Erasmus University Rotterdam
Erasmus School of Economics

Master Thesis Econometrics and Management Science
Operations Research and Quantitative Logistics

Finding cost-effective colorectal cancer screening strategies using multi-objective evolutionary algorithms and the MISCAN-Colon microsimulation model

Name student: Niels Dunnewind
Student ID number: 385959

Supervisor: Prof. Dr. S.I. (Ilker) Birbil
Second assessor: Dr. W. (Wilco) van den Heuvel

External supervisors: Dr. I. (Iris) Lansdorp-Vogelaar
A. (Andrea) Gini

Date final version: January 16, 2020

The content of this thesis is the sole responsibility of the author and does not reflect the view of either Erasmus School of Economics or Erasmus University.

Abstract

Costs and effectiveness of screening strategies for colorectal cancer can be predicted using microsimulation models, such as MISCAN-Colon. Cost-effectiveness analyses use these outcomes to recommend efficient strategies. These studies usually consider a small number of strategies, mostly using only a single screening test and fixed intervals between interventions. By considering more strategies, efficiency can possibly be improved. However, the number of possible strategies is high and the microsimulation models are computationally expensive. Thus, an efficient algorithm is needed to identify efficient strategies. This thesis compares the performance of four multi-objective evolutionary algorithms (NSGA-II, SPEA2, PESA-II and IBEA) on an enumerated test case. First, each algorithm was tuned to perform well on this test case. Performance was then assessed using three unary (ε-Performance, Inverted Generational Distance and Hypervolume) and two binary (Binary Hypervolume and Coverage) multi-objective performance measures. Statistical analysis showed that all measures indicate that NSGA-II performs best on this problem. SPEA2 performed slightly worse, followed by IBEA and finally PESA-II. Inverted Generational Distance and Hypervolume were the most powerful measures, as they were able to find significant differences between each pair of algorithms. Finally, NSGA-II was used to identify efficient strategies for a real case based on the United States scenario. Effectiveness could be improved by 2-22%, depending on the budget. Costs could be reduced by 8-201%, depending on the desired effectiveness. However, the identified strategies are probably too complex to be implemented in practice.

Keywords: colorectal cancer, screening, cost-effectiveness, microsimulation, multi-objective optimization, multi-objective evolutionary algorithms

Contents

Abbreviations

COL	colonoscopy
CRC	colorectal cancer
EA	evolutionary algorithm
FIT	fecal immunochemical test
FS	flexible sigmoidoscopy
IBEA	Indicator-Based Evolutionary Algorithm
ICER	incremental cost-effectiveness ratio
LYG	life-years gained
MISCAN	Microsimulation Screening Analysis
MOEA	multi-objective evolutionary algorithm
NSGA	Non-dominated Sorting Genetic Algorithm
PESA	Pareto Envelope-based Selection Algorithm
POF	Pareto optimal frontier
SPEA	Strength Pareto Evolutionary Algorithm

1 Introduction

Colorectal cancer (CRC) is one of the most common cancers worldwide, causing around 881,000 deaths annually (Bray et al., 2018). Luckily, different screening tests are available to detect the disease, even in an early stage. If detected in time, the undeveloped cancer can be removed relatively easily. However, screening is expensive and comes with certain risks, such as false-positive test results, complications and overdiagnosis. Thus, it should carefully be decided at which ages to screen a population and which test(s) to use.

A screening intervention is defined as a certain test to be performed at a specific age. A set of interventions is referred to as a (screening) strategy. Ideally, strategies are evaluated by randomized controlled trials. However, this approach is expensive and too much time is needed to obtain results. To evaluate strategies without these practical challenges, sophisticated microsimulation models have been developed. These models are able to predict costs and life-years gained (LYG) for any strategy. MISCAN-Colon is one of the leading models, developed by the Department of Public Health of the Erasmus Medical Center.

Several studies (e.g. Wilschut et al., 2011; van den Akker-van Marle et al., 2002) have focused on finding cost-effective strategies, gaining as much LYG as possible with the least costs. As these objectives are conflicting, a single strategy obtaining the highest LYG and lowest costs at the same time most likely does not exist. In contrast, multiple efficient strategies should be identified, together forming the cost-effectiveness curve. Currently, such cost-effectiveness curves are approximated by simulating a relatively small set of predetermined strategies.

Given that an infinite number of strategies exists, the number of evaluated strategies in previous research is small. Also, most of these predetermined strategies are relatively 'simple' in the sense that only one test type is used and the interventions are performed at fixed intervals (e.g. the same test at ages 55, 60 and 65). The actual cost-effective strategies may not be among these strategies, but combine different tests and use varying intervals. Possibly, major public health benefits are missed by only considering simple strategies.

Strategies not considered in previous research will be referred to as 'complex' strategies. As evaluation of all these strategies is computationally too expensive, an efficient optimization algorithm needs to be developed to identify cost-effective strategies. In this thesis, various algorithms from the field of multi-objective optimization are applied to this problem.

1.1 Contributions

Although many cost-effectiveness analyses have been published, a complete mathematical formulation of the used concepts does not exist, to the best of the author's knowledge. Such a formulation is introduced in this thesis, aiming to provide a solid foundation for further research.

Several algorithms were compared on the problem of identifying cost-effective screening strategies. This comparison shows which algorithm and parameter settings can best be applied to such problems. This algorithm (and formulation) can be applied to any cancer site (or other disease), simulation model and/or country.

For the specific case of CRC screening in the United States, this thesis gives an optimized cost-effectiveness curve. Comparing this curve to the cost-effectiveness curve based on simple strategies used in previous research provides insight into the possible gain in cost-effectiveness. Furthermore, the obtained cost-effective strategies are analyzed. Possibly, they can be implemented in practice to attain maximum efficiency. If not, they can be used to inspire and recommend (improvements in) screening policies. For example, the results indicate which tests should be implemented at which ages for maximum effectiveness, given a certain budget.

This thesis also contributes to the literature on comparing multi-objective optimization algorithms. A rigorous procedure, supported with statistical analyses, to compare such algorithms on a specific case and identify the best candidate is demonstrated.

1.2 Outline

The outline of this thesis is as follows. The literature on cost-effectiveness analysis in public health and on multi-objective optimization is reviewed in Section 2. Background information on CRC and MISCAN-Colon is provided in Section 3. The problem is mathematically formulated in Section 4. Section 5 describes the proposed methodology, which consists of four solution approaches and various ways to measure their performance. Section 6 demonstrates this methodology using a case study of the United states. Finally, a conclusion, limitations and suggestions for future research are given in Section 7.

2 Literature review

In this section, relevant literature is reviewed to establish the context in which this thesis was written. First, cost-effectiveness analysis in public health and specifically for (colorectal) cancer screening is further examined. Then, literature on multi-objective optimization is reviewed.

2.1 Cost-effectiveness analysis

Cost-effectiveness analysis is widely used in public health research to compare different policies. Sanders et al. (2016) provided recommendations for conducting such analyses. They advice to summarize results using the incremental cost-effectiveness ratio (ICER). The ICER gives the "ratio of the difference in costs between 2 alternatives to the difference in effectiveness between the same 2 alternatives" (Sanders et al., 2016). Decision makers prefer the strategy with the highest ICER that is still below their willingness-to-pay threshold. The authors recommended to avoid using a single threshold. Furthermore, the societal perspective should be taken: "a viewpoint for conducting a cost-effectiveness analysis that incorporates all costs and health effects regardless of who incurs the costs and who obtains the effects."

Microsimulation models were used in cost-effectiveness analyses for screening for various diseases. For example, Wilschut et al. (2011) found a cost-effectiveness curve for CRC screening by evaluating 50 strategies using MISCAN-Colon, varying the starting and ending age, interval and cut-off value of the considered fecal immunochemical test (FIT). Similarly, van den Akker-van Marle et al. (2002) used MISCAN-Cervix to evaluate nearly 500 strategies for cervical cancer. Rather than financial costs, the number of colonoscopies were used to express societal harm in Knudsen et al. (2016). This study advised the United States Preventive Task Force on CRC screening based on outcomes of MISCAN-Colon and two other microsimulation models, considering 204 strategies. Similar studies considered numbers of predetermined strategies in the same order of magnitude, possibly missing major health benefits. Furthermore, the obtained cost-effectiveness curves consisted of relatively small numbers of strategies, often resulting in large 'jumps' in ICERs, underestimation of ICERs and recommendation of strategies with an ICER well below the willingness-to-pay threshold. O'Mahony et al. (2015) addressed this problem and identified it in many cost-effectiveness analyses of screening for cervical cancer.

To deal with the high number of possible strategies, a few algorithmic approaches for finding cost-effective strategies have been published. Koffijberg et al. (2017) used an evolutionary approach to optimize strategies in a single-objective setting, given a constraint on the number of colonoscopies. By varying the constraint, they identified the Pareto optimal frontier (POF) between LYG and colonoscopy capacity. For cervical cancer, McLay et al. (2010) and Gustafsson and Adami (1992) algorithmically optimized the ages at which to screen the population, given a certain number of available screening interventions. Underwood et al. (2012) used a genetic algorithm to find optimal cut-off values at each age for a prostate cancer test. Erenay et al. (2014) found the optimal risk level at which individuals should undergo a colonoscopy to improve their expected quality-adjusted life-years. While some of these studies considered varying intervals, none of them considered strategies combining different screening tests.

2.2 Multi-objective optimization

In many scenarios, decision makers are interested in multiple objectives. The corresponding problems are often referred to as multi-objective problems or multiple criteria decision making (MCDM) problems. Typically, a single optimal solution does not exist: a set of Pareto optimal solutions should be found. Numerous multi-objective optimization algorithms have been introduced to efficiently find such solutions.

Deb (2014) made a distinction between interactive and noninteractive multi-objective optimization methods. In contrast to the latter methods, the former allow for an active role of the decision maker in the optimization process. If the decision maker can not interfere during the optimization process, a priori methods allow the decision maker to express his preferences (on the trade-off between the objectives) a priori, so that the method can focus on finding solutions that satisfy these preferences. On the other hand, a posteriori methods attempt to find the whole POF, after which the decision maker can choose from the obtained solutions. Deb (2014) also classified the problems to be solved. The objective function for linear problems are fully linear in all decision variables, while this is not true for nonlinear problems for at least one decision variable. Furthermore, an infinite number of optimal solutions typically exists for continuous problems, while this number is finite for combinatorial problems. This thesis focuses on noninteractive a posteriori methods for nonlinear combinatorial problems.

Multi-objective optimization has been applied in many fields, in particular in the field of engineering. For example, Alonso et al. (2009) optimized the design of supersonic aircraft, maximizing their range, while minimizing the noise level of the sonic boom. Okasha and Frangopol (2009) optimized bridge designs, minimizing their life-cycle costs, while maximizing their reliability. Multi-objective optimization techniques are also applied to some problems in health science, for example that of designing intensity modulated radiotherapy. Such a design should ensure that the dose of radiation delivered to tumors is as close as possible to the prescribed dose, while doses delivered to critical organs and normal tissue are minimized (Ehrgott et al., 2010). Drug design and discovery is another medical field in which problems with conflicting objectives are relevant (Nicolaou and Brown, 2013). Despite its many applications, multi-objective optimization is not yet established in the field of public health.

Whereas performance measurement of an algorithm in the single-objective setting is rather trivial, it is more complex to measure performance in the multi-objective setting. Many multi-objective performance measures have been introduced to solve this problem. Zitzler et al. (2003) gave an extensive overview of the 30 most widely used measures. With this many measures and possible conflicting outcomes, a rigorous approach is needed to draw conclusions on the (relative) performance of various algorithms. This thesis builds upon the methodology as demonstrated in Kollat and Reed (2006), who optimized long-term groundwater monitoring design. They defined and enumerated a test case by restricting the search space of their problem. With the POF known, the performance of various algorithms could be assessed by calculating performance measures for multiple runs (with different random seeds). Based on the obtained values (and their distribution), the most promising candidate algorithm was identified and applied to the actual case. However, they did not provide a rigorous statistical analysis to test for significant differences in outcomes of these measures. Such an analysis will be demonstrated in this thesis.

3 Background

This section provides the reader with relevant background information. First, the natural history of CRC and the various ways of screening for this disease are explained. Then, the MISCAN-Colon model is introduced.

3.1 Colorectal cancer

CRC is a neoplasm that occurs in the colon and/or rectum. It is generally accepted that the majority of the CRC cases starts with precursor lesions, most of which are adenomas (Morson, 1974; Vogelstein et al., 1988), while others are sessile serrated polyps (Snover, 2011). After a certain time, these lesions may develop into (preclinical) CRC. This process can take a long time and often starts at a relatively high age. Eventually around 35% of the population develops a lesion (Corley et al., 2013), while only approximately 4.5% develops CRC (Siegel et al., 2017). When symptoms emerge, the cancer can be diagnosed. As soon as the cancer is diagnosed, either by symptoms or by screening, the clinical stage is reached and the cancer is treated. Depending on the stage in which the cancer is detected, the five-year survival rates range from around 10% to 90% (Siegel et al., 2017). The natural history process is summarized in Figure 1.

In the US, both CRC incidence and mortality have declined in the past decades. This trend is attributed to improved treatments and changes in lifestyle, but mostly to screening (Siegel et al., 2017).

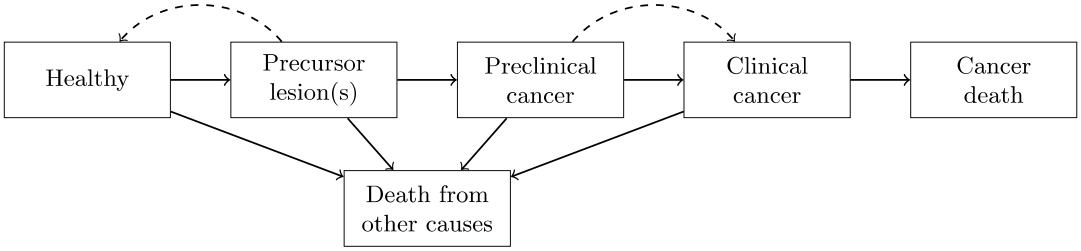


Figure 1: Simplified overview of possible health states and transitions. The solid lines indicate the CRC natural history, while the dashed lines represent the effects of screening. Adapted version from Zauber et al. (2009, Figure 1)

.

Figure 1 shows the possible health states and transitions, where the dashed lines show the possible effects of screening. The most desirable effect is to prevent cancer by removing the precursor lesion(s), which is represented by the transition from the *Precursor lesion(s)* state to the *Healthy* state. If the lesion has already developed into preclinical cancer, screening can detect it before symptoms emerge, which can be beneficial because treatment can start earlier and the cancer can be prevented to reach its next stage. This effect is represented by the transition from the *Preclinical cancer* state to the *Clinical cancer* state.

Despite these positive effects, screening should not be performed too often because it can also have significant unwanted effects, such as overdiagnosis, false-positive results and complications. Overdiagnosis refers to the case in which a cancer is diagnosed with screening that would not have been diagnosed in the absence of screening. For example, if a patient will die from other causes without experiencing any CRC symptoms, it is better not to screen this person.

Preclinical CRC and its precursors can be detected by different screening tests. Table 1 summarizes the tests considered in this thesis. The sensitivity is the probability that the disease is detected by a certain test, given the current state. Specificity refers to the probability of a positive test result, given that the person does not have the disease. A distinction is made between stool-based and endoscopic tests.

Stool-based tests check the stool for signs of CRC, such as blood. These tests are relatively inexpensive (except when combined with a DNA test), easy to perform and safe. However, the probability of a false-negative result is relatively high. Because the tests can not detect the origin of the blood, false-positive results are also relatively common. The fecal immunochemical test (FIT) is a stool-based test that checks the stool sample for hemoglobin. If the amount of hemoglobin found exceeds a certain cut-off value, the test is positive. In this thesis, a number of previously used cut-off values will be used, namely 10, 20 and 40 μg/g. FIT-DNA (or multi-targeted stool DNA testing) is a test that combines a FIT with testing for certain mutated biomarkers in the DNA in the feces.

With endoscopic tests a gastroenterologist inserts a flexible tube with camera into the anus, and the rectum and colon can be inspected visually for lesions. False-positive results occur when the gastroenterologist concludes from a biopsy that the removed lesion was not a precursor of CRC. Precursor lesions seldom remain undetected, thus the probability of false-negative results is modest, especially when compared to stool-based tests. Colonoscopy (COL) is the most commonly used endoscopic test. During this test, detected lesions can directly be removed. Generally, every positive test other than colonoscopy is followed up by a diagnostic colonoscopy. A flexible sigmoidoscopy (FS) is a similar test, but it requires less preparation by the patient and can only inspect the first part of the colon (and the whole rectum).

It is possible to combine certain tests: a FIT (with any cut-off value) or a FIT-DNA can be combined with a FS. These combined tests are treated as separate tests, denoted by FIT10+FS, FIT20+FS, FIT40+FS and FIT-DNA+FS. The stool-based test will be performed first, after which the FS will only be performed if the test is negative. Other combinations are not considered in this thesis.

A (screening) strategy is a sequence of (screening) interventions, with each intervention being a combination of an age and a test to perform at that age. The benefits of a strategy can be expressed in life-years gained (LYG). This measure gives the difference in total life-years of a population in a scenario with and without the strategy. The costs of a strategy can be expressed in monetary terms. Costs that should be considered are costs for screening, surveillance, treatment and complications. Both LYG and costs are discounted. An annual discounting percentage of 3% is most commonly used for both measures.

3.2 MISCAN-Colon

In 1985, the MISCAN (Microsimulation Screening Analysis) model was introduced by the Department of Public Health of the Erasmus Medical Center as a general model for the evaluation of screening for disease (Habbema et al., 1985). It is a microsimulation model, implying that every individual within the simulated population is simulated individually. The model uses semi-Markov processes to simulate transitions between the discrete states, where each state

Table 1: Overview of used screening tests and their characteristics.

Test	Costs[a] ($)		Specificity[b] (%)	Sensitivity[c] (%)				
	Positive	Negative		Adenoma			CRC	
				≤ 5	6-9	≥ 10	Early	Late
Stool-based								
FIT10[d]	39.94	39.94	95.79	0.00	9.60	16.10	65.00	90.00
FIT20[d]	39.94	39.94	97.76	0.00	4.40	13.10	52.00	83.50
FIT40[d]	39.94	39.94	98.70	0.00	2.50	10.30	50.00	82.50
FIT-DNA[e]	530.55	530.55	89.80	0.00	22.00	28.35	96.72	86.36
Endoscopic								
COL[e]	1,656.38	1,332.01	86.00	75.00	85.00	95.00	95.00	95.00
FS[e]	557.76	557.76	87.00	75.00	85.00	95.00	95.00	95.00

[a] From societal perspective, as used in Knudsen et al. (2016).

[b] Per person.

[c] Per lesion.

[d] Specificity and sensitivity obtained from Goede et al. (2013) (this study expresses hemoglobin levels in ng/mL; 1μg/g is equal to 5ng/mL).

[e] Specificity and sensitivity as used in Knudsen et al. (2016).

represents a phase of the disease. It uses continuous time.

Throughout the years, different variants of the MISCAN model have been developed for different diseases. MISCAN-Colon is the variant focusing on CRC (Loeve et al., 1999). It has been calibrated and validated against different data sets (Rutter et al., 2016; Buskermolen et al., 2018) and is one of the leading models to evaluate screening strategies for CRC.

