairwise comparison matrix with SELFIE criteria*

| | Resilience | Physical functioning | Person-centered | Continuity of care | Health care costs | Social care costs |
|---|---|---|---|---|---|---|
| Resilience | 1 | 3 | 4 | 2 | 7 | 7 |
| Physical functioning | $^1/_3$ | 1 | 3 | $^1/_3$ | 5 | 5 |
| Person-centered | $^1/_4$ | $^1/_3$ | 1 | $^1/_5$ | 6 | 6 |
| Continuity of care | $^1/_2$ | 3 | 5 | 1 | 8 | 8 |
| Health care costs | $^1/_7$ | $^1/_5$ | $^1/_6$ | $^1/_8$ | 1 | 1 |
| Social care costs | $^1/_7$ | $^1/_5$ | $^1/_6$ | $^1/_8$ | 1 | 1 |

A set of weights is estimated that is most consistent with the relativities presented in the matrix (i.e., **Table 6**). Although there is consistency in reciprocal judgements, this is not guaranteed for judgements *between* pairs. Consistency needs to be checked using the Consistency Index (calculated using the maximal eigenvalue and dimensions of the matrix), with 0 representing complete consistency. Next, a consistency ratio is calculated which is the Consistency Index divided by the random index (i.e., average consistency index of 500 reciprocal matrices filled with values from the fundamental scale of 1-9). This ratio should be below the threshold of 0.10, which would mean 10% inconsistency compared to the average inconsistency of a random reciprocal matrix. (Rietkötter, 2016)

AHP computer packages can be used to calculate weights that best fit the observed matrix. (Department for Communities and Local Government: London, 2009) Another, more pragmatic, method to calculate the weights on the basis of the matrix is by calculating the geometric mean of each row, total these means, and normalise each geometric mean by dividing by the total (see **Table 7**). This will result in a total weight of 1.

**Table 7**: *SELFIE example of weights corresponding to AHP matrix – geometric mean method*

| | | Geometric mean (GM) | GM/Sum = Weight |
|---|---|---|---|
| Resilience | $(1 \times 3 \times 4 \times 2 \times 7 \times 7)^{1/6}$ | 3.25 | 0.34 |
| Physical functioning | $(^1/_3 \times 1 \times 3 \times {}^1/_3 \times 5 \times 5)^{1/6}$ | 1.42 | 0.15 |
| Person-centeredness | $(^1/_4 \times {}^1/_3 \times 1 \times {}^1/_5 \times 6 \times 6)^{1/6}$ | 0.92 | 0.10 |
| Continuity of care | $(^1/_2 \times 3 \times 5 \times 1 \times 8 \times 8)^{1/6}$ | 2.80 | 0.29 |
| Health care costs | $(^1/_7 \times {}^1/_5 \times {}^1/_6 \times {}^1/_8 \times 1 \times 1)^{1/6}$ | 0.93 | 0.10 |
| Social care costs | $(^1/_7 \times {}^1/_5 \times {}^1/_6 \times {}^1/_8 \times 1 \times 1)^{1/6}$ | 0.29 | 0.03 |
| | **SUM** | 9.61 | 1.00 |

In contrast to the other weighting methods described above, in AHP and the further weight-elicitation methods described below, participants are required to make a trade-off between criteria.

An advantage to eliciting weights using AHP is that clusters of criteria (e.g., to sub-aims) can allow for a series of small sets of pairwise comparisons to be done and this may be pragmatic (less time needed to do so). Furthermore, the method of making pairwise comparisons is usually well accepted in practice. (Department for Communities and Local Government: London, 2009) The matrix also allows for cross checking. Main disadvantages of the method, however, have also been recognised, which are (Department for Communities and Local Government: London, 2009):

- The 1-9 scale cannot be internally consistent: *x* may be scored 3 in relation to *y*, and y 5 in relation to *z*. But the 1-9 scale means that a consistent ranking of *x* relative to *y* (requiring a score of 15) is impossible.
- AHP does not require that preferences be transitive. As a result, AHP is subject to rank reversal.
- The labelled descriptions per point on the 1-9 scale and the links therein have no theoretical foundation.
- Introducing new alternatives can change the relative ranking of some of the original alternatives ('rank reversal').

## 4.4.5. Swing weighting

Swing weighting is a trade-off weighting method, in which the relative importance is determined on the basis of moving from the worst to best score on a scale (full swing). **SMARTS** (Simple Multi-Attribute Rating Techniques with Swing) and **SMARTER** (SMART Exploiting Ranks) are swing weighting techniques. These two methods extend on the more basic SMART weighting technique, which is only based on ranking and not on a 'swing'. In SMART, sequentially, the lowest ranked criterion is eliminated and ranking is done again. SMART is more similar to direct ranking, however, a check is done to determine whether judgements are consistent because ranking is done repeatedly as each lowest ranked criterion is eliminated. (Edwards & Barron, 1994; Thokala et al., 2016; Marsh et al., 2016; Department of Communities and Local Government, 2009)

SMARTS & SMARTER further improve the SMART procedure by ensuring that the derived criteria rankings incorporate the full range of differences among alternatives. The underlying point here is that a criterion that captures more differences among alternatives, should play a greater role. A criterion where there is a 10-fold difference between the best and worst alternative, should have a higher priority than one whether the difference is 2-fold. The advantage of using swing weighting with SMARTS and SMARTER is that actual performance scores or the range therein can be used to elicit weights.

*SELFIE example - SMARTER*: Stakeholders would receive the following description: An integrated care programme for persons with multi-morbidity has been followed up on for one year. Their performance on six criteria was monitored and their outcomes are presented in **Table 8** below.

**Table 8:** *SELFIE example of worst performance information (swing weighting example)*

|  | Performance |
|---|---|
| Resilience | Low |
| Physical functioning | Low |
| Person-centeredness | Low |
| Continuity of care | Low |
| Health care costs | High |
| Social care costs | High |

For this example, the performance is presented in general terms. Other possibilities are to use more quantitative information, for example report-card grades/numbers can be presented.

Next, the stakeholder is asked: If you could change the performance of this care programme on only one of these criteria from the worst performance to the best (presented in **Table 9** below), which would change?

**Table 9:** *SELFIE example of best performance information (swing weighting example)*

|  | Performance |
|---|---|
| Resilience | High |
| Physical functioning | High |
| Person-centeredness | High |
| Continuity of care | High |
| Health care costs | Low |
| Social care costs | Low |

If, for example, the stakeholder decides that it is most important to him/her that the resilience is high as opposed to low, resilience would receive a rank of 1. The exercise is then continued for the remaining criteria.

An important point to note is that swing weighting can be done when empirical performance has already been obtained on alternatives and thus may use the true observed best and worst scores on criteria to create fictive alternatives for the stakeholder to judge. However, these best and worst values can also be determined a priori (e.g., using the literature).

In SMARTS and SMARTER there are two important assumptions (Edwards & Barron, 1994). The first is that of linear functions, whereby 'more' or 'higher' scores pertain to better, and 'less' or 'lower' to worse, or reversed. When it is clear that the middle point is neither best nor worse, the linear approximation assumption holds up. The second important assumption is that of 'conditional monotonicity', which is needed for additive models, this is not held up when at one level of x, more of y is better than less, while at another level of x, less of y is better than more. (Edwards & Barron, 1994) This is also related to the preferential independence precondition for criteria in MCDA described in **Chapter 4.2**. Throughout the development of the SELFIE criteria this requirement was taken into consideration (see WP4 Deliverable Report 1).

In swing weighting there are 2 steps: 1) rank order of the weights, and 2) yielding the weights themselves. The way in which step 1 occurs differs between SMARTS and SMARTER. In step 1, there should be a proposition of the worst possible scenario whereby just one criteria can be changed to the best, this is then done consecutively until there are no criteria left. In SMARTS, direct magnitude estimates are used for this. In this method, the most important criteria that is changed from worst to best, i.e., the one that is changed first, is worth a full 100 points. This criteria becomes the standard to which all other criteria are compared (Department of Communities and Local Government, 2009). Next, the respondent is asked on a scale from 0 to 100, what the weight of a full swing on the second most important criteria would be. If this would be given a score of 50, this would mean that it is half as important as the first criteria. This is done for all criteria, where after the 0-100 scores are normalised into weights (step 2). (Edwards & Barron, 1994)

SMARTER is seen as the shorter and more easily applicable version of SMARTS as it is less demanding for the respondent (Department of Communities and Local Government, 2009). In SMARTER, the stakeholder is also asked to pretend there is an alternative (e.g., care programme) that has the lowest possible score on all criteria. However, in SMARTER no points are given. Instead there is only the ranking of which criteria would be selected first to swing from the worst to the best level (step 1). These ranks are then turned into weights using, for example, the rank ordered centroid method (step 2). A downside to SMARTER as compared to

SMARTS, is the assumption that the difference in ranking is equal between various criteria, i.e., the step from 1st to 2nd is the same as from 6th to 7th. With SMARTS, the relative difference in ranking is taken into consideration with the 100 point scoring.

### 4.4.6. DCE

In a DCE, choice sets with scenarios are presented, usually around 10 per stakeholder (i.e., respondent). The scenarios consist of various alternatives (e.g., care programmes) that systematically differ on criteria performance (i.e., attributes). Stakeholders are asked which scenario they prefer. Weights for each criterion can be statistically derived on the basis of the likelihood that one scenario, with specific criteria performance, is preferred over another. (Marsh et al., 2016; Lanscar & Louviere, 2008)

As with the other weighting methods, for DCE, it is very important that the descriptions used in the choice sets are clear and can link back to the actual performance.

The number of choice sets that need to be presented to stakeholders depends on which type of econometric model will be estimated, how many criteria will be studied, what the attribute levels herein are, and what their properties are (e.g., nominal, interval). In DCEs there are two design possibilities, i.e., full factorial and fractional design. In full factorial designs, stakeholders judge all choice sets (i.e., all combinations of the different attribute levels of the criteria). If for example, in SELFIE we have 10 criteria in the core set with 5 criteria with 2 levels and 5 with 3 levels, this will already result in 7776 possible alternatives ($2^5$ X $3^5$). Subsequently combing these alternatives into choice sets would result in an extremely large number of choice sets that stakeholders would need to judge. It would be impossible to have stakeholders judge this many choice sets. The fractional design presents a sub-set of choice sets to each stakeholder. A disadvantage to this design is that interaction effects between criteria cannot all be calculated, however, the increased efficiency of such a design and pragmatic/feasibility benefits hereof largely outweigh this downside. Within the fractional design there are various specific sub-designs (e.g., orthogonal, Bayesian efficient).

*SELFIE example*: Stakeholders would be asked: which programme do you prefer?

**Table 10**: *SELFIE example of a DCE scenario*

| Criteria | Programme A | Programme B |
|---|---|---|
| Resilience | Average | High |
| Physical functioning | High | Average |
| Person-centeredness | High | High |
| Continuity of care | High | High |
| Health care costs | Low | High |
| Social care costs | High | Average |
| **Selected programme:** | X | |

*Note: These colours are presented for the purpose of this example; there is some debate as to whether this should be done in DCE's.*

As mentioned, it is very important for all weight-elicitation methods to ensure that the descriptions used (e.g., in the DCE choice sets) are clear and link back to the indicators that are used to measure a programmes' performance. When using criteria that are difficult to quantify, elicitation relies heavily on the descriptions given in the questionnaire and even then it remains a subjective approach. A more quantitative approach to defining criteria (e.g., x points on a scale from 1-10) is less reliant on the descriptions, and makes criteria less ambiguous. In SELFIE, we need to ensure that the same descriptions can be interpreted by all 5P stakeholders. The levels in the current example are quite generic; more specific levels, as to be applied in SELFIE, will be presented in **Chapter 6**.

## 4.5. Step 5: Creating a global score

The objective of this step is to create one global score per alternative. For MAUT and outranking methods, this is done by combining the standardised performance scores (obtained in step 3) with the weights (obtained in step 4). Hereafter AHP is described, a method by which preferences for overall alternatives and weights can be derived.

### 4.5.1. Combining performance and weights with MAUT

The most common method for combining weights and performance is the 'Weighted Sum Approach', or 'Additive Aggregation', in which the sum is taken of each criterion's performance multiplied by its weight. (Thokala & Duenas, 2012; Marsh et al., 2016)

$$W_a = \sum_{n=1}^{M} (W_{ai}) \times (P_{ai})$$

$W_a$ = Overall weight of alternative $a$

$M$ = Criterion

$W_{ai}$ = Weight (of alternative $a$) on criteria $i$

$P_{ai}$ = Performance of alternative $a$ on criteria $i$

An important assumption to this method is that of preferential independence. Preferential independence means that the weight on one criterion can be elicited independently from the performance score on another criterion. It implies that decisions can be made by using only the subset of criteria on which the alternatives differ. An emergent property is that importance weights should be based on trade-offs between criteria, i.e., weights should represent the gain in overall value from replacing the worst performance score with the best performance score on a criterion. The ratio of two weights (e.g., $W_i/W_k$) should indicate the change in standardised performance score of criteria $k$ that is required to compensate for a unit loss in standardised performance score of criteria $i$. For example, if $i$ = resilience and $k$ = physical functioning, and

these have weights 0.3 and 0.2, respectively (see **Table 3**), then the weight-ratio is 1.5 (0.3/0.2). This means that a physical functioning improvement in standardised performance score of 1.5 is required to compensate for a unit loss in resilience.

Aggregated performance for the two alternatives in the SELFIE example is presented in **Table 11** below, whereby weights and standardised performance scores are aggregated to calculate an overall performance of the integrated care and usual care programmes. In this example the relative standardised scores are used from **Table 5**. In the **Appendix 4, Table A4.1** the aggregation on the basis of global ranging scaling is shown. For the prior, stakeholder group 1 would prefer integrated above usual care (0.69 vs. 0.64), whereas stakeholder group 2 would prefer usual care above integrated (0.48 vs. 0.47). Subsequently, it is possible to determine whether the aggregated performances differ statistically significantly from one another, taking the confidence intervals of the performance scores and weights into consideration.

**Table 11**: *SELFIE example MAUT aggregating weights and performance*

| Criteria | Weight | | Care alternatives | | Aggregated weight | | | |
| | | | Integrated | Usual | Integrated | | Usual | |
| | S1 | S2 | Standardised performance* | | S1 | S2 | S1 | S2 |
|---|---|---|---|---|---|---|---|---|
| Resilience | .30 | .20 | **0.74** | **0.67** | 0.22 | 0.15 | 0.20 | 0.13 |
| Physical functioning | .20 | .15 | **0.68** | **0.73** | 0.14 | 0.10 | 0.15 | 0.11 |
| Person-centeredness | .15 | .05 | **0.80** | **0.60** | 0.12 | 0.04 | 0.09 | 0.03 |
| Continuity of care | .25 | .05 | **0.77** | **0.63** | 0.19 | 0.04 | 0.16 | 0.03 |
| Health care costs | .05 | .30 | **0.28** | **0.30** | 0.01 | 0.08 | 0.02 | 0.09 |
| Social care costs | .05 | .25 | **0.24** | **0.34** | 0.01 | 0.06 | 0.02 | 0.09 |
| | | | | **Sum** | **0.69** | **0.47** | **0.64** | **0.48** |

Note: *Standardised performance based on relative standardisation (from **Table 5**). S1 = Stakeholder 1, S2 = Stakeholder 2.

### 4.5.2. Combining performance and weights with outranking methods

Like for MAUT, to apply outranking methods, the performance of alternatives on criteria needs to be known, as well as the weights for these criteria. Weights can be elicited with any of the

weight-elicitation methods described above. (Department for Communities and Local Government: London, 2009; Thokala & Duenas, 2012)

Outranking methods start with a pairwise comparison of the performance of all alternatives on all criteria in order to construct a concordance and discordance matrix. The concordance matrix shows the concordance index of every pairwise comparison between alternatives. The concordance index of alternative A versus B is based on the criteria on which A scores better than B, i.e., outranks B. The discordance index of A over B is based on the criteria on which A scores worse, or is outranked, by B. Sometimes an indifference threshold is determined, saying that performance can only be said to differ between alternatives A and B if the difference is larger than *x* (this is taking imprecision into account).

There are different ways to quantify the concordance and discordance indices. In ELECTRE, the concordance index is defined as the ratio of the sum of weights of the criteria for which alternative A is at least as good as alternative B, to the sum of weights in all criteria. The discordance index is calculated as follows: for each criterion where A is outranked by B, the ratio of the difference between the performance scores of A and B to the maximum difference in performance score among all alternative is calculated. The highest ratio found is the discordance index of A versus B. (Thokala & Duenas, 2012)

$$C_{(a,b)} = (\sum_{n=1}^{M} W_d)/(\sum_{n=1}^{M} W)$$

$C_{(a,b)}$ = Concordance index of alternative A over B

$M$ = Total amount of criteria

$W_d$ = Weights of criterion for which alternative A dominates B

$W$ = Weights

$$D_{(a,b)} = \max\{(Va - Vb)/(V\max - V\min)\} \text{ for all criterion where } V_x > V_y$$

D$_{(a,b)}$ = Discordance index of alternative A over B

$V_a$ = Value (i.e., performance score) of alternative A

$V_b$ = Value (i.e., performance score of alternative B

$V_{max}$ = Highest value on a specific criterion across all alternatives (thus only relevant when >2 alternatives are being ranked)

$V_{min}$ = Lowest value on a specific criteria across all alternatives.

Both the concordance and the discordance indices are compared against index thresholds. When the concordance index is greater than the concordance threshold, and the discordance index is less than the discordance threshold, the alternative is said to outrank the others. (Thokala & Duenas, 2012)

In **Table 12** a *SELFIE example* is provided, in which the indifference threshold is set at $\geq$0.05. Since health care costs do not meet this threshold, they are not included in the subsequent analyses.

**Table 12**: *SELFIE example of performance values and weights for outranking (a vs. b vs. c)*

| Criteria | Weights | Integrated care – 1 (a) | Usual care (b) | Integrated care – 2 (c) | $V_{max} - V_{min}$ | A vs. B | C vs. B | A vs. C |
|---|---|---|---|---|---|---|---|---|
| Resilience | .30 | **0.74** | **0.67** | **0.80** | *0.13* | A | C | C |
| Physical functioning | .20 | **0.68** | **0.73** | **0.75** | *0.07* | B | - | C |
| Person-centeredness | .15 | **0.80** | **0.60** | **0.50** | *0.30* | A | B | A |
| Continuity of care | .25 | **0.77** | **0.63** | **0.70** | *0.14* | A | C | A |
| Health care costs | .05 | **0.28** | **0.30** | **0.32** | *0.04* | - | - | - |
| Social care costs | .05 | **0.24** | **0.34** | **0.30** | *0.10* | B | - | C |

Note: Weights are those from Stakeholder 1 (S1) perspective from **Table 3**. Indifference threshold set at $\geq 0.05$, thus health care costs are not included in the analyses. The maximum difference presented in column $V_{max}$-$V_{min}$ is used as the denominator in the Discordance Index calculation.

$C_{(a,b)}$ = (0.30+0.15+0.25)/1 = 0.70
$D_{(a,b)}$ = max { (0.73-0.68)/*0.07* or (0.34-0.24)/*0.10* } = max { (0.71) (1.00) } = 1.00
$C_{(b,a)}$ = (0.20+0.05)/1 = 0.25
$D_{(b,a)}$ = max { (0.74-0.67)/*0.13* or (0.80-0.60)/*0.30* } = max { (0.07) (0.20) } = 0.20


$C_{(c,b)}$ = (0.30+0.25)/1=0.55
$D_{(c,b)}$ = max { (0.60-0.50)/*0.30* } = max { (0.40) } = 0.40
$C_{(b,c)}$ = (0.20+0.05)/1=0.25
$D_{(b,c)}$ = max { (0.80-0.67)/*0.13* or (0.70-0.63)/*0.14* } = max { (1.00) (0.50) } = 1.00


$C_{(a,c)}$ = (0.15+0.25)/1=0.40
$D_{(a,c)}$ = max { (0.80-0.74)/*0.13* or (0.75-0.68)/*0.07* or (0.30-0.24)/*0.10*) } = max { (0.46) (1.00) (0.60) } = 1.00
$C_{(c,a)}$ = (0.30+0.20+0.05)/1=0.55
$D_{(c,a)}$ = max { (0.80-0.50)/*0.30* or (0.77-0.70)/*0.14* } = max { (1.00) (0.50) } = 1.00

Because the weights in this example add up to one, the denominator in the concordance equation is 1.

When there are more than two alternatives being ranked, a matrix of concordance and discordance can be made, see **Tables 13a** and **13b** (for this reason an additional integrated care programme has been added to the example (c)).

**Table 13a:** *SELFIE example of a concordance matrix*

|   | A | B | C |
|---|---|---|---|
| **A** | - | 0.70 > * | 0.40 |
| **B** | 0.25 | - | 0.25 |
| **C** | 0.55 > * | 0.55 > * | - |

**Table 13b:** *SELFIE example of a discordance matrix*

|   | A | B | C |
|---|---|---|---|
| **A** | - | 1.00 | 1.00 |
| **B** | 0.20 < * | - | 1.00 |
| **C** | 1.00 | 0.40 < * | - |

The concordance and discordance indices are compared to the concordance and discordance thresholds ($C^*$ and $D^*$). These are often calculated as the mean of all values in either matrix. In this example, $C^*$ thus equals 0.45 and $D^*$ equals 0.77. When $C_{(a,b)} > C^*$ and $D_{(a,b)} < D^*$, option A is said to dominate option B (marked in **Tables 13 a** and **b** below as > * and < *). Patterns of dominance among alternatives are studied to reach a final decision about the best alternative. Thus in this example, integrated care programme C dominates usual care programme B. There is an indication that integrated care programme C also dominates integrated care programme A. Programme A and programme B seem not to differ from one-another – as in the concordance matrix programme A seems to dominate programme B, but in the discordance matrix programme B dominates programme A.

### 4.5.3. Analytical Hierarchy Process (AHP)

Unlike the combination of performance scores and criteria weights in the methods described above, it is also possible to use Analytical Hierarchy Process (AHP) as an overall MCDA method in which performance, weights, and aggregated performance-weights for overall alternatives are determined simultaneously in the same procedure. The overall AHP method is very similar to using AHP for solely determining weights for criteria (see **Chapter 4.4.4.**). Either weights can first be assigned to criteria, where after priorities for alternatives are determined (Top-Down valuation) or vice versa (Bottom-Up valuation). In the bottom-up method, first, two alternatives are evaluated with regard to a specific criterion – the comparison is made in terms of their abilities to meet the criteria using the one to nine index. Then, two criteria are compared with respect to their importance. After the consistency ratio (see page 50) has been calculated, criteria weights and local alternative priorities are determined. The latter shows the preference for alternatives with respect to certain criteria. Similar as for MAUT, overall alternative preferences (also called global alternative preferences) are then used to calculate an aggregated preference for an alternative. (Dolan, 2010; Department for Communities and Local Government: London, 2009; Rietkötter, 2016)

*SELFIE example*: To what extent is programme A preferred / prioritised over programme B with respect to meeting the **resilience** criterion?

    1 = Equal

    3 = Moderately more preferred

    5 = Strongly more preferred

    7 = Very strongly more preferred

    9 = Overwhelmingly more preferred

Whereby 2, 4, 6, and 8 represent 'shades' of the above judgements.

This is then done for each pair of alternatives (in this case only two, A and B) for each criterion. To help stakeholders answer these question, a table might be presented such as below (**Table**

**14**), where the strength (i.e., intensity) of preference is filled in in the last column, after stakeholders have reviewed the information in the previous column. In this table the actual performance of the alternatives on each criterion is shown in their natural scale. Furthermore, a hypothetical minimally or maximally acceptable score is presented for each criterion – this may also influence how stakeholders perceive the preference. For example, if both meet this score (e.g., resilience) preferences might be less strong than when one does and the other does not (e.g., person-centeredness).

**Table 14**: *SELFIE example of an AHP comparison of alternatives on criteria*

| | Integrated care A | Usual care B | Preferred | Difference | | *Compared to min/max* | | Strength of preference (A vs. B) |
|---|---|---|---|---|---|---|---|---|
| **Criteria** | | | | Absolute | Ratio | A | B | |
| Resilience *(0-100 scale)* *Min = 60.00* | 70 | 63 | A | 7 | 1.11 | 10 | 3 | 5 |
| Physical functioning *(0-100 scale)* *Min = 55.00* | 60.50 | 65.00 | B | 4.50 | 1.07 | 5.50 | 10.0 | 2 |
| Person-centeredness *(1-5 scale)* *Min = 3.25* | 4 | 3 | A | 1 | 1.33 | 0.75 | -0.25 | 7 |
| Continuity of care *(1-7 scale)* *Min = 4* | 5.5 | 4.5 | A | 1 | 1.22 | 1.5 | 0.5 | 6 |
| Health care costs *Max = 8800* | 9200 | 9000 | B | 200 | 1.02 | -400 | -200 | 1/5 |
| Social care costs *Max = 1500* | 1500 | 1300 | B | 200 | 1.15 | 0 | -200 | 1/4 |

In order to transform the 1-9 scores into weights, again matrices (e.g., like for AHP weights in **Table 6**) are used – if there are $m$ alternatives (e.g., programmes) and $n$ criteria, then $n$ separate $m$ x $m$ matrices have to be created. The matrix of alternative A as compared to B for resilience is simply:

| | A | B |
|---|---|---|
| A | 1 | 5 |
| B | $1/5$ | 1 |

Thus it is not necessary when there are only two alternatives to create all these matrices. Usually software would be used to calculate the weights, but for the example now we again use the geometric mean method, as was done for the AHP weights in **Table 7**. The local priorities for the programmes are as follows:

**Resilience**

*Programme A vs. programme B = $(5)^{1/2}$ = 2.23 / (2.23+0.45) = 0.83*

*Programme B vs. programme A = $(1/5)^{1/2}$ = 0.45 / (0.45+2.23) = 0.17*

**Physical functioning**

*Programme A vs. programme B = $(2)^{1/2}$ = 1.41 / (1.41+0.71) = 0.67*

*Programme B vs. programme A = $(1/2)^{1/2}$ = 0.71 / (0.71+1.41) = 0.33*

**Person-centeredness**

*Programme A vs. programme B = $(7)^{1/2}$ = 2.65 / (2.65+0.38) = 0.87*

*Programme B vs. programme A = $(1/7)^{1/2}$ = 0.38 / (0.38+2.65) = 0.13*

**Continuity of care**

*Programme A vs. programme B = $(6)^{1/2}$ = 2.45 / (2.45+0.41) = 0.86*

*Programme B vs. programme A = $(1/6)^{1/2}$ = 0.41 / (0.41+2.45) = 0.14*

**Health care costs**

*Programme A vs. programme B = $(1/5)^{1/2}$ = 0.45 / (0.45+2.23) = 0.17*

*Programme B vs. programme A = $(5)^{1/2}$ = 2.23 / (2.23+0.45) = 0.83*

**Social care costs**

*Programme A vs. programme B = $(1/4)^{1/2}$ = 0.50 / (0.50+2) = 0.20*

*Programme B vs. programme A = $(4)^{1/2}$ = 2 / (2+0.50) = 0.80*

These local priorities can then be transformed to global priorities per criteria, given the criteria weight:

*Local priority (A) * Criteria weight = Global priority (A)*

*Local priority (B) * Criteria weight = Global priority (B)*

Thus each alternative gets a global priority for each criterion. These are then summed over the alternative to determine the overall preference for that particular alternative. The sum of these values across alternatives should equal 1.