The model requires demographic, natural history and screening parameters. Demographic parameters describe the population under consideration with, for example, birth and life tables. The simulation of the natural history of CRC is dependent on the natural history parameters, which include distribution parameters of dwelling times, localization and incidence. Screening parameters describe the different screening tests. These inputs will be discussed for the US case study in Section 6.

Furthermore, input on which tests to use at what age (i.e. the screening strategy) is required. This part of the input constitutes the decision variables for this thesis and is formalized in Section 4.

For each individual in the simulated population, the model first simulates the life history without CRC. Then, CRC lesions are simulated using a semi-Markov process with transitions between discrete states as shown in Figure 2. Some individuals develop adenomas, each of which initializes in the state *Adenoma ≤ 5 mm*. The adenoma can grow over time and become malignant, modelled as a transition to the state *Preclinical cancer stage I*. As soon as symptoms emerge, a transition from the preclinical state to the clinical state at the corresponding stage is made. In the end, the cancer leads to death, unless a person dies of other causes first or treatment was able to halt further progress. Finally, the provided screening and surveillance strategies are applied to each individual. Currently, only the adenoma-carcinoma sequence is incorporated in the model.

By simulating the life histories with and without screening, two ‘parallel universes’ are created. This enables the model to calculate the number of LYG by screening. Other relevant

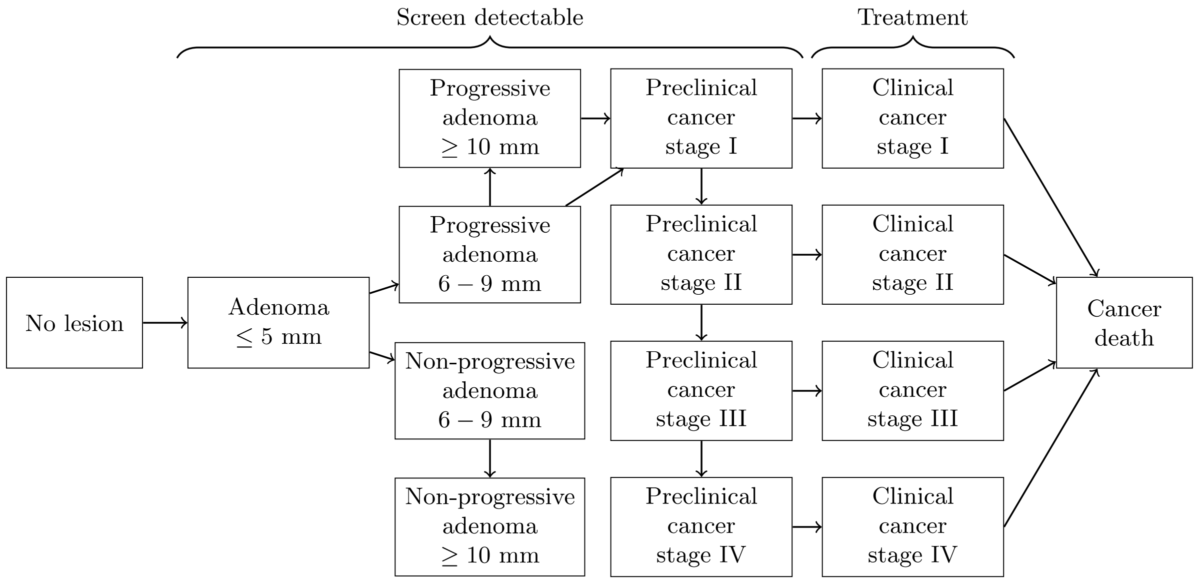

Screen detectable
Treatment
No lesion
Adenoma ≤ 5 mm
Progressive adenoma ≥ 10 mm
Progressive adenoma 6 − 9 mm
Non-progressive adenoma 6 − 9 mm
Non-progressive adenoma ≥ 10 mm
Preclinical cancer stage I
Preclinical cancer stage II
Preclinical cancer stage III
Preclinical cancer stage IV
Clinical cancer stage I
Clinical cancer stage II
Clinical cancer stage III
Clinical cancer stage IV
Cancer death

Figure 2: States and transition of a single lesion in the MISCAN-Colon model. It is indicated in which states it is detectable by screening and in which states the cancer is treated. Adapted version of Knudsen et al. (2016, Figure 1).

outputs include numbers of performed screening tests, prevented CRCs and CRC deaths. For a more extensive description of the model, see Loeve et al. (1999).

4 Problem formulation

In this section, a mathematical problem formulation is given. First, the strategies are formalized. Then, their corresponding objective values are discussed. Dominance relationships between strategies and special sets of strategies are then defined. Finally, mathematical formulations of concepts used in cost-effectiveness analyses are provided.

4.1 Strategies

Screening strategies are represented by $x \in \Theta$, with Θ being the set of all possible strategies. Let A be the ages at which a screening intervention can take place. The set T represents the possible screening tests to use in an intervention. Possible combinations of tests at the same age are considered as a separate, unique test. So, at most one test can be used at any age. As it is also possible to not use any test at a certain age, the number of possible strategies $|\Theta|$ is given by $(|T|+1)^{|A|}$. The search space Θ can be formalized as

$$\Theta = \{x \in \mathbb{N}^{|A|} \ : x_a \in T \cup \{0\}, \ \forall a \in A\}, \tag{1}$$

where $x_a = 0$ indicates that no test is performed at age a.

Consider the following example case. Colonoscopy (COL) and FIT20 can be used as screening tests and thirteen possible ages are considered. In this setup, the number of possible strategies already amounts to $(2+1)^{13} = 1{,}594{,}323$. An example strategy x' in which colonoscopy is performed at the first possible age and FIT20 at the fourth and sixth ages can be represented by

$$x' = \begin{bmatrix} \text{COL} & 0 & 0 & \text{FIT20} & 0 & \text{FIT20} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

4.2 Objectives

For screening strategy x, the (discounted) costs and LYG per individual are given by $C(x)$ and $Y(x)$, respectively. The (multi-objective) problem can be formulated as

$$\min_{x \in \Theta} \begin{bmatrix} C(x) \\ -Y(x) \end{bmatrix}. \tag{2}$$

The values and distributions of $C(x)$ and $Y(x)$ are unknown, but can be expressed as the expectations of the simulation replications outcomes $D(x,\omega)$ and $Z(x,\omega)$, respectively, where $\omega \in \Omega$ is a sample path. A sample path ω describes a simulation replication and can be summarized by the seeds for the random number generator and the size of the simulated population. The simulation replications outcomes (of which the distributions are unknown) are assumed to be unbiased estimators of $C(x)$ and $Y(x)$:

$$C(x) = \mathrm{E}_{\omega \in \Omega}[D(x,\omega)], \tag{3}$$

$$Y(x) = \mathrm{E}_{\omega \in \Omega}[Z(x,\omega)]. \tag{4}$$

For a given sample path ω', the estimators are

$$\hat{C}(x) = D(x, \omega'), \tag{5}$$

$$\hat{Y}(x) = Z(x, \omega'). \tag{6}$$

Functional forms of (5) and (6) do not exist: their values can only be obtained by simulation.

4.3 Approximation sets

To compare strategies in a multi-objective setting, the concept of Pareto dominance is used. If a strategy x_1 is not worse in any of the objectives and better in at least one objective compared to another strategy x_2, x_1 is said to dominate x_2. Such a relation is denoted by $x_1 \succ x_2$. For the problem at hand, this can be formalized as

$$x_1 \succ x_2 \iff (C(x_1) \leq C(x_2) \wedge Y(x_1) \geq Y(x_2)) \wedge (C(x_1) < C(x_2) \vee Y(x_1) > Y(x_2)), \tag{7}$$

for any $x_1, x_2 \in \Theta$. If the right-hand side of (7) does not hold true, x_1 does not dominate x_2, which is denoted by $x_1 \not\succ x_2$. Figure 3 shows a number of example strategies in the objective space. Here, x_6 is dominated by x_2, because $C(x_2) < C(x_6) \wedge Y(x_2) > Y(x_6) \implies x_2 \succ x_6$ by (7), but x_2 does not dominate x_7, i.e. $x_2 \not\succ x_7$.

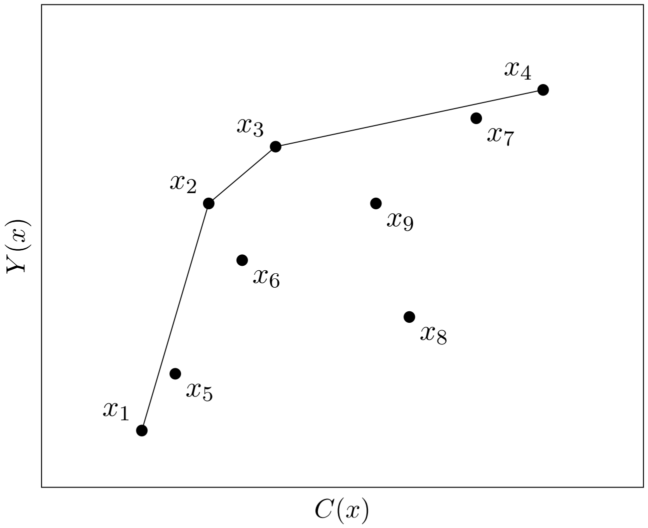


Figure 3: Plot of a number of example strategies in the objective space to illustrate the concepts of Pareto efficiency and cost-effectiveness. The solid line represents the cost-effectiveness curve.

As the objectives in (2) are conflicting, a single strategy optimising both of them simultaneously most likely does not exist. Instead, the solution to such a multi-objective problem is a set of multiple strategies. Each strategy in this set should not be dominated by any of the other strategies. Such a set is defined as an approximation set (Zitzler et al., 2003). For any set of strategies $A \subseteq \Theta$, the corresponding approximation set is given by

$$\psi(A) = \{x_1 \in A : (\nexists x_2 \in A : x_2 \succ x_1\}. \tag{8}$$

All possible approximation sets are gathered in the set Ψ:

$$\Psi = \{A \subseteq \Theta : x_1 \not\succ x_2 \wedge x_2 \not\succ x_1 \ \forall x_1, x_2 \in A\}. \tag{9}$$

The strategies in an approximation set are said to be incomparable. A valid approximation set in the example is $A_1 = \{x_2, x_3, x_7\}$. In contrast, $A_2 = \{x_6, x_8, x_9\}$ is not a valid approximation set, because $x_6 \succ x_8$. However, an approximation set can be obtained from A_2 by applying (8), which yields $\psi(A_2) = \{x_6, x_9\}$.

Like dominance relationships have been defined for strategies, similar definitions have been introduced for approximation sets. An approximation set $A \in \Psi$ dominates $B \in \Psi$ if all strategies in B are dominated by at least one strategy in A:

$$A \succ B \iff \forall x_2 \in B : (\exists x_1 \in A : x_1 \succ x_2). \tag{10}$$

For the example approximation sets A_1 and A_2, it holds that $A_1 \succ A_2$, because $x_2 \succ x_6$ and $x_3 \succ x_9$.

The approximation set $P \in \Psi$ exists exactly of all strategies not dominated by any other possible strategy: the Pareto optimal set. This set is also referred to as the the Pareto (optimal) front (POF) and can be formalized as $P = \psi(\Theta)$. Note that no approximation set exists that dominates P. In the example it is assumed that the shown strategies are the only possible strategies, i.e. $\Theta = \{x_i : i = 1, \ldots, 9\}$. Six of these strategies are not dominated by any other strategies, thus $P = \psi(\Theta) = \{x_1, x_5, x_2, x_3, x_7, x_4\}$.

4.4 Cost-effectiveness

In cost-effectiveness analysis, not all strategies on the POF are considered to be cost-effective. With respect to a set of strategies $A \subseteq \Theta$, a strategy $x \in A$ is cost-effective if and only if no convex combination of (the objective values of) two strategies $x_1, x_2 \in A$ exists that dominates x. This thesis introduces the following mathematical formulation of the set of all cost-effective strategies in a set of strategies $A \subseteq \Theta$:

$$\begin{aligned}\xi(A) = \{x \in A : \big(\nexists x_1, x_2 \in A, w \in [0,1] :\\ &((wC(x_1) + (1-w)C(x_2) < C(x) \wedge wY(x_1) + (1-w)Y(x_2) \geq Y(x)) \vee \\ &(wC(x_1) + (1-w)C(x_2) \leq C(x) \wedge wY(x_1) + (1-w)Y(x_2) > Y(x)))\big)\},\end{aligned} \tag{11}$$

where w is the coefficient weighing (the objective values of) each two strategies. Note that $\xi(A) \subseteq \psi(A)$ and $\xi(A) = \xi(\psi(A)) = \psi(\xi(A))$ for all $A \subseteq \Theta$. Furthermore, the least expensive strategy and the most effective strategy considered are always cost-effective. The line connecting all the cost-effective strategies (sorted by either LYG or costs) is referred to as the cost-effectiveness curve (or: cost-effectiveness frontier). Figure 3 show this curve as a solid line for the considered example.

Given a reference strategy $x' \in \Theta$, the ICER of strategy $x \in \Theta$ is given by

$$\sigma(x; x') = \frac{C(x) - C(x')}{Y(x) - Y(x')}. \tag{12}$$

Thus, it is equal to the inverse of the slope between x and x'.

When assessing the cost-effectiveness for a set of strategies $A \subseteq \Theta$, the ICER is set to ∞ for all strategies that are not cost-effective. The cost-effective strategy with the least costs is assigned an ICER of $-\infty$. The ICER for the remaining strategies (i.e. all cost-effective strategies except the least expensive) is calculated using (12), with the previous cost-effective strategy as reference. Thus, for any reference set of strategies $A \subseteq \Theta$ and strategy $x \in A$, the ICER is given by

$$\rho(x; A) = \begin{cases} \sigma(x; \underset{x' \in \xi(A): Y(x') < Y(x)}{\arg\max} Y(x')), & x \in \xi(A), Y(x) > \min_{x' \in \xi(A)} Y(x'), \\ -\infty, & x \in \xi(A), Y(x) = \min_{x' \in \xi(A)} Y(x'), \\ \infty, & x \notin \xi(A). \end{cases} \tag{13}$$

For example, given the set of example strategies in Figure 3, the ICER is not equal to ∞ for x_1, x_2, x_3 and x_4. In particular, the ICER of x_1 is $-\infty$. The ICER of x_4 is in this case equal to the inverse of the slope of the line between x_3 and x_4. Note that, for $A \subseteq \Theta$ and $x_1, x_2 \in \xi(A)$, it holds that

$$\rho(x_1; A) > \rho(x_2; A) \iff Y(x_1) > Y(x_2) \wedge C(x_1) > C(x_2). \tag{14}$$

For any set of strategies $A \subseteq \Theta$ and willingness-to-pay threshold $\lambda \in \mathbb{R}$, the recommended strategy $z(\lambda; A)$ is the (cost-effective) strategy with the highest ICER within the budget:

$$z(\lambda; A) = \underset{x \in A: \rho(x; A) \leq \lambda}{\arg\max} \rho(x; A). \tag{15}$$

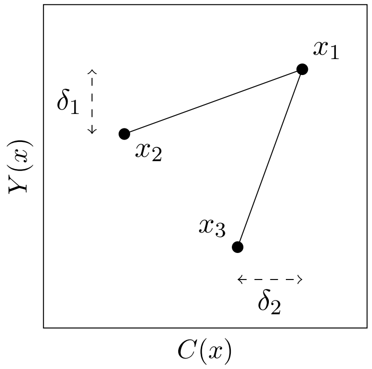


Figure 4: Illustration of the proof showing that the ICER $\rho(x; A)$ and the recommended strategy $z(\lambda; A)$ heavily depend on the reference set A. Strategies are represented by dots. The ICERs $\rho(x; A_1)$ and $\rho(x; A_2)$ are equal to the inverse of the slope of the lines, for $A_1 = \{x_1, x_2\}$ and $A_2 = \{x_1, x_3\}$. The difference between $Y(x_1)$ and $Y(x_2)$ is indicated by δ_1, while δ_2 represents the difference between $C(x_1)$ and $C(x_3)$.

Note that $\rho(x; A)$ and therefore also $z(\lambda; A)$ are heavily dependent on the strategies in A. Assuming strategies exist for any objective values, it can be proven that the difference in ICER of the same strategy, calculated using two different reference sets of strategies, can be arbitrary large. This implies the choice of the set of considered strategies potentially greatly influences the conclusion of a cost-effectiveness analysis. Furthermore, it shows caution should be taken when interpreting the ICERs of such an analysis. Consider the strategies x_1, x_2 and x_3 as shown in Figure 4. Let $A_1 = \{x_1, x_2\}$, $Y(x_2) = Y(x_1) - \delta_1$ and $C(x_2) < C(x_1)$. Then, for any positive value of ε_1, a value for δ_1 can be found so that $\rho(x_1; A_1) > \varepsilon_1$:

$$\exists \delta_1 > 0 : \rho(x_1; A_1) > \varepsilon_1, \qquad \forall \varepsilon_1 > 0. \tag{16}$$

Similarly, it can be shown that $\rho(x_1; A)$ can be arbitrary close to zero. Let $A_2 = \{x_1, x_3\}$, $C(x_3) = C(x_1) - \delta_2$ and $Y(x_3) < Y(x_1)$. Then, it holds that

$$\exists \delta_2 > 0 : \rho(x_1; A_2) < \varepsilon_2, \qquad \forall \varepsilon_2 > 0. \tag{17}$$

In conclusion, the difference $\rho(x_1; A_1) - \rho(x_1; A_2)$ can be arbitrary large:

$$\exists \delta_1, \delta_2 > 0 : \rho(x_1; A_1) - \rho(x_1; A_2) > \varepsilon, \qquad \forall \varepsilon > 0. \tag{18}$$

5 Methodology

This section describes the used methodology. First, multi-objective evolutionary algorithms are described and four of these algorithms are proposed as solution approaches. Then, various multi-objective measures are proposed to evaluate the performances of these algorithms.

5.1 Solution approaches

Evolutionary algorithms (EAs) are metaheuristic optimization algorithms that develop a population of solutions. One of the most well-known evolutionary algorithms is the genetic algorithm, inspired by natural selection. It initializes with a first generation of solutions (strategies), each with its own fitness (costs and LYG). Next generations are iteratively generated by creating offspring, i.e. combining elements (interventions) of the solutions of the previous generation. When creating a new solution, certain mutations may occur, affecting one or more elements of the solution.

In multi-objective optimization, a single solution optimizing all objectives usually does not exist. Instead, a set of Pareto optimal solutions should be found. EAs are logical candidates to solve these problems, as they work with a population of solutions at any time. Thus, the concept of EAs was extended by introducing multi-objective evolutionary algorithms (MOEAs). The aim of these algorithms is twofold. In the first place, the POF obtained by an algorithm should be as close to the actual POF as possible. Furthermore, the found solutions should be evenly spread over the POF. Multiple MOEAs have been introduced in the last decades, including many genetic algorithms: an extensive overview is given by Konak et al. (2006).

MOEAs are especially useful for expensive multi-objective optimization problems, because the solutions in a generation can be evaluated in parallel. The problem at hand is categorical. While not all optimization algorithms can (easily) handle categorical variables, genetic algorithms can be used to solve such problems. Therefore, MOEAs (in particular genetic algorithms) are used in this thesis. Before describing the four chosen MOEAs, the remainder of this section discusses the general characteristics of these algorithms.