**Table 15**: *SELFIE example of AHP alternative preferences, local to global priority*

|  | Local priority | | Global priority | | Criteria weight* |
|---|---|---|---|---|---|
|  | **A** | **B** | **A** | **B** | |
| Resilience | 0.83 | 0.17 | 0.30 | 0.06 | 0.36 |
| Physical functioning | 0.67 | 0.33 | 0.11 | 0.05 | 0.16 |
| Person-centeredness | 0.87 | 0.13 | 0.09 | 0.01 | 0.10 |
| Continuity of care | 0.86 | 0.14 | 0.27 | 0.04 | 0.31 |
| Health care costs | 0.17 | 0.83 | 0.005 | 0.025 | 0.03 |
| Social care costs | 0.20 | 0.80 | 0.006 | 0.024 | 0.03 |
| **Overall preferred alternative** | | | 0.78 | 0.21 | |

*Note: *These weights stem from the AHP weight-elicitation example presented in **Table 7.***

In this example alternative A is strongly preferred over B.

An advantage of the AHP method for performance and criteria weighting / prioritising, is that the natural range of a scale can be presented to the stakeholders before they answer the question as to how much programme A is preferred over programme B. Furthermore, more descriptive information on performance could also be presented to stakeholders, e.g., in the case of missing information, to allow them to make a comparison and decide their preferences.

## 4.6. Step 6: Sensitivity analyses

When examining the results, it is important to take uncertainty throughout the MCDA process into consideration.

- Levels and sources of uncertainty should thus be reported. Types of uncertainty include (Marsh et al., 2016; IJzerman et al., 2011):
  - Stochastic uncertainty:
    - Heterogeneity, such as the variability in performance in different subgroups (e.g., gender)
    - Imprecision, such as parameter uncertainty in both performance and weights
  - Structural uncertainty about the assumptions of the MCDA design, for example:
    - Disagreement on weight method
    - Disagreement on the value tree
    - Level of experts' knowledge when determining weights and/or performance

- To deal with uncertainty, either an uncertainty criteria can be included in the MCDA or sensitivity analyses can be done.
  - Uncertainty criteria: This is the risk that the benefits captured by the criterion will not be attained; one can see this as a negative penalty score. The EVIDEM framework can be used to support this.
  - Sensitivity analyses, which can be differentiated into two types:
    1. Deterministic sensitivity analyses: This analysis involves changing the value of a single parameter (e.g., performance score or weight) and keeping the values of the others constant. Multiple analyses can be done, for example with worst, best, and most likely scenarios. Benefits to this method are that it is easy to understand. Drawbacks include that only a few discrete options

are looked into, and independence between inputs (e.g., weights, performance) is assumed.

2. Probabilistic sensitivity analyses (or Monte Carlo simulation): In this method, uncertainty is modelled by including a range of possible values drawn from probability distributions. This can provide different probabilities of different outcomes occurring. There are different types of distributions that can be chosen (e.g., normal, lognormal, uniform, triangular). Probability distributions can be specified for both performance scores and criteria weights. Monte Carlo simulation can then be done by which values are sampled at random from the selected probability distribution. Countless iterations can be done to derive the probability distribution of possible outcomes. The main argument for using this method is that results show the likelihood of different outcomes.

### Application in SELFIE

There is a debate on how to include uncertainty levels into analyses. In SELFIE, it is likely that sensitivity analyses will be used instead of including an uncertainty criterion. Capturing all uncertainty into one measure is difficult and has methodological consequences (e.g., is preferential independence still guaranteed?). Instead, sensitivity analyses can be done using both deterministic and probabilistic methods. Deterministic (univariate) sensitivity analyses can for example be used to assess the impact of excluding a criterion and setting a criterion weight to the minimum or maximum value of a possible range on the MCDA results.

Probabilistic sensitivity analyses will be used to assess the impact of uncertainty in all parameters on the MCDA results simultaneously. By parameters we refer to both criteria weights and performance scores. To do this, a large number of Monte-Carlo simulations will be performed based on the predefined distribution of each parameter. In order to display uncertainty of the MCDA results, we will further explore the possibility of graphical presentation in Conditional Multi-attribute Acceptability Curves (CMACs). This is a novel method that we aim to develop in SELFIE in order to facilitate the decision-making process

based on the MCDA results. The purpose of CMACs is thus similar to that of Cost-Effectiveness Acceptability Curves (CEACs) in Cost-Utility frameworks (Al, 2013). The vertical y-axis of the CMAC displays the probability of the intervention to be accepted as the preferred alternative against the comparator. In the case of SELFIE this pertains to the likelihood that the integrated care programme has the highest global (or total) score (as calculated in Step 5, **Chapter 4.5.1.**), while the budget impact stabs below a set threshold. The x-axis displays different levels of budget available to be allocated to either intervention or comparator, for the treatment of a given population. A CMAC is illustrated in **Figure 2** below.

**Figure 2:** *Example Conditional Multi-attribute Acceptability Curve (CMAC)*
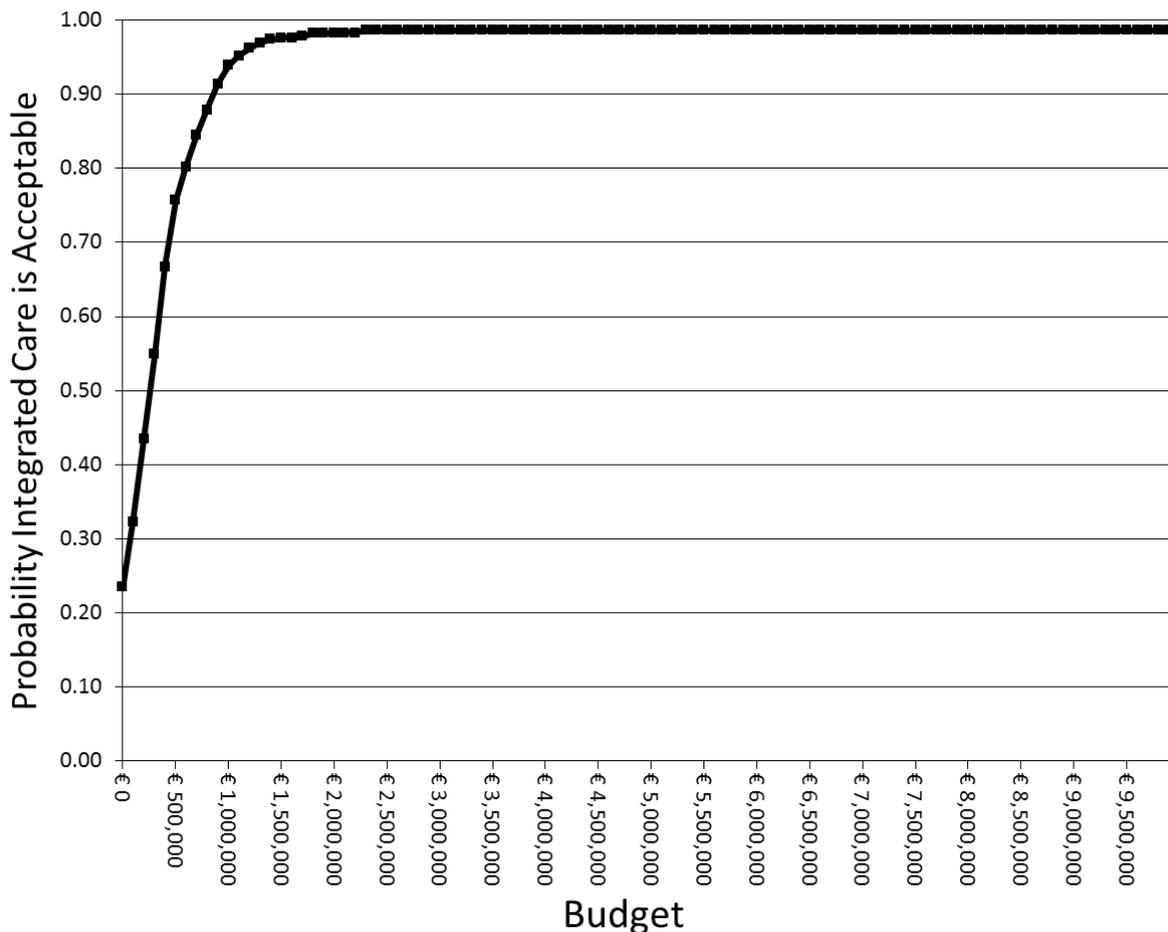
The probability of the intervention to be accepted as the preferred alternative is conditional to:

a)  the Net Global Score = the difference in Global Score (or total / aggregated and weighted performance) between the intervention and comparator

b)  the intervention's Affordability = which compares the product of mean costs per person in the intervention against a specific level of available budget

These two factors are integrated into the formula below, resulting in the probability of acceptability:

$$P = \frac{\Sigma \left( \begin{array}{l} 1 \; if \; NGS > 0 \; AND \; n \times C_A < \lambda \\ 0 \; if \; NGS < 0 \; OR \; n \times C_A > \lambda \end{array} \right)}{M}$$

$P$ = probability of intervention acceptability, plotted against the y-axis in **Figure 2**

$NGS$ = Net Global Score

$n$ = target population

$C_A$ = mean cost per person in the intervention group

$\lambda$ = level of available budget

$M$ = number of Monte-Carlo simulations

In the example illustrated in **Figure 2**, the probability that the integrated care programme is effective and affordable, is 76% at a budget of 500,000 euros and this increases to 94% at a budget of 1 million euros.

The Conditional Multi-attribute Acceptability Cure is a new concept and has the potential to inform decision-makers about the uncertainty of the MCDA results in a meaningful and accessible way. However, it still needs to be developed further and tested in the evaluation of the SELFIE programmes.

## 4.7. Step 7: Examine results

The MCDA should facilitate the decision-making process of current and future integrated care programmes by informing stakeholders who are involved about their preferences in a structured way. The methodological decisions made throughout the MCDA process should be transparent. When examining the results of an MCDA it is thus beneficial to show the outcomes of the different steps that resulted in the aggregate performance of alternatives. This demonstrates where an alternative scored well and where it did not. The weights themselves are an important finding, especially in the context of SELFIE, where it is important that those stakeholders who make the final decision about reimbursement, continuation, extension and/or wider implementation of a programme are well informed about the preferences of each of the 5Ps.

# 5. Selecting the MCDA and weighting methods in SELFIE

In order to select the overall MCDA method and weighting methods in SELFIE, the primary goals of the MCDA in SELFIE have been made explicit, as follows:

- In SELFIE we will **conduct 17 empirical evaluations** in which integrated care programmes are compared to usual care / a control group.
- The main aim of using MCDA in SELFIE is to support a **broad empirical evaluation** of these programmes by measuring the performance of a range of different outcomes, or criteria, and combining these performance scores with their relative importance weights in order to obtain an aggregated value.
- The main aim of MCDA is **not to rank these** 17 programmes, **but to value each** integrated care programme against its comparator (i.e., usual care / control group).
- The MCDA will be done **to structure and inform the deliberate decision-making process** on reimbursement, continuation, extension, and/or wider implementation of integrated care programmes.
- The MCDA supports **the transparency, consistency, accountability, credibility, and acceptability** of the decision-making because the alternatives (the integrated care programme and its comparator) are assessed on the basis of explicit aims, for which measurable criteria are established and weighted.

The following secondary goals have also been made explicit:

- To contribute to **methodological innovations** in evaluating integrated care in multi-morbidity. We can meet this aim by
  - Applying MCDA to evaluate integrated care for multi-morbidity
  - Comparing different weighting methods
  - Comparing different evaluation methods (e.g., MCDA as compared to traditional cost-effectiveness evaluations)
- To compare the **preferences of different stakeholder groups** (i.e., weights from the 5P stakeholder groups).

- To compare the **preferences** of the same stakeholder groups **in different countries** (i.e., weights from the 8 countries).
- To provide information to build a user friendly reusable **MCDA online-tool.**


As is extensively described in **WP4 Deliverable Report 1**, and in **Chapter 4.2.**, a core set of criteria will be used in all 17 programme evaluations. These criteria were defined in terms of general outcome-concepts. The core set includes: physical functioning, psychological well-being, enjoyment of life, social relationships and participation, resilience, person-centeredness, continuity of care, and total health- and social care costs. The evaluations will also include programme-type specific criteria. Specifically, programmes have been divided into one of four types: population health management programmes, programmes targeting frail elderly, palliative or oncological care programmes, and programmes targeting persons with problems in multiple life domains. See **Appendix 2** for an overview of the criteria.

Below the choices for the MCDA method and specific weight-elicitation methods that best meet the primary and secondary goals in SELFIE are presented.

## 5.1. Overall MCDA method selection

In order to address our primary goal, to conduct broad empirical and methodologically sound evaluations, in SELFIE the primary MCDA method that will be used is Multi-Attribute Utility Theory (MAUT). In this chapter we will justify this choice. The consequence hereof is that weights will be elicited in a separate process from performance scores on programmes. The choice for the weight-elicitation in SELFIE method will be described in **Chapter 5.2** below. As it is debated as to whether or not costs should be among the criteria included in MCDA, this issue is addressed in **Chapter 5.3**.

MAUT is currently the most frequently used MCDA method. (Marsh et al., 2016) An important reason to use MAUT is that it is intuitively easy for decision makers to understand. Furthermore, MAUT is well-founded in scientific theory. According to classic utility theory individuals seek to maximise utility with the available resources. Utility is 'the property in any object [...] to produce benefit, advantage, pleasure, good or happiness' or 'to prevent the happening of mischief, pain, evil or unhappiness' (Bentham, 1879). Utility theory is the theoretical basis for efficient allocation of scarce resources. Allocative efficiency is about producing those types and amounts of outputs that people value most. Preference is a common operationalisation of utility. In our SELFIE MCDA, we aim to determine the [relative] preference for the multiple criteria (i.e., attributes, outcome-concepts) of integrated care programmes. We use a wide range of criteria valued by society. We apply a multi-attribute utility function to obtain an overall value for the programmes and their comparators.

Another advantage to using MAUT is that criteria-weights can be determined in a parallel trajectory to the empirical data collection of the programmes' performance. As explained in WP Deliverable Report 1, criteria were defined in terms of general outcome-concepts grouped by the Triple Aim. Hence, the weights will link to general outcome-concepts, which allow them to be reused in future MCDAs. This can also allow for evaluations using slightly different indicators to measure the same outcome-concepts, which enhances the reusability of the MCDA framework. Using MAUT thus allows us to address the reusability goal in SELFIE, and is in many ways practically necessary. Namely, some of the programmes that are

76

being evaluated have already been collecting data for years. It would be infeasible and of less value to ask these programmes to start measuring new and different indicators when they have been using certain indicators for years. The advantage of having a longer follow-up and continue historical analyse with the same indicators outweigh the disadvantage of having to map the existing indicators to the outcome-concepts that we are including in the MCDA.

Possible challenges in using MAUT include meeting its two methodological assumptions, i.e., preferential independence and obtaining weights by trade-offs whereby weights represent the gain in overall value from replacing the worst with the best performance on a criterion (described in **Chapter 4.5**). Furthermore, the choice for MAUT implies that we have to obtain the performance scores separately from the weight-elicitation. This can be time-consuming. Fortunately, in SELFIE, we have planned to conduct extensive empirical evaluations of the integrated care programmes in order to be able to collect performance scores and/or obtain these from existing data registries. This gives us the unique opportunity to apply MAUT methods. (Dolan, 2010; Thokala et al., 2016)

Considering the strengths and limitations, and possibilities and goals of the SELFIE project, **we have chosen to use MAUT as the primary method of conducting the MCDAs.**

## 5.2. Weight-elicitation method selection

In order to address the primary goal of conducting MCDAs of the 17 programmes, using MAUT, weights will be elicited for a core set of criteria and for programme-specific types of criteria. More details on the criteria to be included in the MCDAs can be found in WP4 Deliverable Report 1. For the core set of criteria a Discrete Choice Experiment (DCE) method will be used across all eight SELFIE partner countries, amongst the 5P stakeholder groups. In order to elicit weights for the programme- type sets of criteria, swing weighting will be used. Using both DCE and swing weighting will also allow us to address our secondary goals, namely to compare weighting methods, compare viewpoints across the 5P stakeholder groups and across the eight SELFIE countries. Generating weights for the core set using DCE will result in high quality weights that will be of great value to the reusability thereof in the online MCDA tool, another secondary SELFIE goal.

The theoretically best-founded weights are those based on techniques that take account of the entire potential range of performance of alternatives. Furthermore, methods are preferred that force stakeholders to trade criteria off against one-another, as opposed to merely rating a single criterion. (Marsh et al., 2016) These conditions are best met by AHP, swing weighting and DCEs. We opted for the latter two methods and not for AHP because DCEs and swing weighting take the full range of criteria-levels into account. This is important for the future reusability of the MCDA framework, because if we would only elicit weights for the performance as actually observed in the empirical evaluations of our selected programmes, we could not apply these weights in future evaluations where the range may be different. Furthermore, as with using AHP as an overall method, rank-reversal is an issue in this method as well as the relatively arbitrary 9 point scale.

In DCE weights are obtained for specific criteria-levels (in DCE terms 'attribute levels'), like 'severely limited', 'moderately limited', and 'hardly or not at all' limited in physical functioning. DCE is the theoretically soundest value-based method founded in random utility theory (Thurstone, 1927; Manski, 1977;), as expanded upon by McFadden (McFadden, 1974) and in Lancaster's economic theory of value (Lancaster, 1966). For this reason, **for the core set,**

**weights will be elicited using DCE methods.** Each SELFIE partner country will use the same DCE to obtain weights for the core set, this will allow for comparisons to be made. The weights for the core set, defined at a conceptual level, will thus be obtained using a state of the art method, amongst 5 key stakeholders in integrated care for multi-morbidity. This will be of great value in the reusability of the MCDA.

Because in SELFIE different types of integrated care programmes are being evaluated, programme-type specific criteria sets were also developed. It is infeasible to include these additional criteria in the DCE because of the practical constraint that this would result in too many criteria/attributes. This would subsequently require an immense amount of DCE questionnaire respondents which is simply not possible from the 5P stakeholder groups we are interested in. Furthermore, this would result in many different DCEs across partner countries and programmes. For this reason the DCE for the core set will be uniform across countries, but **each country will include an additional weight-elicitation for their programme-type specific criteria. This will be done using swing weighting.**

Swing weighting is another advanced value-based method. Swing weighting, like DCE, forces respondents to make trade-offs between criteria, taking their criteria level ranges explicitly into account (Department of Communities and Local Government, 2009). A difference between DCE and swing weighting, is that in swing weighting weights are obtained for criteria as a whole and not for criteria-levels. Swing weighting also allows more criteria to be included simultaneously, tackling some practical feasibility issues. In the swing weighting exercise it is necessary that the core set and programme-specific criteria are both used in order to have a complete picture of the relative importance of the different criteria. This is also necessary to ensure that the weights all add up to 1. Further, this will allow for the weight-elicitation methods to be compared. Namely, the weights for the core set from the swing weighting exercise can be rescaled excluding the additional criteria, and compared to those derived from the DCE.

In **Chapter 4.4.5.**, swing weighting was introduced. The advantage of using swing weighting with SMARTS and SMARTER is that actual performance scores or the range therein can be used to elicit weights. In the case of SELFIE, however, we are conducting the weight-

elicitation before the performance scores are obtained. This means that in the swing weighting elicitation we will include a hypothetical range, as will also be done in defining the DCE levels. The principle in swing weighting is to determine the relative importance of criteria, considering their full range possibility; namely, their worst to their best. **In SELFIE we will work with SMARTER.** SMARTER is seen as the shorter and more easily applicable version of SMARTS; it is less demanding for the respondent (Department of Communities and Local Government, 2009). In SMARTER no points have been allocated, only a ranking of which criteria selected to swing from the worst to the best level need to be given. In SELFIE we have decided to work with SMARTER to ensure that the questionnaire does not become too extensive or complex for respondents.

In **Table 16** below we indicate which programme-type specific criteria each partner is including in the swing weighting method besides the DCE for the core set of criteria. This is determined by what types of programmes they are evaluating (also see **Appendix 1**).

**Table 16:** *Variation of weight-elicitation methods across countries and criteria*

| Country | Core set | Swing weighting programme-type specific criteria | | | |
|---|---|---|---|---|---|
| | | Frail elderly | Population health management | Palliative/ Oncological | Multiple problems |
| *Netherlands* | **DCE** | X | | | X |
| *Austria* | **DCE** | X | | | X |
| *Croatia* | **DCE** | X | | X | |
| *Germany* | **DCE** | X | X | | |
| *Hungary* | **DCE** | | | X | |
| *Norway* | **DCE** | X | | | X |
| *Spain* | **DCE** | X | X | | |
| *UK* | **DCE** | X | X | | |

## 5.3. Costs included in the SELFIE MCDA

A debated topic among MCDA experts, is whether or not to include costs as a criterion in the analysis. Total costs can be included as a single criterion or alternatively, different types of costs can be included as separate criteria. In SELFIE, overall health- and social care costs are included as a criterion in the core set. Furthermore, specific sub-categories of costs are also included as criteria in the programme-type specific sets. Below we briefly describe the debate and the rationale for our decision to include costs in the SELFIE MCDAs.

Those who are against including costs in MCDA, argue that costs do not adequately capture the opportunity costs of alternative uses of resources. (Claxton et al., 2015) They also argue that MCDA creates a new composite score of benefit and that the main question to be answered is what the opportunity costs are of one unit of additional benefit on that composite score. In other words, how much additional money can be spent at maximum for one unit of this composite score?

Those who are in favour of including costs argue that each MCDA will result in a different composite score, dependent on what criteria are included. This seems to make it impossible to determine some type of threshold for such a unit of improvement. Furthermore, it is widely recognised that MCDA is not a solution to the challenge of estimating opportunity costs. (Marsh et al., 2016) Supporters of including costs also argue that by including costs in the weight-elicitation respondents explicitly trade costs off against the other criteria in the analysis. Costs are then included just as other criteria in the overall MCDA, making their relative contribution throughout the decision-making process explicit. This is seen as being equivalent to estimating willingness-to-pay values for benefits. (Marsh et al., 2016)

**In SELFIE costs will be included in the MCDA itself**. The main arguments for this are listed below:
- It is very difficult to estimate the foregone health gains of investing in integrated care as opposed to something else (i.e., the opportunity cost) because integrated care is such a complex intervention with dynamic loops, many interactions, and non-linear outcomes.

Thus we prefer to assume that the costs included in the MCDA approximate the foregone outcomes from investing in the integrated care programmes (i.e., the opportunity cost) rather than to try estimating the real opportunity costs.

- We will not be able to develop a composite measure in SELFIE for which a threshold is determined. Thus there is room to include costs in the MCDA itself.

- Integrated care programmes are trying to achieve the Triple Aim, reducing costs is one of these aims, and hence it cannot be seen separately from other primary outcomes (i.e., criteria).

- In the weight-elicitation, we want the 5P stakeholders, the decision-makers, to make choices and trade-offs that are as similar as possible to 'real world' decisions. In such real-world decisions costs are also included in trade-offs in decision-making.

In SELFIE, in most cases the decision context is whether to continue or roll-out piloted integrated care programmes, and thus that the decision to finance integrated care has already been taken. Hence, the question is whether the particular integrated care programme evaluated generates sufficient benefits over the comparator to justify allocation of resources to that particular programme. Hereby, we acknowledge that usual care (i.e., the comparator alternative in SELFIE) increasingly encompasses elements of integrated care, but at a very slow pace for persons with multi-morbidity. Therefore, the question addressed in SELFIE is how our MCDA composite scores which in themselves include costs, compare between integrated care programmes and the comparator.

# 6. Weight-elicitation questionnaire – approach

As described above, the weight-elicitation questionnaire will encompass both DCE and swing weighting sections. The DCE section pertains solely to the core set of criteria, and the swing weighting section to the core set as well as programme-type specific criteria. The questionnaire will consist of the following sections: introduction, instructions DCE, 6 DCE scenarios, demographic questions, 6 DCE scenarios, general health questions, 6 DCE scenarios, swing weighting instructions, swing weight exercise, conclusion.

The criteria for both the DCE and swing weighting are described more extensively in **WP4 Deliverable Report 1**, also see **Appendix 2** for an overview.

In this chapter we describe the development of the weight elicitation questionnaire, the DCE design, the recruitment of respondents, and the general data collection procedure. These steps are described in the following order:

- Defining core set criteria levels
- Defining best- and worst levels for the programme-type specific criteria
- DCE design
- Determining priors for the DCE
- Respondents
- Procedure (piloting, translating, updating design)

## 6.1. Defining core set criteria levels

For the design of the DCE, the core set of criteria needed to be clearly defined in a way that can be understood by each of the 5P stakeholder groups. Furthermore, levels within the criteria needed to be determined and defined. In SELFIE, three levels per criteria are defined, ranging across 'good', 'average', and 'poor'. The core set of criteria and their levels are presented in **Tables 17(1) – 17(3)**. For an overview of the core set see **Appendix 2** and for more details and background information on the criteria and sources for the definitions of the levels, see **Appendix 5**.