The algorithms considered in this thesis use binary tournament selection to select parent strategies. This selection procedure randomly samples (with or without replacement) two strategies from the current generation. Based on one or multiple fitness values assigned to these strategies by the algorithm, one of these strategies is chosen as parent. A so-called mating pool is filled with the selected parents.

When the mating pool is filled, pairs of parents are used to create offspring (i.e. new strategies). Multiple methods exist to do so. The k-point crossover operator splits the parents at k random points and recombines the obtained blocks to form two children. Single-point crossover (C_{sp}) and two-point crossover (C_{tp}) are frequently used special cases with $k = 1$ and $k = 2$, respectively. The uniform crossover (C_u) creates offspring by deciding for each age if the first (second) child strategy inherits the test on that age from the first (second) or the second (first) parent strategy with equal probabilities.

After a child strategy has been created, it is randomly mutated. This enables the algorithm to diversify and escape local minima. The used mutation operator M_r^p considers each age of the strategy separately. With a probability of r it changes the decision variable corresponding to each age. If no intervention is scheduled, it adds an intervention with a random test. In case an intervention is already scheduled, the operator removes it with a probability of p and otherwise changes its test to a randomly selected other test.

Besides the regular population (the internal population), some MOEAs store certain solutions in an archive (the external population). This archive is updated in each generation by adding new well performing solutions and possibly removing the worst solutions. In general, the archive is larger than the internal population.

A number of MOEAs was selected from the literature. The main criterion was that these algorithms should be able to find batches of strategies to simulate, instead of finding strategies one by one. In this way, multiple strategies can be simulated in parallel, which is desirable because of the high computational effort needed for a single simulation (approximately one minute for a simulated population size of 2.5e6 on a single processor core of an average consumer computer). Furthermore, the computational complexity of these algorithms is not important, as the speed of the optimization is governed by the simulations. Thus, the computational overhead of the algorithms is negligible. In the following, the selected algorithms are described. All of these algorithms differ fundamentally in the way they assign fitness values to strategies.

5.1.1 NSGA-II

Srinivas and Deb (1994) introduced Non-dominated Sorting Genetic Algorithm (NSGA). NSGA-II is an improved version by Deb et al. (2002) and is currently one of the leading MOEAs.

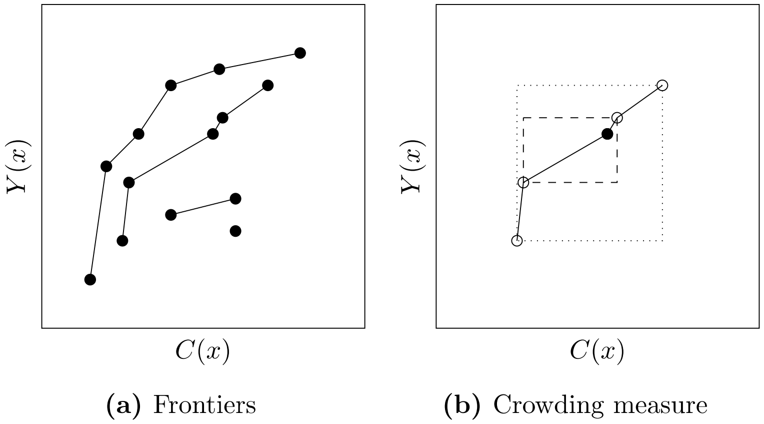
In each generation t, the algorithm divides the strategies in the current population P_t into multiple frontiers. The first frontier $F_{t,1} \subseteq P_t$ consists of the strategies not dominated by any other strategy in P_t. The second frontier $F_{t,2}$ holds the strategies not dominated by any other strategy in the remainder of the population (i.e. $P_t \setminus F_{t,1}$), et cetera. This can be formulated using the following recursive relation for generation t and frontier i:

$$F_{t,i} = \begin{cases} \psi(P_t), & i = 1, \\ \psi(P_t \setminus \bigcup_{j=1}^{i-1} F_{t,j}), & i = 2, 3, \ldots, \end{cases} \tag{19}$$

Because the strategies in each frontier are incomparable, any non-empty frontier is an approximation set, i.e. $\{F_{t,i} : i \in \mathbb{N}_{\geq 1}, F_{t,i} \neq \emptyset\} \subseteq \Psi$ for all generations t. Figure 5 shows a population of strategies. Strategies in the same frontier are connected with solid lines. Note that the fourth frontier consists of a single strategy.

The algorithm initializes the next generation with n strategies from the current generation by copying the strategies from each frontier, starting with the first frontier and resuming in ascending order. If the next frontier does not fit entirely in the next generation, the strategies in the least crowded areas are selected to fill the generation. This is done using a crowding measure, expressed as the distance of the strategy, which is calculated as follows per frontier. All distances are initialized at zero. Then, the following procedure is carried out for each objective:

1. The strategies in the frontier are sorted in ascending order.



(a) Frontiers **(b)** Crowding measure

Figure 5: Illustrations of the concepts used in NSGA-II. Each dot represents a strategy in the objective space.

2. Strategies with the extreme objective values (i.e. the boundary points) are assigned an infinite distance.
3. The distances of all other strategies are incremented by the difference in objective values of the previous and next strategy, normalized by dividing by the absolute difference in extreme values.

Figure 5b illustrates this concept for a strategy in the second frontier, represented by a solid dot. The width and height of the dashed rectangle are divided by the width and height of the dotted rectangle, respectively. The crowding measure equals the sum of these values.

Finally, the generation (currently containing n strategies) is used to create n offspring strategies. This is done using binary tournament selection, where the frontier of a strategy is used to determine its performance. Ties are broken by choosing the strategy with the highest distance.

5.1.2 SPEA2

SPEA2 (Zitzler et al., 2001) is the improved version of Strength Pareto Evolutionary Algorithm (SPEA) (Zitzler and Thiele, 1999). Besides the regular population P_t of size n, it keeps an archive $\overline{P}_t$ of size $\overline{n}$, with t being the generation counter.

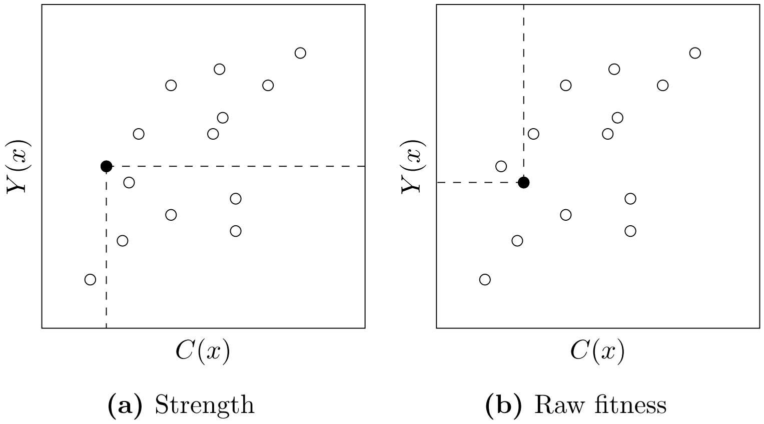
The calculation of the fitness involves multiple steps. First, the strength $S(x)$ of each strategy x in the current generation P_t and archive $\overline{P}_t$ is calculated as the number of strategies it dominates:

$$S(x) = |\{x' \in P_t \cup \overline{P}_t : x \succ x'\}|. \tag{20}$$

The raw fitness $R(x)$ of strategy x is then defined as the sum of the strengths of all strategies it is dominated by:

$$R(x) = \sum_{x' \in P_t \cup \overline{P}_t : x' \succ x} S(x'). \tag{21}$$

Consider the example shown in Figure 6. The highlighted strategy in Figure 6a dominates five strategies. Its strength is therefore equal to 5. The highlighted strategy in Figure 6b is dominated by this strategy and this strategy only. Thus, its raw fitness is equal to 5.



(a) Strength **(b)** Raw fitness

Figure 6: Illustrations of the concepts used in SPEA2. Each dot represents a strategy in the objective space.

The raw fitness might be equal for certain strategies. Therefore, a density is added to penalize strategies in crowded areas. This is done by calculating for each strategy the normalized Euclidean distance to each other strategy in the objective space. After sorting the list of distances relative to strategy x in ascending order, the k-th element, denoted by σ_x^k, is used to calculate the density $D(x)$:

$$D(x) = \frac{1}{\sigma_x^k + 2}. \tag{22}$$

The value of k is often chosen to be $\lfloor \sqrt{n + \overline{n}} \rfloor$. To ensure $0 < D(x) < 1$, 2 is added in the denominator. Finally, the raw fitness and density are summed to obtain the fitness $F(x)$:

$$F(x) = R(x) + D(x). \tag{23}$$

Note that for non-dominated strategies in $P_t \cup \overline{P}_t$ it holds that $F(x) < 1$. These strategies are copied to the next archive $\overline{P}_{t+1}$:

$$\overline{P}_{t+1} = \{x \in P_t \cup \overline{P}_t : F(x) < 1\}. \tag{24}$$

If $|\overline{P}_{t+1}| = \overline{n}$, the construction of the archive is finished. In case $|\overline{P}_{t+1}| < \overline{n}$, the archive is filled with the $\overline{n} - |\overline{P}_{t+1}|$ dominated strategies in $P_t \cup \overline{P}_t$ with the lowest fitness. Else, $|\overline{P}_{t+1}| > \overline{n}$ and $|\overline{P}_{t+1}| - \overline{n}$ members of the archive need to be removed. This is done iteratively by removing strategy x with the smallest distance to another individual, denoted by $x \leq_d x'$ for $x, x' \in \overline{P}_{t+1}$, where

$$\begin{aligned} x \leq_d x' \iff & \left(\forall\, 0 < k < |\overline{P}_{t+1}| : \sigma_x^k = \sigma_{x'}^k\right) \quad \vee \\ & \left(\exists\, 0 < k < |\overline{P}_{t+1}| : \left((\forall\, 0 < l < k : \sigma_x^l = \sigma_{x'}^l) \wedge \sigma_x^l < \sigma_{x'}^l\right)\right). \end{aligned} \tag{25}$$

Binary tournament selection with replacement is performed on the archive $\overline{P}_{t+1}$ to fill the mating pool with n strategies. Offspring and mutation operators are then applied to the mating pool to create P_{t+1} of size n.

5.1.3 PESA-II

The Pareto Envelope-based Selection Algorithm (PESA) (Corne et al., 2000) also maintains an archive and uses it to select parents from. The algorithm divides the objective space using a (regular) grid. It then calculates the current amount of strategies in each cell of the grid (the squeeze factor) and uses it as the fitness value for those strategies. Binary tournament selection is used to fill the mating pool, after which offspring and mutation operators are used to create the next generation. The objective values for each strategy in this new generation are evaluated. Then, the archive is iteratively updated with the new generation. If a strategy is not dominated by any member of the archive, it enters the archive and any dominated archive members are removed. In case this leaves the archive overfull, a random archive member with the highest squeeze factor is removed.

The concept of the grid and squeeze factors is illustrated in Figure 7. Each black dot represents a strategy in the population in a certain generation. The squeeze factor of a strategy is equal to the total number of strategies in the corresponding grid cell. For example, the squeeze factor of the strategies in the hatched cell is equal to 3.

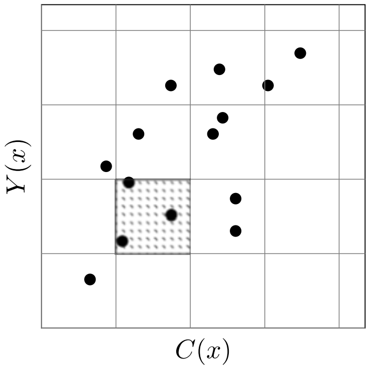


Figure 7: Illustration of the grid and squeeze factor concepts used in PESA-II. As the hatched grid contains three strategies, the squeeze factor of all of these strategies is equal to 3.

Corne et al. (2001) introduced an improved version of the algorithm: PESA-II. Instead of using tournament selection on the strategies, it picks two cells (containing at least one strategy) and compares their squeeze factors. A parent is then randomly picked from the hypercube with the lowest squeeze factor. This solves the problem of the relative low probability in PESA of selecting strategies in uncrowded cells.

5.1.4 IBEA

While both MOEAs and performance measures are widely used concepts in multi-objective optimization, not much effort had been made to combine both until Zitzler and Künzli (2004) introduced the Indicator-Based Evolutionary Algorithm (IBEA). It uses a binary performance measure I (introduced in Section 5.2.2) to calculate fitness values in each generation as follows:

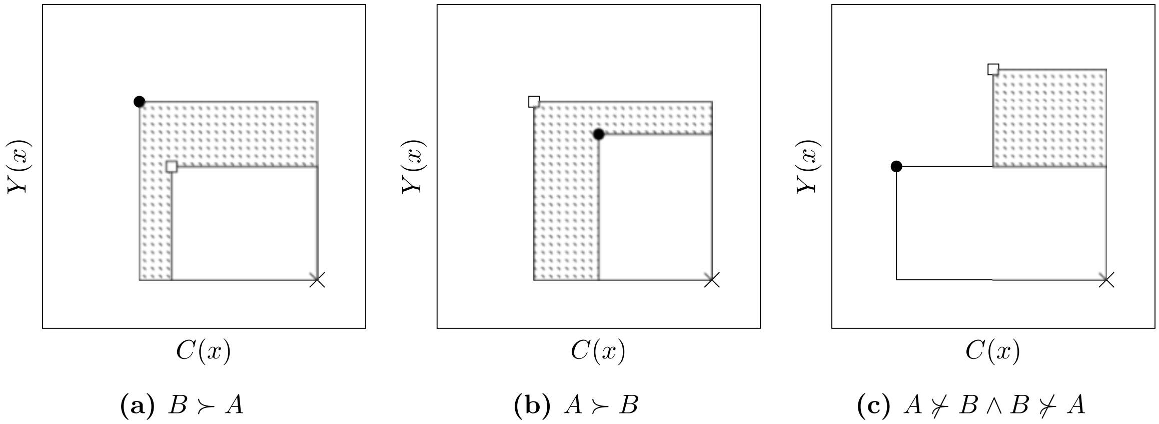
$$F(x) = \sum_{x' \in P \setminus \{x\}} -e^{-I(\{x\},\{x'\})/\kappa}, \qquad \forall\, x \in P, \tag{26}$$

for the current population P. The fitness scaling factor κ should be greater than 0. One of the proposed measures (I_{HD}) is a variant of the Binary Hypervolume indicator I_{HV2} (Section 5.2.2). It uses the unary Hypervolume indicator I_{HV} (Section 5.2.1):

$$I_{HD}(A, B) = \begin{cases} I_{HV}(A) - I_{HV}(B), & B \succ A, \\ I_{HV}(A \cup B) - I_{HV}(B), & \text{otherwise.} \end{cases} \tag{27}$$

This measure gives the hypervolume dominated by the strategies in A, but not by the strategies in B. In case A is dominated by B, the measure is equal to the negative hypervolume dominated by the strategies in B, but not by the strategies in A. Note that, in this algorithm, A and B each always contain exactly one strategy.

Figure 8 shows the three possible scenarios regarding the dominance relation between the strategies in A and B. In Figure 8a, $B \succ A$ and $I_{HD}(A, B)$ is equal to the hatched area multiplied by -1. In case $A \succ B$ (Figure 8b) or $A \not\succ B \wedge B \not\succ A$ (Figure 8c), $I_{HD}(A, B)$ is equal to the hatched area.



(a) $B \succ A$ **(b)** $A \succ B$ **(c)** $A \not\succ B \wedge B \not\succ A$

Figure 8: Illustrations of the possible scenarios regarding the dominance relationship between the strategies in approximation set A and B ($|A| = 1$, $|B| = 1$) for the calculation of $I_{HD}(A, B)$ as used in IBEA. The strategies in A and B are represented by open squares and black dots, respectively. The cross represents the hypervolume reference point.

Any other dominance preserving (see Zitzler and Künzli, 2004) binary performance measure can be used in this algorithm. As the values of such measures can vary a lot for different problems, an adaptive version is also described in Zitzler and Künzli (2004). Both the objective values and the performance measure values are normalized to render the fine-tuning of parameters for each specific problem unnecessary. While the objective values are mapped to values in $[0, 1]$, the performance measure values are mapped to the $[-1, 1]$ range. To achieve the latter, the maximum absolute value c, given the strategies in the current population P, needs to be calculated:

$$c = \max_{x, x' \in P} |I(\{x\}, \{x'\})|. \tag{28}$$

Using the normalized objective values, the fitness values are then calculated as follows:

$$F(x) = \sum_{x' \in P \setminus \{x\}} -e^{-I(\{x\}, \{x'\})/(c \cdot \kappa)}, \qquad \forall\, x \in P. \tag{29}$$

Using this approach, the same reference point and value of κ can be used for all problems.

After calculating the fitness of all strategies in the current population P, members of the population are removed iteratively until its size is equal to n. In each such iteration, the individual x' with the lowest fitness value is removed, i.e. $F(x') \leq F(x)\ \forall x \in P$. In case multiple individuals have been assigned this fitness value, a random individual from this group is removed. Then, the fitness values $F(x)$ of the remaining strategies are updated to $F'(x)$:

$$F'(x) = F(x) + e^{-I(\{x\},\{x'\})/(c \cdot k)}, \qquad \forall\ x \in P. \tag{30}$$

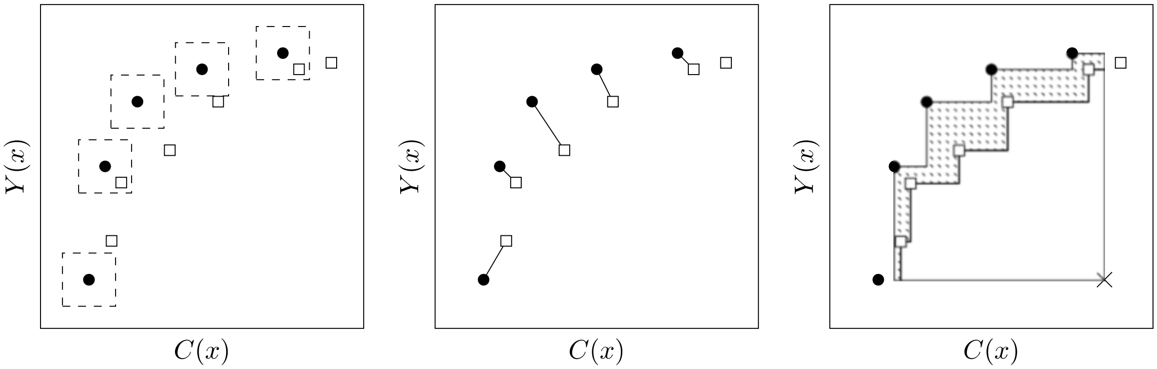
Binary tournament selection with replacement is performed on the remaining population P of size n to fill the temporary mating pool P'. The n new strategies created by applying offspring and mutation operators to P' are added to the population P, increasing its size temporarily to $2n$. In this thesis, the adaptive IBEA-II with the I_{HD} measure is used.

5.2 Performance measurement

The performance of multi-objective algorithms can be compared using various performance measures, of which Zitzler et al. (2003) gave an extensive overview. A distinction can be made between unary and binary measures. In this section, the performance measures used in this thesis are described.

5.2.1 Unary

Unary measures give an indication of the performance of a single approximation set A, i.e. each unary measure I is a function $I : A \to \mathbb{R}$ mapping an approximation set to a real value. These measures indicate to what extent obtained approximation sets have converged to the POF (convergence) and/or are evenly spread out in the objective space (diversity). While some only measure convergence or diversity, others are able to measure both. In this section, it is assumed that the POF, P is known. Three popular unary measures were selected from literature and are described in the following sections. Figure 9 illustrates these measures.



(a) ε-Performance **(b)** Inverted Generational Distance **(c)** Hypervolume

Figure 9: Illustration of unary performance measures. Black dots represent the (optimal) strategies in P. The strategies in approximation set A are represented by open squares.

ε-Performance

In the optimal situation, an algorithm finds all optimal strategies. However, it might be sufficient to find suboptimal strategies with slightly worse objective values. The ε-Performance measure (I_ε) (Kollat and Reed, 2005) defines a hypercube in the objective space centered around each strategy in P. If the objective values of a strategy in an approximation set A fall within such a hypercube, the strategy is matched to the strategy within the hypercube. Each strategy in P can be matched to only a single strategy in A. If multiple strategies in A are within a hypercube, only the strategy with the minimal normalized Euclidean distance is matched. $I_\varepsilon(A)$ is calculated as the ratio of strategies in P which have been matched with a strategy in A. The size ε of the hypercubes is defined by the user. Note that $I_\varepsilon \in [0, 1]$. A value of zero indicates that not any strategy could be matched, while a value of one indicates all strategies in P are matched with a strategy in A.

Figure 9a illustrates the concept of ε-Performance. Here, the hypercubes are given by the dashed squares around the optimal strategies (represented by black dots). Two of these hypercubes contain a strategy belonging to the approximation set A (represented by open squares). Thus, $I_\varepsilon(A) = \frac{2}{5} = 0.4$ in this case.

Inverted Generational Distance

The Inverted Generational Distance (I_{IGD}) (Sierra and Coello, 2005) gives the average distance between the objective values of each strategy in P and its nearest neighbour in the approximation set A. This can be expressed using the following equation:

$$I_{IGD}(A) = \frac{\sum_{i \in P} \min_{j \in A} d_{ij}}{|P|}, \tag{31}$$

where d_{ij} is the distance between strategies i and j, and $|P|$ is the cardinality of P. The Euclidean distance is used in this case. Prior to the calculation of the distance, the objective values are normalized to ensure each objective is weighted equally. The extreme values of P are used to do so. Note that $I_{IGD} \in [0, \infty)$ and $I_{IGD}(A) = 0 \iff P \subseteq A$.

In the example in Figure 9b, the lines represent the distances between each optimal strategy and the nearest strategy in A. I_{IGD} is equal to the average length of these lines. Note that not all strategies in A necessarily contribute to this measure.

Hypervolume

Every approximation set dominates a certain area in the objective space. Consider the case of a two-dimensional minimization problem. The dominated area is infinitely large for any approximation set, as (∞, ∞) is always dominated. When bounded using a certain reference point, the remaining area is referred to as the Hypervolume (I_{HV}) of the approximation set. The Hypervolume is the most popular performance measure in recent literature (Riquelme et al., 2015) and is denoted by I_{HV}.

If the POF P is known, $I_{HV}(P)$ can be calculated. The measure I_{HV}^P represents the difference in Hypervolume between P and the approximation set A:

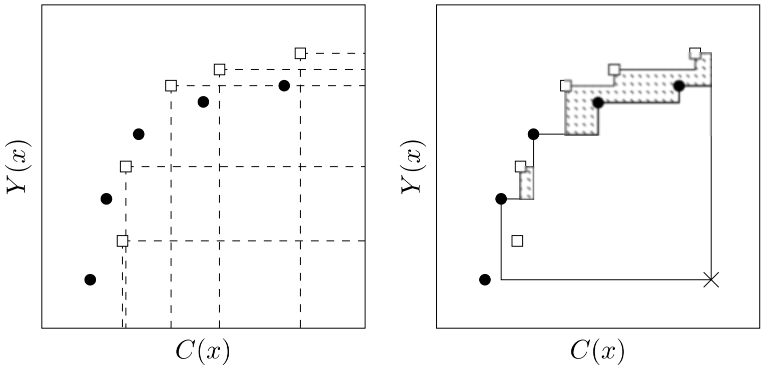
$$I_{HV}^P(A) = I_{HV}(P) - I_{HV}(A). \tag{32}$$

Note that $I_{HV}, I_{HV}^P \in [0, I_{HV}(P)]$ and $I_{HV}^P(A) = 0 \iff P \subseteq A$, i.e. a value of zero indicates that the whole POF is found.

The cross in Figure 9c represents the reference point chosen in this example. Then, the white area represents the Hypervolume of A. $I_{HV}^P(A)$ is represented by the hatched area. Note that both a strategy in A and in P do not affect $I_{HV}^P(A)$.

5.2.2 Binary

Binary measures give an indication of the performance of an approximation set A relative to another approximation set B. Such a measure I is a function $I : (A, B) \rightarrow \mathbb{R}$ mapping two approximation sets to a real value. Note that $I(A, B)$ is not necessarily equal to $I(B, A)$. The remainder of this section describes the two binary measures used in this thesis. Figure 10 illustrates these measures.



(a) Coverage **(b)** Binary Hypervolume

Figure 10: Illustration of binary performance measures. Approximation sets A and B are represented by open squares and black dots, respectively.

Coverage

The Coverage performance measure (I_C) was proposed by Zitzler and Thiele (1998). It gives the ratio of the strategies in B dominated by at least one strategy in A. For any $A, B \in \Psi$, the measure can be calculated as

$$I_C(A, B) = \frac{|\{x_2 \in B : (\exists x_1 \in A : x_1 \succ x_2)\}|}{|B|}. \tag{33}$$

Note that $I_C \in [0, 1]$. If $I_C(A, B) = 0$, none of the strategies in B is dominated by the strategies in A, while $I_C(A, B) = 1$ if all are dominated.

Figure 10a shows the areas dominated by the strategies in A using dashed lines. Two of the strategies in B are within these areas. Thus, $I_C(A, B) = \frac{2}{5} = 0.4$ in this case. Furthermore, $I_C(B, A) = \frac{1}{5} = 0.2$.

Binary Hypervolume

Zitzler (1999) proposed a binary variant of the Hypervolume measure: the Binary Hypervolume (I_{HV2}). This measure also requires the user to choose a reference point. It gives the hypervolume covered by A, but not by B:

$$I_{HV2}(A, B) = I_{HV}(A \cup B) - I_{HV}(B), \tag{34}$$

for any $A, B \in \Psi$. Note that $I_{HV2} \in [0, \infty)$ and $I_{HV2}(A, B) = 0 \implies I_C(A, B) = 0$.

The chosen reference point in the example in Figure 10b is indicated by the cross. The area enclosed by solid lines represents the area dominated by B, bounded by the reference point. Then, the hypervolume dominated by A, but not by B, is given by the hatched area.

6 Case study

The United States scenario was used as a case study for the proposed approach. Knudsen et al. (2016) recently used MISCAN-Colon (and two other similar models) to inform the United States Preventive Services Task Force on CRC screening, using a hypothetical cohort of unscreened 40-years-olds. An adherence of 100% was assumed for all tests. The same assumptions were made for this case study and the same surveillance rules and calibrated parameters were used. Following the US recommendations, both costs and LYG were discounted 3% annually.

In Section 6.1, the performances of the algorithms as described in Section 5.1 are compared on a small test case. This test case was fully enumerated, so that the optimum was known and the strategies found by the algorithms could be compared with the optimal strategies to assess their performances. As the objective values of all strategies were known after enumeration, no further simulations were required for the comparison of the performances of the algorithms on the test case. Based on the performances on the test case, the most promising algorithm was chosen to approximate a solution to the real case in Section 6.2.

While normally simulations are run with a population of at least 1.0e7 individuals, a population size of 2.5e6 was chosen to limit the computation time. Common Random Numbers were used to ensure that all simulations share the same individuals. In this way, while the absolute results might not be sufficiently accurate for interpretation, the cost-effectiveness of the strategies can be compared, despite the relatively low simulated population size. The final results for the real case were simulated again with a population size of 1.0e7.

6.1 Test case

In the test case, only colonoscopy and FIT20 are considered as screening tests. These tests were chosen because they are dissimilar: FIT20 is a relatively cheap stool-based test, while colonoscopy is a relatively expensive endoscopic test. 13 possible screening ages are considered, ranging from 51 to 75, with intervals of two years. This range includes many ages often considered for screening (e.g. Knudsen et al., 2016). With these settings, 1,594,323 strategies are possible.

Enumeration

The test case was fully enumerated to find the optimal strategies. The Extreme-scale Model Exploration with Swift (EMEWS) framework (Ozik et al., 2016) was used to run all simulations efficiently in parallel on the Bebop supercomputer at the Argonne National Laboratory. Figure 11 shows the obtained costs and LYG as gray dots. The scenario without screening is indicated by a cross.

Performance measures

As all possible strategies had been simulated, the performance of an algorithm could be evaluated on the test case. The Pareto optimal set consists of 381 strategies. To prevent certain areas of the POF to influence the unary performance measures I_ε and I_{IGD} too much, a trimmed version of the Pareto optimal set was used as reference set. The following procedure, based on

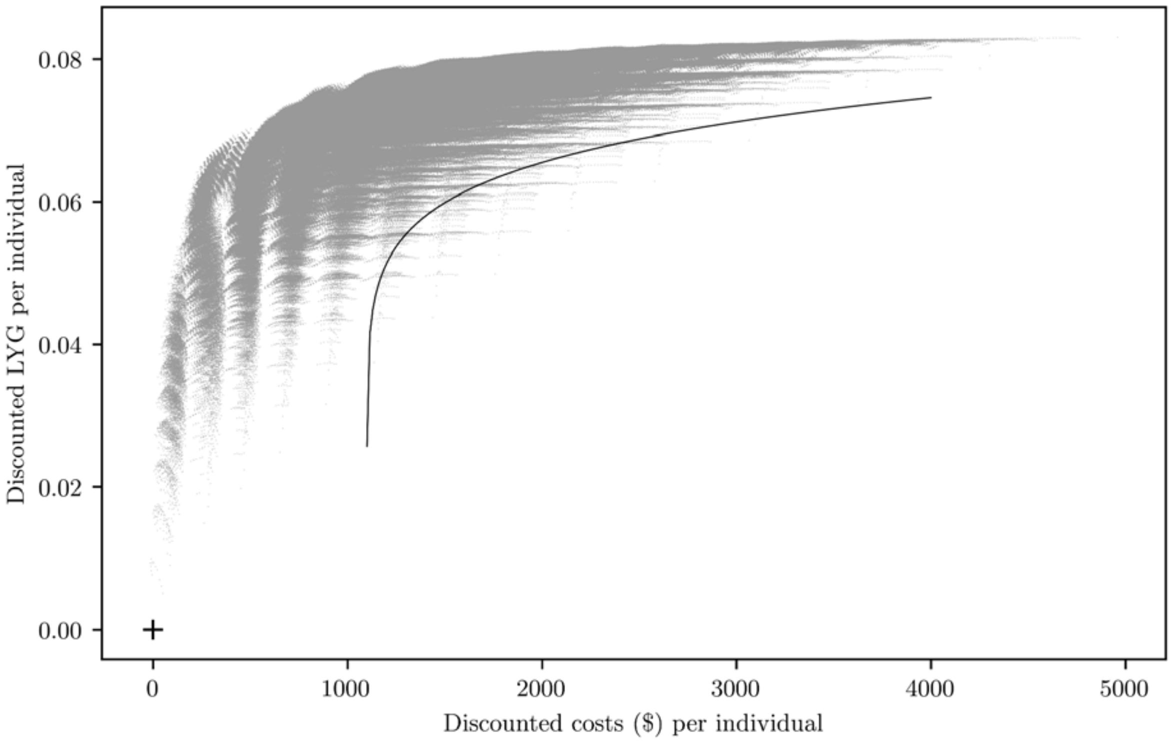


Figure 11: Enumeration of the test case. Each gray dot represents a strategy. The cross represents the scenario without screening. Strategies under the curve are used to initialize the algorithms.

the concept of ε-dominance (Laumanns et al., 2002), was used to do so. The objective space was divided into hypercubes of size (5, 0.0001). In hypercubes containing multiple Pareto optimal strategies, only the strategy with the minimum normalized Euclidean distance to the upper left corner of the hypercube was preserved. Furthermore, strategies with costs higher than $3,500 per individual were removed. This resulted in a reference set containing 148 strategies. The entire POF was used for the calculation of I_{HV}^{P}. I_{ε} was implemented with ε=(5, 0.0001). For I_{HV}^{P}, (4,000, -0.01) was used as the reference point.

Algorithm settings

The performance of an algorithm is dependent on its settings (i.e., parameters and operators). The relation between each setting and the performance varies per problem. Therefore, different settings of each algorithm were evaluated to find which yield the highest performance on the test case. Table 2 summarizes the chosen best performing settings for each algorithm.

For all algorithms and performance measures, C_{tp} appeared to be the superior crossover operator and clearly outperformed the other considered operators (C_{sp} and C_b). The crossover probability appeared to be of less importance, as varying this parameter did not result in significant changes in the performance measures. However, this probability should be sufficiently high to ensure the algorithm does not simply function like a local search procedure. A probability of 0.9 was chosen for all algorithms. For the mutation operator, the parameter r was chosen to be $0.5/|A|$. Thus, it mutates on average 0.5 decision variables per strategy. This value scales

Table 2: Settings (parameters and operators) and their chosen best performing values (on the test case) for each algorithm.

Setting	NSGA-II	SPEA2	PESA-II	IBEA
Crossover operator	C_{tp}	C_{tp}	C_{tp}	C_{tp}
Crossover probability	0.9	0.9	0.9	0.9
Mutation operator	$M^{0.75}_{0.5/\lvert A\rvert}$	$M^{0.75}_{0.5/\lvert A\rvert}$	$M^{0.75}_{0.5/\lvert A\rvert}$	$M^{0.75}_{0.5/\lvert A\rvert}$
Population size	200	80	80	80
Archive size	-	300	100	-
Grid size	-	-	(25, 0.0005)	-
Scale parameter	-	-	-	0.001
Reference point	-	-	-	(2, 2)

with the number of ages considered, so it will perform similarly in the real case. Furthermore, a value of p of 0.75 gave the best and the most stable results. Thus, if the operator mutates a decision variable corresponding to an age at which an intervention is already scheduled, the intervention is removed with a probability of 0.75. Else, it changes the test to a randomly selected other test.

NSGA-II showed a rapid increase in performance for all performance measures when using a small population size. However, its performance stalls relatively soon. A large population size, in contrast, results in a slow start but works better in the long run. A size of 200 appeared to be the well-functioning middle-way.

For SPEA2, a large archive yields a stable performance. However, an archive that is too large slows the development of the algorithm down. A relatively small population size is desirable, because the archive will be updated more frequently, However, for the problem at hand, the population size should be sufficiently large to be able to simulate many strategies simultaneously. A population size of 80 and an archive size of 300 were chosen.

PESA-II should be configured so that the squeeze factor is not one for all or most strategies. It should also not be the case that most archived strategies are within only a few grids. This can be achieved by choosing a balanced combination of the archive size and grid size. Grids of size (25, 0.0005) and an archive with 100 strategies appeared to work well. Furthermore, the best results were obtained using a population size of 80.

For IBEA, the recommended reference point of $(2, 2)$ (Zitzler and Künzli, 2004) works well on this case and was therefore chosen. In a small population, the strategies in the upper left corner are assigned a relatively high fitness to. Thus, the population size should be sufficiently large to allow the algorithm to explore strategies in the tails (i.e. the lower left and upper right corner). A size of 80 was chosen. The scale parameter should be sufficiently large (>0.0001) to prevent numerical problems in the calculations of equations (29) and (30). A value of 0.001 was chosen.

Algorithm comparison

Using the chosen settings, the performance of the algorithms could be compared. Each algorithm was run 100 times on the test case with different random seeds and initial strategies. After 10,000 'simulations' (or, more generally: function evaluations), the runs were terminated.

For a single run per algorithm, the strategies found by each algorithm (i.e. the function evaluations) are shown in Figure A.1 in Appendix A. PESA-II seems to evaluate only a few expensive strategies. Whereas the other algorithms found the left part of the POF rather well, IBEA evaluates only a few cheap strategies. NSGA-II and SPEA2 seem to behave similarly. Obviously, well-founded conclusions can not be drawn based on such plots. To draw further conclusions on the performance and stability, a more rigorous analysis is now given, exposing nuances by using performance measures and multiple runs.

Figure 12 shows the unary performance measures for all algorithms as a function of the number of simulations. NSGA-II and SPEA2 exhibit similar behaviour for all measures and are relatively stable. PESA-II is less stable and shows lower performance after any number of function evaluations. IBEA develops relatively fast, but ultimately reaches worse values than NSGA-II and SPEA2.

Figure 13 gives boxplots of the final values for each algorithm and unary performance measure. For each measure, a statistical analysis was carried out to draw conclusions on the differences in final performance. The test statistics and p-values are included in Appendix B. The distributions of the I_ε values seem to be normal. However, the distributions of I_{IGD} and I_{HV}^P appear to be skewed. These observations are confirmed by a Shapiro-Wilk test for normality (Table B.1). Log-transformations were not able to normalize all distributions of I_{IGD} and I_{HV}^P. As normality can not always be assumed, a conservative non-parametric approach was taken by using the Kruskal-Wallis test as an omnibus test. This test rejects the null-hypothesis that the four sets of outcomes originate from the same distribution for all measures, with p-values below 0.001 (Table B.2). As there is a statistically significant difference between the algorithms, pairwise post-hoc tests can be performed to find the best performing algorithm. Conover's-Test was used to do so. As multiple inferences are drawn from the same data, the multiple testing problem arises. The Hölm procedure, as recommended by Derrac et al. (2011), was used to adjust the obtained p-values to account for the increased probability of type-I errors due to this problem (Table B.4). Using a significance level of 0.05, it can be concluded that NSGA-II outperforms the other algorithms on all unary performance measures. PESA-II is outperformed by all other algorithms. No significant difference could be found between SPEA2 and IBEA for I_ε, although SPEA2 outperforms IBEA on I_{HV} and I_{IGD}.