**Table 17 (1):** *Health and well-being*

**Physical functioning**
- Defined as acceptable physical health and being able to do daily activities without needing assistance (e.g., getting dressed, setting down and getting up from a chair, taking your medications)
  1. **Severely** limited in physical health and activities of daily living
  2. **Moderately** limited in physical health and activities of daily living
  3. **Hardly or not at all** limited in physical health and activities of daily living

**Psychological problems**
- Defined as the occurrence of stress, worrying, listlessness, anxiety, and feeling down.
  1. **Always, or mostly**, being stressed, worried, listless, anxious, and down.
  2. **Regularly** being stressed, worried, listless, anxious, and down.
  3. **Seldom, or never,** being stressed, worried, listless, anxious, and down.

**Enjoyment of life**
- Defined as having pleasure and happiness in life
  1. **Not, or barely, having** pleasure and happiness in life
  2. **Having some** pleasure and happiness in life
  3. **Having a lot of** pleasure and happiness in life

**Social relationships & participation**
- Defined as having meaningful connections with others as desired
  1. **Not, or barely, having** meaningful connections with others
  2. **Having some** meaningful connections with others
  3. **Having a lot of** meaningful connections with others

**Resilience**
- Defined as the ability to recover from or adjust to difficulties and to restore ones equilibrium
  1. **Poor** ability to recover, adjust, and restore equilibrium
  2. **Fair** ability to recover, adjust, and restore equilibrium
  3. **Good** ability to recover, adjust, and restore equilibrium

**Table 17(2):** *Experience with care*

**Person-centeredness**
- Defined as care that care matches an individual's needs, capabilities and preferences and jointly making informed decisions
  **1. Not, or barely,** person-centred; i.e., care does not match an individual's needs, capabilities, and preferences, and decisions are not made jointly.
  **2. Somewhat** person-centred; i.e., care somewhat matches an individual's needs, capabilities, and preferences, and some decisions are made jointly.
  **3. Highly** person-centred; i.e., care matches an individual's needs, capabilities, and preferences, and decisions are made jointly.

**Continuity of care**
- Defined as good collaboration, smooth transitions between caregivers, and no waste of time
  **1. Poor** collaboration, transitions, and timeliness
  **2. Fair** collaboration, transitions, and timeliness
  **3. Good** collaboration, transitions, and timeliness

**Table 17 (3):** *Costs*

**Costs per participant per year**
- Defined as the total health- and social care costs per participant in the programme, per year
  **1. 5000 euros** per participant per year
  **2. 2000 euros** per participant per year
  **3. 500 euros** per participant per year

## 6.2. Defining best- and worst levels for the programme-type specific criteria

For the swing weighting weight-elicitation exercise only the worst and best end-points need to be defined. For the four types of programmes that are distinguished in SELFIE, we describe the criteria and their levels below in **Tables 18(1) – 18(4)**. For more details on the programme-specific criteria and their levels, see **Appendix 6.1-6.4**. Please note that for the purpose of the weight-elicitation questionnaire some terminology has been simplified, e.g., 'living at home' is now called 'long-term institution admissions'.

**Table 18 (1):** *Population health management*

| |
|---|
| **Activation & engagement** |
| • Defined as taking on the role of manging one's own health and care<br>**Worst: Unsuccessful in** managing your own health and care<br>**Best: Actively taking on the role** of managing your own health and care |
| **Avoidable hospital admissions** |
| • Defined as the number of hospital admissions that could have been avoided with better care<br>**Worst: 15 out of 100** hospital admissions could have been avoided with better care (avoidable hospital admissions)<br>**Best: 5 out of 100** hospital admissions could have been avoided with better care (avoidable hospital admissions) |
| **Hospital re-admissions** |
| • Defined as the number of persons who are re-admitted to a hospital within 30 days of their prior hospital discharge.<br>**Worst: 10 out of 100** persons are re-admitted to hospital within 30 days of hospital discharge.<br>**Best: 5 out of 100** persons are re-admitted to hospital within 30 days of hospital discharge. |

**Table 18 (2):** *Programmes targeting frail elderly*

| |
|---|
| **Autonomy**<br>• Defined as remaining in charge and making own decisions on how one lives his/her own life<br>   **Worst: Not, or barely**, in charge or making own decisions<br>   **Best: In** charge and making own decisions |
| **Burden of medication**<br>• Defined the amount of burden medicines are (considering for example administering/taking the medicines, side effects, understanding their purpose and why they're being taken, worries about interaction between medicines, and expenses)<br>   **Worst: High** burden of medication<br>   **Best: No, or low,** burden of medication |
| **Burden of informal caregiving**<br>• Defined the stress of informal caregiving due to the energy it costs, the little time it leaves for own interests and recovery and the sadness over the fate of the supported person<br>   **Worst: High** burden of informal caregiving<br>   **Best: Low** burden of informal caregiving |
| **Long-term institution admissions**<br>• Defined as the number of participants admitted to long-term institution care (for example a nursing home), during the programme<br>   **Worst: 20 out of 100** participants are admitted to a long-term institution<br>   **Best: 10 out of 100** participants are admitted to a long-term institution |
| **Falls leading to hospital admissions**<br>• Defined as the number of frail elderly that is admitted to an emergency room or hospital because of a fall<br>   **Worst: 10 out of 100** of participants have a fall that results in hospital admission<br>   **Best: 5 out of 100** of participants have a fall that results in hospital admission |

**Table 18 (3):** *Palliative care / Oncological programmes*

| |
|---|
| **Mortality**<br>• Defined as the change in 3-month life expectancy.<br><span style="color:red">**Worst: No improvement** in life expectancy</span><br><span style="color:green">**Best: Improvement** in life expectancy</span> |
| **Pain and other symptoms**<br>• Defined as physical symptoms like pain, fatigue, nausea/vomiting, short of breath, appetite loss, constipation/diarrhoea, insomnia<br><span style="color:red">**Worst: A lot of** pain and symptoms</span><br><span style="color:green">**Best: No, or little,** pain and symptoms</span> |
| **Burden of informal caregiving**<br>• Defined the stress of informal caregiving due to the energy it costs, the little time it leaves for own interests and recovery and the sadness over the fate of the supported person<br><span style="color:red">**Worst: High** burden of informal caregiving</span><br><span style="color:green">**Best: Low** burden of informal caregiving</span> |
| **Compassionate care**<br>• Defined as care that is provided in a warm, sensitive and dignified way with sympathy and respect<br><span style="color:red">**Worst: Not, or barely,** compassionate</span><br><span style="color:green">**Best: Very** compassionate</span> |
| **Timely access to care**<br>• Defined the time between referral and start of treatment or care of interest<br><span style="color:red">**Worst: A long** time between referral and start of treatment/care</span><br><span style="color:green">**Best: A short time** between referral and start of treatment/care</span> |
| **Preferred place of death**<br>• Defined as the number of participants that pass away in the location of their preference (e.g., at home).<br><span style="color:red">**Worst: 50 out of 100** participants pass away in preferred location</span><br><span style="color:green">**Best: 75 out of 100** participants pass away in preferred location</span> |


**Table 18 (4):** *Programmes targeting persons with problems in multiple life domains*

| |
|---|
| **Financial independence**<br>• Defined as being financially in control to meet basic needs with little or no debts.<br><span style="color:red">**Worst:** Financially **not in control** with growing debts</span><br><span style="color:green">**Best:** Financially **in control**; finances well-managed</span> |
| **Contact with justice system**<br>• Defined as contact with justice system, such as with criminal justice services, nights in police cell, and court attendance<br><span style="color:red">**Worst: Regular contact** with justice system</span><br><span style="color:green">**Best: No, or rare, contact** with justice system</span> |

## 6.3. DCE design

With eight criteria and three levels per criterion a full factorial design would lead to an infeasible large number of choice scenarios (see **Chapter 4.4.6.**). Therefore, computer generated efficient designs are commonly used in DCEs. In SELFIE, Bayesian efficient design algorithms will be used to optimise the DCE design. This will be done based on the D-efficiency criterion, which minimises the generalised variance of the parameter estimates based on a pre-specified model. The pre-specified model includes the specification of the model type and expected weight variation (i.e., priors, see **Chapter 6.4.**). Creating a D-efficient model is an iterative process that eventually results in the design with highest efficiency, i.e., the lowest generalised variance. To maximise the D-efficiency of the designs while accommodating respondent heterogeneity, a DCE design format commonly referred to as a heterogeneous DCE design (Sandor & Wedel, 2005) will be used. These designs consist of several sub-designs and each participating respondent is only asked to complete a randomly chosen sub-design. Accordingly, the [cognitive] burden for individual respondents is reduced because they only need to fill in a relatively limited amount of choice tasks; this in contrast to when a full factorial design is applied.

The heterogeneous DCE design will be optimised based on a conditional logit main-effects utility function. Furthermore, the optimisation criterion will be calculated as the weighted average Bayesian D-efficiency of 18 choice tasks per sub-design, with ⅓ of the weight assigned to the population efficiency and ⅔ of the weight assigned to the individual efficiencies of the sub-designs. In total, there will be 10 different sub-designs and in each sub-design there will be attribute level overlap (see e.g., Madalla et al. 2003; Kessels et al. 2012). This will be used to reduce the overall complexity of the choice tasks and improve response efficiency (see **Table 10** in **Chapter 4.4.6.** where there is overlap in person-centeredness and continuity of care). In the actual choice task, initially half of the attributes' levels will be restricted to the same level between the two alternatives. Throughout the process, we can adjust the number of attributes for which the levels are restricted. We may choose to set the majority of attribute levels as restricted in the first few choice scenarios to make these easy for the respondents, and

consecutively make less restricted and thus the tasks more difficult as respondents become used to the type of questions at hand. In order to reduce the repetitiveness, the choice scenarios will be presented in three blocks of six in the questionnaire. In between these blocks, general demographic and health questions will be asked.

The prior information that is required to optimise the design will be updated after a pilot run of approximately 50 respondents to further maximise statistical efficiency. Finally, sample size calculations based on De Bekker-Grob and colleagues (2015) have shown that a sample size of 150-200 respondents will be sufficient to obtain significant results.

## 6.4. Determining priors for the DCE

As described above, prior preference information can be used to create an efficient design. Initially, these priors are determined on the basis of literature. According to good research practice in DCE, the designs will be updated one or more times once the data collection has started. The design is re-estimated using the answers from the first circa 50 respondents that have already completed the questionnaire (see **Chapter 6.3.** above).

To determine the initial priors, PubMed and Google searches were conducted to find weight-elicitation or preference studies that focused on more than one criterion from the Triple Aim: health and well-being, experience, and costs. In total eight studies were included to inform the priors. The main outcome measures in these studies were mapped onto the SELFIE core set criteria, and their main findings are summarised in **Table 19**.

**Table 19:** *Input for determining priors*

| | Study reference | Mapping of outcomes – SELFIE criterion or Triple Aim category | Main findings |
|---|---|---|---|
| I | Comparison of preferences for **end-of-life care** among patients with advanced cancer and their caregivers: A discrete choice experiment (Malhotra et al., 2015) | • Severity of pain, expected length of survival ≈ **physical functioning**<br>• Quality of health-care experience ≈ **experience**<br>• Expected costs, source of payment ≈ **costs** | • Quality of care experience [*experience*] seems slightly less important than pain and survival [*physical functioning*]<br>• Expected costs and sources of payment [*costs*] seems of similar importance to quality of care experience [*experience*]<br>• Pain and survival [*physical functioning*] seem more important than expected costs and sources of payment [*costs*] |
| II | Patients' preferences for attributes related to **health care services** at hospitals in Amhara Region, northern Ethiopia: a discrete choice experiment (Berhane & Enquselassie, 2015) | • Physician & nursing communication ≈ **person-centeredness**<br>• Continuity of care ≈ **continuity of care** | • Continuity of care [*continuity of care*] was more important than the communication facets [*patient centeredness*] |
| III | **Health system goals**: A discrete choice experiment to obtain societal valuations (Franken & Koolman, 2013) | • Average level of health ≈ **health/well-being**<br>• Patient experience / process outcomes ≈ **experience**<br>• Financial fairness ≈ **costs** | • Health much more important than experience<br>• Health and costs of similar importance, health slightly more important<br>• Financial fairness [*costs*] more important than experience |

| | Study reference | Mapping of outcomes – SELFIE criterion or Triple Aim category | Main findings |
|---|---|---|---|
| IV | KCE Report – Incorporating societal preferences in **reimbursement decisions** – relative importance of decision criteria according to Belgian citizens (Cleemput et al., 2014) | • Quality of life (QoL) ≈ **health/well-being**<br>• Discomfort of treatment ≈ **physical functioning**<br>• Impact on disease-related public expenditures per patient ≈ **total health- and social care costs** | • Current treatment:<br>  ○ General population: QoL and discomfort equally important *[general health/well-being & physical functioning]* (log-likelihood method); QoL more important than discomfort (coefficient range method).<br>  ○ Decision-makers: QoL more important than discomfort *[general health/well-being vs. physical functioning]* (both methods).<br>• New treatment:<br>  ○ General population: QoL much more important than public expenditures and discomfort [general health/well-being vs. costs and vs. physical functioning] (log-likelihood method – similar for coefficient range method).<br>  ○ Decision-makers: QoL much more important than public expenditures and discomfort [general heath/well-being vs. costs and vs. physical functioning] (both methods). |

| | Study reference | Mapping of outcomes – SELFIE criterion or Triple Aim category | Main findings |
|---|---|---|---|
| V | Discussion paper – Explorations of the Effect of **Experience on Preference**: Two Health-Care Case Studies (Neuman & Neuman, 2006) (**Women who gave birth & women with breast cancer**) | • Information transferred to patient, attitude of staff ≈ **person-centeredness**<br>• Travel time to hospital ≈ **continuity of care** | • Women who gave birth:<br>　o Attitude of staff much more important than information,<br>　o Information *[person-centeredness]* more important than travel time *[continuity]*<br>• Women diagnosed with breast cancer:<br>　o Information and attitude similar,<br>• Information and attitude *[person-centeredness]* more important than travel time *[continuity]* |
| VI | Using discrete choice experiments to estimate a preference-based measure of outcome – an application to **social care for older people** (Ryan et al., 2006) | • Personal care ≈ **physical functioning**<br>• Social participation and involvement ≈ **social relationships and participation** | • Personal care very important *[physical functioning]*, social relationships slightly less important *[social relationships and participation]* |
| VII | Involving the public in **priority setting**: a case study using discrete choice experiments (Watson et al., 2012) | • Health gain ≈ **health / well-being**<br>• Patient involvement in own care / shared decision-making ≈ **person-centeredness**<br>• Management of care / teamwork ≈ **continuity of care** | • Teamwork *[continuity]* seems to be of more importance than shared decision-making *[person-centeredness]* |

| | Study reference | Mapping of outcomes – SELFIE criterion or Triple Aim category | Main findings |
|---|---|---|---|
| VIII | Development of the Assessment of Burden of COPD (ABC) index: experienced burden of disease and the prediction of healthcare consumption (Goossens et al., submitted) | • Fatigue, limitations in physical activities, limitations in daily activities ≈ **physical functioning**<br>• General mental problems – feeling depressed, listless, tense, worrying ≈ **psychological well-being**<br>• Limitations in social activities ≈ **social relationships and participation** | • Fatigue and limitations in physical activities [*physical functioning*] more important than limitations in social activities [*social relationships and participation*] and more important than general mental problems [*psychological well-being*]<br>• Limitations in social activities [*social relationships and participation*] seem slightly more important than general mental problems [*psychological well-being*]<br>• These differences seem smaller in the general population as compared to a patient population. |

The information from **Table 19** was grouped to create an order of importance, or a ranking, in the criteria per study. See **Table 20** for an overview, now organised according to the SELFIE core set of criteria. When no differences were observed, the same ranking was given.

**Table 20:** *Ranking of SELFIE core set criteria according to literature*

| | Study reference | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Attributes** | I | II | III | IV | V | VI | VII | VIII |
| *Health / well-being* | 1 | | 1 | 1 | | | | |
| Physical functioning | | | | 2 | | 1 | | 1 |
| Psychological well-being | | | | | | | | 3 |
| Enjoyment of life | | | | | | | | |
| Social relationships & participation | | | | | | 2 | | 2 |
| Resilience | | | | | | | | |
| *Experience* | 2 | | 3 | | | | | |
| Person-centeredness | | 2 | | | 1 | | 2 | |
| Continuity of care | | 1 | | | 2 | | 1 | |
| *Costs* | 2 | | 2 | 2 | | | | |
| Total health and social care costs | | | | | | | | |

Hereafter, on the basis of the input above, a range in criteria levels is determined. The larger the range, the more 'important' the criteria. This range is shown in the list below, from most to least important:

- Physical functioning 0.60
- Social relationships & participation 0.45
- Psychological well-being 0.40
- Enjoyment of life 0.40
- Resilience 0.40
- Costs 0.35
- Continuity of care 0.30
- Person-centeredness 0.25

Next, priors are proposed for levels within each criterion. Each criteria has three levels, poor – average – good. These are operationalised in detail in **Chapters 6.1.** and **6.2.** In the list of priors for the levels that is presented below, we have used simple level-labels. In all cases, except for costs, higher and better values are desired – these have positive attribute weights as the worst level is coded as the reference. For costs, low costs is the reference level, making the other attribute levels negative.

- Physical functioning 0.60
    - Poor (reference)
    - Average (0.30)
    - Good (0.60)
- Social relationships & participation 0.45
    - Poor (reference)
    - Average (0.25)
    - Able to maintain (0.45)
- Psychological well-being 0.40
    - Poor (reference)
    - Average (0.20)
    - Good (0.40)
- Enjoyment of life 0.40
    - Poor (reference)
    - Average (0.20)
    - Good (0.40)
- Resilience 0.40
    - Poor (reference)
    - Average (0.20)
    - Good (0.40)
- Costs 0.35
    - Poor (-0.35) (high costs)
    - Average (-0.15) (average costs)
    - Good (reference) (low costs)
- Continuity of care 0.30
    - Poor (reference)
    - Average (0.15)
    - Good (0.30)
- Person-centeredness 0.25
    - Poor (reference)
    - Average (0.10)
    - Good (0.25)

## 6.5. Identifying respondents – the 5P stakeholders

SELFIE partners were asked to reach out to each of the 5P stakeholder groups for the weight-elicitation questionnaire. To identify specific groups of persons and organisations for each of the 5Ps, definitions of these stakeholders were developed and recommendations were given for ways to reach out to these groups, see **Table 21** below.

**Table 21:** *5P stakeholder respondents of the SELFIE weight-elicitation questionnaire*

| |
|---|
| **Patients:** Defined as persons with multi-morbidity. It is possible to cover the broad concept of multi-morbidity, including specific disease comorbidities, frail persons, and persons with problems in multiple life domains. |
| *Potential sources*:<br>• Patient federations<br>• Attendants of the National Stakeholder Workshop (WP8)<br>• Attendants of the focus group (for WP4 – task 2)<br>• National cohort studies that include information on health and can be reached out to for additional data collection. |
| **Partners & Informal caregivers:** Defined as persons that support the care of individuals with multi-morbidity, this can be family (children, spouse), neighbours, or friends for example. |
| *Potential sources*:<br>• Informal caregiver federations or support groups<br>• Via patients that participate in the weight-elicitation questionnaire (closing question 'Do you have an informal caregiver? If so, would he/she be interested in filling in a similar questionnaire? If so, please provide his/her email address …).<br>• Attendants of the National Stakeholder Workshop |
| **Professionals (care providers):** Defined as persons that regularly provide care for persons with multi-morbidity in their daily work. Considering the broad definition of multi-morbidity applied, this can include nurses, GPs, specialists (e.g., geriatricians, elderly care physicians), social workers, or other professional groups that for example take on the role of case manager or care coordinator. |
| *Potential sources:*<br>• Professional/occupational union organisations<br>• Care centre staff (e.g., hospitals, primary care clinics) – via a contact person there or human resources that can email these employees with the questionnaire link<br>• Newsletters/magazines of interest to these professional groups<br>• Organisations/Attendants of the National Stakeholder Workshop<br>• Professional care providers from the partner country who are members of the International Foundation of Integrated Care (IFIC), Joint-Action CHRODIS or other international organisations |

**Policy makers/advisors:** Defined as persons that have some level of influence on policies regarding integrated care and/or multi-morbidity. This can for example include persons from a regional government/municipality position, or from a more macro-level position. This group can include policy-maker advisors.

*Potential sources:*
- Ministry employees
- Municipality employees
- Governmental organisations/associations
- Regional/National/Municipal policy advisory organisations
- Organisations/Attendants of the National Stakeholder Workshop

**Payers:** Defined as persons that have some level of influence regarding financial decisions surrounding integrated care and/or multi-morbidity. This can include budget-holders in municipalities or government, social care budget holders, or health care budget holders. It can include persons working at health insurance companies that purchase care programmes.

*Potential sources:*
- Health insurance company employees
- Municipalities
- Health- and social care budget holders
- Organisations/Attendants of the National Stakeholder Workshop

## 6.6. Procedure

The weight-elicitation questionnaire is currently under development by the EUR team and will be an online questionnaire. Sawtooth Software is being used to administer the questionnaire. This software is specifically meant for online DCE surveys and allows for an easy integration of our efficient DCE design and convenient processing of the survey data.

The questionnaire will be piloted amongst several persons with multi-morbidity in the Netherlands. These persons will fill in the questionnaire using a 'think-aloud' method whereby they tell the researcher who is either present via telephone or face-to-face their thoughts about the questionnaire. The researcher will take notes. On the basis of these pilot questionnaire notes, the questionnaire will be adapted. The adapted version will be translated to English. Hereafter, each SELFIE partner will translate the questionnaire. The following translation instructions are used:

I.   Two persons independently from one-another translate the English questionnaire into their native language. These persons should have a high level of English proficiency in order to be able to interpret the English questionnaire correctly.

II.  Hereafter, these two persons compare the native language questionnaire and discuss discrepancies to create one 'ideal' version.

III. Two *different* persons independently from one-another translate the native language questionnaire back to English. These persons should be native speakers of the non-English language, and should be highly proficient in English. However, their English translations will not be used for actual data collection.

IV.  Hereafter these two persons, and those involved in the first step, compare the new English version to the original English questionnaire. Where there are discrepancies the team should go back to the native language questionnaire and make adaptations.

The goal is that the questionnaires are identical content-wise which may inevitably require some grammatical (e.g., sentence order) adaptations to be made.
Hereafter, the EUR team will prepare each partners' online questionnaire and share the link with the partner. The partners are then required to disseminate the questionnaire amongst the 5P stakeholder groups.

After the questionnaire has been filled in by circa 50 persons in each country and in each 5P stakeholder group, initial analyses will be conducted. These analyses will be used to refine the initial priors (as presented in Table 19) and hence to create a more efficient and more relevant DCE design. Hereafter, data collection can continue until the target sample size has been reached (expected to be approximately 150-200 persons from each 5P stakeholder group per country have filled in the questionnaire). In case the target sample size cannot be reached, country or stakeholder groups may be pooled.

# 7. Study designs for the case study evaluations

In SELFIE we aim to apply study designs that allow for the intervention, integrated care programmes, to be compared to a control situation in the most robust way possible. We aim to measure performance indicators at least twice in both groups. **Thus, we will estimate the performance after follow-up (e.g., 6- or 12-months) using information from baseline scores as well.** This is especially important if there are baseline differences in the indicators between the two groups, i.e., the groups have a different starting point. It is important to be aware and acknowledge that we elicit weights for criteria defined on absolute scales rather than in terms of change. These weights are combined with performance estimates in which we incorporate baseline information (as described in Step 5, **Chapter 4.5.**).

There are different ways to analyse performance indicators throughout follow-up between two groups. One option is to conduct analyses that take baseline values of the performance indicator into account through covariance. This can, for example, be done by estimating the performance indicator at follow-up for both integrated care and control group by including the baseline score of the performance indicator of both groups in the analyses. Such a model can also include more advanced statistical methods that ensure further comparability of the groups, e.g., by weighting for propensity scores, and/or adjusting for other confounders. Thus in such an analysis a performance indicator score at follow-up would be obtained for both the integrated care and control group. This analysis does not generate change values, but the analysis does take the baseline performance indicator score into account, and thus includes the covariance between the baseline and follow-up score. Especially in the case that persons in the integrated care and control group differ at baseline with regard to scores on performance indicators (e.g., their starting level of resilience is different), this is a favoured method. (Twisk, 2003)

Another way of looking at changes in performance indicator through time in both groups is by calculating the difference in performance indicator score and relating this to a threshold difference. For example, is a difference of 0.01% between baseline and follow-up on an indicator at all relevant? There are various ways of determining what constitutes a relevant change, one such example is the Minimal Important Difference (MID). One option is to define the MID as 1.5 of the standard deviation around the baseline mean of the

performance indicator per group, as a result of which the MID could even differ between the two groups (integrated care and control). Subsequently, we could express performance at follow-up as the extent to which / how many MIDs were achieved.

Calculating a difference score between follow-up and baseline and using this as the performance indicator (e.g., an absolute change scores), even when relating this to an MID, is usually not preferred. Instead, the covariance method takes better account of regression to the mean. Regression to the mean is an issue when, for example, a control group starts out quite high and the intervention group starts out quite low, after which both groups are most likely to regress to the mean. The control group at a certain point simply cannot get better, and the intervention group cannot get worse. (Twisk, 2003)

Below different study designs are presented that aim to collect repeated data in two comparable groups.

## 7.1. Overview of possible study designs

In this report, three main types of study designs are distinguished, each with sub-categories that may be of interest in SELFIE:

- **Experimental**
    - Randomised controlled trial (RCT)
    - Cluster RCT
- **Quasi-experimental** (i.e., natural experiments)
    - Stepped-wedge design
    - Propensity score matching
    - Instrumental variable
    - Regression discontinuity design
    - Difference-in-difference
- **Observational**
    - Cohort study
    - Case control study
    - Interrupted time series

Several main sources were used to distinguish and describe these study designs: Tsiachristas et al., 2016; Schelvis et al., 2015; ZonMw Startdocument, 2015; Langendam et al., 2013; White & Sabarwal, 2014; Katz, 2010; Penford & Zhang, 2013; Rosenbaum & Rubin, 1983.

The ease of attributing causality is the highest with experimental designs, and the lowest with observational designs. The listed sub categories are in principle study designs, but also methods to try to attribute effects to an exposure (e.g., care programme). It is also possible to use multiple methods and do so post-hoc.

Important issues to consider in study designs are causality and confounding, see **Figure 2** below.

**Figure 2:** *Exposure, outcome, confounding*

Exposure → Outcome

E.g., integrated care intervention
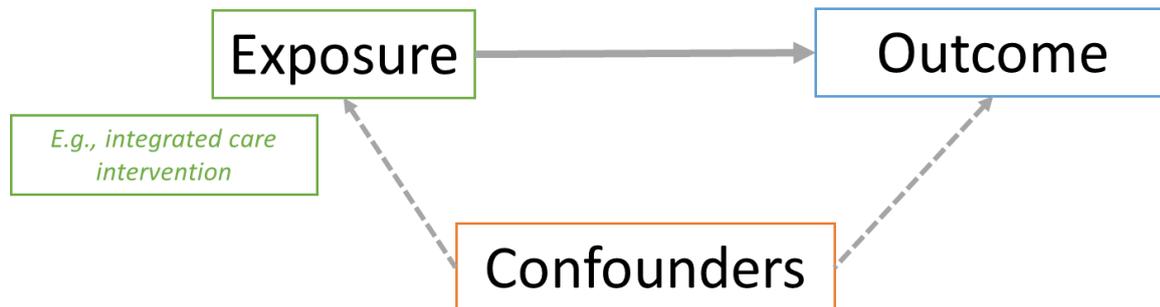
Confounders

In order for confounding to occur, three conditions must be present:

1.  The confounding factor must be associated with **both** the exposure and the outcome.

2.  The confounding factor must be **distributed unequally** among the groups being compared.

3.  A confounder can**not** be an intermediary step in the **causal pathway** from the exposure to the outcome.

**Effect modification** occurs when the magnitude of the effect of an exposure on the outcome differs depending on the level of a third variable.

In **experimental** designs there is random allocation of persons/clusters to an intervention or control group and groups are followed parallel to one-another. This design is often chosen because it is the most robust design to infer causality. It is important to note that in integrated care that is still difficult due to the complexity of the intervention and organisational level contamination. How 'usual' is usual care? (E.g., every form of care is integrating) Furthermore, in experimental designs there are issues related to transferability of findings into real-life practice that need to be considered.

**Quasi-experimental** designs are also known as natural experiments. Like experimental designs, the purpose is to test causal hypotheses – can we attribute an outcome to an exposure? However, unlike experimental designs, there is *no randomisation*. Instead, allocation to exposed (e.g., integrated care programme) vs. unexposed (e.g., usual care) is due to self-selection and/or administrator selection. The goal is to have the unexposed

comparator group as similar as possible to the exposed group at baseline. This in turn can facilitate the attribution of differences in outcomes to differences in exposure. In such a design there is a higher risk of confounding than in experimental designs, but there are often less issues to do with transferability. There are different methods available to deal with observed and unobserved confounding between groups in this design, namely:

- **Observed**: regression adjustment and propensity score matching methods;
- **Unobserved**: difference-in-difference, instrumental variables, and regression discontinuity methods.

In **observational** designs there is no group allocation, mere observation. This method is often selected due to financial reasons, when it is difficult to identify appropriate participants, there are concerns about generalisability, and there are ethical issues. The main limitation to this design is that it is difficult to attribute causality.

## 7.2. Experimental designs

In Randomised Controlled Trials (**RCTs**), there is patient-level randomisation to exposed (e.g., integrated care) and unexposed (e.g., usual care) groups. Next, there is prospective parallel-group comparison between exposed and unexposed groups. RCTs are not possible or desirable in complex interventions when controlling all factors/context is impossible and contamination between groups is likely, if randomisation (withholding an intervention in the control group) is considered unethical, or if the population in the RCT is not representative for real-world population. In deciding whether or not to conduct an RCT, a trade-off needs to be made between internal and external validity.

In **cluster-RCTs**, there is group-level randomisation to exposed (e.g., integrated care) and unexposed (e.g., usual care) groups. Clusters of patients in a group (e.g., neighbourhood, department, GP-practice) are randomised. Randomising a group (instead of an individual) prevents contamination (how persons undergoing the intervention influence those not undergoing the intervention) and increases the logistic feasibility of implementing the intervention. For example, if a GP is trained to provide integrated and collaborative care for 50% of his/her patients (exposed), it is highly likely that he/she will subconsciously transfer these skills to the other 50% of patients (that are supposed to be unexposed). This makes the difference between exposure and un-exposure smaller and will reduce the difference in outcome between the two groups (e.g., satisfaction with care). The power of the design is determined by the number of clusters. If too few clusters are included, controlling by chance for all factors that might differ between groups is impeded and internal validity is compromised.

**Figure 3:** *RCTs and Cluster-RCTs*
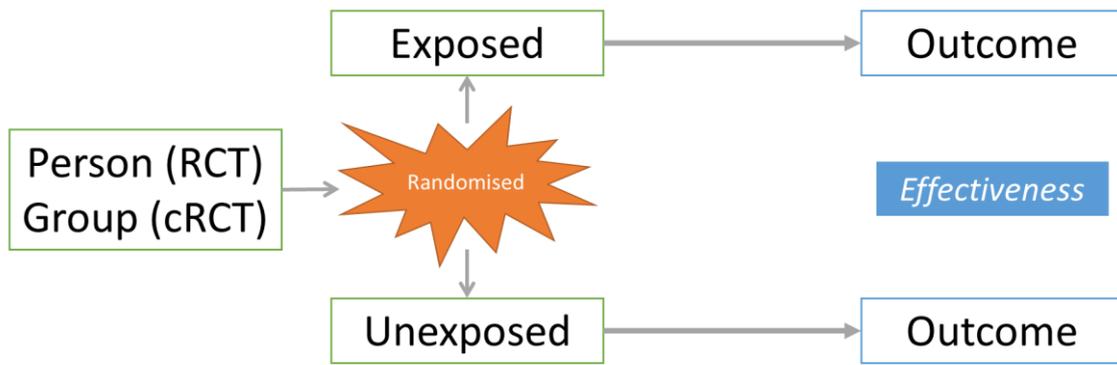
## 7.3. Quasi-experimental designs

There are many different designs and methodologies that are considered quasi-experimental, below the stepped-wedge design, propensity score matching, instrumental variable analyses, regression discontinuity designs, and difference-in-difference analyses will be described.

The **stepped-wedge design** is a modification of the individual or cluster RCT in which the exposure (e.g., intervention) is sequentially rolled-out to all subjects over consecutive time periods. In this method, selection bias is prevented by randomising the order by which subjects receive the exposure (e.g., intervention). Because all subjects receive the intervention ethical issues of withholding the intervention are solved. Changes in intervention based on lessons learned in previous step are possible before the next step. Through this design it is possible to measure different types of effects: both short and long-term effects, fade out effects, and the natural course of the condition under study.

**Figure 4:** *Stepped-wedge design*
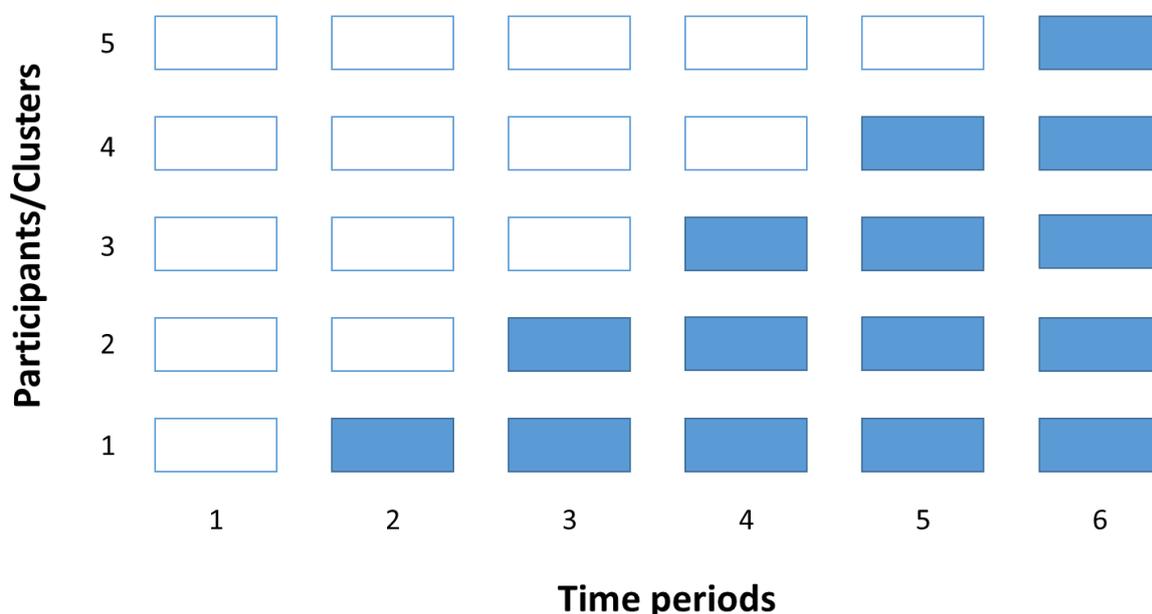


In **Figure 4** the stepped-wedge design is shown, whereby each block represents data collection. Shaded blocks represent intervention periods, and blank blocks represent control periods. For this design, larger sample sizes are needed to ensure enough statistical power. Furthermore, there may be a higher burden on participants and researchers due to

necessity to repeatedly collect data. Thus the design is most feasible if data can be (partly) routinely collected at the appropriate time intervals in a reliable and valid way. The statistical analyses in such a design are more complex, since it is necessary to include a random coefficient for cluster and a fixed effect coefficient for time.

**Propensity score matching** is a means to match exposed to unexposed persons. In a perfect scenario: each exposed individual is matched to one or more unexposed individuals who are identical on all relevant observable characteristics. Since this is impossible, the goal is to form pairs of exposed and unexposed individuals who have a similar likelihood of being exposed, i.e., have a similar value of the propensity score.

**Figure 5:** *Matching exposed and unexposed groups in propensity score matching*



*Some determinants can also be considered confounders.*

Propensity scores can be estimated using logistic regression analyses modelling the exposure as the dependent variable and the potential determinants of being exposed as independent variables. Further, propensity scores can also be used to stratify analyses, adjust analyses using the propensity as a covariate, and to perform inversed probability weighting. An assumption in propensity score matching is that the average characteristics of the exposed vs. unexposed group are similar. There are various methods to achieve one-to-one matching in propensity score matching:

- *Nearest neighbour matching*: matches an unexposed individual with the closest propensity score of an exposed individual. This can be done with or without replacement of the unexposed individual.
- *Nearest neighbour matching within a specified calliper distance*: similar as above but with the further restriction that the absolute difference in the propensity scores of

matched individuals must be below some pre-specified threshold (the calliper distance) (e.g,. equal to 0.2 of the standard deviation of the logit of the propensity score).

- *Kernel weighting*: each individual in the exposed group gets a weight of 1 and each unexposed patient gets a weight that depends on the distance of their propensity score from the propensity score of the exposed patients. Higher weights represent better matches. There are different types of Kernel weighting (e.g., Gaussian type, Epanechnikov type, bi-weight type, uniform type, tricube type).

With propensity score matching it is possible to use inverse probability weighting. This method uses weights based on the propensity score to create a synthetic sample in which the distribution of measured baseline covariates is independent of treatment assignment. Here a subject's weight is equal to the inverse of the probability of receiving the treatment that the subject actually received.

In **instrumental variable analyses** it is hypothesised that an instrumental variable (IV) can 'randomise' by adjusting for measured and unmeasured confounding. This IV should be a proxy for the exposure (e.g., intervention), but cannot have a direct association to the outcome (only via the exposure), and also cannot be associated with unmeasured confounders. The ratio of the effect of the IV on the outcome on the one hand, and of the instrumental variable on the exposure on the other hand, shows the *true* effect of the exposure on the outcome. It is important to note that choosing the correct instrumental variable is a challenge. Such analyses are often used in trials when there is non-compliance and/or for intention to treat analyses – to counter loss-to-follow-up.

**Figure 6:** *Instrumental variable analyses*
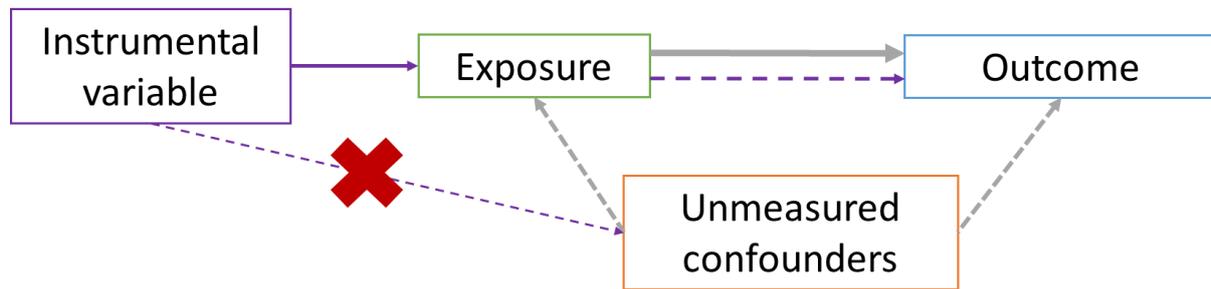


There are four assumptions of the instrumental variable

1. IV is not associated with the outcome except through the exposure
2. IV is strongly associated with the exposure (e.g., intervention vs. control group assignment)
3. IV is not associated with baseline characteristics known to be associated with the outcome.
4. IV has a monotonic relationship with exposure: everyone who would be exposed with a low score predicting exposure, would also get exposure with a high score predicting intervention.

Assumptions 2 and 3 can be empirically tested, assumptions 1 and 4 should be assumed on the basis of theoretical considerations. A strong association between IV and exposure makes for a better instrumental variable.

After testing the assumptions, analyses are conducted in 2 steps (similar to PSM):

1. The IV and any other predictors of the exposure, predict the likelihood of being exposed for each subject.
2. These predicted values then replace the actual intervention assignment for each subject and are used to estimate the impact of the intervention on the outcome.

Usually estimates using the IV to predict the outcome are less precise than when the exposure itself is used to predict the outcome.

**Figure 7a:** *Instrumental variable analysis example 1*
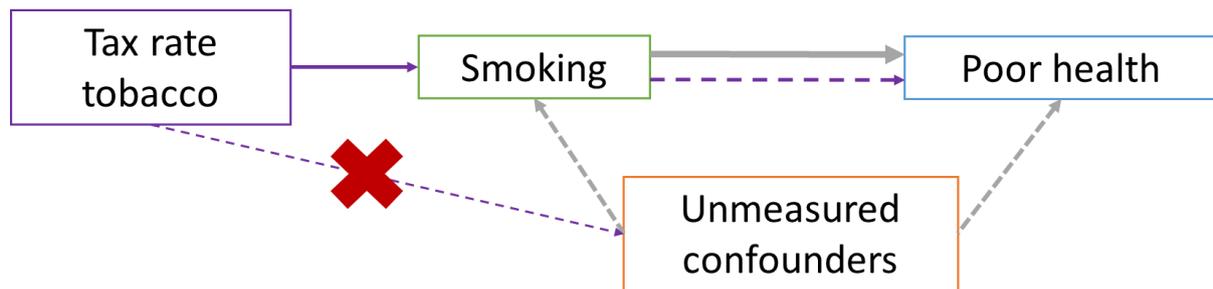


**Figure 7b:** *Instrumental variable analysis example 2*
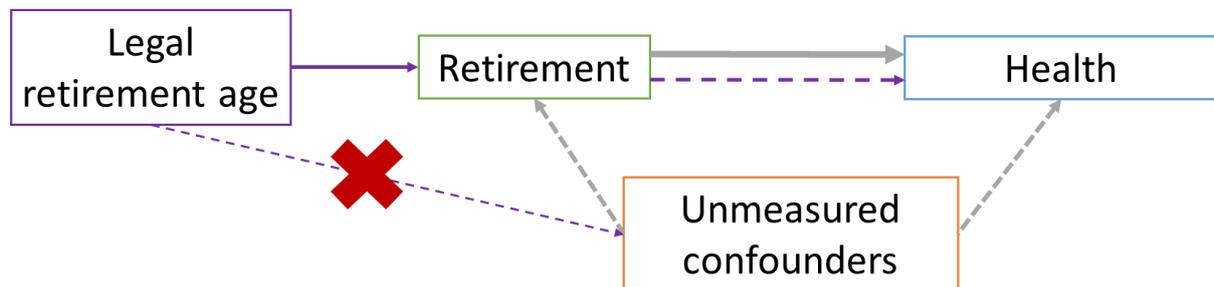


**Regression discontinuity design (RDD)** is a method can be used when there is some kind of criterion that must be met before subjects can be exposed (i.e., included in an intervention). For example, there is a minimal age of 50 for free breast cancer screening and a minimal BMI of 25 for a weight-loss intervention. Of course persons well over and under 50 or 25 differ greatly from one-another and these criterion are then related to the outcome. However, persons that are just above or below this cut-off criterion most likely to do not differ so much: 49-50 vs. 50-51 and 24.5-25.0 vs. 25.1-25.5.

In RDD, the exposure (e.g., intervention group) is assigned to subjects that score [just] above a certain 'cut-off point' (on a continuous variable); and the un-exposure (e.g., control group) is made up of individuals that score [just] below that cut-off point. The main assumption in RDD is that individuals that just score on either side of the 'cut-off point' or threshold, belong to the same population. Allocation of these subjects to either exposed or unexposed is therefore considered 'random' – assuming that subjects cannot manipulate the threshold value. Intervention effects are estimated by comparing the outcomes of the group that meets the criterion with the group that just not meets the criterion using

regression techniques: non-parametric (local linear regression) or parametric (polynomial regression).

In RDD the assumption that individuals around the threshold are similar is often debatable. The design requires larger sample sizes than an RCT to achieve sufficient statistical power. Furthermore, it requires data collection for all individuals *considered* for intervention (also those who are not included in the intervention).

An extended example in the case of a remedial education programme: The selection criterion for eligibility to participate in the programme is a pre-intervention test score, with a threshold of 60. The outcome variable is a post-intervention test score. The scatter plot shows that these two variables are related. There is a positive relationship between pre- and post-intervention test scores. Children with a pre-intervention test score of below 60 received the remedial classes. The sample used for the analysis is taken from just either side of the threshold – those included have pre-intervention test scores in the range of 50 to 70, i.e., 10 units either side of the threshold. The fitted regression line has a 'jump'; this is the discontinuity. The size of this jump (which is 10) is the impact of the programme – that is, the remedial education programme increases test scores by 10 points on average. (White & Sabarwal, 2014)

**Figure 8:** *Regression discontinuity design*
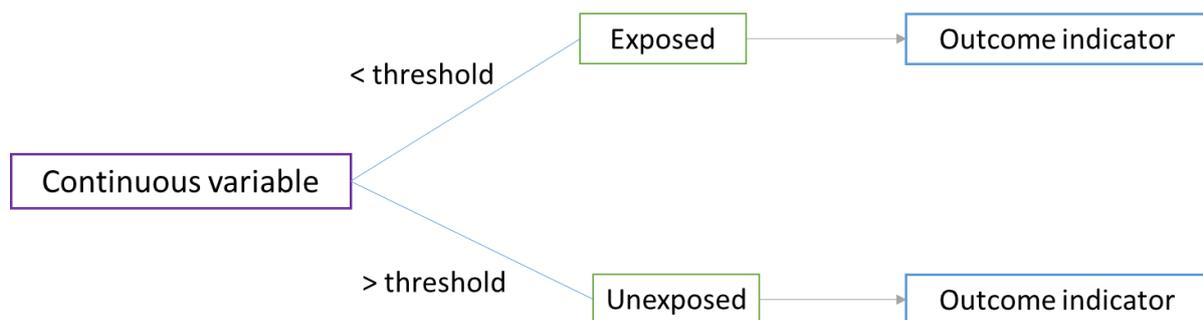


**Difference-in-difference** methods are also known as 'double difference' methods. In this method changes in the outcome over time between two groups (i.e., exposed and unexposed) are compared to estimate an effect. This way the difference at baseline between the groups is 'removed'. The method assumes that the outcomes of interest in the unexposed group follow the same trend over time as the outcomes would do in the exposed group in the absence of the intervention; thus that there is no confounder that influences

one group and not the other. The assumption of parallel trends between intervention and control groups can be relaxed in a differential trend model.

**Figure 9a:** *Difference-in-difference analyses – parallel trajectory assumption*
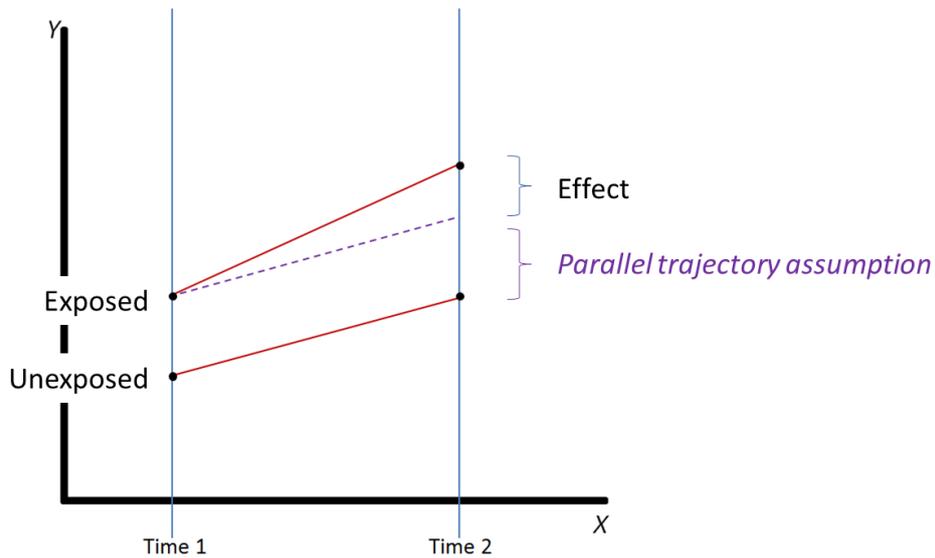


**Figure 9b:** *Difference-in-difference analyses – comparison to single difference*
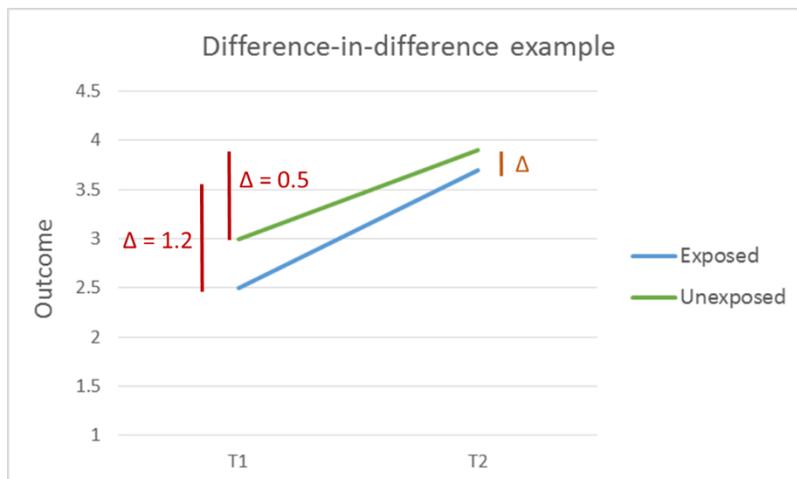
## 7.4. Observational designs

The starting point in **cohort studies**, is the selection of a study population, or a cohort. Information is obtained to determine which persons in this cohort are exposed to a specific factor, and which are not (e.g., which persons participated in an integrated care programme, and which did not). Allocation of participants to either group (exposed vs. unexposed) is not influenced by the investigator. In such a study, outcomes can be collected prospectively or retrospectively.

**Figure 10:** *Cohort studies*
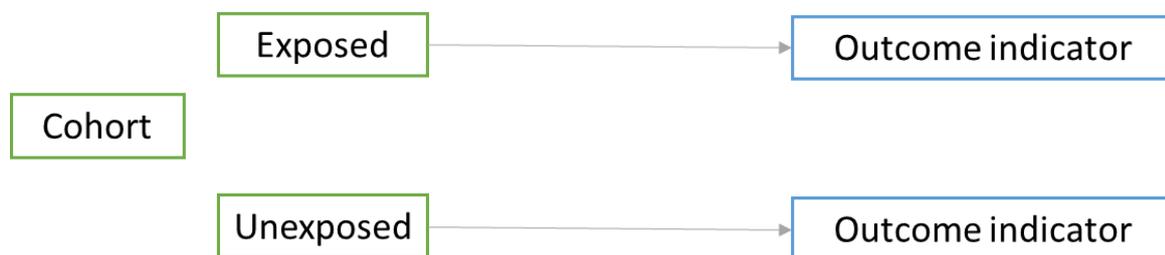


As opposed to cohort studies, **case-control studies** start off with the identification of 'cases' and 'controls' (outcome present or not present). The next step is then to go back and look at which cases and which controls were exposed vs. unexposed (i.e., participated in an integrated care programme vs. usual care). There is then also the possibility to do a **nested case-control study**, in which a subset of controls are matched to cases. X controls are matched to a single case based on certain characteristics, making this a more efficient model.

**Figure 11:** *Case-control studies*
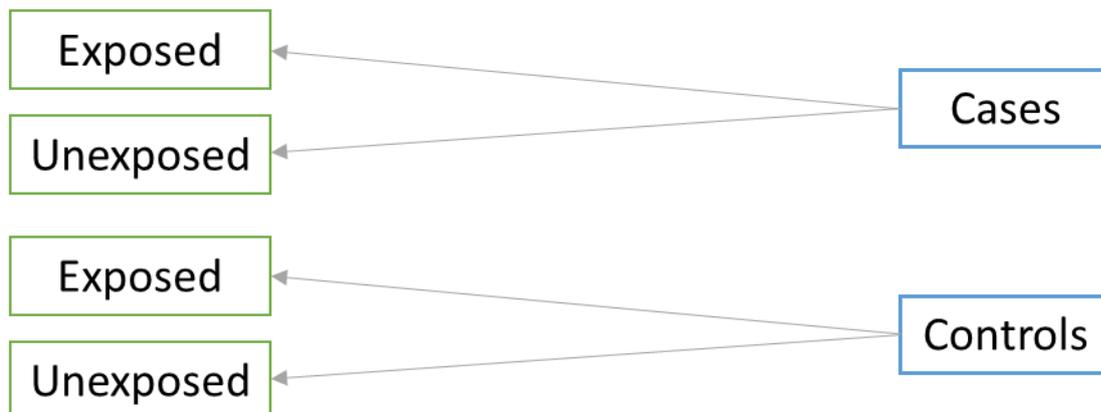


In **Figure 11**:

- **Exposed**: Has determinant of interest (e.g., integrated care programme)
- **Unexposed**: Does not have determinant of interest (e.g., usual care)
- **Cases**: Have outcome of interest (e.g., high quality of life)
- **Controls**: Do not have outcome of interest (e.g., high quality of life)

A relevant analyses technique for observational study designs is the **interrupted time-series analyses**. For such analyses repeated measurements are performed before and after exposure at population level in order to detect whether the intervention has a greater effect than the underlying secular trend (e.g., economic, market, or demographic trend). Next, we need to determine whether the intervention had a larger effect than any underlying trend – this is estimated by comparing the trend in the outcome after the intervention to the trend in the pre-intervention period. This is a relevant design when using routinely collected data (e.g., insurance data) and when looking at macro interventions. Some advantages to this method include that it allows for control over secular trends, the ability to evaluate outcomes at population-level, a clear graphical presentation of results, and the ability to do stratified analyses and look at intended and unintended consequences of interventions. Some disadvantages to this method include that multiple measurement waves are needed (circa 8 before and after exposure (i.e., implementation)).