The binary performance measures are calculated on the approximation sets based on all 10,000 simulations. A random pairing of the runs was made to obtain 100 pairs for each of the six possible combinations of two algorithms. The measures were than calculated for each pair of approximation sets. Table 3 gives the resulting values for Coverage and Binary Hypervolume. For both measures, a paired difference test was used to statistically analyze the difference for each combination of algorithms. A Shapiro-Wilk test was used to assess the normality of the differences (Table B.3). For I_C, the null hypothesis of normality could not be rejected for any differences using a significance level of 0.05. However, this hypothesis is rejected for most differences in I_{HV2}. Therefore, the non-parametric Wilcoxon signed-rank was used for I_{HV2} and the paired two-sample t-test was used for I_C. The results can be found in Table B.4. For I_{HV2}, no significant difference was found between NSGA-II and SPEA2, while I_C indicates that NSGA-II outperforms SPEA2. Both algorithms significantly outperform PESA-II and

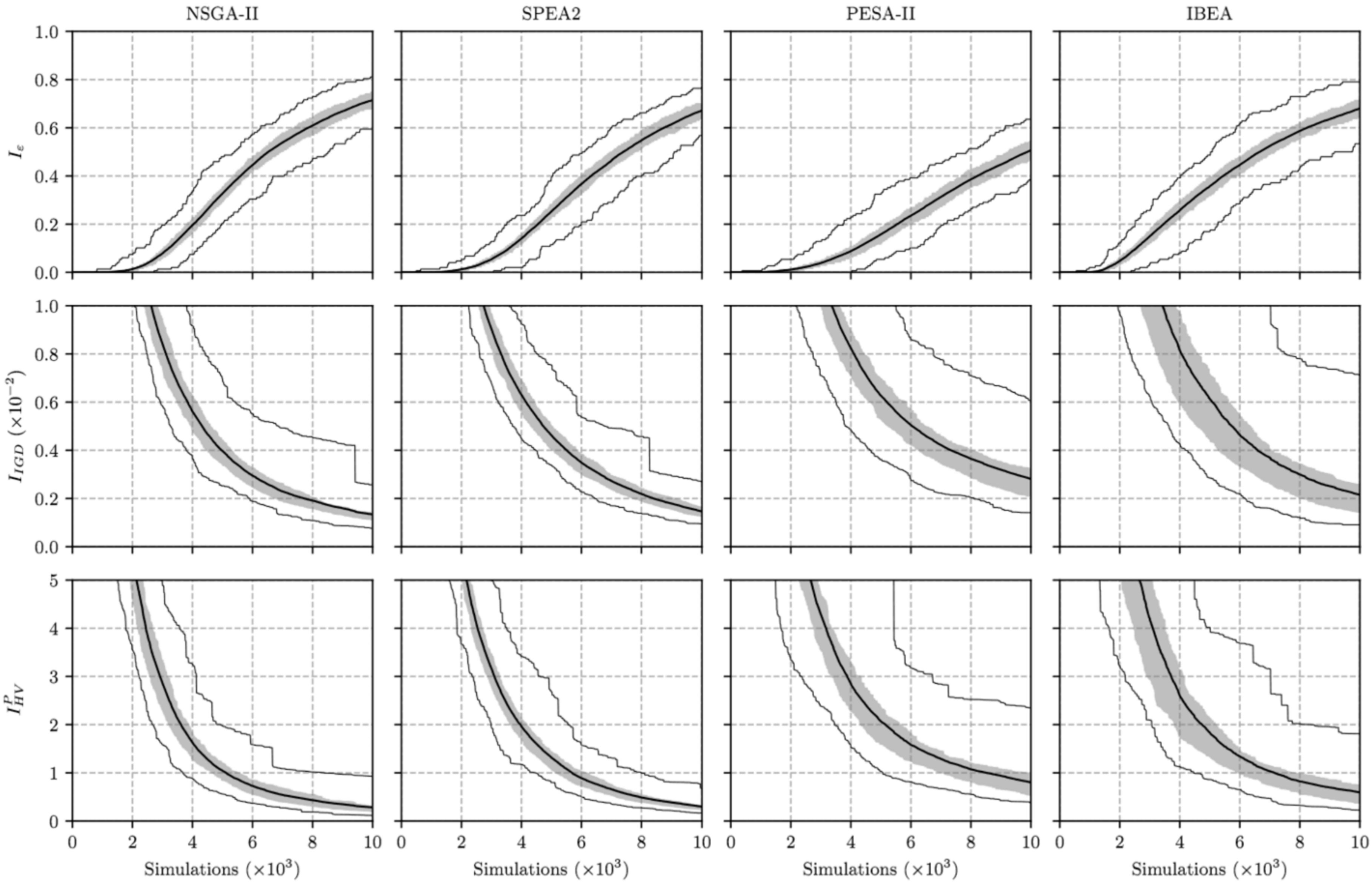


Figure 12: Mean performance over time of the algorithms on the test case after 100 runs with the range between the first and third quartiles as shaded area and the minimum and maximum values as thin lines. I_ε is applied with ϵ=(5, 0.0001) and I_{HV}^P with reference point (4,000, -0.01).

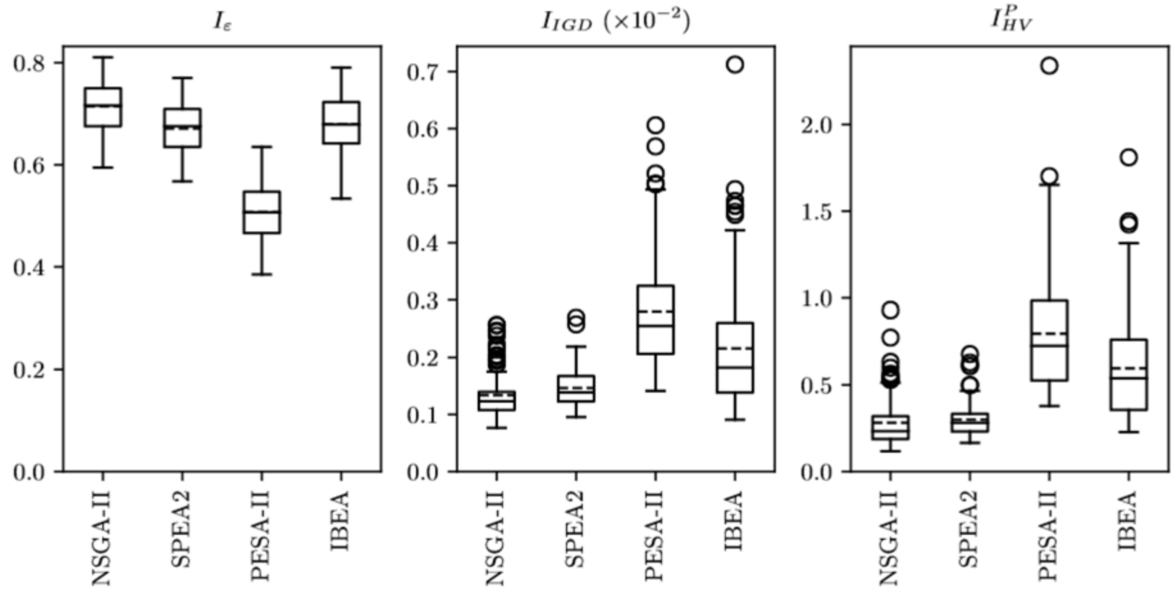


Figure 13: Box plots of the unary performance measures in 100 runs for each algorithm after 10,000 simulations on the test case. The means are indicated by dashed lines and outlier by circles. I_ε is applied with $\epsilon = (5, 0.0001)$ and I_{HV}^P with reference point (4,000, -0.01).

Table 3: Binary performance measures for all permutations of two algorithms. The measure is calculated on the approximation sets based on all 10,000 simulations per run. 100 runs are compared. The values are means over all runs, with the standard deviations in parentheses.

(a) Coverage $I_C(A, B)$ with A vertically and B horizontally.

	NSGA-II	SPEA2	PESA-II	IBEA
NSGA-II	-	0.2179 (0.0567)	0.3609 (0.0843)	0.2320 (0.0642)
SPEA2	0.2208 (0.0529)	-	0.3677 (0.0836)	0.2264 (0.0667)
PESA-II	0.1528 (0.0546)	0.1643 (0.0561)	-	0.1731 (0.0679)
IBEA	0.2093 (0.0496)	0.2137 (0.0555)	0.3446 (0.0802)	-

(b) Binary Hypervolume $I_{HV2}(A, B)$ with A vertically and B horizontally. (4,000, -0.01) is used as the reference point.

	NSGA-II	SPEA2	PESA-II	IBEA
NSGA-II	-	0.1737 (0.0851)	0.4729 (0.2847)	0.3726 (0.1680)
SPEA2	0.1684 (0.1023)	-	0.4752 (0.2792)	0.3840 (0.1771)
PESA-II	0.1116 (0.0944)	0.1228 (0.0649)	-	0.2774 (0.1628)
IBEA	0.1236 (0.0944)	0.1396 (0.0691)	0.3752 (0.2824)	-

IBEA. Among PESA-II and IBEA, IBEA is the best performing algorithm based on the binary measures.

As can be concluded from Table B.4, NSGA-II is not outperformed by any of the other algorithms on any of the unary or binary measures. SPEA2 is outperformed by NSGA-II on all measures except I_{HV2}, and it dominates PESA-II and IBEA on most measures. IBEA outperforms PESA-II on all measures. This leaves PESA-II as the worst performing algorithm: it is outperformed by all other algorithms on all measures with a statistically significant difference. In conclusion, NSGA-II is the best performing algorithm for the problem at hand. Therefore, this algorithm will be used to solve the real case.

Zitzler et al. (2001) also found NSGA-II and SPEA2 to have competitive performance when compared on different test problems. They concluded that these algorithms outperform SPEA and PESA. Kollat and Reed (2006) compared NSGA-II, SPEA2, ε-NSGA-II and εMOEA on a "four-objective long-term groundwater monitoring design" problem. However, they found that NSGA-II was significantly outperformed by the other algorithms considered. This difference in conclusion might be explained by the different number of objectives, the usage of real decision variables or simply the fact that the problem differs fundamentally. Anyway, it underlines such results can not be generalized and the best performing algorithm should be identified per problem and formulation.

6.2 Real case

For the real case, all ten tests as defined in Section 3.1 (Table 1) are considered. The used ages range from 40 to 90, with intervals of 1 year. With these 10 tests and 51 ages, approximately 1.29e53 strategies are possible in the real case. Obviously, even with extremely high computing power this number is too high for enumeration. Therefore, NSGA-II is used to approximate the solution to this problem. The algorithm settings as used in the previous section are used. It is initialized with 200 of the 251 'simple' strategies as summarized in Table 4, selected using the NSGA-II procedure as described in Section 5.1.1.

A network of computers was used to be able to simulate all strategies in a generation simultaneously. The search was terminated after 1,227 generations and a total of 215,272 unique simulations. To analyze to what extent the algorithm had converged, statistics for the POF after each generation were calculated and plotted in Figure 14. The hypervolume increased substantially in the first 100 generations, after which it grew consistently further, albeit at a lower rate. In the first phase, up to 41 strategies were added to the POF in a single generation and up to 35 were removed. The number of added and removed strategies decreased as the algorithm progressed, but, even after 1,200 generations, strategies are still added and removed in most generations. However, these changes represent minor improvements, as the hypervolume barely increases after generation 500.

Pareto optimal frontier

As all simulations ran by the algorithm shared the same random seed, it might be that the algorithm found strategies that only work well for the specific population of 2.5e6 individuals corresponding to this seed. Furthermore, this population size is too small for the absolute

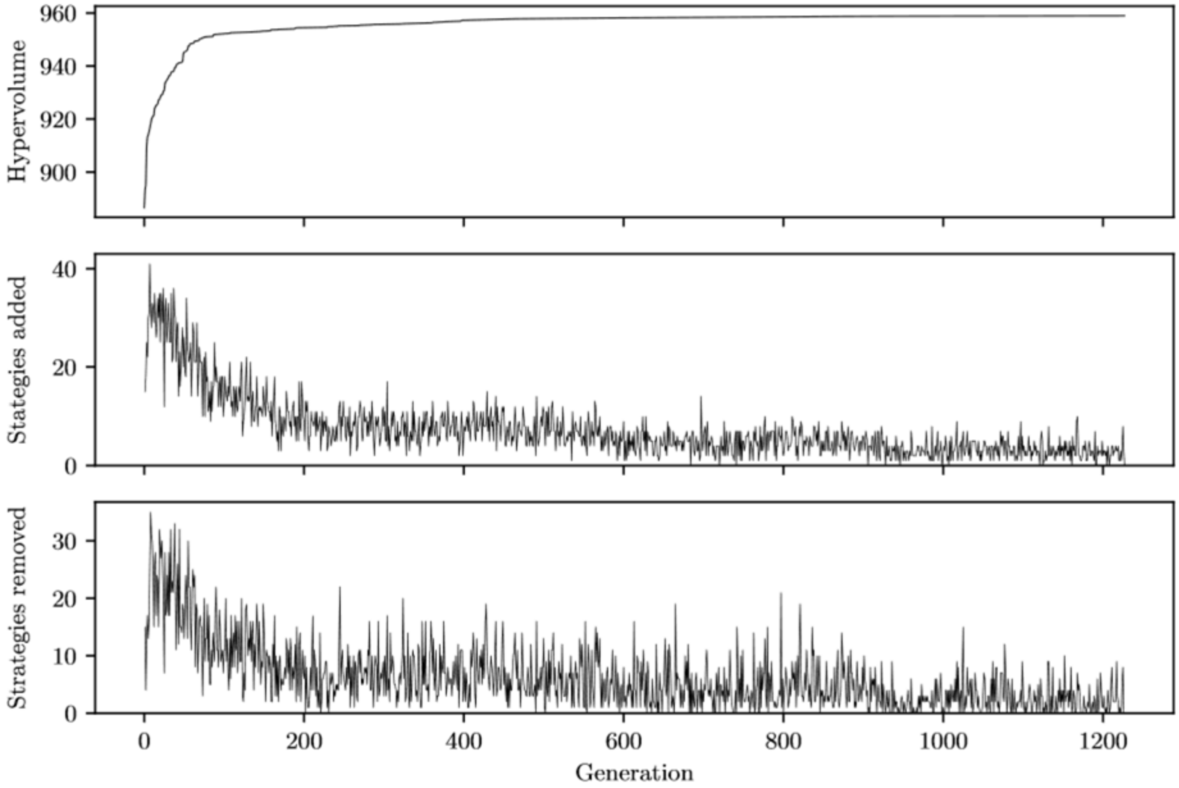


Figure 14: Statistics of the POF per NSGA-II generation on the real case. (10,000, 0) was used as the reference point for the calculation of the hypervolume. The number of strategies added and removed to the POF in each generation is also given.

interpretation of the outcomes. For these reasons, the 1,420 found Pareto optimal strategies were simulated again with a population size of 1.0e7. A different random seed was chosen for (and shared by) these simulations. This resulted in a Pareto optimal set containing 843 strategies. Figure 15 shows the resulting costs and LYG. The simple strategies considered in Knudsen et al. (2016) are summarized in Table 4. Their costs and LYG (based on simulations with a population size of 1.0e7) are also plotted in Figure 15 to be able assess the possible improvement in cost-effectiveness. Clearly, many complex Pareto optimal strategies have been found compared to the number of simple Pareto optimal strategies.

The left side of the POF (i.e. LYG smaller than 0.08) seems to be rather well approximated by the simple strategies. However, the upper and right auxiliary plots in Figure 15 reveal that, especially for the lowest costs, a relative improvement in LYG of up to 22% is possible. For an effectiveness of 0.035 LYG, costs can be reduced by 201%: from $32 to -$32. The possible relative improvements in costs and LYG are smaller further along this part of the POF. The POF 'bends' in the area with LYG higher than 0.08 and costs lower than $3,000. This area is not well covered by the simple strategies, allowing the possible improvement in LYG to increase up to 9%. The possible cost reduction reaches 60% for a desired effectiveness of at least 0.085 LYG. For costs higher than $3,000, almost no simple strategies exist, while many complex Pareto optimal strategies have been found. Effectiveness can be improved by up to 9% and costs can be reduced by up to 86%.

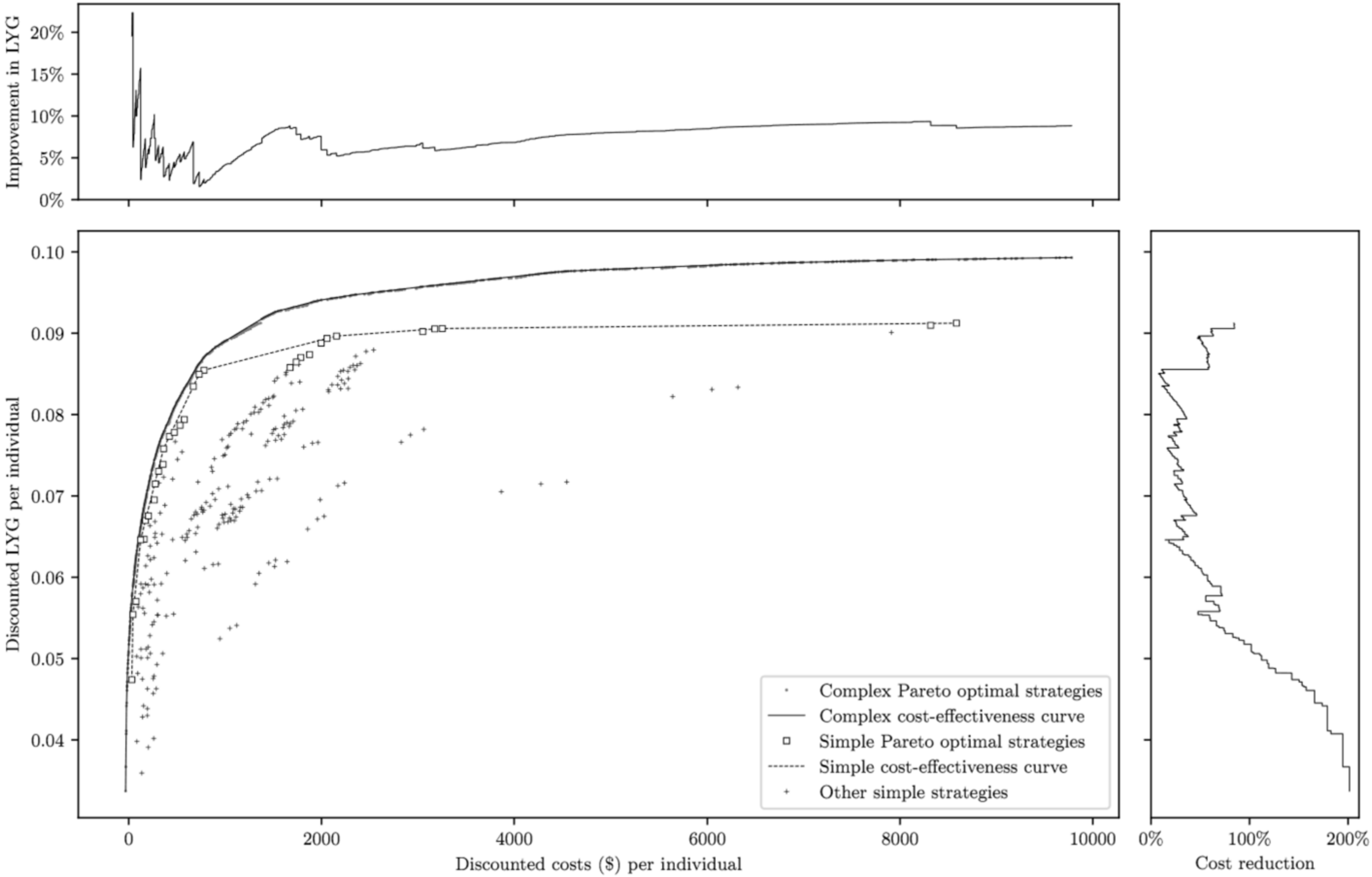


Figure 15: The central plot shows the simple and complex Pareto optimal strategies. The other simple strategies are also shown. The cost-effectiveness curves connect the cost-effective strategies. The upper plot shows the possible relative improvement in LYG, given certain costs. The right plot shows the possible relative cost reduction, given a desired minimum effectiveness.