**Figure 12a:** *Interrupted time-series analyses*
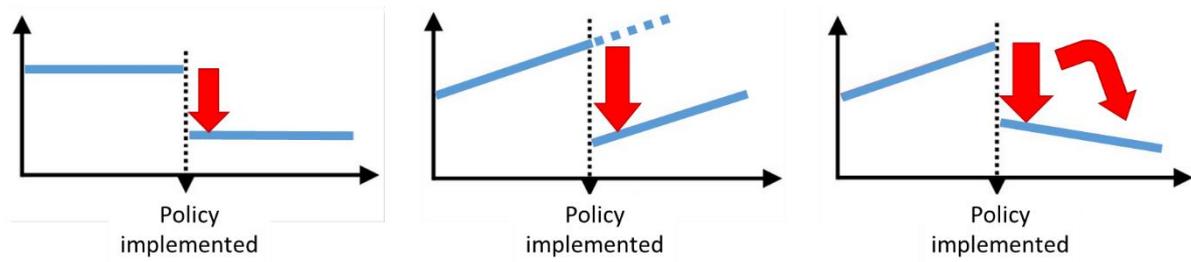


If we take **Figure 12a** and an example in which a sugar tax is implemented in New York state and we want to know the effects thereof on BMI (exaggerated for the case of the example), we can look in neighbouring state New Jersey (see **Figure 12b**).

**Figure 12b:** *Interrupted time-series analyses – example*
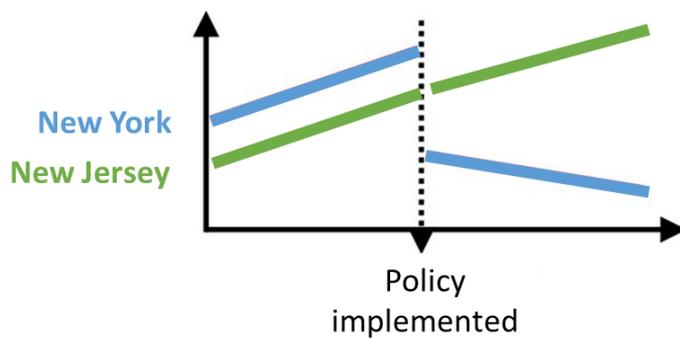
# 8. Proposed study design

In **Table 22** below the key features of the 17 planned case studies in SELFIE are presented. In the remainder of this chapter a short summary of the planned evaluations per programme per country is presented. It is important to note that these are preliminary plans and that throughout WP5 these will take further form.

**Table 22:** *Key elements of the 17 programme evaluations*

| Programme / Country | Target group | Type and focus of care (e.g., prevention, primary, secondary, tertiary, palliative, health- and/or social care, population health management) | Key study design features (e.g., prospective/retrospective, cohort, RCT, randomised/non-randomised, empirical questionnaire data collection, data linkage, matching techniques, participation selection) |
|---|---|---|---|
| **The Netherlands** | | | |
| *Proactive Primary Care Approach for Frail Elderly (U-PROFIT)* | Frail elderly | Prevention-focused, primary health care and social care, screening, case management, multi-disciplinary team meetings | (1) Prospective cohort study, regression discontinuity design, propensity score matching, empirical questionnaire data collection, linking questionnaire to registry data<br>(2) Retrospective re-analysis cluster-RCT through additional informed consent, linkage of questionnaire with registry data, propensity score matching |
| *Care Chain Frail Elderly (CCFE)* | Frail elderly | Primary health care and social care, case finding, multi-disciplinary team meetings | Prospective cohort study, non-randomised intervention and control group, propensity score matching, empirical questionnaire data collection, linking questionnaire to registry data |
| *Better Together in Amsterdam North (BSiN)* | Persons with problems in multiple life domains | Social care and primary health care, case finding, triage, case management | Quasi-experimental design, non-randomised intervention and control group, empirical questionnaire data collection, linking questionnaire to registry data, propensity score matching |
| **Austria** | | | |
| *Health Network Tennengau* | Population of the Tennengau region | Coordination of health and social care providers, case management, counselling services | (1) Quasi-experimental controlled design, population-level administrative claims data, propensity score matching<br>(2) Observational design, empirical questionnaire data, linkage of questionnaire data with individual administrative claims data |
| *Sociomedical Centre Liebenau* | Persons with problems in multiple life domains | Primary health care and social care, focus on vulnerable groups, addiction treatment | Quasi-experimental controlled design, empirical questionnaire data, linkage of questionnaire data with individual administrative claims data, propensity score matching |

| Country<br><br>Programme | Target group | Type and focus of care (e.g., prevention, primary, secondary, tertiary, palliative, health- and/or social care, population health management) | Key study design features (e.g., prospective/retrospective, cohort, RCT, randomised/non-randomised, empirical questionnaire data collection, data linkage, matching techniques, participation selection) |
|---|---|---|---|
| **Croatia** | | | |
| *GeroS Model* | Geriatric patients with multi-morbidity | Primary health care and social care (institutional or non-institutional) | Prospective cohort study, empirical questionnaire data collection, linking questionnaire to GP information system data, health insurer cost data, social care data, post-hoc statistical matching techniques |
| *Palliative care Model* | Palliative care patients | Primary health care (treatment by Mobile multidisciplinary specialist palliative care team - MMSPCT or usual care) and social care | Prospective cohort study, empirical questionnaire data collection, linking questionnaire to GP information system data, health insurer cost data, social care data, post-hoc statistical matching techniques |
| **Germany** | | | |
| *Casaplus* | Frail elderly | Case-management, prevention –focused, primary and social care | (1) Quasi-experimental regression discontinuity design, non-randomised intervention and control group, propensity score matching, claims data analyses<br>(2) Before-after controlled study, empirical questionnaire data collection |
| *Gesundes Kinzigtal (GK)* | Population health management | Population health management, health- and social care | (1) Quasi-experimental controlled study, non-randomised intervention and control group, claims data analyses<br>(2) Observational (trend) study, empirical questionnaire data collection |
| **Hungary** | | | |
| *OnkoNetwork* | Patients with new suspect of a solid tumour | The primary aim of OnkoNetwork is to improve the clinical outcomes through better pathway management | Prospective and retrospective comparative longitudinal, non-interventional cohort studies, with multivariate regression models to control for patient heterogeneity in the target population. Regression analyses will be restricted to pre-selected cancer types. |

| Country<br><br>Programme | Target group | Type and focus of care (e.g., prevention, primary, secondary, tertiary, palliative, health- and/or social care, population health management) | Key study design features (e.g., prospective/retrospective, cohort, RCT, randomised/non-randomised, empirical questionnaire data collection, data linkage, matching techniques, participation selection) |
|---|---|---|---|
| *Palliative Care Consult Service* | Patients with palliative care needs | managing home-based hospice-palliative care, psychosocial support, mental care, pain relief or other symptom management | Prospective and retrospective cohort studies with multivariate regression models to control for patient heterogeneity in the target population. Questionnaire data, cost of provider and national payers' database will be collected |
| **Norway** | | | |
| *Learning Networks* | Frail elderly | Primary health care and social care, case management, multi-disciplinary team meetings, rehabilitation | Prospective cohort study, propensity score matching. Retrospective cohort study, propensity score matching. |
| *MAR Bergen* | Persons with problems in multiple life domains | Secondary health care, primary health care, triage, social care, multi-disciplinary team meetings | Prospective cohort study, propensity score matching. Retrospective cohort study, propensity score matching. |
| **Spain** | | | |
| *Area Integral de Salut de Barcelona-Esquerra (AISBE)* | Population health management | Management of Chronic Complex Patients (CCP) requiring specialised care. | Quasi-experimental design, non-randomised intervention and control group, empirical questionnaire data collection, linking questionnaire to registry data, propensity score matching |
| *Badalona Serveis Assistencials (BSA)* | Frail elderly | Integration between home-based healthcare and social support allowing provision of innovative services that promote independent living and proactive care. | Quasi-experimental design, non-randomised intervention and control group, empirical questionnaire data collection, linking questionnaire to registry data, propensity score matching |

| Programme | Country | Target group | Type and focus of care (e.g., prevention, primary, secondary, tertiary, palliative, health- and/or social care, population health management) | Key study design features (e.g., prospective/retrospective, cohort, RCT, randomised/non-randomised, empirical questionnaire data collection, data linkage, matching techniques, participation selection) |
|---|---|---|---|---|
| **The UK** | | | | |
| *Salford Integrated Care Programme (SICP) / Salford Together* | | Originally 65+, now all adults | Prevention-focused, self-management, primary health care and social care, mental health care, secondary care, case management, multi-disciplinary team meetings, organisational change, population health management | Quasi-experimental designs, difference-in-differences, regression discontinuity, routinely collected data, population-level analysis, intervention-level analysis, impact of additional organisational changes to support service delivery changes |
| *South Somerset Symphony Programme* | | Primary focus on 3 or more conditions, but all adults | Health coaching, self-management, primary health care, secondary care, case management, multi-disciplinary team meetings, organisational change, population health management | Quasi-experimental designs, difference-in-differences, regression discontinuity, routinely collected data, population-level analysis, intervention-level analysis |

## 8.1. The Netherlands

### 8.1.1. Proactive Primary Care Approach for Frail Elderly (U-PROFIT)

**Care programme**

The U-PROFIT care approach for frail elderly consist of 2 steps: 1) U-PRIM screening, and 2) the U-CARE nurse-led care approach. The U-PRIM screening of Electronic Medical Records (EMR) allows primary care centres to identify potentially frail elderly on the basis of polypharmacy, multi-morbidity, and/or a lack of contact with the GP practice (consultation gap). An elderly care nurse is in charge of using U-PRIM and contacting potentially frail elderly for further holistic assessment. In U-CARE the elderly care nurse goes on to provide integrated and tailored care, by taking the findings from the holistic assessment and the preferences of the frail elderly [and informal caregiver] to create an individualised care plan. The needed care is provided in collaboration with the GP and other relevant disciplines. The approach was initially implemented in the form of a cluster-randomised controlled trial (cluster-RCT) and is currently being implemented in eight primary care centres in Utrecht.

**Study design**

The main evaluation to be conducted is a prospective cohort study applying a regression-discontinuity design (RDD). Data collected in 2016-2017 amongst frail elderly receiving the U-PROFIT approach care, as it is being implemented in the eight primary care centres in Utrecht, will be used. At the time that the evaluation will begin (winter/spring 2017), the included frail elderly have filled in a baseline questionnaire that mostly pertains to health and well-being. At 12-month follow-up we will have them fill in an expanded questionnaire that includes: experiences with care and self-report care utilisation. Furthermore, a [more extensive] informed consent will be obtained in order to get permission to link questionnaire to care utilisation- and cost data (e.g., GP EMR data, health insurer data, and national health information system registry data (*Dutch*: *Vektis*).

Data will also be collected amongst a control group of frail elderly from several of the eight primary care centres. These control frail elderly will be selected using a regression

discontinuity design. This is a design that can be used when there is some criterion that must be met before persons can be included in an intervention (i.e., the U-PROFIT approach). In an RDD, the intervention is assigned to persons, i.e., frail elderly, that fall just above the cut-off point of this criterion, and the control situation is assigned to persons that fall just below this cut-off point. The idea is that this cut-off point, or threshold, is relatively arbitrary and random, and thus that comparing the groups is justified. In the case of U-PROFIT, there is an age cut-off point for inclusion, namely 60 years. Considering that the main inclusion criteria are based on the extent of frailty, falling just below or above this age cut-off point can be construed as relatively arbitrary. The control group of frail elderly will be selected in a uniform way as the intervention group, namely on the basis of a two-step frailty screening: U-PRIM screening of EMR data and with the Groningen Frailty Index (GFI) questionnaire. However, whereas in the intervention group the U-PRIM software makes a selection of persons aged 60 and over, in the control group this selection will be set between 55-60 years. These control frail elderly will be asked to fill in a baseline and 12-month follow-up questionnaire. They will also be asked to fill in the same informed consent. In order to make the frail elderly in the intervention- and control group more comparable, propensity score weights will be applied in the evaluation analyses.

The collected data in the prospective cohort RDD will be brought together in a Multi-Criteria Decision Analysis (MCDA). In an MCDA multiple outcomes can be included, in the current study pertaining to health/well-being, experience, and costs. Further, outcomes are weighted on the basis of the chosen perspective that is taken. In the SELFIE project, 5 sets of weights, from 5 stakeholder perspectives, will be obtained in a parallel study and applied to the data collected: Patients, Partners (i.e., informal caregivers), Professionals, Payers, and Policy makers.

A secondary evaluation will also be done using data from the cluster-RCT of the U-PROFIT approach that ran from 2010-2013. During the cluster-RCT data was collected in two groups implementing different components of the U-PROFIT intervention (i.e., i. only U-PRIM screening and ii. U-PRIM screening and U-CARE nurse-led care) and a control group of frail elderly on health- and well-being, satisfaction with care, and on self-report care utilisation. Two PhD theses were written on the findings of the cluster-RCT; findings were mixed. The relatively

short follow-up period of 12-months may have played a role in this. For this reason, the data will be re-analysed and expanded upon with a longer follow-up of cost data. The re-analyses will use an MCDA framework, thus applying the SELFIE weight-sets to determine whether this affects overall findings. For this secondary evaluation we will need to reach out to elderly from the trial to ask them to sign a novel, more extensive, informed consent that will make it possible to link their questionnaire data from the cluster-RCT phase to longer follow-up care utilisation, i.e., cost, data from registries (i.e., *Vektis*).

## 8.1.2. Care Chain Frail Elderly (CCFE)

**Care programme**

The CCFE is currently being implemented in three care groups in Southeast Brabant: PoZoB, DOH, and SGE. All General Practices (GPs) affiliated with DOH and SGE are implementing the CCFE. However, in care group PoZoB there are GPs not doing so yet. The CCFE targets frail elderly living at home, with complex care needs that require case management to prevent or restrain loss of control. The care programme is based on six key elements: case finding, poly-pharmacy, care coordination, multidisciplinary care with individual care planning, transfer care and organising a community network. Professionals involved are the GP, nurse practitioner specialised in elderly care, and the district nurse. In some cases the elderly care physician and/or a case manager specialised in dementia are also involved.

**Study design**

A prospective cohort study design will be applied to compare frail elderly included in the CCFE and a control group of frail elderly receiving usual care. As the programme is being implemented in daily practice, no randomisation will take place. Instead, the two groups will be matched to one-another using post-hoc statistical matching techniques. Frail elderly for both the intervention and control groups will be contacted for participation in the evaluation via their GPs. Intervention group frail elderly will be recruited via GPs from all three care groups implementing the approach; control group frail elderly will be recruited via GPs from care group PoZoB that are not yet implementing the CCFE.

At baseline (T0), 6 months follow up (T1) and 12 months follow up (T2) questionnaire data will be collected amongst frail elderly in both groups by trained researchers. At the end of the trial period this questionnaire data will be linked to health- and social care use and cost data. At baseline, participants will sign an informed consent for the use of their data for research purposes and for the linking of data (e.g., questionnaire data, GP information system data, health insurer cost data, Vektis data).

The study population consists of an intervention and control group. The intervention group consists of elderly that are identified as frail by their GP; affiliated with one of the three care groups (i.e., SGE, DOH, PoZoB) and that is implementing the CCFE, from one of the three care groups) and that are [about to be] included in the CCFE. Frailty is thereby defined as older persons with care- and case complexity that leads to loss of control over the person's own life, which requires case management and multidisciplinary care. The control group consists of elderly that are identified as frail by their GP; affiliated with one care group (i.e., PoZoB) that is not implementing CCFE. Frail elderly are identified using the same case finding approach and the same definition of frailty that is used in the intervention group. However, case finding is not part of the CCFE nor does it result in the CCFE trajectory of care. The case finding process will be primarily done for the purpose of identifying frail elderly viable for the current evaluation. A secondary purpose is to allow the GPs to 'practice' with case finding and start to work with this method, as in the future it will also become usual care for them.

Frail elderly in the intervention group receive care according to the 'Care Chain Frail Elderly' (CCFE). These elderly receive multidisciplinary care and are assigned a case manager. For the GPs implementing CCFE, this constitutes usual frail elderly care. In the control group, the CCFE has not been implemented yet, and there is no set care programme for frail elderly; they receive usual care (in their own context). The current evaluation in the context of the SELFIE project will evaluate the (cost-) effectiveness of the CCFE on the Triple Aim: health/well-being, experience and costs. These are operationalised by the following concepts: physical functioning, psychological well-being, enjoyment of life, social relationships and participation, resilience, autonomy, burden of medication, per-centeredness, continuity of care, falls, living situation and total health and social care costs.

### 8.1.3. Better Together in Amsterdam North (BSiN)

**Care programme**

Providers from each of the organisations involved in the BSiN alliance can request that an individual be 'triaged' to determine whether they can be included in the BSiN programme and what type of care they need. Triage entails holistically assessing the individual and discussing him/her in a multidisciplinary team meeting to determine the level of care that is needed and which organisation can best provide it. The Self-Sufficiency Matrix (SSM) is used for assessment. This helps determine problem areas and needs in the following life domains: finances, daily activities, housing, relationships at home, mental health, physical health, addiction, activities of daily living, social network, social participation and justice. Four quadrants are distinguished after triage: 1) self-sufficient (no care needed), 2) care coordination needed, 3) client support needed, and 4) case management needed. Each of the organisations in the alliance can provide staff who will be trained to work as case managers and work thereafter as such on a part-time basis alongside their regular work. In the case management quadrant, care is integrated and coordinated, an individualised care plan is drawn up together with the person, and progress is routinely monitored by the case manager. The person of interest is actively involved, and a focus is placed on his/her own abilities in solving problems. A typical case management trajectory takes six months to one year.

**Study design**

This study is an extension of the current ongoing evaluation study performed by research organisation TNO (Dutch Trial Register number: NTR5068). We will make use of the data that has been collected between September 2013 and December 2016 and we will collaborate with research organisation TNO to continue the data collection in 2017-2018. The method that will be used for the evaluation is a quasi-experimental study design. Participants in the case management group are recruited by care providers and welfare workers in Amsterdam North and the control group participants are identified via a regional monitoring questionnaire study. Persons are selected from this monitor dataset and matched to the intervention group using propensity square matching. Face-to-face interviews are held with the participants in both the

intervention and control groups at baseline, 6 months, and 12 months to collect questionnaire data. The outcomes in the evaluation are measures related to (i) health/well-being (physical functioning, psychological well-being, enjoyment of life, social participation and relationship, resilience, and autonomy), (ii) experience of care (person-centeredness and continuity of care), and (iii) costs (of health care, social care, and welfare). In SELFIE we will use the collected data to perform an MCDA.

## 8.2. Austria

### 8.2.1. Health Network Tennengau

**Care programme**

The Health Network Tennengau (HNT) is a bottom-up network comprised of social and health service providers and voluntary organisations. The Tennengau region is a rural area in the state of Salzburg with approximately 54,000 inhabitants. The HNT has its origins in a pilot project for medical nursing home care introduced in 1995 and has since gradually evolved into a comprehensive network. Since 2003, it has been organised as a not-for-profit cooperative that currently has 29 member institutions.

Service delivery within the programme is divided into 6 modules: (1) the core of the network, which coordinates the entire programme, (2) a counselling service for the elderly in municipalities, (3) cooperation and exchange of information with GPs and specialists, (4) geriatric remobilisation at the hospital in Abtenau, (5) cooperation with inpatient care for the elderly and with social and psychological services, and (6) discharge management and transition care at the public hospital in Hallein.

**Study design**

Two major types of study designs will be used in the empirical evaluation of the HNT: a quasi-experimental controlled study design (study design 1) and an observational study design (study design 2).

Study design 1 primarily aims at investigating resource utilisation and healthcare costs in the Tennengau region compared to similar regions. Pseudonymised administrative data from the regional health insurance fund of the state of Salzburg will be used as the main data source. The database covers various patient characteristics, extensive information regarding outpatient service use, as well as more general information regarding inpatient service use. In addition to the SELFIE outcomes on resource utilisation and costs required for the MCDA, further potentially interesting parameters will possibly be investigated. The intervention group comprises all persons who live in the Tennengau region and are insured at the regional health

insurance fund of the state of Salzburg (possibly with an age-specific restriction). The control group is a sample of persons living in different rural regions in the state of Salzburg, which are similar to the Tennengau region. The sample will also be drawn from those insured at the regional health insurance fund of the state of Salzburg. In order to compare outcomes in the intervention group with the control group, matched pairs of individuals will be formed by means of propensity score matching.

Study design 2 will focus on patient-reported health, well-being and experience of care in patients using services provided by the HNT, and provide some additional information on resource utilisation in these patients. Two data sources will be used: (1) questionnaire data and (2) administrative claims data linked to the questionnaire data. The questionnaires will be distributed to clients of the HNT with the help of service providers involved in the HNT together with on-site aid by a member of the IHS research team. In addition to gathering information on the SELFIE outcomes on health, well-being and experience of care, the questionnaire will be used to ask patients' consent to use their claims data from the administrative database of the respective health insurance fund. This allows investigating resource utilisation specifically of clients of the HNT. The questionnaire will also include a question regarding which services of the HNT the client is using, so that questionnaire data as well as claims data can be analysed separately for different initiatives of the HNT. The intervention group consists of persons who use at least one of the several services provided by the HNT. These persons will be recruited via personal contact in practices of physicians participating in the HNT, as well as through the hospital Hallein. Forming an adequate control group is problematic in the context of the questionnaire survey, since health insurance funds are not allowed to provide personal data on insured persons without the latter's consent. Alternative ways of recruiting (a) control group(s) are still being explored. Data will be collected at two different points in time. The first wave of data collection is planned for mid-2017, the second wave for mid-2018.

The data collected by means of both study designs will be analysed using several statistical tests, and will subsequently feed into a multi-criteria decision analysis (MCDA). The MCDA approach allows incorporating perspectives of multiple relevant stakeholder groups on the Triple Aim of improved health and well-being, improved care experience and reduced costs.

## 8.2.2. Sociomedical Centre Liebenau

**Care programme**

The Sociomedical Centre Liebenau (SMC) is a bottom-up pioneer model providing health and social care predominantly to vulnerable and disadvantaged groups. It is situated in the socially-deprived Liebenau district in the Austrian city of Graz and was founded by three physicians in 1984. It is currently organised as a collaboration between a group practice and the Association for Practical Social Medicine.

All activities at the SMC are based on a "social health and medicine approach" with a focus on inequality with respect to health and social status. The programme's target clientele are persons with physical and mental disorders and/or social problems. The SMC is operated by a multidisciplinary team of physicians, social workers and social pedagogues. This team provides both medical care (e.g., primary medical care, psychotherapy, addiction treatment, health promotion) and social care (e.g., various counselling services, community work). A continuous holistic assessment of the patient with an emphasis on social aspects is central to the SMC's work, along with appropriate target group communication and joint case conferences.

**Study design**

The evaluation of the SMC will be based on a quasi-experimental controlled approach and will specifically focus on services provided to drug users. The primary aim of the study design is to investigate patient-reported health, well-being and experience of care as well as resource utilisation in drug users receiving services provided by the SMC in comparison to a control group.

Two data sources will be used: (1) questionnaire data and (2) administrative claims data linked to the questionnaire data. The questionnaires will be distributed with the help of the SMC and the social services of Graz and/or Vienna, respectively, together with on-site aid by a member of the IHS research team. In addition to gathering information on the SELFIE outcomes on health, well-being and experience of care, the questionnaire will be used to ask patients' consent to use their claims data from the administrative database of the respective health

insurance fund, so that patients' healthcare resource utilisation can also be investigated. The intervention group consists of drug users receiving services provided by the SMC who will be recruited via personal contact at the SMC as well as via referral from service providers involved in the SMC (e.g., physicians, social workers). The control group consists of drug users who do not receive services provided by the SMC. The latter will be recruited via personal contact at addiction aid centres (in Graz or Vienna), pharmacies and/or via referral from physicians providing substitution therapy. In order to compare outcomes in the intervention group with the control group, matched pairs of individuals will be formed by means of propensity score matching.

The data collected by means of the described study design will be analysed using several statistical tests, and will subsequently feed into a multi-criteria decision analysis (MCDA). The MCDA approach allows incorporating perspectives of multiple relevant stakeholder groups on the Triple Aim of improved health and well-being, improved care experience and reduced costs.

## 8.3. Croatia

### 8.3.1. The GeroS Model

**Care programme**

The GeroS Model is an integrated care model for geriatric patients with multi-morbidity. The programme is currently still in its pilot phase and has therefore not yet been fully implemented at the national level. The system will be inseparably linked to the Central Health Information System of the Republic Croatia (CEZIH), as a subsystem. GeroS is designed as a 15-module system for the monitoring and evaluation of health needs and functional ability of insured persons over the age of 65, and in particular of geriatric patients receiving corresponding care. One purpose of the programme is the computerisation of all health and social care records of this age group. GeroS involves primary healthcare providers, specialists, hospitals, long-term care providers, homes for the elderly and centres for social care. Individual care plans, a person-centred attitude, proactive care and risk stratification are part of the programme. Management of polypharmacy and clinical guideline interaction will also be provided once the programme has been fully implemented. The programme is based on a holistic assessment of the multi-morbid person and informal caregivers. GeroS aims to achieve rationalization of geriatric health and social care. It ensures that relevant information and changes are promptly and effectively available to interested parties involved in the social and health care process, in accordance with their responsibilities in the system. Self-management interventions are supported by indirect empowerment from the Reference Centre for Health Gerontology through educational material and guidelines. New professional roles will be developed from already established professions and must be created with the prefix "geronto" in order to provide specific care to geriatric persons. The monitoring system for the programme has been provided. The financing and payment scheme is based on the national mandatory health insurance through the Croatian Health Insurance Fund for health care and social services in homes for elderly persons through the social care system.