Table 4: Simple strategies as in Knudsen et al. (2016) (excluding strategies using test types not considered in this thesis).

Test	Start ages	Stop ages	Intervals (years)	Unique strategies
FIT10	45, 50, 55	75, 80, 85	1, 2, 3	27
FIT20	45, 50, 55	75, 80, 85	1, 2, 3	27
FIT40	45, 50, 55	75, 80, 85	1, 2, 3	27
FIT-DNA	45, 50, 55	75, 80, 85	1, 3, 5	27
COL	45, 50, 55	75, 80, 85	5, 10, 15	20
FS	45, 50, 55	75, 80, 85	5, 10	15
FS, FIT10[a]	45, 50, 55	75, 80, 85	5_2, 5_3, 10_1, 10_2	36
FS, FIT20[a]	45, 50, 55	75, 80, 85	5_2, 5_3, 10_1, 10_2	36
FS, FIT40[a]	45, 50, 55	75, 80, 85	5_2, 5_3, 10_1, 10_2	36

[a] This strategy uses FS and FIT. The corresponding intervals are separated by an underscore. If both tests are scheduled at the same age, a combined test type is used (e.g. FS+FIT10).

Strategies

To visualise the interventions used by the found strategies, the POF is divided into seven sections based on costs. For each section, the average number of times each test is used is reported in Figure 16. Strategies are divided into five parts based on age.

Along the whole POF, FIT10 is intensively used in all strategies. In line with this finding, Wilschut et al. (2011) compared the cost-effectiveness of FIT strategies with different cut-offs for the Dutch population and concluded a cut-off value of 10μg/g[1] is optimal. They argued that most studies on cost-effectiveness identified the most sensitive stool-test considered as optimal. For cheaper strategies, FIT10 is especially used between ages 60 and 69. As strategies get more expensive, FIT10 is used more often at both lower and higher ages. This is in line with the stool-based screening programs implemented in England and Finland, which start at age 60 and end at age 69 (Zavoral et al., 2014). More intensive programs start at age 50 and end at age 69 (England, Scotland and Spain) or 74 (Netherlands and France) (Zavoral et al., 2014).

Interestingly, optimal strategies use the less sensitive FIT20 and FIT40 almost exclusively between ages 40 and 49 and between ages 70 and 90. This observation can be explained as follows. The prevalence of CRC among younger individuals (aged 40 to 49) is relatively low. Therefore, using a low cut-off value will result in many false-positive test results and consequently high costs, while the effectiveness is minimal. Increasing the cut-off to 20μg/g or 40μg/g will significantly decrease the number of false-positives, while only barely decreasing the effectiveness. While the prevalence of CRC among older individuals is relatively high, their life expectancy is relatively low. Thus, sensitive tests are likely to cause overdiagnosis by detecting lesions that will not develop into cancer within the remaining lifetime. Less sensitive tests are therefore more appropriate between ages 70 and 90. This result implies strategies could be improved by varying the FIT cut-off values depending on age. However, this could also be a spurious result, caused by the stochastic nature of the algorithm. In that case, the less sensitive FITs will be removed or replaced by FITs with a lower cut-off value if the algorithm is ran for a longer time.

Starting from costs of approximately $1,000, colonoscopies are introduced: first between

[1] Wilschut et al. (2011) expressed hemoglobin levels in ng/mL. 1μg/g is equal to 5ng/mL.

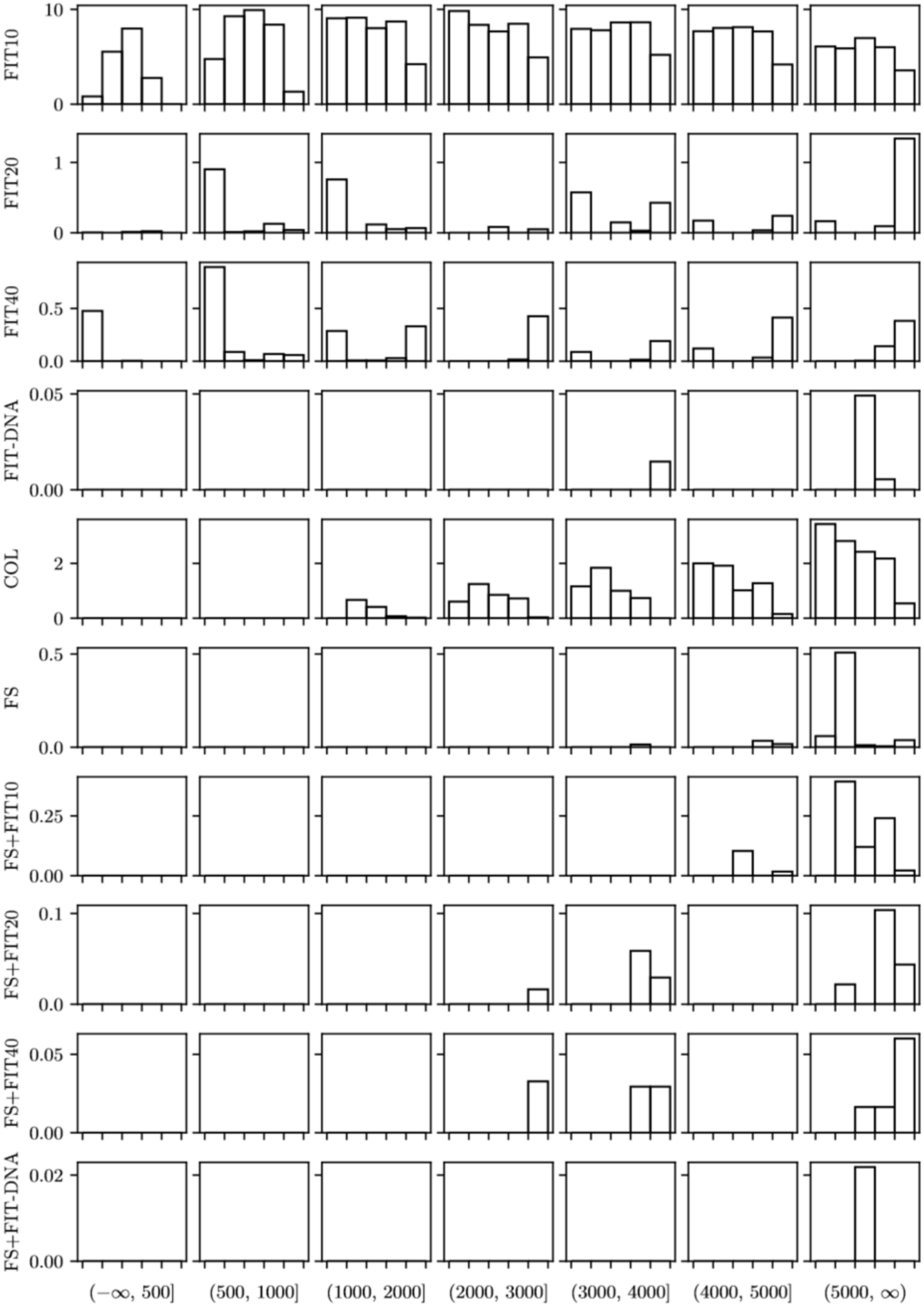


Figure 16: Interventions in a strategy on the found POF, per test type and discounted costs per individual range. Each bar plot shows the average number of interventions in the age intervals $[40, 50)$, $[50, 60)$, $[60, 70)$, $[70, 80)$ and $[80, 90]$, respectively.

ages 50 and 79 and later also before age 50 (and some after 79). These strategies still intensively use FIT. Thus, simple colonoscopy strategies could possibly be improved by adding FIT interventions. This is supported by Lane et al. (2010), who found that using FITs between every two colonoscopies in a surveillance program speeds up the detection of lesions. It should be noted that only high-risk individuals participate in such a program (whereas this thesis focuses on the general population) and cost-effectiveness was not considered. Similarly, Knudsen et al. (2012) found that negative colonoscopies should be followed up with less intense screening tests, such as the FIT.

Optimal strategies rarely use FIT-DNA, FS and combinations of FS and FIT. Strategies using these tests often have an ICER way above any reasonable willingness-to-pay threshold. Thus, these tests should not be used in most cases.

Despite the improved efficiency of the found strategies, caution should be taken when generalizing these results to the real world. In modelling studies, perfect adherence is usually assumed. Knudsen et al. (2016) reasoned that "identifying model-recommendable strategies based on imperfect adherence could result in selection of strategies with short intervals to make up for suboptimal population-level adherence; it could also lead to overscreening for those individuals who adhere to recommendations, potentially at the cost of unnecessary risks and burden." In reality, however, adherence rates are far from perfect (Subramanian et al., 2004). Increasing the complexity of screening will likely cause confusion among the population and result in even lower adherence rates.

Cost-effectiveness

Of the strategies on the POF, 58 are cost-effective. These are given in Table C.1 in Appendix C. Only 12 simple strategies are cost-effective. Figure 15 shows the corresponding cost-effectiveness curves.

As discussed in Section 2, one should beware of 'kinks' in the cost-effectiveness curve. Whereas at least one 'kink' can clearly be observed in the simple cost-effectiveness curve in Figure 15, the complex cost-effectiveness curve is significantly more 'smooth'. Therefore, ICERs are likely to be more accurate, whereas the ICERs for the simple cost-effective strategies are likely to be underestimated. Figure 17 shows the maximum attainable LYG (and the corresponding costs) for willingness-to-pay thresholds up to 150,000, both for the simple strategies as well as for the found complex cost-effective strategies. This figure clearly shows that decision makers have significantly more options if they would consider complex instead of simple strategies. For these reasons, the proposed algorithmic approach of finding cost-effectiveness curve is recommended for future analyses.

The knowledge of the approximate location of the optimal cost-effectiveness curve (and POF) can be used in future cost-effectiveness analyses. It enables researchers to assess to what extent the performance of the strategies in their research differs from the optimum. If the observed difference is not satisfactorily small in certain areas of the objective space, they might attempt to minimize this difference by adding strategies that are sufficiently simple. The efficient strategies found in this thesis can serve as inspiration for such new strategies for CRC screening.

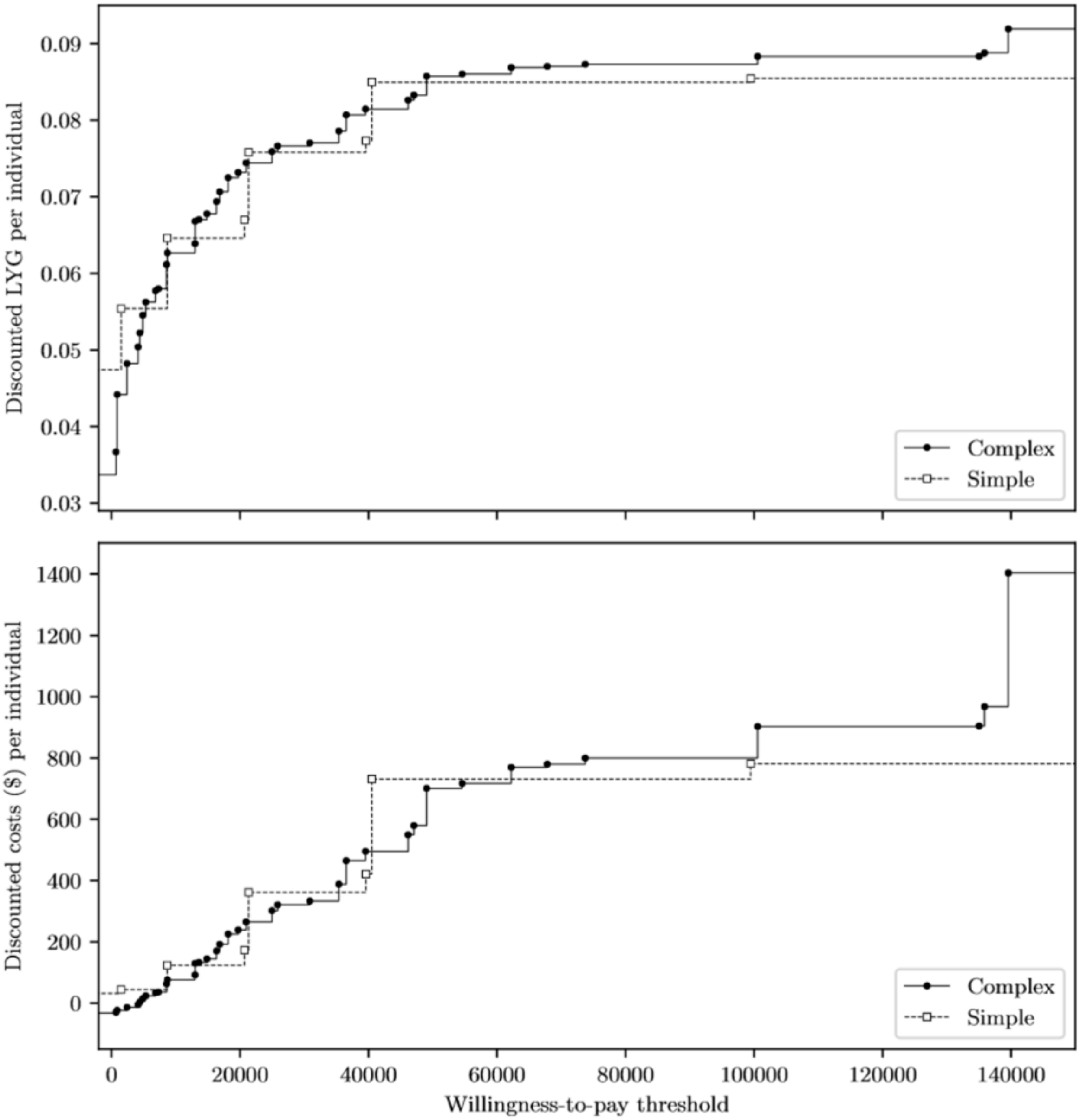


Figure 17: LYG and costs of the cost-effective strategies, with the corresponding ICER on the horizontal axis, for simple and complex strategies. The lines indicate the maximum attainable LYG and costs given a willingness-to-pay threshold. Note that the graph continues to the right.

7 Conclusion

In this thesis, a methodology is proposed to algorithmically find cost-effective strategies for CRC screening, using the MISCAN-Colon microsimulation model to predict costs and LYG. The class of multi-objective evolutionary algorithms appeared to be appropriate for problems of this nature. The specific case of CRC in the United States was used as a case study to be able to systematically compare the performance of the algorithms within this class and tune their parameters. Various performance measures were chosen from literature for these comparisons. The statistical procedure to compare the performance of the considered algorithms as demonstrated in this thesis was able to identify significant differences. Inverted Generational Distance and Hypervolume turned out to be the most useful measures, as significant differences were found between all algorithms. NSGA-II was identified as the best performing algorithm and was therefore applied to an actual case to obtain the final results.

For the specific case of screening for CRC in the United States, the cost-effectiveness curve was optimized. Major improvements are possible relative to the strategies considered in previous research. Effectiveness could be improved by 2-22%, depending on the budget. Costs could be reduced by 8-201%, depending on the desired effectiveness. No 'kinks' could be observed in the optimized cost-effectiveness curve, solving the problem of underestimated ICERs. Furthermore, relatively many cost-effective strategies were identified. These improvements allow decision makers to choose from a wider range of screening strategies, for which ICERs are accurately estimated. Carolyn Rutter, member of the Cancer Intervention and Surveillance Modeling Network (CISNET) and co-author of Knudsen et al. (2016), commented as follows: "We currently estimate cost-effectiveness curves by comparing the cost and effectiveness of a relatively small subset of policy alternatives. Because of this, we need to infer the dominance of some strategies. This thesis shows how we can eliminate the need to infer dominance, allowing us to determine the most effective strategy at any given cost" (personal communication, December 18, 2019).

Cost-effective strategies intensively use FIT, mostly with a cut-off value of 10μg/g. For lower budgets, interventions with this test are mainly implemented between ages 60 and 69. As the budget rises, gradually implementing these interventions at lower and higher ages is cost-effective. FIT interventions with a higher cut-off value are also incorporated in optimal strategies, but to a lesser extent and almost exclusively at both relatively low and high ages. Starting from a budget of $1,000 (discounted) per individual, the usage of colonoscopies as a screening test becomes cost-effective. A pattern similar to that of FIT10 is observed regarding the ages at which this test is implemented along the cost-effectiveness curve. FIT-DNA, FS and FS combined with a FIT are less prevalent in the optimal strategies.

To achieve optimal efficiency in CRC screening, the strategies found in this thesis should be implemented in practice. However, these strategies are more complex compared to strategies currently used due to the usage of different tests and the lack of a fixed interval between interventions. This increased complexity is likely to cause confusion among the population and resistance among doctors, despite the possible improvement in cost-effectiveness. Decision makers will have to decide whether the benefits of the complex strategies outweigh their harms.

7.1 Limitations

As in other studies on multi-objective optimization, a test case was used to assess the performance of different algorithms and parameter settings. Based on this test case, an algorithm and parameter settings were identified as optimal. It was assumed that this would also be the optimal choice for the real case. However, this extrapolation of the results on the test case to the real case might not be justified. Furthermore, parameter settings were tuned by visual inspection of their effect on the performance measures. A systematic factorial design might be more appropriate.

MISCAN-Colon uses many parameters that determine, among other things, the growth of the disease and the costs associated with screening and treatment. The values used for these parameters are highly uncertain over the considered time interval. Therefore, a (probabilistic) sensitivity analysis is often performed to assess the outcomes under different parameter values. This thesis optimizes the POF assuming a specific set of parameter values, while other values might yield substantially different results.

The accuracy of the outcomes are limited by the accuracy of the simulation model. Thus, the possible increase in cost-effectiveness might in reality differ significantly from the predictions. Furthermore, the algorithms might have overfitted on the simulation model: possibly the obtained optimal strategies are strategies that yield favorable results due to the specific structure and assumptions of the simulation model. In other words, it is unsure if optimizing the input of the simulation model is equivalent to finding strategies that work well in reality.

Finally, the obtained strategies are possibly not feasible in reality due to the limited capacities of hospitals. This is especially problematic for colonoscopies, as for example shown for the US by Vijan et al. (2004).

7.2 Suggestions for future research

In cancer screening research, the concept of risk-based or personalized screening is currently upcoming. Obviously, this makes the problem at hand even more complex. Methods as proposed in this thesis can be used to determine an optimal stratification of the population and/or find optimal strategies for each stratum.

In this thesis, three cut-off values for FIT where considered. However, as this is a quantitative test, any cut-off value can be used. It might be interesting to consider the cut-off value as a real decision variable, enabling a further optimization of the POF.

Surveillance was implemented in MISCAN-Colon following the US recommendations. By assuming a fixed policy, the problem arises that the screening can be more intense than the surveillance, in particular for the more expensive strategies. A solution would be to incorporate the surveillance policy in the decision variables. By enlarging the solution space in this way, more cost-efficient solutions can likely be found.

As mentioned in the limitations, obtained strategies might be infeasible in practice due to capacities of hospitals. Multiple methods have been proposed to incorporate constraints in MOEAs. The capacities of hospitals can be implemented as such constraints.

In case of expensive function evaluations, meta-models are often trained to (partly) replace the actual function. These are often implemented in the form of artificial neural networks or

Gaussian processes. In fact, the whole class of surrogate-assisted multi-objective evolutionary algorithms is dedicated to MOEAs aided in some way be a meta-model. Simulation models for the evaluation of cancer screening or generally computationally expensive. Therefore, such an algorithm might easily outperform any of the algorithms as used in this thesis.