**Study design**

A prospective observational cohort study will be performed, with a 12-month follow-up period, during 2017-2018. The goal will be to evaluate how the "GeroS Model", specifically four modules affect health and well-being, experience of care, resource utilisation and costs in comparison to usual care (corresponding the 'Triple Aim'). The four modules that will be looked at are:

i. **Four degrees of geriatric health care**: holistic assessment on the basis of physical mobility and psychological independence that determines the type of care needed, as divided into four degrees of geriatric health care needed, ranging from minimal level – in the residential part of the home, to the maximum level for infirmary in the home.

ii. **Nutritional Risk Screening**

iii. **Record Sheets 1 and 4**: This is a nurse documentation system, which allows for guidelines for individual geriatric care and a person-centred attitude. The focus here is also on monitoring and evaluation. Health- and social care needs are brought to light via these record sheets.

iv. **Non-institutional care and activities**, as provided by the Gerontology Centres in the local community. These are commonly placed within elderly homes and provide immediate non-institutional care. They also focus on primary prevention.

The study population consists of geriatric patients, from three Croatian counties (Grad Zagreb, either Karlovačka or Zagrebačka (to be determined), Koprivničko-križevačka županija) exposed or unexposed to the specific parts of the Geros Model (i.e., the four aspects described above). In Grad Zagreb (around 800,000 inhabitants in 2015) two homes for the elderly (provides institutional and non-institutional care to around 680 geriatric persons) with primary care centres will be involved (exposed cohort). Karlovačka (128,749 inhabitants in 2011) or Zagrebačka županija (317,642 inhabitants in 2011) and Koprivničko-križevačka (115 582 inhabitants in 2011) have not yet implemented the GeroS Model (unexposed cohort).

Patients were considered for enrolment if they are geriatric patients with multi-morbidity (65 years and older) and have a life expectancy of more than 12 months. Patients who are not able to give answers on questionnaires (have a diagnosis of dementia or are

unresponsive or nonverbal), patients unlikely to survive more than 12 months based on their clinicians' judgments and those who do not want to signed an informed consent, will be excluded. Follow-up will be discontinued when the patient dies, is permanently transferred to another institution or no longer wants this treatment. Patients can withdraw participation from this study at any time. The primary care physician must be informed as well. The eligible geriatric patient with multi-morbidity (identified from two sources: receiving social institutional or non-institutional care) will be asked by a responsible member for the conducting of the study whether they would be willing to hear more about the study. The informed consent form will be used to provide information about the study. Patients who agree to hear more about the study are given detailed information, as consistent with informed consent procedures by the responsible staff at the home for the elderly / Gerontology centre. Patients who agree to participate will sign written informed consent (for the use of their data for research purposes and for linking of data, i.e., questionnaire data, GP information system data, health insurer cost data, social care data) and fill in the standardised questionnaire - at the beginning of the study period (T0) with the help of a trained research assistant (interviewer). This is important due to expected difficulties of collecting data from geriatric patients and to assure confidentiality of patient answers without any influence on the further care received. All recruitment efforts regarding eligibility, participation, and the reasons for ineligibility, opting out or non-participation (if reasons are offered) will be documented using a standardised template. The sample size required for detecting the anticipated difference on two or three outcomes (i.e., quality of life, pain or nausea) is obtained from the literature data: the sample size of each cohort could range from 110 to 200 patients.

The same standardised questionnaire on outcomes related to health/well-being, experiences with care, resource utilisation and costs, will be again filled after a 6-month (T1) and 12-month period (T2). Due to expected difficulties of collecting data from geriatric patients, a trained research assistant (interviewer) will help patient to fill in the questionnaire at each wave. Part of the data related to resource utilisation and costs will be extracted from patient health records and/or will be given by care providers.

Multi-Criteria Decision Analysis (MCDA) and Cost-effectiveness (or CUA) analysis will be performed as well. The results of CEA will be compared to the results of the MCDA.

### 8.3.2. The Palliative Care Model

**Care programme**

The Palliative Care Model is an integrated care programme specifically designed for palliative care patients; it is based on the National Healthcare Strategy and the Strategic Plan for Palliative Care 2014–2016. It is estimated that a minimum of 20% of cancer patients and 5% of non-oncological patients need palliative care in the last year of their life. Stakeholders and resources exist in four key domains: the healthcare system, social welfare system, education/higher educational system, and civil society. Currently, the Palliative Care System is not fully integrated in the national healthcare system; it is based on pilot-level projects in different parts (counties) of Croatia. Full implementation on the national level is not complete yet, this is awaiting changes in the legal framework. The Committee for Palliative Care, appointed by the Ministry of Health, is responsible for the further activities and suggestions on the further implementation of palliative care and development of a new Strategic Plan for Development of Palliative Care in Croatia for the period after 2017 as well. Sustainable systematic data collection and monitoring, quality indicators, and a registry of palliative patients are necessary at the national level. Palliative care is organised on three levels: home care (provided by primary healthcare centres staff: family medicine physicians, community nurses, mobile palliative care team, and other social care); extended palliative care (provided by social services); and hospital care. Coordination Centres for Palliative Care coordinate care between hospitals, ambulatory palliative care, specialized palliative care teams, mobile specialist palliative care teams at the primary care level, and social care, providing vertical, horizontal and intersectoral cooperation and collaboration, at the county and national levels. Informal caregivers/volunteers are included, as well as Institution/Offices for renting of medical aids/devices. Mobile multidisciplinary specialist palliative care teams, an example of a new innovative role in the Croatian palliative care model, are established at the county level (at the primary care level). The palliative care programme includes different kind of professionals, such

as physicians, nurses at different health institutions, community care nurses, psychologists, social workers, volunteers and family, mobile teams, university staff, and priests. Better support of the care process, improved continuity, comprehensiveness, and coordination of care for multi-morbid persons is achieved through linkage of three levels of healthcare, social care, and informal caregivers. The financing and payment scheme is based on national mandatory health insurance through the Croatian Health Insurance Fund (HZZO) for health care and different social services, like in homes for elderly persons, or through the social care system. Beyond regular funding, the HZZO provided additional funds for ten mobile palliative team pilot projects in primary healthcare centres. Supplementary funding is also provided by some counties.

**Study design**

A prospective observational cohort study will be performed, with a 6-month followed-up period, during 2017-2018. The goal is to evaluate how the "Palliative care Model", specifically treatment by the mobile multidisciplinary specialist palliative care team (MMSPCT), affect health and well-being, experience of care, resource utilisation and costs in comparison to usual care (corresponding the 'Triple Aim'). Sustainable and permanent MMSPCTs are already implemented in three Croatian counties only (Grad Zagreb, Istarska and Primorsko-goranska županija).

The study population consists of palliative care patients sample (around 60%-80% of cancer patients and 20%-40% of non-cancer patients), from six Croatian counties (Grad Zagreb, Istarska, Primorsko-goranska, Karlovačka, Koprivničko-križevačka and Zagrebačka) with around 5800 palliative care patients, exposed or unexposed to the treatment by MMSPCT. Grad Zagreb (around 800,000 inhabitants in 2015) has 2 MMSPCTs; Istarska županija (208,440 inhabitants in 2011) has 1 MMSPCT and Primorsko-goranska (296,123 inhabitants in 2011) has 4 MMSPCT teams. Karlovačka (128,749 inhabitants in 2011), Koprivničko-križevačka (115,582 inhabitants in 2011) and Zagrebačka županija (317,642 inhabitants in 2011) have not yet implemented sustainable and permanent MMSPCTs.

Patients will be considered for enrolment if they are palliative care patients (according the Supportive and Palliative Care Indicators Tool-SPICTTM and ICD-10: Z51.5), 18 years or older, with life expectancy ranging from 1 to 6 months. Patients and/or families who refuse further care by the MMSPCT or usual care, who are not able to give answers on questionnaires (have a diagnosis of dementia or are unresponsive or nonverbal), patients unlikely to survive more than 3-months based on their clinicians' judgments and those who do not want to sign the inform consent will be excluded.

Treatment by the MMSPCT or family physician will be discontinued when the patient dies, is permanently transferred to the secondary health care institutions or no longer wants this treatment. Patients can withdraw participation from this study at any time. The sample size required in order to detect the anticipated difference on two or three outcomes (i.e., place of death, quality of life, pain or nausea) is obtained from the literature data: the sample size of each group will range from 150 to 200 patients.

At the first visit by the MMSPCT or the family physician the eligible patient and/or family will be asked whether they would be willing to hear more about the study. The informed consent form will be used to provide information about the study. Patients who agree to hear more about the study are given detailed information as consistent with the informed consent procedures. Patients who agree to participate will provide written informed consent (for the use of their data for research purposes and for linking of data, i.e., questionnaire data, GP information system data, health insurer cost data, social care data) and fill the standardised questionnaire - at the beginning of the study period (T0) with help of a trained research assistant (interviewer). This is important due to expected difficulties of collecting data from palliative care patients and to assure confidentiality of patient answers without any influence on the further care received. This process could also be done at the county hospital level before discharge of palliative patient and his referral to MMSPCT or the family physician, by a member of the hospital palliative care team. All recruitment efforts regarding eligibility, participation, and the reasons for ineligibility, opting out or non-participation (if reasons are offered) will be documented using a standardised template.

The same standardised questionnaire, on outcomes related to health/well-being, experiences with care, resource utilisation and costs, will be again filled after 1-month (T1) and 3-month period (T2). Due to expected difficulties of collecting data from palliative care patients, a trained research assistant (interviewer) will help patient to fill the questionnaire. Part of the data related to resource utilisation and costs will be extracted from patient health records and/or will be given by care providers. For 6-month overall mortality rate these health records will also be consulted. Remuneration to participating patients will not be provided.

Collected data will be used for Multi-Criteria Decision Analysis (MCDA) and Cost-effectiveness (or CUA) analysis. The results of CEA will be compared to the results of the MCDA.

## 8.4. Germany

### 8.4.1. Casaplus

**Care programme**

In order to improve and reorganise health care services for elderly people with multiple chronic diseases the Casaplus case management programme was founded in April 2007. It addresses people older than 55 years, with multiple chronic conditions, and at high risk for hospital admissions within the next 12 months. The programme has a single point of entry, defined as a process in which persons insured with collaborating sickness funds are identified with a specifically developed screening software calculating their risk of hospitalization. The overall aim of the programme is to provide comprehensive, easy accessible and high quality case management. More precisely, the programme aims at reducing avoidable hospital admissions through preventive case management. Trained case managers assess, inform, support, and monitor the well-being of the enrolled elderly, multi-morbid persons. As part of the Casaplus care team, nursing professionals visit enrolled patients at home to assess possible risks of falling, social risks, and risks of malnutrition. Additionally, other professionals and informal care givers are consulted if necessary.

**Study design**

Casaplus will mainly be evaluated using a quasi-experimental regression discontinuity design (RDD). The intervention group consists of patients eligible for participation in Casaplus. These patients have to meet the inclusion criterion of a predicted risk for hospital admissions within the next 12 months of ≥ 0.5 and have to be 55 years or older. To avoid selection bias, both participants and non-participants (who would have been eligible), will be included in the intervention group (intention-to-treat). The control group consists of persons insured in the BKK (Betriebskrankenkassen) insurance funds aged between 50 and 55 years which is right below the official inclusion criteria (aged 55 years and above). The inclusion criteria of an estimated likelihood for hospitalization of ≥ 0.5 also applies for the control group. Both the control and the intervention group will be recruited via claims data provided by the statutory

141

BKK health insurance funds. This data will also be used for assessing the outcomes for the evaluation. Intervention effects are estimated by comparing the outcomes of the intervention group with the control group using regression techniques. If it turns out that there are significant differences between intervention and control group before the intervention started, propensity score matching could be used to ensure comparability between both groups.

In a further evaluation it is planned to evaluate patient reported outcomes such as experience of care and subjective well-being. The study design is planned as an uncontrolled before-after study measuring patient satisfaction before and after the enrolment to the programme using survey data of participating patients. However, the details of this evaluation are currently under discussion with the programme managers.

### 8.4.2. Gesundes Kinzigtal (GK)

**Care programme**

The Gesundes Kinzigtal (GK) model was founded in 2005 and is situated in the State of Baden-Württemberg, in the rural area of Southwest Germany. The GK model pursues a population-based approach that organises care across all health service sectors and indications. The GK model is designed around the "Triple Aim" approach: improving the health of the population in the Kinzigtal region, improving the individuals experience of care and at the same time reducing the per capita costs of care. The overall aim is to foster patient self-management and enhance shared decision-making with individual care plans and shared goal setting agreements between the physicians and the patients. After a person enrols in the GK model, a comprehensive check-up, including e.g.: medication, medical history, resources etc. by a GP follows. If a person is classified as being at risk, an individual treatment plan is developed accordingly. Furthermore, based on the completion of a questionnaire regarding their general health situation, further treatment goals are jointly agreed upon. GK offers tailor-made care programmes for enrolled persons. The system-wide access to electronic health records (HER) enables information exchange, transparency and an improvement of the quality of care. The population health is improved by a shared health gain approach and a shared savings contract.

**Study design**

The overall methodological approach for evaluating the GK is a quasi-experimental controlled study. The aim is to evaluate whether GK leads to changes in population health, quality of care and resources utilisation for people living in the Kinzigtal region compared to a control group living in other comparable regions. The study draws on pseudonymized health insurers' administrative claims and diagnosis data provided by the AOK Baden-Württemberg for the years 2005-2015/16. The base year of the study, 2005, is the first year of the intervention. The intervention group is a cohort of all insured of the AOK Baden-Wuerttemberg living in the intervention region (Kinzigtal) between 2006 and 2015. Insured do not have to be enrolled in the GK to be part of the intervention group as the effect of the population based programme GK on the whole population living in the target area will be evaluated. The control group will be recruited via a sample drawn from insured persons in AOK Baden-Wuerttemberg who are not living in the Kinzigtal (max. 500,000 insured). Regarding the statistical analysis there are two potential solutions to compare outcomes between intervention and control group. One option is to compare outcomes using regression analyses and controlling for potential confounders. The second alternative is 1:1 propensity score matching and comparing outcome variables using Pearson chi-square and t-tests.

For evaluating patient reported outcomes, such as satisfaction with care, subjective well-being, and patient reported experience with care an observational study design (trend study) will be used. The intervention group consists of a random sample (n ~ 3.000) of all enrolled participants in the GK. Several surveys will be used to compare outcomes of the intervention group with persons who are not living in the Kinzigtal region. The "Weisse Liste" survey will be used to compare the outcome "patient satisfaction with ambulatory care physicians" of the GK population with the German population. EQ-5-D valuation studies for Germany will be used to compare quality of life with the German population.

## 8.5. Hungary

### 8.5.1. OnkoNetwork

**Care programme**

OnkoNetwork is a patient pathway management system based on clinical protocols, which ensures quality and equal access for cancer patients to timely cancer diagnosis and treatment initiation. The target group of OnkoNetwork is comprised of adult patients with new suspect or new diagnosis of solid tumours in the catchment area of Complex Oncology Centre at Kaposvár, with or without chronic comorbidities. The programme offers priority status to both cancer-related and comorbidity-related healthcare services for the enrolled patients. Selected non-physician staff members (assistants and administrators) were trained and empowered in all departments of the Centre to schedule the necessary diagnostics, and to record relevant data in a custom patient path monitoring and management IT system (OncoLogistic). OnkoNetwork Office was established with the following tasks: patient path management and monitoring in the diagnostic phase; collection and overview of documents supporting the Tumour Board; and continued patient path management after Tumour Board decision. Two supervisor physicians have been appointed to negotiate with department physicians on OnkoNetwork related issues that could not be solved by the OnkoNetwork Office. The supervisor physician may request clarification and justification for missing or delayed diagnostics, without interfering with medical decisions of department physicians.

**Study design**

The study design involves prospective collection of patient reported outcomes as well as a retrospective analysis of medical records. Results from prospective and retrospective data collection will be combined and further assessed in a Multi-Criteria Decision Analysis and in cost-effectiveness analyses, conforming the harmonized requirements of the SELFIE project and also supporting evidence-based national decisions on the programme's scale-up and/or reimbursement.

Prospective data collection is planned for all patient reported items of the SELFIE questionnaire pertaining to health/well-being, experience, and costs, in a prospective, longitudinal, non-interventional comparative cohort study. After receiving their informed consent, the SELFIE questionnaire will be distributed to all enrolled patients at the first patient visit following the Tumor Board meeting. Patients will be asked to fill and return questionnaires within a few days by mail or in person, before starting cancer treatment phase. The investigated patient population in the prospective study relates to pre-selected types of solid tumors, as this allows for matched comparisons by primary tumor site. Primary tumor locations have been selected based on incidence in Hungary, frequency of regional progression or distant metastases at the time of diagnosis, and poor 5-year survival in regional or distant metastases. The intervention group to be investigated consists of all adult patients with new suspect or new diagnosis of lung, colorectal, pancreas, or stomach cancer. The intervention group patients will be recruited from two locations: 1) the catchment area of the Móritz Kaposi County Hospital in the period of April - September 2017, and 2) from the catchment area of Bács-Kiskun County Hospital in the period of December 2017 - March 2018 (assuming that OnkoNetwork will be introduced there in August – November 2017). The control group is defined as adult patients with new suspect or new diagnosis of lung, colorectal, pancreas, or stomach cancer in the catchment area of the Bács-Kiskun County Hospital, occurring in the time period of April - July 2017, when OnKoNetwork is not yet being implemented. The collected data follows the structure and content of the SELFIE questionnaire developed for palliative care/oncology programmes; supplemented with patient characteristics to control for these covariates in multivariate regression models.

The retrospective analysis has a longitudinal, comparative cohort study design. Study cohorts are defined retrospectively in the Hospital's medical database. Patients will not be recruited, informed, or contacted in any way in this retrospective non-interventional study. The intervention cohort is defined as the OnkoNetwork target population occurring in the period of December 2015 – November 2016 in the catchment area of the Móritz Kaposi County Hospital. A historic control cohort is selected from the same hospital in the period of September 2014 – August 2015. We plan to control for the relevant patient characteristics by multivariate

regression models in populations matched only by primary tumour site. To compensate for the heterogeneity of the investigated population, all statistical models will be specific to cancer type (lung, pancreas, or stomach) and will be adjusted to cancer stage at baseline, patient age, and gender. Sub-group analyses in both the intervention and control groups will be conducted based on cardiovascular, hypertension and diabetes comorbidities. The analyses will also consider the available socio-economic status parameters and smoking habits. Retrospective data collection is planned for the assessment of timely access to care, patient survival, costs claimed by the Hospital towards the national healthcare payer, and exploratory clinical outcomes such as tumour size.

## 8.5.2. Palliative Care Consult Service (PCCS)

**Care programme**

This programme is a local project at the Medical Centre of the University of Pécs. The Palliative Care Consulting Service (PCCS) team was established to offer interdisciplinary help to clinicians and to coordinate the complex care of terminally ill patients within the acute hospital and across healthcare providers. PCCS provides professional support for mainly cancer patients and their family members. The programme started in 2013 and has provided different health and social care services to almost 1000 patients.

Patient care is provided by a dedicated, trained and interdisciplinary team that is in close collaboration with other professionals from the hospital. The team is available upon request for their services and for consultations with the treating physicians and/or specialists. At the bedside of the patient the treating physician and members of the mobile team consult about the condition and the reasons for providing palliative care. The reason for requesting consultation has to be selected in the system which can be 1) managing home-based hospice-palliative care, 2) transferring the patient to another hospice institute 3) start the palliative care at the ward: psychosocial support, mental care, pain relief or other symptom management. The coordinator of the mobile team defines the further hospice/palliative somatic care and the psychological/social care after discussion with the treating physician, the patient and the family members.

**Study design**

A prospective comparative cohort study will be applied that aims to support a comprehensive evaluation of PCCS using the MCDA framework developed by the SELIFE consortium. The enrolled population in this study has malignant cancer requiring complex levels of support at the Internal Medicine Clinic and the Oncology Clinic. Inclusion criteria are malignancies (patient diagnosed with ICD C-code) and weak performance status (based on Karnofsky scale) recorded in hospital medical system. Exclusion criteria were also defined in order to achieve higher comparability of groups: short length of stay ≤ 3 days, hospitalisation was for routine chemotherapy, patient is able to work. The intervention group consists of patients receiving palliative care from the PCCS team during the hospital stay. The control group receives usual care, patients in this group do not have an interaction with the PCCS team. SELFIE questionnaire data will be collected with the help of trained medical professionals on the day of discharge. Data will be obtained about physical functioning, psychological well-being, pain and other symptoms, general health, patient-centeredness and compassionate care. Presumably, PCCS will have impact on these indicators during hospital stay. Further data on enjoyment of life, social relationships and participation, resilience, continuity of care and care utilisation will be obtained on the phone or personally by a trained researcher. These indicators will be captured with a two-week delay (post discharge) to observe the after-effect of the PCCS team's activity. Informed consent by all involved patients will be filled in upon discharge with the help of dedicated medical professionals. Data collection will be conducted between April 2017 - March 2018. A supplementary cost data collection will also be applied from both national health care payer and provider perspective. The former will include the total payment of the National Healthcare Fund in a 90-day period post discharge obtained from the payer's database. The latter will include the direct cost of care including diagnostic imaging, laboratory test, other interventions, pharmacy utilisation, hospitalization cost (ICU and normal ward) and the cost of the PCCS team's intervention obtained from hospital database. To control for covariates such as patient characteristics or physicians requested the PCCS team, multivariate regression models will be used in the prospective study.

A retrospective cohort study will also be conducted. The study aims to provide more generalizable information for hospital management to support decision-making on the implementation and operation of hospital-based palliative care consult programmes in Hungary by increasing the number of patients involved for cost data analyses. Due to the limited availability of good quality data, inclusion is restricted to patients with distant metastases. Exclusion criteria are short length of stay (≤ 3 days) and previous hospitalization for routine chemotherapy. Similar definitions of intervention and control groups will be obtained as the prospective study. Direct cost of care including diagnostic imaging, laboratory test, other interventions, pharmacy utilisation, hospitalisation cost (ICU and normal ward) and cost of PCCS team's intervention will be measured retrospectively. Additionally, average length of stay average length of ICU stay will be obtained. Data will be generated from the hospital medical system. To control for covariates such as patient characteristics or physicians requested the PCCS team, multivariate regression models will be used in the retrospective study.

## 8.6. Norway

### 8.6.1. Learning network for whole, coordinated and safe pathways (Learning networks)

**Care programme**

Learning networks are developed in collaboration between KS (Norwegian association of local and regional authorities) and the Norwegian Knowledge Centre for the Health Services. The aim is to improve pathways in the municipalities for older patients with recognised functional impairment, being new users of municipal home care service or short term stay in nursing home. Assessment of patient functioning, focus on ability rather than impairment, follow up by what matters to the patient, a designated nurse-coordinator, and early involvement of patients' GP are core elements of the programme. The programme has been implemented in 35 municipalities so far (2013-2015). The core team (GP and nurse-coordinator) is extended with other professionals (e.g., physiotherapist, social worker) if appropriate. So far, no scientific evaluation of the Learning network has been conducted, but the idea and practice of generic patient pathways is based on previous empirical research and development of the "Orkdal model".

**Study design**

The empirical evaluation will be conducted using two study designs. A retrospective cohort study based on data from national registers will aim at determining whether Learning networks contribute to improved patient health and well-being and to improved resource utilisation. A prospective cohort study will be based on data from surveys, aiming at determining whether Learning networks contribute to improved health and well-being and to improved experience of care.

For both study designs we will use data from 10 municipalities in the Agder counties comprised by the Learning networks during the period 2013-2015 (intervention group). Control groups will be established in 10 similar municipalities that are not yet enrolled in the Learning networks.

Data for the retrospective study will be obtained from Statistics Norway (Population, income and education registers), the Norwegian Health Directorate (Norwegian Patient Register and Control and Payment of Health Reimbursement Database), and the Norwegian Institute of Public Health (Norwegian Prescription Database). We will compare key outcome variables for patients in intervention municipalities and 10 control municipalities (propensity score matching), controlling for demographic and socioeconomic patient characteristics (sex, age, housing, income, education, employment, social assistance), comorbidities and hospital type. Outcome variables will include hospitalizations, mortality, use of primary and secondary health care, and use of prescription medicine.

For the prospective study the same intervention municipalities will be approached for a survey among eligible patients, and a comparable patient group in the control municipalities (propensity score matching). We will also use information (schemes) already collected by the municipalities, i.e., functional assessment at baseline and after 5 weeks (PSFS or COPM scale), patients' experience with follow-up (interview or survey), and degree of implementation of programme among eligible patients.

The collected data from the cohort studies will be brought together in a Multi-Criteria Decision Analysis (MCDA). In an MCDA multiple outcomes can be included, in the current study pertaining to health/well-being, experience, and costs. Further, outcomes are weighted on the basis of the chosen perspective that is taken. In the SELFIE project, 5 sets of weights, from 5 stakeholder perspectives, will be obtained in a parallel study and applied to the data collected: Patients, Partners (i.e., informal caregivers), Professionals, Payers, and Policy makers.

## 8.6.2. Medically Assisted Rehabilitation (MAR Bergen)

**Care programme**

MAR is an interdisciplinary specialised treatment programme for opioid addiction, where requisition of addictive medicine in a fixed dose (substitution treatment) is a partial measure in an overall rehabilitation pathway. The programme was developed in collaboration between psychiatric and addiction specialist healthcare services and municipal health and social care. The MAR Bergen model is of particular interest by its emphasis on low-threshold access. The

patients are in a phase where survival is the key objective, and it aims at patient support to promote change of opioid intakes. The patients have typically several diagnoses, mental and physical, in addition to addiction, and they often have problematic living conditions. The substitution treatment (i.e., requisition of addictive medicine in a fixed dose) is care-oriented, and considered as an 'entrance ticket' to further treatment. There has so far been no systematic evaluation of MAR Bergen, however, some medical consequences of the intervention have been investigated.

**Study design**

We will use two different research design approaches depending on possibility of finding relevant data for key outcome variables; a retrospective cohort study based on data from national registries, and prospective data collection through surveys among MAR patients. The retrospective cohort study will provide data for analysing mortality, use of health care services (hospitalisations, primary care) and use of prescription drugs. Data will be collected from Statistics Norway (Population, income and education registers), the Norwegian Health Directorate (Norwegian Patient Register and Control and Payment of Health Reimbursement Database), and the Norwegian Institute of Public Health (Norwegian Prescription Database). Some data are available from annual status reports from the Norwegian Centre for Addiction Research (Seraf). This will provide possibilities to do a before-after study with control group. Propensity score matching or similar matching procedure will be used to match MAR patients in Bergen to similar patients in the MAR population in Oslo.

A prospective study will be performed using surveys including patients registered as MAR patients in Bergen and Oslo. Information collected though surveys is related to outcome variables that are not available in public registers. Patients' perceptions of own health and satisfaction with health care services will be investigated through questionnaires distributed twice to MAR patients in Bergen, and to a control group (Oslo). A sample of MAR-patients will be drawn from patients registered with the programme. We will also collect patient background information in the questionnaires. Questionnaires will be completed when visiting a MAR clinic. No before-after analyses will be done, as data will be collected in patients already

'enrolled'. However, longitudinal analyses are possible because the outcomes are repeatedly assessed. Data collection will be done in cooperation with another survey being conducted by the University of Bergen to the same patient group.

The collected data from the cohort studies will be brought together in a Multi-Criteria Decision Analysis (MCDA). In an MCDA multiple outcomes can be included, in the current study pertaining to health/well-being, experience, and costs. Further, outcomes are weighted on the basis of the chosen perspective that is taken. In the SELFIE project, 5 sets of weights, from 5 stakeholder perspectives, will be obtained in a parallel study and applied to the data collected: Patients, Partners (i.e., informal caregivers), Professionals, Payers, and Policy makers.

## 8.7. Spain

### 8.7.1. Management of complex chronic patients (CCP) requiring specialised care

**Care programme**

The protocol applied in the AISBE programme addresses five aims. Firstly, implementation of two integrated care interventions using a collaborative adaptive case management approach (ACM): i) Community-based management of CCP (Home hospitalisation & transitional care); and, ii) Integrated care for patients under long-term oxygen therapy (LTOT). The second aim is the adoption of information and communication technologies (ICT) required to support collaborative ACM. Thirdly, the programme evaluates the impact of enhanced clinical health risk assessment and stratification. The fourth aim is assessment of healthcare value generation of the services, both during the deployment phase and after regional scale-up of the novel services. Finally, the current study generates a roadmap for regional adoption of the CCP program. The protocol emerges from previous studies that have generated solid evidence on efficacy of the interventions as well as their high potential for health value generation.

**Study design**

The evaluation of the CCP protocol is closely aligned with the activities developed within the frame of the Nextcare (Innovation in Integrated Care Services for Chronic Patients) program. It is a regional deployment initiative in Catalonia (ES) belonging to the RIS3 (Research and Innovation Strategies for Smart Specializations) programme (2016-2019).

The main AISBE study is to assess patients included in the home hospitalization & transitional care programme (intervention group) at the Hospital Clinic. They will be recruited sequentially, one out of three consecutive patients (1:3 ratio), during the study period. The control group will include the same number of patients following identical criteria (1:3 ratio) among those admitted in the Hospital Sagrat Cor within the same healthcare sector. The two groups will be matched by age, sex and GMA (adjusted morbidity grouper) scoring.

The integrated care intervention on patients under LTOT will be compared with standard care through registry data obtained from the Catalan Health Surveillance System

(CHSS). At programme level, we will use key performance indicators (KPI) addressing four dimensions, namely: (i) clinical aspects, (ii) maturity of implementation, (iii) impact of the intervention on the healthcare system including cost analysis; and, (iv) interactions between healthcare and social support following local guidelines generated by the PIAISS (programme aiming at fostering integration between healthcare and social support at regional level). In SELFIE we will use the collected data to perform an MCDA.

### 8.7.2. Badalona Serveis Assistencials (BSA)

**Care programme**

Badalona Serveis Assistenials (BSA) serves a population of 236,000 citizens living in a suburban area of Barcelona. The programme puts the person at the centre, integrating healthcare and social services and providing 24-7-365 emergency support. The target candidates for inclusion in the programme are frail elderly citizens often with several chronic disorders. They usually live at home showing unmet needs for care and support which in turn makes them at high risk of exclusion due to illness or disability of any kind. The BSA services also coach informal (family members, friends, neighbours) and formal (professionals) caregivers providing assistance to the patient on a regular basis. The core skills are management of multi-morbidity and cognitive impairments. The main aims of the programme are to promote independent living, offering support to prevent institutionalisation and avoidable hospitalisation.

The programme is composed of various medical services – Case Management Nurse (CMN), Medical Attention at Home (AtDom), Hospital at Home (HaH), Nursing Homes Attention (NH AtDom) and Palliative Attention (PADES) – which are integrated with social care services – Telecare, cleaning services, meals at home, home care services, GPS localizer and home rehabilitation. The uniqueness of the programme is the integration between healthcare and social support allowing provision of innovative services that promote independent living and proactive care.

**Study design**

We are planning to assess patients included in the home-care programme (intervention group) at BSA. They will be recruited sequentially, one out of three consecutive patients (1:3 ratio), during the study period. The control group (standard care) will include the same number of patients following identical criteria (1:3 ratio) among those treated by the other main provider (ICS, Institut Catalan of Health) in the same healthcare sector. The two groups will be matched by age, sex and GMA scoring.

At programme level, we will use key performance indicators (KPI) addressing four dimensions, namely: (i) clinical aspects, (ii) maturity of implementation, (iii) impact of the intervention on the healthcare system including cost analysis; and, (iv) interactions between healthcare and social support following local guidelines generated by the PIAISS (programme aiming at fostering integration between healthcare and social support at regional level). In SELFIE we will use the collected data to perform an MCDA.

## 8.8. The United Kingdom

### 8.8.1. Salford Integrated Care Programme (SICP) / Salford Together

**Care programme**

Salford's programme has been (until recently, when this was expanded to all adults) exclusively aimed at the elderly (over 65s), and consists of three broad interventions: 1. Multidisciplinary groups (MDGs) – case management of the highest-risk patients by neighbourhood groups (this intervention is most likely to be directly relevant to multi-morbid patients); 2. Community assets – investment in 'community assets' (e.g., community groups, religious groups, charities, sports groups) to promote social interaction and active lifestyle; 3. Centre of contact (/health coaching) – a centralised telephone hub to help with navigating services and self-management. Further organisational changes towards developing an Accountable Care Organisation (ACO) have taken place from July 2016 onwards. These organisational changes are what the programme managers estimate will contribute most significantly to their primary aims in terms of reducing secondary care admissions and costs (above and beyond the service delivery changes detailed above, which have been implemented prior). By investing in programmes like community assets which have the opportunity to impact on all citizens in the local area, and organisational changes which aim to shift the accountability of the health and care system to put the person at the centre and thus better incentivise prevention-oriented tasks, the programme can be considered a population health management model, serving all health and risk levels of adults within its boundaries.

**Study design**

Population health management models are complex interventions, with multiple components and targeting a heterogeneous population at different levels of risk. They aim to change an entire health system for those in the geographical area that they cover. The complexity science literature suggests that complex adaptive systems, like health systems, exhibit certain properties that are important to consider when evaluating. Most importantly, these systems exhibit emergence, that "the whole is more than the sum of its parts", i.e., evaluating

effectiveness of each component of the intervention individually will not teach us about overall effectiveness of the programme as a whole, as the effectiveness is not simply additive. In addition, the major component of the Salford integrated care programme that applies to complex multi-morbid patients is MDG case management, and is very widespread in England (and internationally) making it difficult to obtain a control group for this specific intervention at this population level. However, what makes this programme unique is the additional components that are presently applied to the slightly lower risk population with the aim of being preventative, i.e., the addition of community assets and centre for contact on top of MDG case management. Therefore, our analysis focuses primarily on assessing the programme as a whole at its higher intended level, i.e., the population-level that it is aimed at (as the programme is delivered to a geographically defined population), comparing to control groups from outside of the area that is implementing it.

*Method 1: Population-level analysis*

We will use national routinely collected datasets (the GP Patient Survey [GPPS] for patient experience data, and Hospital Episode Statistics [HES] for admissions and cost data) to compare outcomes from Salford Clinical Commissioning Group (CCG) compared to other CCG areas in England (propensity score matched where necessary). Quasi-experimental methods will allow us to compare the difference from before to after the implementation of the integrated care programme in Salford and control areas. Where possible, we will use a regression discontinuity design using the inclusion age of 65+ for integrated care as the arbitrary cut-value, where the data allows us to meet the requirements of this analysis method. Where necessary we will substitute or complement this approach with a difference-in-differences analysis. When possible, we will use subgroup analysis to look at effects of the programme on different stratifications of multi-morbid patients. In addition, we will compare outcomes of the programme when only service delivery changes were rolled out, to when these were later supported by organisational integration changes.

*Method 2: Intervention-level analysis*

The main intervention which likely applies to complex multi-morbid patients in Salford is the MDG case management component. As highlighted above, it is difficult to measure impact of this component comparing to other populations in England, as the intervention is extremely widespread. However, the intervention was rolled out in three waves within Salford, meaning we can utilise this natural experiment to attempt to look at the impact of this specific intervention. To build on the previous CLASSIC evaluation that has taken place in the area, we will also look at effects of the programme on different stratifications of multi-morbid patients. We will use the CLASSIC questionnaire for patient experience data (where possible, and complemented by practice-level data from the GPPS if necessary), and HES data for utilisation and cost.

## 8.8.2. South Somerset Symphony Programme

**Care programme**

The initial focus of South Somerset's programme was primarily on multi-morbidity, and the programme consists of two broad service delivery interventions (with a commonality of 'health coaching'): 1. Complex care hubs – an 'extensivist' GP model with GPs located in a hospital hub and managing, as part of a multidisciplinary hub team, the most complex patients (i.e., patients with 3 or more multiple chronic conditions, although this has been adapted, as in practice this broad definition was found to be inadequate for predicting complexity, with a 'Symphony score' incorporating clinical data plus clinical judgement now used in addition to target patients); 2. Enhanced primary care (EPC) – co-location of health coaches in all (but one, which chose not to take part) GP practices in the area, to assist with disease self-management and prevention. The EPC model has been rolled out to 18 of the 19 GP practices in the area over three waves. The area is also particularly keen to implement organisational integration changes, including formation of an Accountable Care Organisation (ACO) model, and have taken steps towards doing this with Vanguard funding and formation of a limited company (located in the hospital) which has taken over the contracting of a small number of the GP practices in the area to date. The programme can therefore be considered a population health management model, serving all health and risk levels within its geographical boundaries.

**Study design**

As another population health management model, like for the analysis detailed for the Salford programme, our analysis focuses primarily on assessing the programme at its higher intended level, i.e., the population-level that it is aimed at (as the programme is delivered to a geographically defined population), comparing to control groups from outside of the area that is implementing it.

*Method 1: Population-level analysis*

We will use national routinely collected datasets (the GP Patient Survey [GPPS] for patient experience data, and Hospital Episode Statistics [HES] for admissions and cost data) to compare outcomes from South Somerset Clinical Commissioning Group (CCG) compared to other CCG areas in England (propensity matched where necessary). We will use quasi-experimental methods such as difference-in-differences and regression discontinuity design (attempting to use distance to geographical boundary as a running variable) to compare population-level outcomes before and after the integrated care programme was implemented for the population in South Somerset to a control population that was not exposed to the programme. When possible, we will use subgroup analysis to look at effects of the programme on different stratifications of multi-morbid patients.

*Method 2: Intervention-level analysis*

The South Somerset programme consists of two broad interventions that both relate to multi-morbid patients (at different risk levels). For assessing the complex care hub, we will explore whether the programme management can identify the patients that have been through the care hub to date, and compare outcomes with propensity score-matched patients (from within the area, who we know have not received this additional MDT case management). For assessing the EPC model individually, we will use the phased GP practice roll-out (rolled out to practices in three waves) as a natural experiment. For both interventions, we will also look at effects of the programme on different stratifications of multi-morbid patients. We will use the locally available data supplied by the site to conduct the analyses.

# 9. Concluding remarks

In this deliverable report the Multi-Criteria Decision Analysis framework that will be applied in SELFIE was described. A general introduction to MCDA was given, and the standard steps followed in an MCDA were described. We subsequently described our chosen MCDA method and weight-elicitation method in SELFIE: multi-attribute utility theory (MAUT) and discrete choice experiments and swing weighting, respectively. Thereafter, we described in detail the design of the weight-elicitation questionnaire. Furthermore, possible study designs were described that can be used to obtain performance scores of integrated care programmes as compared to usual care. Lastly, we presented brief summaries of the proposed study designs per country and per programme in SELFIE.

As described in **Chapter 1** of this report, the work done in WP4 forms the foundation for the work to be done in WP5. In WP5 the proposed study designs will be made definite and formalised in ethics committee requests. Subsequently, data collection will take place both for the weights as well as for the performance scores. At the end of WP5 these weights and performance scores will be brought together in the MCDA framework proposed in the current report.

# 10. Appendix

## 10.1. Appendix 1: 17 integrated care programmes

| | |
|---|---|
| *Netherlands (NL)* | **1.** Proactive Primary Care Approach for Frail Elderly (U-PROFIT)<br>**2.** Care Chain Frail Elderly (previously called KOMPLEET)<br>**3.** Better together in Amsterdam North (BSiN) |
| *Austria (AT)* | **4.** Health Network Tennengau (Gesundheitsnetzwerk Tennengau)<br>**5.** Sociomedical Centre Liebenau (Sozialmedizinisches Zentrum Liebenau) |
| *Croatia (HR)* | **6.** GeroS System<br>**7.** Palliative Care System |
| *Germany (DE)* | **8.** Casaplus<br>**9.** Gesundes Kinzigtal |
| *Hungary (HU)* | **10.** Onconetwork<br>**11.** Palliative Care Consulting Service (Mobile) Team |
| *Norway (NO)* | **12.** Learning network<br>**13.** Medically Assisted Rehabilitation (MAR) Bergen |
| *Spain (ESP)* | **14.** Badalona Serveis Assistencials (BSA)<br>**15.** Barcelona Esquerre (AISBE) |
| *The UK* | **16.** South Somerset Symphony Programme<br>**17.** Salford – Salford Integrated Care Programme (SICP)/ Salford Together |

## 10.2. Appendix 2: SELFIE outcomes

| Outcomes for integrated care for persons with multi-morbidity | | | | |
|---|---|---|---|---|
| | **Core set outcomes** | **Programme-type specific outcomes** | | |
| | | *Population health management* | *Frail elderly* | *Palliative and oncology* | *Problems in multiple life domains* |
| **Health & well-being** | Physical functioning | Activation & engagement | Autonomy | Mortality | Self-sufficiency |
| | Psychological well-being | | | Pain and other symptoms | |
| | Social participation/ relationships | | | | |
| | Resilience | | | | |
| | Enjoyment of life | | | | |
| **Experience** | Person-centeredness | | Burden of medication | Compassionate care | |
| | Continuity of care | | Burden of informal caregiving | Timely access to care | |
| | | | | Preferred place of death | |
| | | | | Burden of informal caregiving | |
| **Costs** | Total health- and social care costs | Ambulatory care sensitive hospital admissions | Living at home | | Justice costs |
| | | Hospital re-admissions | Falls leading to hospital admissions | | |

## 10.3. Appendix 3: Methods to standardise performance scores

**1. Relative standardisation**

Using this method, the standardised scores are dependent on the observed performance values of the alternatives being compared. Referring back to **Table 3**, the performance scores have been standardised using the formula below, in **Table A3.1**.

$$S_{aj} = \frac{x_{aj}}{\left(x_{aj}^2 + x_{bj}^2\right)^{1/2}} \qquad\qquad S_{bj} = \frac{x_{bj}}{\left(x_{bj}^2 + x_{aj}^2\right)^{1/2}}$$

$S_{aj}$ = standardisation of the performance score ($x$) of alternative $a$ on criteria $j$ relative to the combined performance score ($x$) of both alternatives $a$ and $b$ on criteria $j$.

$S_{bj}$ = standardisation of the performance score ($x$) of alternative $b$ on criteria $j$ relative to the combined performance score ($x$) of both alternatives $a$ and $b$ on criteria $j$.

$x$ = performance score (on the natural range/scale)
$a$ = alternative a
$b$ = alternative b
$j$ = criteria j

Whereby for reverse coded criteria (which should attain a lower standardised score the higher the non-standardised score) we have:

$$S*_{aj} = 1 - S_{aj}$$

**Table A3.1**: *SELFIE example of relative standardising performance scores*

| | | Alternatives | |
| | | Integrated care | Usual care |
|---|---|---|---|
| | | **Performance** : **Standardised** | |
| Resilience | Indicator a *(0-100)* | 70 : **0.74** | 63 : **0.67** |
| Physical functioning | Indicator b *(0-100)* | 60.50 : **0.68** | 65.00 : **0.73** |
| Person-centeredness | Indicator c *(1-5)* | 4 : **0.80** | 3 : **0.60** |
| Continuity of care | Indicator d *(1-7)* | 5.5 : **0.77** | 4.5 : **0.63** |
| Total health- and social care costs | Indicator e (*health care costs – € per capita per year*)* | 9200 : 0.72 = **0.28** | 9000 : 0.70 = **0.30** |
| | Indicator f (*social care costs – € per capita per month*)* | 1500 : 0.76 = **0.24** | 1300 : 0.66 = **0.34** |

*Note: *Because for costs higher values are worse, 1 minus the standardised performance score is taken.*

## 2. Ranging standardisation

Using this method, the standardised scores are dependent on the range of the natural scale, and not on the observed performance values of the alternatives being compared. For this form of standardisation, a selection first needs to be made between two methods of determining the reference points (i.e., range), global or local. In global scaling the lowest value (e.g., 0) is the 'absolute worst' and the highest value (e.g., 1 or 100) the 'absolute best'. In local scaling the lowest value (e.g., 0.2) is the performance score of the current alternative that scores worst, and the highest value (e.g., 0.9) of the alternative that scores best. An advantage of using global scaling is that new alternatives can later be added (e.g., if another integrated programme for multi-morbidity will be additionally evaluated). A disadvantage, however, is that judgments of the absolute worst and best need to be made, which for example, for costs is an estimation. The formula and example below in **Table A3.2** use global scaling, thus absolute worst and best are taken as the natural range. For the costs this is an estimation.

$$S_{aj} = \frac{x_{aj}}{(Rn \,/\, Rs)}$$

$S_{aj}$ = standardisation of the performance score ($x$) of alternative $a$ on criteria $j$ on the global scale 0-1.

$x$ = performance score (on natural range/scale)

$a$ = alternative a

$R$ = range

$n$ = natural range/scale (min-max)

$s$ = standardised scale (global, 0-1)

**Table A3.2**: *SELFIE example of global ranging standardisation of performance scores*

| | Alternatives | |
|---|---|---|
| | *Integrated care* | *Usual care* |
| **Indicator** *(Natural range)* | **Performance : Standardised** | |
| Indicator a *(0-100 (best))* | 70 : **0.70** | 63 : **0.63** |
| Indicator b *(0-100 (best))* | 60.50 : **0.61** | 65.00 : **0.65** |
| Indicator c *(1-5 (best))* | 4 : **0.80** | 4 : **0.60** |
| Indicator d *(1-7 (best))* | 5.5 : **0.78** | 4.5 : **0.64** |
| Indicator e *(health care costs – € per capita p/y)* (↑ worse) Global min = 0, max = 30000 p/y) * | 9200 : 0.31 = **0.69** | 9000 : 0.30 = **0.70** |
| Indicator f *(social care costs – € per capita p/m)* (↑ worst) Global min = 0, max = 5000 p/m)* | 1500 : 0.30 = **0.70** | 1300 : 0.26 = **0.74** |

Note: *Because for costs higher values are worse, 1 minus the standardised performance score is taken.

Both relative and ranging standardisation methods have advantages and disadvantages. In relative standardisation, as with local ranging standardisation, standardised performance scores will change if alternatives are added. With global ranging standardisation a theoretical worst and best is included, that may in actuality not be relevant (e.g., do scores of 0 on the Barthel Index ever occur or have meaning?). The differences between the alternatives using relative as opposed to ranging standardisation are always at least greater because the difference in two performance scores can never be larger than the natural range. Thus using relative standardisation may increase the discriminative ability of the MCDA. In the unlikely situation

that within a case study the same criterion is measured with a different indicators, with different ranges, ranging standardisation still allows for a meaningful comparison to be made.

As the process of standardisation does not influence how data needs to be collected, in SELFIE, we can use both methods, calculate what differences are found, and explore why this is the case.

## 10.4. Appendix 4: Additional tables MCDA examples

**Appendix Table A4.1:** *SELFIE example MAUT aggregating weights and performance*

| | Weight | | Care alternatives | | Aggregated weight | | | |
| | | | Integrated | Usual | Integrated | | Usual | |
| Criteria | S1 | S2 | Standardised performance* | | S1 | S2 | S1 | S2 |
|---|---|---|---|---|---|---|---|---|
| Resilience | .30 | .20 | 0.70 | 0.63 | 0.21 | 0.14 | 0.19 | 0.13 |
| Physical functioning | .20 | .15 | 0.61 | 0.65 | 0.12 | 0.09 | 0.13 | 0.10 |
| Person-centeredness | .15 | .05 | 0.80 | 0.60 | 0.12 | 0.04 | 0.09 | 0.03 |
| Continuity of care | .25 | .05 | 0.78 | 0.64 | 0.20 | 0.04 | 0.16 | 0.03 |
| Health care costs | .05 | .30 | 0.69 | 0.70 | 0.03 | 0.21 | 0.04 | 0.21 |
| Social care costs | .05 | .25 | 0.70 | 0.74 | 0.04 | 0.18 | 0.04 | 0.19 |
| | | | | Sum | 0.72 | 0.70 | 0.65 | 0.69 |

*Note: *Standardised performance based on global ranging scaling (from **Table 6**). S1 = Stakeholder 1 (e.g., patient), S2 = Stakeholder 2 (e.g., payer).*

## 10.5. Appendix 5: Additional details defining levels – core set

**Table A5.1:** *SELFIE DCE criteria and criteria levels –* **Physical functioning**

<table>
<tr><td colspan="3" align="center"><strong>Physical functioning</strong></td></tr>
<tr>
<td>Questionnaire definition:</td>
<td colspan="2">Acceptable physical health and being able to do daily activities without needing assistance</td>
</tr>
<tr>
<td rowspan="2">SELFIE questionnaire:</td>
<td><em>Population health management & multi-problems:</em><br>General physical health (SF-36, physical functioning domain)</td>
<td><em>Frail elderly & palliative/oncology:</em><br>Activities of Daily Living (ADL) (Katz-15)</td>
</tr>
<tr>
<td><em>The following items are about activities you might do during a typical day. Does your health now limit you in these activities? If so, how much?</em><br><em>-Yes, limited a lot</em><br><em>-Yes, limited a little</em><br><em>-No, not limited at all</em><br><em>*Vigorous activities, moderate activities, lifting/carrying, climbing several/one flight of stairs, bending/kneeling/stooping, walking (more than a mile/ several blocks/one block), bathing or dressing</em><br><em>***Level of limitation (a lot – a little)</em></td>
<td><em>The following items are about activities of daily living. Please mark for each question whether you need help to do these activities.</em><br><em>-Yes</em><br><em>-No</em><br><em>*taking a bath or shower, getting dressed, toileting, sitting down and getting up from a chair, incontinence products, eating, using the telephone, shopping, preparing a meal, taking care of your house, travelling, taking your medications, handling your finances, brushing your hair or shaving, help walking about</em><br><em>***Whether someone is dependent or not (yes – no)</em></td>
</tr>
<tr>
<td>DCE levels:</td>
<td colspan="2"><strong>DCE attribute presented:</strong> Physical functioning<br><strong>*DCE definition presented:</strong> Defined as acceptable physical health and being able to do daily activities without needing assistance (e.g., getting dressed, setting down and getting up from a chair, taking your medications)<br>    1. <strong><u>Severely</u> limited</strong> in physical health and activities of daily living<br>    2. <strong><u>Moderately</u> limited</strong> in physical health and activities of daily living<br>    3. <strong><u>Hardly or not at all</u> limited</strong> in physical health and activities of daily living</td>
</tr>
</table>

*Differs slightly from questionnaire definition.

**Table A5.2:** *SELFIE DCE criteria and criteria levels –* **Psychological well-being**

| Psychological well-being | |
|---|---|
| Questionnaire definition: | Absence of stress, worrying, listlessness, anxiety, and feeling down |
| SELFIE questionnaire: | Mental Health Inventory (MHI-5) |
| | *For the following questions, please tick the box that best describes how things have been for you during the past month.* <br> *-All of the time* <br> *-Most of the time* <br> *-A good bit of the time* <br> *-Some of the time* <br> *-A little of the time* <br> *-None of the time* <br> *\*Happy person, felt calm and peaceful, very nervous person, felt downhearted and blue, felt so down in the dumps that nothing could cheer you up* <br> *\*\*\*Frequency over past month* |
| DCE levels: | **\*DCE attribute presented:** Psychological problems <br> **\*DCE definition presented:** The occurrence of stress, worrying, listlessness, anxiety, and feeling down. <br>     1. **Always, or mostly**, being <u>stressed</u>, worried, listless, anxious, and down. <br>     2. **Regularly** being <u>stressed</u>, worried, listless, anxious, and down. <br>     3. **Seldom, or never,** being <u>stressed</u>, worried, listless, anxious, and down. |

*Differs slightly from questionnaire definition and original attribute title.