As noted by for example Kollat and Reed (2006), MOEAs seem to perform better on some binary or integer problems when implemented as a real problem. This might be because more sophisticated operators (e.g. simulated binary crossover (Deb and Agrawal, 1995)) can be used when dealing with real decision variables. It might be worth considering formulating the problem differently to make use of these findings.

Genetic algorithms form only a subclass of the metaheuristics. For other metaheuristics, such as ant colony optimization and particle swarm optimization, multi-objective variants have also been developed that could be applied to this problem.

In this thesis, only generational MOEAs were considered because of computational reasons. If the simulation model to be optimized can be run in parallel, steady-state algorithms could be interesting alternatives. Options include MOEA/D (Zhang and Li, 2007), ε-MOEA (Reed et al., 2003), PAES (Knowles and Corne, 1999) and SMS-EMOA (Beume et al., 2007).

Based on the performance on the test case, a population size was chosen for each algorithm. This choice was often a decision on the trade-off between a fast start and a high performance in the long run. However, best of both worlds could possibly be achieved by dynamically adjusting the population size during a run. For example, Reed et al. (2003) proposed the ε-NSGA-II: an adapted version of NSGA-II with a dynamic population size.

References

Alonso, J. J., LeGresley, P., and Pereyra, V. (2009). Aircraft design optimization. *Mathematics and Computers in Simulation*, 79(6):1948–1958.

Beume, N., Naujoks, B., and Emmerich, M. (2007). SMS-EMOA: Multiobjective selection based on dominated hypervolume. *European Journal of Operational Research*, 181(3):1653–1669.

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., and Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 68(6):394–424.

Buskermolen, M., Gini, A., Naber, S. K., Toes-Zoutendijk, E., de Koning, H. J., and Lansdorp-Vogelaar, I. (2018). Modeling in colorectal cancer screening: assessing external and predictive validity of MISCAN-Colon microsimulation model using NORCCAP trial results. *Medical Decision Making*, 38(8):917–929.

Corley, D. A., Jensen, C. D., Marks, A. R., Zhao, W. K., de Boer, J., Levin, T. R., Doubeni, C., Fireman, B. H., and Quesenberry, C. P. (2013). Variation of adenoma prevalence by age, sex, race, and colon location in a large population: implications for screening and quality programs. *Clinical Gastroenterology and Hepatology*, 11(2):172–180.

Corne, D. W., Jerram, N. R., Knowles, J. D., and Oates, M. J. (2001). PESA-II: Region-based selection in evolutionary multiobjective optimization. In *Proceedings of the 3rd Annual Conference on Genetic and Evolutionary Computation*, pages 283–290, San Francisco, California. Morgan Kaufmann Publishers Inc.

Corne, D. W., Knowles, J. D., and Oates, M. J. (2000). The Pareto envelope-based selection algorithm for multiobjective optimization. In *International conference on Parallel Problem Solving from Nature*, pages 839–848, Berlin, Heidelberg. Springer.

Deb, K. (2014). Multi-objective optimization. In Burke, E. K. and Kendall, G., editors, *Search Methodologies*, chapter 15, pages 403–449. Springer.

Deb, K. and Agrawal, R. B. (1995). Simulated binary crossover for continuous search space. *Complex Systems*, 9(2):115–148.

Deb, K., Pratap, A., Agarwal, S., and Meyarivan, T. (2002). A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation*, 6(2):182–197.

Derrac, J., García, S., Molina, D., and Herrera, F. (2011). A practical tutorial on the use of nonparametric statistical tests as a methodology for comparing evolutionary and swarm intelligence algorithms. *Swarm and Evolutionary Computation*, 1(1):3–18.

Ehrgott, M., Güler, Ç., Hamacher, H. W., and Shao, L. (2010). Mathematical optimization in intensity modulated radiation therapy. *Annals of Operations Research*, 175(1):309–365.

Erenay, F. S., Alagoz, O., and Said, A. (2014). Optimizing colonoscopy screening for colorectal cancer prevention and surveillance. *Manufacturing & Service Operations Management*, 16(3):381–400.

Goede, S. L., van Roon, A. H., Reijerink, J. C., van Vuuren, A. J., Lansdorp-Vogelaar, I., Habbema, J. D. F., Kuipers, E. J., van Leerdam, M. E., and van Ballegooijen, M. (2013). Cost-effectiveness of one versus two sample faecal immunochemical testing for colorectal cancer screening. *Gut*, 62(5):727–734.

Gustafsson, L. and Adami, H.-O. (1992). Optimization of cervical cancer screening. *Cancer Causes & Control*, 3(2):125–136.

Habbema, J., Van Oortmarssen, G., Lubbe, J. T. N., and Van der Maas, P. (1985). The MISCAN simulation program for the evaluation of screening for disease. *Computer Methods and Programs in Biomedicine*, 20(1):79–93.

Knowles, J. and Corne, D. (1999). The pareto archived evolution strategy: A new baseline algorithm for pareto multiobjective optimisation. In *Proceedings of the 1999 Congress on Evolutionary Computation*, pages 98–105, Washington, DC. IEEE.

Knudsen, A. B., Hur, C., Gazelle, G. S., Schrag, D., McFarland, E. G., and Kuntz, K. M. (2012). Rescreening of persons with a negative colonoscopy result: results from a microsimulation model. *Annals of Internal Medicine*, 157(9):611–620.

Knudsen, A. B., Zauber, A. G., Rutter, C. M., Naber, S. K., Doria-Rose, V. P., Pabiniak, C., Johanson, C., Fischer, S. E., Lansdorp-Vogelaar, I., and Kuntz, K. M. (2016). Estimation of benefits, burden, and harms of colorectal cancer screening strategies: modeling study for the US Preventive Services Task Force. *Jama*, 315(23):2595–2609.

Koffijberg, H., Coupe, V., Ijzerman, M. J., Degeling, K., and Greuter, M. J. (2017). From evaluation to optimization: using a meta-model to maximize the benefits of colorectal screening accounting for capacity constraints. *Value in Health*, 20(9):A757.

Kollat, J. B. and Reed, P. M. (2005). The value of online adaptive search: a performance comparison of NSGAII, ε-NSGAII and εMOEA. In *International Conference on Evolutionary Multi-Criterion Optimization*, pages 386–398, Berlin, Heidelberg. Springer.

Kollat, J. B. and Reed, P. M. (2006). Comparing state-of-the-art evolutionary multi-objective algorithms for long-term groundwater monitoring design. *Advances in Water Resources*, 29(6):792–807.

Konak, A., Coit, D. W., and Smith, A. E. (2006). Multi-objective optimization using genetic algorithms: A tutorial. *Reliability Engineering & System Safety*, 91(9):992–1007.

Lane, J. M., Chow, E., Young, G. P., Good, N., Smith, A., Bull, J., Sandford, J., Morcom, J., Bampton, P. A., and Cole, S. R. (2010). Interval fecal immunochemical testing in a colonoscopic surveillance program speeds detection of colorectal neoplasia. *Gastroenterology*, 139(6):1918–1926.

Laumanns, M., Thiele, L., Deb, K., and Zitzler, E. (2002). Combining convergence and diversity in evolutionary multiobjective optimization. *Evolutionary Computation*, 10(3):263–282.

Loeve, F., Boer, R., van Oortmarssen, G. J., van Ballegooijen, M., and Habbema, J. D. F. (1999). The MISCAN-COLON simulation model for the evaluation of colorectal cancer screening. *Computers and Biomedical Research*, 32(1):13–33.

McLay, L. A., Foufoulides, C., and Merrick, J. R. (2010). Using simulation-optimization to construct screening strategies for cervical cancer. *Health Care Management Science*, 13(4):294–318.

Morson, B. (1974). The polyp-cancer sequence in the large bowel. *Proceedings of the Royal Society of Medicine*, 67(6 Pt 1):451–457.

Nicolaou, C. A. and Brown, N. (2013). Multi-objective optimization methods in drug design. *Drug Discovery Today: Technologies*, 10(3):e427–e435.

Okasha, N. M. and Frangopol, D. M. (2009). Lifetime-oriented multi-objective optimization of structural maintenance considering system reliability, redundancy and life-cycle cost using GA. *Structural Safety*, 31(6):460–474.

Ozik, J., Collier, N. T., Wozniak, J. M., and Spagnuolo, C. (2016). From desktop to large-scale model exploration with Swift/T. In *2016 Winter Simulation Conference (WSC)*, pages 206–220, Arlington, Virginia. IEEE.

O'Mahony, J. F., Naber, S. K., Normand, C., Sharp, L., O'Leary, J. J., and de Kok, I. M. (2015). Beware of kinked frontiers: a systematic review of the choice of comparator strategies in cost-effectiveness analyses of human papillomavirus testing in cervical screening. *Value in Health*, 18(8):1138–1151.

Reed, P., Minsker, B. S., and Goldberg, D. E. (2003). Simplifying multiobjective optimization: An automated design methodology for the nondominated sorted genetic algorithm-II. *Water Resources Research*, 39(7):1196.

Riquelme, N., Von Lücken, C., and Baran, B. (2015). Performance metrics in multi-objective optimization. In *2015 Latin American Computing Conference (CLEI)*, pages 1–11. IEEE.

Rutter, C. M., Knudsen, A. B., Marsh, T. L., Doria-Rose, V. P., Johnson, E., Pabiniak, C., Kuntz, K. M., Van Ballegooijen, M., Zauber, A. G., and Lansdorp-Vogelaar, I. (2016). Validation of models used to inform colorectal cancer screening guidelines: accuracy and implications. *Medical Decision Making*, 36(5):604–614.

Sanders, G. D., Neumann, P. J., Basu, A., Brock, D. W., Feeny, D., Krahn, M., Kuntz, K. M., Meltzer, D. O., Owens, D. K., Prosser, L. A., et al. (2016). Recommendations for conduct, methodological practices, and reporting of cost-effectiveness analyses: second panel on cost-effectiveness in health and medicine. *Jama*, 316(10):1093–1103.

Siegel, R. L., Miller, K. D., Fedewa, S. A., Ahnen, D. J., Meester, R. G., Barzi, A., and Jemal, A. (2017). Colorectal cancer statistics, 2017. *CA: A Cancer Journal for Clinicians*, 67(3):177–193.

Sierra, M. R. and Coello, C. A. C. (2005). Improving PSO-based multi-objective optimization using crowding, mutation and ε-dominance. In *International Conference on Evolutionary Multi-Criterion Optimization*, pages 505–519, Berlin, Heidelberg. Springer.

Snover, D. C. (2011). Update on the serrated pathway to colorectal carcinoma. *Human Pathology*, 42(1):1–10.

Srinivas, N. and Deb, K. (1994). Muiltiobjective optimization using nondominated sorting in genetic algorithms. *Evolutionary Computation*, 2(3):221–248.

Subramanian, S., Klosterman, M., Amonkar, M. M., and Hunt, T. L. (2004). Adherence with colorectal cancer screening guidelines: a review. *Preventive Medicine*, 38(5):536–550.

Underwood, D. J., Zhang, J., Denton, B. T., Shah, N. D., and Inman, B. A. (2012). Simulation optimization of PSA-threshold based prostate cancer screening policies. *Health Care Management Science*, 15(4):293–309.

van den Akker-van Marle, M. E., van Ballegooijen, M., van Oortmarssen, G. J., Boer, R., and Habbema, J. D. F. (2002). Cost-effectiveness of cervical cancer screening: comparison of screening policies. *Journal of the National Cancer Institute*, 94(3):193–204.

Vijan, S., Inadomi, J., Hayward, R., Hofer, T., and Fendrick, A. (2004). Projections of demand and capacity for colonoscopy related to increasing rates of colorectal cancer screening in the United States. *Alimentary Pharmacology & Therapeutics*, 20(5):507–515.

Vogelstein, B., Fearon, E. R., Hamilton, S. R., Kern, S. E., Preisinger, A. C., Leppert, M., Smits, A. M., and Bos, J. L. (1988). Genetic alterations during colorectal-tumor development. *New England Journal of Medicine*, 319(9):525–532.

Wilschut, J. A., Hol, L., Dekker, E., Jansen, J. B., Van Leerdam, M. E., Lansdorp-Vogelaar, I., Kuipers, E. J., Habbema, J. D. F., and Van Ballegooijen, M. (2011). Cost-effectiveness analysis of a quantitative immunochemical test for colorectal cancer screening. *Gastroenterology*, 141(5):1648–1655.

Zauber, A. G., Knudsen, A. B., Rutter, C. M., Lansdorp-Vogelaar, I., Savarino, J. E., van Ballegooijen, M., and Kuntz, K. M. (2009). Cost-effectiveness of CT colonography to screen for colorectal cancer. Technical report, Agency for Healthcare Research and Quality (US), Rockville (MD).

Zavoral, M., Suchanek, S., Majek, O., Fric, P., Minarikova, P., Minarik, M., Seifert, B., and Dusek, L. (2014). Colorectal cancer screening: 20 years of development and recent progress. *World Journal of Gastroenterology*, 20(14):3825.

Zhang, Q. and Li, H. (2007). MOEA/D: A multiobjective evolutionary algorithm based on decomposition. *IEEE Transactions on Evolutionary Computation*, 11(6):712–731.

Zitzler, E. (1999). *Evolutionary Algorithms for Multiobjective Optimization: Methods and Applications*. PhD thesis, Swiss Federal Institute of Technology.

Zitzler, E. and Künzli, S. (2004). Indicator-based selection in multiobjective search. In *International Conference on Parallel Problem Solving from Nature*, pages 832–842, Berlin, Heidelberg. Springer.

Zitzler, E., Laumanns, M., and Thiele, L. (2001). SPEA2: Improving the Strength Pareto Evolutionary Algorithm. *TIK-Report*, 103.

Zitzler, E. and Thiele, L. (1998). Multiobjective optimization using evolutionary algorithms—a comparative case study. In *International Conference on Parallel Problem Solving from Nature*, pages 292–301, Berlin, Heidelberg. Springer.

Zitzler, E. and Thiele, L. (1999). Multiobjective evolutionary algorithms: a comparative case study and the strength Pareto approach. *IEEE Transactions on Evolutionary Computation*, 3(4):257–271.

Zitzler, E., Thiele, L., Laumanns, M., Fonseca, C. M., and Da Fonseca, V. G. (2003). Performance assessment of multiobjective optimizers: An analysis and review. *IEEE Transactions on Evolutionary Computation*, 7(2):117–132.

Appendix A Function evaluations

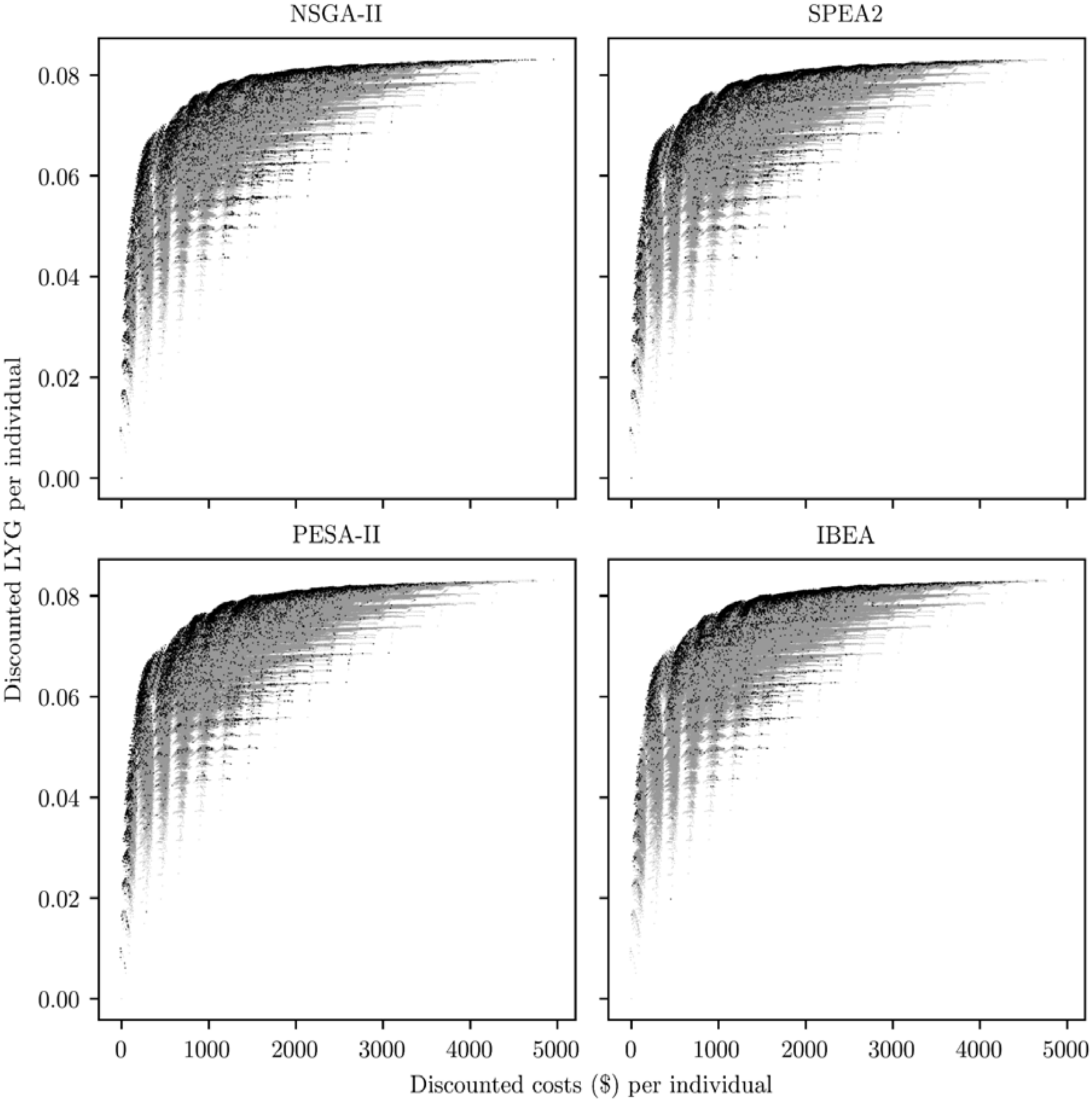


Figure A.1: Evaluated strategies by each algorithm on the test case in a single run with a limit of 10,000 function evaluations, represented in black. The gray dots represent all possible strategies in the test case.

Appendix B Statistical analysis algorithm comparison

Table B.1: Shapiro-Wilk test for normality results on the unary performance measures for each algorithm. Based on 100 runs (with 10,000 function evaluations) per algorithm on the test case.

Measure	Algorithm	Test statistic	p-value
I_ε	NSGA-II	0.9863	0.3908
	SPEA2	0.9748	0.0516
	PESA-II	0.9794	0.1201
	IBEA	0.9859	0.3664
I_{IGD}	NSGA-II	0.8372	< 0.0001
	SPEA2	0.9278	< 0.0001
	PESA-II	0.8989	< 0.0001
	IBEA	0.8416	< 0.0001
I_{HV}	NSGA-II	0.8123	< 0.0001
	SPEA2	0.8622	< 0.0001
	PESA-II	0.8854	< 0.0001
	IBEA	0.8897	< 0.0001

Table B.2: Kruskal-Wallis test results for the unary measures. Based on 100 runs (with 10,000 function evaluations) for each of the four algorithms on the test case.

Measure	Test statistic	p-value
I_ε	237.9162	< 0.0001
I_{IGD}	180.9480	< 0.0001
I_{HV}	228.2555	< 0.0001

Table B.3: Shapiro-Wilk test for normality results on the difference in binary performance measures for each combination of two algorithms. Based on 100 runs (with 10,000 function evaluations) per algorithm on the test case.

		I_C		I_{HV2}	
Algorithm	Compared to	Test statistic	p-value	Test statistic	p-value
NSGA-II	SPEA2	0.9820	0.1908	0.9773	0.0810
	PESA-II	0.9857	0.3537	0.9071	< 0.0001
	IBEA	0.9880	0.5096	0.9642	0.0081
SPEA2	NSGA-II	0.9820	0.1908	0.9773	0.0810
	PESA-II	0.9891	0.5961	0.9409	0.0002
	IBEA	0.9873	0.4556	0.9286	< 0.0001
PESA-II	NSGA-II	0.9857	0.3536	0.9071	< 0.0001
	SPEA2	0.9891	0.5960	0.9409	0.0002
	IBEA	0.9927	0.8716	0.9851	0.3227
IBEA	NSGA-II	0.9880	0.5096	0.9642	0.0081
	SPEA2	0.9873	0.4556	0.9286	< 0.0001
	PESA-II	0.9927	0.8716	0.9851	0.3227

Table B.4: Pairwise comparison tests of the four algorithms on each of the five performance measures, based on 100 runs (with 10,000 function evaluations) per algorithm on the test case. Conover's test with Hölm adjusted p-values was used for the unary measures (I_ε, I_{IGD} and I_{HV}^P). The paired two sample t-test was used for I_C and the Wilcoxon signed-rank test was used for I_{HV2}. Using a significance level of 0.05, statistically significant better and worse performance is indicated by ▲ and ▽, respectively. A dash (–) indicates no significant difference was observed. The p-values are also given.

Algorithm	Compared to	I_ε		I_{IGD}		I_{HV}		I_C		I_{HV2}	
NSGA-II	SPEA2	▲	< 0.0001	▲	0.0011	▲	0.0372	▲	0.0001	–	0.4852
	PESA-II	▲	< 0.0001	▲	< 0.0001	▲	< 0.0001	▲	< 0.0001	▲	< 0.0001
	IBEA	▲	< 0.0001	▲	< 0.0001	▲	< 0.0001	▲	0.0005	▲	< 0.0001
SPEA2	NSGA-II	▽	< 0.0001	▽	0.0011	▽	0.0372	▽	0.0001	–	0.4852
	PESA-II	▲	< 0.0001	▲	< 0.0001	▲	< 0.0001	▲	< 0.0001	▲	< 0.0001
	IBEA	–	0.2070	▲	< 0.0001	▲	< 0.0001	–	0.6851	▲	< 0.0001
PESA-II	NSGA-II	▽	< 0.0001	▽	< 0.0001	▽	< 0.0001	▽	< 0.0001	▽	< 0.0001
	SPEA2	▽	< 0.0001	▽	< 0.0001	▽	< 0.0001	▽	< 0.0001	▽	< 0.0001
	IBEA	▽	< 0.0001	▽	< 0.0001	▽	< 0.0001	▽	< 0.0001	▽	< 0.0001
IBEA	NSGA-II	▽	< 0.0001	▽	< 0.0001	▽	< 0.0001	▽	0.0005	▽	< 0.0001
	SPEA2	–	0.2070	▽	< 0.0001	▽	< 0.0001	–	0.6851	▽	< 0.0001
	PESA-II	▲	< 0.0001	▲	< 0.0001	▲	< 0.0001	▲	< 0.0001	▲	< 0.0001

Appendix C Cost-effective complex strategies

Table C.1: Cost-effective complex strategies and their costs, LYG and ICER.

ICER	Costs	LYG	Interventions (age(s):test)
$-\infty$	-32.24	0.033,707	61-64:FIT10, 66:FIT10
718.84	-30.09	0.036,696	61-64:FIT10, 66-67:FIT10
889.32	-23.43	0.044,183	57-58:FIT10, 61:FIT10, 63-66:FIT10
2,427.27	-13.62	0.048,224	56:FIT10, 58:FIT10, 60-61:FIT10, 63-64:FIT10, 66-67:FIT10
4,142.97	-4.61	0.050,399	56:FIT10, 58:FIT10, 60-62:FIT10, 64:FIT10, 66-68:FIT10
4,417.15	3.49	0.052,233	56:FIT10, 58:FIT10, 60-62:FIT10, 64:FIT10, 66-69:FIT10
4,885.90	14.79	0.054,547	54:FIT10, 56:FIT10, 58:FIT10, 60:FIT10, 62-63:FIT10, 65-68:FIT10
5,342.32	23.91	0.056,254	54:FIT10, 56:FIT10, 58:FIT10, 60-61:FIT10, 63-64:FIT10, 66-69:FIT10
6,864.50	34.02	0.057,726	54:FIT10, 56:FIT10, 58:FIT10, 60-61:FIT10, 63-64:FIT10, 66-70:FIT10
7,316.68	36.05	0.058,003	54:FIT10, 56:FIT10, 58:FIT10, 60-61:FIT10, 63-64:FIT10, 66-68:FIT10, 70-71:FIT10
8,585.79	63.09	0.061,153	52:FIT10, 55-56:FIT10, 58:FIT10, 60:FIT10, 62-64:FIT10, 66-70:FIT10
8,719.97	76.33	0.062,671	52:FIT10, 54:FIT10, 56:FIT10, 58:FIT10, 60:FIT10, 62-64:FIT10, 66-71:FIT10
13,015.87	92.17	0.063,888	52:FIT10, 54:FIT10, 56:FIT10, 58-60:FIT10, 62-64:FIT10, 66-71:FIT10
13,030.10	130.12	0.066,801	50:FIT10, 53:FIT10, 55-56:FIT10, 58:FIT10, 60-61:FIT10, 63-64:FIT10, 66-72:FIT10
13,621.09	133.10	0.067,020	50:FIT10, 53:FIT10, 55-56:FIT10, 58:FIT10, 60-61:FIT10, 63-67:FIT10, 69-71:FIT10, 73:FIT10
14,871.78	144.39	0.067,778	50:FIT10, 53:FIT10, 55-56:FIT10, 58:FIT10, 60-62:FIT10, 64-68:FIT10, 70-73:FIT10
16,351.83	170.37	0.069,367	50:FIT10, 52:FIT10, 54:FIT10, 56-57:FIT10, 59-60:FIT10, 62-64:FIT10, 66-70:FIT10, 72-74:FIT10
16,862.64	192.15	0.070,659	47:FIT40, 50:FIT10, 53:FIT10, 55-56:FIT10, 58:FIT10, 60-61:FIT10, 63-64:FIT10, 66-74:FIT10
18,162.84	225.57	0.072,499	45:FIT40, 50:FIT10, 52:FIT10, 54:FIT10, 56-57:FIT10, 59-60:FIT10, 62-70:FIT10, 72-74:FIT10
19,738.36	239.03	0.073,181	45:FIT40, 50:FIT10, 52:FIT10, 54:FIT10, 56-57:FIT10, 59-60:FIT10, 62-70:FIT10, 72-74:FIT10, 76:FIT10
20,977.95	265.23	0.074,430	45:FIT40, 48:FIT10, 51:FIT10, 53-54:FIT10, 56-57:FIT10, 59-60:FIT10, 62-64:FIT10, 66-72:FIT10, 74-75:FIT10
24,995.36	302.03	0.075,902	45:FIT40, 48:FIT10, 50:FIT10, 53-54:FIT10, 56:FIT10, 58-73:FIT10, 76:FIT10
25,878.97	321.01	0.076,635	45:FIT40, 48:FIT10, 50:FIT10, 52:FIT10, 54:FIT10, 56:FIT10, 58-73:FIT10, 75:FIT10, 78:FIT10
30,887.45	333.41	0.077,037	45:FIT40, 48:FIT10, 50:FIT10, 52:FIT10, 54:FIT10, 56-57:FIT10, 59-74:FIT10, 77:FIT20, 79:FIT10
35,392.71	388.22	0.078,586	45:FIT40, 48:FIT10, 50:FIT10, 52-53:FIT10, 55-76:FIT10
36,541.25	465.22	0.080,693	40:FIT40, 45:FIT40, 46:FIT10, 50-51:FIT10, 53-56:FIT10, 58-71:FIT10, 73-74:FIT10, 76-77:FIT10, 79:FIT10
39,544.10	495.34	0.081,454	40:FIT40, 45-46:FIT10, 50-51:FIT10, 53:FIT10, 55-73:FIT10, 75-77:FIT10, 79:FIT10
46,175.23	549.61	0.082,630	40:FIT40, 45:FIT40, 46:FIT10, 49-51:FIT10, 53-56:FIT10, 58-72:FIT10, 74-80:FIT10
47,076.06	579.67	0.083,268	40:FIT40, 45-46:FIT10, 49-51:FIT10, 53-56:FIT10, 58-76:FIT10, 78-79:FIT10, 81:FIT10
49,057.27	701.09	0.085,743	40:FIT40, 42:FIT20, 45-46:FIT10, 48:FIT10, 50-52:FIT10, 53:FIT40, 54-74:FIT10, 76-79:FIT10, 81:FIT10
54,577.27	717.14	0.086,038	40:FIT40, 42:FIT20, 45-46:FIT10, 48:FIT10, 50-76:FIT10, 78-79:FIT10, 81:FIT10
62,221.22	769.59	0.086,880	40:FIT40, 42:FIT20, 43:FIT40, 45-46:FIT10, 48:FIT10, 50-73:FIT10, 75-78:FIT10, 80-82:FIT10
67,824.24	780.05	0.087,035	40:FIT40, 42:FIT20, 43:FIT40, 45-46:FIT10, 48:FIT10, 50-76:FIT10, 78-82:FIT10
73,735.78	799.76	0.087,302	40:FIT40, 42-43:FIT20, 45-46:FIT10, 48:FIT10, 50-82:FIT10
100,556.15	903.02	0.088,329	41:FIT10, 42:FIT20, 43:FIT10, 45-46:FIT10, 48-77:FIT10, 79-82:FIT10
135,018.77	904.58	0.088,340	41:FIT10, 42:FIT20, 43:FIT10, 45-46:FIT10, 48-78:FIT10, 80-81:FIT10, 83:FIT10
135,861.46	967.85	0.088,806	41:FIT10, 42:FIT20, 43:FIT10, 45:FIT40, 46-82:FIT10, 84-85:FIT10
139,565.30	1,403.08	0.091,925	40-41:FIT20, 42-43:FIT10, 44:FIT40, 45-59:FIT10, 60:COL, 64-68:FIT10, 69:FIT20, 70-82:FIT10, 84:FIT10
168,136.62	1,523.20	0.092,639	40-59:FIT10, 60:COL, 62-63:FIT10, 65-83:FIT10, 85:FIT10

Continued on next page

Continued from previous page

ICER	Costs	LYG	Interventions (age(s):test)
271,649.63	1,552.58	0.092,747	40-59:FIT10, 60:COL, 62-68:FIT10, 69:FIT20, 70-83:FIT10, 85:FIT10, 87:FIT10, 89:FIT20
327,587.64	1,911.92	0.093,844	40-53:FIT10, 54:COL, 56-66:FIT10, 67:COL, 69-70:FIT10, 72-74:FIT10, 76-80:FIT10, 82-85:FIT10, 87:FIT10
351,420.47	1,954.36	0.093,965	40-53:FIT10, 54:COL, 56-66:FIT10, 67:COL, 68:FIT10, 69:FIT20, 70-83:FIT10, 85:FIT10, 87:FIT10
360,750.93	1,998.84	0.094,088	40-53:FIT10, 54:COL, 55-64:FIT10, 65:COL, 67-68:FIT10, 70-83:FIT10, 85:FIT10, 87:FIT40
534,709.51	2,318.84	0.094,687	40-53:FIT10, 54:COL, 56-59:FIT10, 60:COL, 62-66:FIT10, 67:COL, 68:FIT10, 69:FIT20, 70-83:FIT10, 85:FIT10, 87:FIT10
611,795.49	2,358.10	0.094,751	40-53:FIT10, 54:COL, 56-59:FIT10, 60:COL, 61-66:FIT10, 67:COL, 68:FIT10, 69:FIT20, 70-82:FIT10, 83:FIT20, 84-85:FIT10, 86:FIT40, 87:FIT10
707,200.59	3,046.48	0.095,724	40-48:FIT10, 49:COL, 51-53:FIT10, 54:COL, 56-59:FIT10, 60:COL, 61-66:FIT10, 67:COL, 68-85:FIT10, 87:FIT10
760,646.58	3,613.26	0.096,469	40:COL, 42:FIT20, 43-53:FIT10, 54:COL, 56-59:FIT10, 60:COL, 61-66:FIT10, 67:COL, 68-71:FIT10, 72:COL, 73-74:FIT10, 76-80:FIT10, 81:FIT40, 82-85:FIT10, 87:FIT10
775,003.74	4,425.94	0.097,518	40:COL, 42:FIT20, 43-47:FIT10, 48:COL, 49:FIT20, 50-53:FIT10, 54:COL, 55-59:FIT10, 60:COL, 61-64:FIT10, 65:COL, 67-71:FIT10, 72:COL, 73-82:FIT10, 84-85:FIT10, 87:FIT10, 88-89:FIT20
1,123,860.83	4,538.29	0.097,618	40:COL, 42-47:FIT10, 48:COL, 49-53:FIT10, 54:COL, 55-59:FIT10, 60:COL, 61-64:FIT10, 65:COL, 67-71:FIT10, 72:COL, 74-78:FIT10, 79:COL, 82:FIT10, 83:FIT20, 85:FIT10, 87:FIT10
1,834,866.04	4,841.79	0.097,783	40:COL, 42-47:FIT10, 48:COL, 49-53:FIT10, 54:COL, 55-59:FIT10, 60:COL, 61-63:FIT10, 64:COL, 66:FIT10, 67:COL, 69-71:FIT10, 72:COL, 73-78:FIT10, 79:COL, 80-82:FIT10, 83:FIT20, 84-85:FIT10, 86:FIT40, 87:FIT10
2,060,461.09	6,148.59	0.098,418	40:COL, 41-42:FIT10, 43:COL, 45-47:FIT10, 48:COL, 49-51:FIT10, 52:FS, 53:FIT10, 54:COL, 55-59:FIT10, 60:COL, 61-63:FIT10, 64:COL, 65-66:FIT10, 67:COL, 69-71:FIT10, 72:COL, 73-74:FIT10, 75:FIT20, 76-78:FIT10, 79:COL, 81-83:FIT10, 84:FIT40, 85:FIT10, 87:FIT10, 89:FIT20
2,654,273.41	6,682.96	0.098,619	40:COL, 41-42:FIT10, 43:COL, 45-47:FIT10, 48:COL, 49-51:FIT10, 52:FS, 53:FIT10, 54:COL, 55-57:FIT10, 58:COL, 60:COL, 61-63:FIT10, 64:COL, 65-66:FIT10, 67:COL, 69:FIT10, 70:FS+FIT40, 71:FIT10, 72:COL, 74-78:FIT10, 79:COL, 80-82:FIT10, 83:FIT20, 84-85:FIT10, 86:FIT40, 87:FIT10, 89:FIT20
2,737,698.23	6,803.18	0.098,663	40:COL, 41-42:FIT10, 43:COL, 45-47:FIT10, 48:COL, 49-51:FIT10, 52:FS, 53:FIT10, 54:COL, 55-57:FIT10, 58:COL, 60:COL, 61-63:FIT10, 64:COL, 65-66:FIT10, 67:COL, 68-69:FIT10, 70:COL, 71:FIT10, 72:COL, 73-78:FIT10, 79:COL, 81-85:FIT10, 87:FIT10, 88-89:FIT20
2,769,494.58	6,852.20	0.098,680	40:COL, 41-42:FIT10, 43:COL, 45-47:FIT10, 48:COL, 49-51:FIT10, 52:FS, 53:FIT10, 54:COL, 55-57:FIT10, 58:COL, 60:COL, 61-63:FIT10, 64:COL, 65-66:FIT10, 67:COL, 68-69:FIT10, 70:COL, 72:COL, 74-78:FIT10, 79:COL, 80-81:FIT40, 82:COL, 84-85:FIT10, 87:FIT10, 89:FIT20
3,685,015.36	7,544.87	0.098,868	40:COL, 41-42:FIT10, 43:COL, 45-47:FIT10, 48:COL, 50:COL, 51:FIT10, 52:FS+FIT10, 53:FIT10, 54:COL, 55-57:FIT10, 58:COL, 60:COL, 61-63:FIT10, 64:COL, 65-66:FIT10, 67:COL, 68-69:FIT10, 70:COL, 71:FIT10, 72:COL, 73-76:FIT10, 77:COL, 79-80:FIT10, 81:COL, 83:FIT10, 85:FIT10, 87:FIT10, 88-89:FIT20
3,783,715.96	8,144.19	0.099,027	40:COL, 41-42:FIT10, 43:COL, 45-47:FIT10, 48:COL, 50:COL, 51:FIT10, 52:FS+FIT10, 53:FIT10, 54-55:COL, 56-57:FIT10, 58:COL, 59:FIT10, 60:COL, 61-63:FIT10, 64:COL, 65-66:FIT10, 67:COL, 68-69:FIT10, 70:COL, 71:FIT10, 72:COL, 74-76:FIT10, 77:COL, 79:FS+FIT20, 80:FIT10, 81:COL, 83:FIT10, 85:FIT10, 87:FIT10, 88-89:FIT20
5,886,401.85	9,718.75	0.099,294	40:COL, 41-42:FIT10, 43:COL, 45-47:FIT10, 48-50:COL, 51:FIT10, 52:FS+FIT20, 53:FIT10, 54-55:COL, 56-57:FIT10, 58:COL, 59:FIT10, 60:COL, 61-62:FIT10, 63:FS+FIT-DNA, 64:COL, 65-66:FIT10, 67-68:COL, 69:FIT10, 70:COL, 71:FIT10, 72:COL, 73:FIT40, 74:FIT10, 75:FS+FIT10, 76:FIT10, 77:COL, 78:FIT10, 79:COL, 80:FIT10, 81:COL, 82:FIT10, 84:FIT10, 87:COL, 89:FIT20

Continued on next page

Continued from previous page

ICER	Costs	LYG	Interventions (age(s):test)
19,741,415.82	9,780.77	0.099,297	40:COL, 41-42:FIT10, 43:COL, 45-47:FIT10, 48-50:COL, 51:FIT10, 52:FS+FIT20, 53:FIT10, 54-55:COL, 56-57:FIT10, 58:COL, 59:FIT10, 60:COL, 61-62:FIT10, 63:FS+FIT-DNA, 64:COL, 65-66:FIT10, 67-68:COL, 69:FIT10, 70:COL, 71:FIT10, 72:COL, 73-74:FIT10, 75:FS+FIT10, 76:FIT10, 77:COL, 79:COL, 80:FIT10, 81:COL, 82:FIT10, 83:FS+FIT10, 84-85:FIT10, 87:COL, 89:FIT20