**Table A5.3:** *SELFIE DCE criteria and criteria levels –* **Enjoyment of life**

| Enjoyment of life | | |
|---|---|---|
| Questionnaire definition: | Having pleasure and happiness in life | |
| SELFIE questionnaire: | *Population health management, frail elderly, multi-problems:* <br> ICECAP-O: Enjoyment & pleasure domain | *Palliative/oncology:* <br> Q-LES-Q, short version |
| | *-Can have all* <br> *-Can have a lot* <br> *-Can have a little* <br> *-Cannot have any* <br> *… of the enjoyment and pleasure in life that I want* <br> *\*\*\*Quantity* | *During the past month, how much of the time have you felt satisfied with your life?* <br> *-Never at all or never* <br> *-Rarely* <br> *-Sometimes* <br> *-Often or most of the time* <br> *-Frequently or all of the time* <br> *\*\*\*Frequency over time, how often.* |
| DCE levels: | **DCE attribute presented:** Enjoyment of life <br> **DCE definition presented:** Having pleasure and happiness in life <br>     1. **Not, or barely, having** <u>pleasure</u> and happiness in life <br>     2. **Having some** <u>pleasure</u> and happiness in life <br>     3. **Having a lot of** <u>pleasure</u> and happiness in life | |

**Table A5.4:** *SELFIE DCE criteria and criteria levels –* **Social relationships & participation**

| Social relationships & participation | |
|---|---|
| Questionnaire definition: | Having meaningful connections with others as desired |
| SELFIE questionnaire: | Impact on Participation and Autonomy: social life & relationships domain |
| | *The next questions are about the quality and frequency of your social relationships. We would like to know whether your health problems or disabilities affect your relationships.*<br>*-Very good*<br>*-Good*<br>*-Fair*<br>*-Poor*<br>*-Very Poor*<br>*\*talking to people close to me on equal terms, quality of my relationships, respect, relationships with acquaintances, respect I receive from acquaintances, chances of having an intimate relationship, chances of seeing people*<br>*\*\*\*How good things are (also chances) (very good – very poor)* |
| DCE levels: | **DCE attribute presented:** Social relationships & participation<br>**DCE definition presented:** Having meaningful connections with others as desired<br>    1. **Not, or barely, having** meaningful connections with others<br>    2. **Having some** meaningful connections with others<br>    3. **Having a lot of** meaningful connections with others |

**Table A5.5:** *SELFIE DCE criteria and criteria levels –* **Resilience**

| Resilience | |
|---|---|
| Questionnaire definition: | The ability to recover from or adjust to difficulties and to restore ones equilibrium. |
| SELFIE questionnaire: | Brief Resilience Scale (BRS) |
| | *Please respond to each item by marking one box per row*<br>*-Strongly disagree*<br>*-Disagree*<br>*-Neutral*<br>*-Agree*<br>*-Strongly agree*<br>*\*Bounce back, making it through stressful events, long to recover from a stressful event, hard for me to snap back when something bad happen, come through difficult times with little trouble, to take a long time to get over set-backs in my life*<br>*\*\*\*Extent of agreement.* |
| DCE levels: | **DCE attribute presented:** Resilience<br>**DCE definition presented:** The ability to recover from or adjust to difficulties and to restore one's equilibrium<br>    1. **Poor** ability to recover, adjust, and restore equilibrium<br>    2. **Fair** ability to recover, adjust, and restore equilibrium<br>    3. **Good** ability to recover, adjust, and restore equilibrium |

**Table A5.6:** *SELFIE DCE criteria and criteria levels –* **Person-centeredness**

| Person-centeredness | |
|---|---|
| Questionnaire definition: | Care that care matches an individual's needs, capabilities and preferences and jointly making informed decisions |
| SELFIE questionnaire: | The Person Centred Coordinated Care Experiences Questionnaire (P3CEQ), Experience of Person Centred Care domain. |
| | *First items about extent of agreement to statements:*<br>*-Not at all*<br>*-To some extent*<br>*-More often than not*<br>*-Always*<br>*Next items about having support (not at all – always) and information (not enough – too much)*<br>****Extent of agreement, not at all – always, not enough – too much.* |
| DCE levels: | **DCE attribute presented:** Person-centeredness<br>**DCE definition presented:** Care that care matches an individual's needs, capabilities and preferences and jointly making informed decisions<br>    **1. Not, or barely,** person-centred; i.e., care does not match an individual's needs, capabilities, and preferences, and decisions are not made jointly.<br>    **2. Somewhat** person-centred; i.e., care somewhat matches an individual's needs, capabilities, and preferences, and some decisions are made jointly.<br>    **3. Highly** person-centred; i.e., care matches an individual's needs, capabilities, and preferences, and decisions are made jointly. |

**Table A5.7:** *SELFIE DCE criteria and criteria levels –* **Continuity of care**

| Continuity of care | |
|---|---|
| Questionnaire definition: | Good collaboration, smooth transitions between caregivers, and no waste of time |
| SELFIE questionnaire: | Nijmegen Continuity Questionnaire (NCQ), Team and cross-boundary continuity domain and Client Perceptions of Coordination Questionnaire (CPCQ) |
| | *The following statements are about your experience of the cooperation between your care providers (e.g., between general practitioner and nurse practitioner or between general practitioner and medical specialists). Please indicate to what extent you agree/disagree with the statement or whether it is not applicable (N/A).*<br>*-Strongly agree*<br>*-Agree*<br>*-Neutral*<br>*-Disagree*<br>*-Strongly disagree*<br>*-N/A*<br>*\*Transfer info, work together, well connected, know what's going, have to wait too long.*<br>*\*\*\*Extent of agreement.* |
| DCE levels: | **DCE attribute presented:** Continuity of care<br>**DCE definition presented:** Good collaboration, smooth transitions between caregivers, and no waste of time<br>     **1. Poor** collaboration, transitions, and timeliness<br>     **2. Fair** collaboration, transitions, and timeliness<br>     **3. Good** collaboration, transitions, and timeliness |

**Table A5.8:** *SELFIE DCE criteria and criteria levels –* **Total health and social care costs**

| Total health and social care costs | |
|---|---|
| Questionnaire definition: | Total health and social care costs per participant |
| SELFIE questionnaire: | *iMTA questionnaire / registry data* |
| DCE levels: | **\*DCE attribute presented:** Costs<br>**\*DCE definition presented:** Total health- and social care costs per participant in the programme, per year<br>     **1. 5000 euros** per participant per year<br>     **2. 2000 euros** per participant per year<br>     **3. 500 euros** per participant per year |

*Differs slightly from questionnaire definition and original attribute title.

## 10.6. Appendix 6: Additional details defining levels – programme-type specific criteria

### 10.6.1. Population health management

**Table A6.1.1:** *SELFIE swing weighting criteria and criteria levels –* **Activation and engagement**

| Activation and engagement | |
|---|---|
| Questionnaire definition: | Taking on the role of managing one's own health and care |
| SELFIE questionnaire: | *Please indicate to what extent you agree/disagree with the statement or whether it is not applicable (N/A).* <br> *-Strongly disagree* <br> *-Disagree* <br> *-Agree* <br> *-Strongly agree* <br> *-N/A* <br> *\*responsible for managing my health condition, active role in my own health care,* <br> *confident that I can take actions, know what each of my prescribed medications does, confident that I can tell when I need to go get medical care and when I can handle a health problem myself, confident I can tell my health care provider concerns, confident that I can follow through on medical treatments I need to do at home, understand the nature and causes of my health condition, know the different options available, maintain the lifestyle changes, prevent further problems, confident I can figure out solutions, confident that I can maintain lifestyle changes* <br> *\*\*\*Extent of agreement to statements about management of own situation* |
| Swing – worst <br> Swing – best | **Attribute presented:** Activation & engagement <br> **Definition presented:** Taking on the role of managing one's own health and care <br> **Worst:** Unsuccessful in managing your own health and care <br> **Best:** Actively taking on the role of managing your own health and care |

**Table A6.1.2:** *SELFIE swing weighting criteria and criteria levels –* **Ambulatory care sensitive hospital admissions**

| Ambulatory care sensitive hospital admissions | |
|---|---|
| Questionnaire definition: | Proportion of hospital admissions that could have been avoided by better ambulatory care |
| SELFIE questionnaire: | *Registry data.* *\*\*\*Proportion of emergency room hospital admissions. \*\*\*Also defined as the proportion that is AVOIDABLE.* |
| Swing – worst Swing – best | **\*Attribute presented:** Avoidable hospital admissions **\*Attribute definition:** Number of hospital admissions that could have been avoided with better care. **Worst:** 15 out of 100 hospital admissions could have been avoided with better care **Best:** 5 out of 100 hospital admissions could have been avoided with better care *Main references: Tian et al., 2012; Blunt, 2013; Weeks et al., 2016.* |

*Differs slightly from questionnaire definition and original attribute title.

**Table A6.1.3:** *SELFIE swing weighting criteria and criteria levels –* **Re-admissions**

| Re-admissions | |
|---|---|
| Questionnaire definition: | Proportion of persons who are re-admitted to hospital within 30 days of hospital discharge. |
| SELFIE questionnaire: | *Registry data* |
| Swing – worst Swing – best | **\*Attribute presented:** Hospital re-admissions **\*Attribute definition:** Number of persons who are re-admitted to a hospital within 30 days of their prior hospital discharge. **Worst:** 10 out of 100 of persons are re-admitted to hospital within 30 days of hospital discharge. **Best:** 5 out of 100 persons are re-admitted to hospital within 30 days of hospital discharge. *Main references: Fingar & Washington, 2015; Nolte et al., 2012* |

*Differs slightly from questionnaire definition and original attribute title.

## 10.6.2. Programmes targeting frail elderly

**Table A6.2.1:** *SELFIE swing weighting criteria and criteria levels –* **Autonomy**

| Autonomy | |
|---|---|
| Questionnaire definition: | Remaining in charge and making own decisions on how one lives his/her own life |
| SELFIE questionnaire: | *Pearlin Mastery scale*<br>*-Strongly disagree – Disagree – Agree – Strongly Agree*<br>*\*Solve problems, pushed around in life, little control, do things you set your mind to, feeling helpless, future depends on me, changing things in life* |
| Swing – worst<br>Swing – best | **Attribute presented:** Autonomy<br>**Attribute definition:** Remaining in charge and making own decisions on how one lives his/her own life<br>**Worst:** Not, or barely, in charge or making own decisions<br>**Best:** Fully in charge and making own decisions |

**Table A6.2.2:** *SELFIE swing weighting criteria and criteria levels –* **Burden of medication**

| Burden of medication | |
|---|---|
| Questionnaire definition: | The amount of burden medicines are (considering e.g., administering/taking the medicines, side effects, understanding their purpose/why they're being taken, worries about interaction between medicines, and expenses) |
| SELFIE questionnaire: | *Living with Medicines Questionnaire (LMQ)*<br>*0-10 scale*<br>*\*How much of a burden do you feel your medicines are to you? (considering e.g administering/taking the medicines, side effects, understanding their purpose/why you're taking them, worry about interaction between medicines, expenses)* |
| Swing – worst<br>Swing – best | **Attribute presented:** Burden of medication<br>**Attribute definition:** The amount of burden medicines are (considering for example administering/taking the medicines, side effects, understanding their purpose and why they're being taken, worries about interaction between medicines, and expenses)<br>**Worst:** High burden of medication<br>**Best:** No, or low, burden of medication |

**Table A6.2.3:** *SELFIE swing weighting criteria and criteria levels –* **Burden of informal caregiving**

| Burden of informal caregiving | |
|---|---|
| Questionnaire definition: | The stress of informal caregiving due to the energy it costs, the little time it leaves for own interests and recovery and the sadness over the fate of the supported person |
| SELFIE questionnaire: | *Hours (or possibly CARER-QoL)* |
| Swing – worst<br>Swing – best | **Attribute presented:** Burden of informal caregiving<br>**Attribute definition:** The stress of informal caregiving due to the energy it costs, the little time it leaves for own interests and recovery and the sadness over the fate of the supported person<br>**Worst:** High burden of informal caregiving<br>**Best:** Low burden of informal caregiving |

**Table A6.2.4:** *SELFIE swing weighting criteria and criteria levels –* **Long-term institution admissions**

| Long-term institution admissions | |
|---|---|
| Questionnaire definition: | Proportion of participants admitted to long-term institutional care (e.g., nursing home) during the programme |
| SELFIE questionnaire: | *Registry* |
| Swing – worst<br>Swing – best | **Attribute presented:** Long-term institution admissions<br>**Attribute definition:** Number of participants admitted to long-term institution care (for example a nursing home), during the programme<br>**Worst:** 20 out of 100 participants are admitted to a long-term institution<br>**Best:** 10 out of 100 participants are admitted to a long-term institution<br>*Based on SELFIE case study estimations.* |

**Table A6.2.5:** *SELFIE swing weighting criteria and criteria levels –* **Falls**

| Falls | |
|---|---|
| Questionnaire definition: | Proportion of frail elderly that is admitted to an emergency room or hospital because of a fall |
| SELFIE questionnaire: | *Registry* |
| Swing – worst<br>Swing – best | **\*Attribute presented:** Falls leading to hospital admission<br>**Attribute definition:** Number of participants that are admitted to an emergency room or hospital because of a fall<br>**Worst:** 10 out of 100 of participants have a fall that results in hospital admission<br>**Best:** 5 out of 100 of participants have a fall that results in hospital admission<br>*Main reference: Kannus et al., 2005;* |

*Differs slightly from original attribute title.

## 10.6.3. Palliative care / Oncological programmes

**Table A6.3.1:** *SELFIE swing weighting criteria and criteria levels –* **Mortality**

| Mortality | |
|---|---|
| Questionnaire definition: | Overall mortality rate in 3-months |
| SELFIE questionnaire: | *Registry* |
| Swing – worst<br>Swing – best | **\*Attribute presented:** Life expectancy<br>**\*Attribute definition:** Change in 3-month life expectancy.<br>**Worst:** No improvement in life expectancy<br>**Best:** Improvement in life expectancy |

\*Differs slightly from original attribute title and from questionnaire definition.

**Table A6.3.2:** *SELFIE swing weighting criteria and criteria levels –* **Pain and other symptoms**

| Pain and other symptoms | |
|---|---|
| Questionnaire definition: | Physical symptoms like pain, fatigue, nausea/vomiting, dyspnea, appetite loss, constipation/diarrhea, insomnia |
| SELFIE questionnaire: | *Quality of Life Questionnaire-Core 15-Palliative Care (QLQ-C15-PAL)*<br>*-Not at all*<br>*-A little*<br>*-Quite a bit*<br>*-Very much*<br>*\*Short walk, staying in bed, help with eating/dressing, short of breath, pain, sleep, weak, appetite, nausea, constipated, tired, pain – ADL, tense, depressed* |
| Swing – worst<br>Swing – best | **Attribute presented:** Pain and other symptoms<br>**Attribute definition:** Physical symptoms like pain, fatigue, nausea/vomiting, short of breath, appetite loss, constipation/diarrhoea, insomnia<br>**Worst:** A lot of pain and symptoms<br>**Best:** Little or no, pain and symptoms |

**Table A6.3.3:** *SELFIE swing weighting criteria and criteria levels –* **Compassionate care**

| Compassionate care | |
|---|---|
| Questionnaire definition: | Care is provided in a warm, sensitive and dignified way with sympathy and respect |
| SELFIE questionnaire: | *The Schwartz Center Compassionate Care Scale*<br>*-1-10 point scale, with anchors at : Not at all successful, Somewhat successful, Very successful*<br>*\*Sensitive, caring, compassionate, understanding emotional needs, consider effects on you/family, listen to you, understandable information, trust, involvement in decision-making, comfortable enough to discuss all issues, treat you as a person not disease, respect, communicate sensitively, spend enough time with you.* |
| Swing – worst<br>Swing – best | **Attribute presented:** Compassionate care<br>**\*Attribute definition:** Defined as care that is provided in a warm, sensitive and dignified way with sympathy and respect<br>**Worst:** <u>Not, or barely,</u> compassionate<br>**Best:** <u>Very</u> compassionate |

*Differs slightly from questionnaire definition.

**Table A6.3.4:** *SELFIE swing weighting criteria and criteria levels –* **Timely access to care**

| Timely access to care | |
|---|---|
| Questionnaire definition: | Number of days between referral and start of treatment or care of interest |
| SELFIE questionnaire: | *Registry* |
| Swing – worst<br>Swing – best | **Attribute presented:** Timely access to care<br>**\*Attribute definition:** Time between referral and start of treatment or care of interest<br>**Worst:** <u>A long time</u> between referral and start of treatment/care<br>**Best:** <u>A short time</u> between referral and start of treatment/care |

*Differs slightly from questionnaire definition.

**Table A6.3.5:** *SELFIE swing weighting criteria and criteria levels –* **Preferred place of death**

| Preferred place of death | |
|---|---|
| Questionnaire definition: | Proportion of patients dying at home or the preferred place |
| SELFIE questionnaire: | *Registry* |
| Swing – worst<br>Swing – best | **Attribute presented:** Preferred place of death<br>**\*Attribute definition:** Number of participants that pass away in the location of their preference (e.g., at home).<br>**Worst:** <u>50 out of 100</u> participants pass away in preferred location<br>**Best:** <u>75 out of 100</u> participants pass away in preferred location<br>*Main reference: De Roo et al., 2014* |

*Differs slightly from questionnaire definition.

## 10.6.4. Programmes targeting persons with problems in multiple life domains

**Table A6.4.1:** *SELFIE swing weighting criteria and criteria levels – **Self-sufficiency***

| Self-sufficiency | |
|---|---|
| Questionnaire definition: | Financially in control to meet basic needs with little or no debts |
| SELFIE questionnaire: | *Self-sufficiency matrix (adapted)*<br>*-No income*<br>*-Inadequate income and/or spontaneous or inappropriate spending*<br>*-Can meet basic needs with subsidy; appropriate spending*<br>*-Can meet basic needs and manage debt without assistance*<br>*-Income is sufficient, well managed; has discretionary income and is able to save* |
| Swing – worst<br>Swing – best | **\*Attribute presented:** Financial independence<br>**Attribute definition:** Financially in control to meet basic needs with little or no debts.<br>**Worst:** <u>Financially not in control</u> with growing debts<br>**Best:** <u>Financially in control;</u> finances well-managed |

*Differs slightly from original attribute title.

**Table A6.4.2:** *SELFIE swing weighting criteria and criteria levels – **Total justice costs***

| Total justice costs | |
|---|---|
| Questionnaire definition: | Total costs to the justice system per participant |
| SELFIE questionnaire: | *Contact with criminal justice services; contacts with police; nights in police cell/prison, psychiatric assessment, court attendance (all over 3 month period)* |
| Swing – worst<br>Swing – best | **\*Attribute presented:** Contact with justice system<br>**\*Attribute definition:** Contact with justice system, such as with criminal justice services, nights in police cell, and court attendance<br>**Worst:** <u>Regular contact</u> with justice system<br>**Best:** <u>No, or rare, contact</u> with justice system |

*Differs from original attribute title and questionnaire definition.

# 11. References

de Bekker-Grob EW, Donkers B, Jonker MF, et al. Sample size requirements for discrete-choice experiments in healthcare: a practical guide. Patient. 2015;8:373-384.

Belton V, Stewart T. Multiple criteria decision analysis: an integrated approach. UK: Kluwer Academic Printing; 2002.

Bentham, J. An Introduction to the Principles of Morals and Legislation. 1879.

Berhane A, Enquselassie F. Patients' preferences for attributes related to health care services at hospitals in Amhara Region, northern Ethiopia: a discrete choice experiment. Patient Prefer Adherence. 2015;9:1293-1301.

Blunt I. QualityWatch Focus on preventable admissions: Trends in emergency admissions for ambulatory care sensitive conditions, 2001 to 2013. The Health Foundation and the Nuffield Trust; 2013.

Byeong SA. Compatible weighting method with rank order centroid: Maximum entropy ordered weighted averaging approach. European Journal of Operational Research. 2011;212:552-559.

Claxton K, Sculpher M, Palmer S, Culyer AJ. Causes for concern: Is NICE failing to uphold its responsibilities to all NHS patients? Health Econ. 2015;24(1):1-7.

Cleemput I, Devriese S, Kohn L, et al. KCE Report – Incorporating societal preferences in reimbursement decisions – relative importance of decision criteria according to Belgian citizens. KCE report. 2014.

Craig P, Dieppe P, Macintyre S, et al. Developing and evaluating complex interventions: the new Medical Research Council guidance. BMJ. 2008;337:a1655.

De Roo ML, Miccinesi G, Ontwuteaka-Philipsen BD, et al. Actual and preferred place of death of home-dwelling patients in four European countries: Making sense of quality indicators. PLOSone. 2014. 9:e93762.

Department of Communities and Local Government. Multi-criteria analysis: a manual. London, 2009.

Devlin NJ, Sussex J. Incorporating multiple criteria in HTA: Methods and processes. Office of Health Economics, London, 2011.

Dolan JG. Multi-criteria clinical decision support: A primer on the use of multiple criteria decision-making methods to promote evidence-based, patient-centered healthcare. Patient. 2010;3:229-248.

Edwards W, Barron FH. SMARTS and SMATER: Improved Simple Methods for Multiattribute Utility Measurement. Organ Behav Hum Dec. 1994;60:306-325.

Fingar K, Washington R. Trends in Hospital Readmissions for Four High-Volume Conditions, 2009-2013. HCUP Statistical Brief #196. Rockville, MD: Agency for Healthcare Research and Quality: 2015.

Franken M, Koolman X. Health system goals: a discrete choice experiment to obtain societal valuations. Health Policy. 2013;112:28-34.

Goossens L, Rutten-van Mölken MPMH, Boland MRS, et al. Development of the Assessment of Burden of COPD (ABC) index: experienced burden of disease and the prediction of healthcare consumption. Submitted, 2016.

IJzerman M, Groothuis-Oudshoorn K, Hummel M. Sensitivity analysis in Multi-Criteria Decision (MCDA) models for benefit-risk assessment. ISPOR Conference Poster. 2011.

Kannus P, Sievänen H, Palvanen M, et al. Prevention of falls and consequent injuries in elderly people. 2005. Lancet;366:1885-1893.

Katz, MH. Evaluating Clinical and Public Health Interventions: A Practical Guide to Study Designs and Statistics. Cambridge University Press, 2010.

Keeney RL, Raiffa H. Decisions with Multiple Objectives: Preferences and Value Tradeoffs. Cambridge, UK: Cambridge University Press; 1993.

Kessels R, Bradley J, Goos P. A comparison of partial profile designs for discrete choice experiments with an application in software development. MyIDEAS. No. 2012004. 2012.

Langendam MW, Hooft L, Heus P, et al. Alternatieven voor Randomized Controlled Trials in onderzoek naar de effectiviteit van interventies: Eindrapport. 2013. Dutch Cochrane Center. [*DUTCH*]

Lanscar E, Louviere J. Conducting discrete choice experiments to inform healthcare decision-making: a user's guide. Pharmacoeconomics. 2008;26:661-677.

Lancaster KJ. A new approach to consumer theory. J Polit Econ. 1966;74:132-157.

Maddala T, Philips KA, Johnson FR. An experiment on simplifying conjoint analysis designs for measuring preferences. Health Econ. 2003;12:1035-1047.

Malhotra C, Faroqui MA, Kanesvaran R, et al. Comparison of preferences for end-of-life care among patients with advanced cancer and their caregivers: A discrete choice experiment. Palliative Medicine. 2015;29:842-850.

Manski CF. Structure of Random Utility Models. Theor Decis. 1977;8:229-254.

Marsh K, IJzerman M, Thokala P, et al. Multiple Criteria Decision Analysis for health care decision-making – emerging good practices: Report 2 of the ISPOR MCDA Emerging Good Practices Task Force. Value in Health. 2016;19:125-137.

Marsh K, Lanitis T, Neasham D, et al. Assessing the value of healthcare interventions using multi-criteria decision analysis: A review of the literature. PharmacoEconomics. 2014;32:345-365.

McFadden D. Conditional logit analysis of qualitative choice behavior. In: Zarembka P, editor. Frontiers of Economics. New York: Academic Press; 1974. P. 105-142.

Neuman E, Neuman S. Explorations of the Effects of Experience on Preferences: Two Health-Care Case Studies. Discussion Paper Series, IZA. 2006.

Nolte E, Roland M, Guthrie S, et al. Preventing emergency admissions to hospital: A scoping review. RAND technical report. RAND Corporation; 2012.

Penfold RB, Zhang F. Use of interrupted time series analysis in evaluating health care quality improvements. Acad Pediatr. 2013;13:S38-S44.

Rietkötter L. Master thesis: Ending the war in multi-criteria decision analysis: Taking the best from two worlds: The development and evaluation of guidelines for the use of MACBETH in multi-criteria group decision-making for the assessment of new medical products, 2016.

Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika 1983; 70: 41-55.

Ryan M, Netten A, Skåtun D, et al. Using discrete choice experiments to estimate a preference-based measure of outcome – an application to social care for older people. J Health Econ. 2006;25:927-944.

Sándor Z, Wedel M. Heterogeneous conjoint choice designs. J Marketing Res. 2005;42:210-218.

Schelvis RMC, Oude Hengel KM, Burdorf A, et al. Evaluation of occupational health interventions using randomized controlled trial: challenges and alternative research designs. SJWEH. 2015;41:491-503.

Thokala P, Devlin N, Marsh K, et al. Multiple Criteria Decision Analysis for health care decision-making – an introduction: Report 1 of the ISPOR MCDA emerging good practices task force. Value in Health. 2016;19:1-13.

Thokala P & Duenas A. Multiple Criteria Decision Analysis for Health Technology Assessment. Value in Health. 2012;5:1172-1181.

Thurstone LL. A law of comparative judgment. Psychol Rev. 1927;34:273.

Tian Y, Dixon A, Gao H. DATA briefing: Emergency hospital admissions for ambulatory care-sensitive conditions: identifying the potential for reductions. The King's Fund; 2012.

Tsiachristas A, Stein Vk, Evers S, Rutten-van Mölken M. Economic evaluation of integrated care: a research agenda (working paper). Health Economics Special Interest Group International Foundation for Integrated Care. 2016.

Twisk, JWR. Applied longitudinal data analysis for epidemiology: A practical guide. Cambridge, UK; Cambridge University Press: 2003.

Watson V, Carnon A, Ryan M, et al. Involving the public in priority setting: a case study using discrete choice experiments. J Public Health (Oxf). 2012;34:253-260.

Weeks WB, Ventelou B, Paraponaris A. Rates of admission for ambulatory care sensitive conditions in France in 2009-2010: trends, geographic variation, costs, and an international comparison. Eur J Health Econ;17:453-470.

White H, Sabarwal S. Quasi-Experimental Design and Methods. UNICEF – Methodological Briefs Impact Evaluation No. 8. 2014.

ZonMw Startdocument 'Alternatieven voor RCT bij evaluatie van effectiviteit van interventies!?': Vertrekpunt voor de ZonMw Invitational Conference over 'Alternatieven voor RCTs'. 2015. [*DUTCH*]