J. Benefit Cost Anal. 2017; 8(1):1–23
doi:10.1017/bca.2017.1
© Society for Benefit-Cost Analysis, 2017

Nicolai Kristensen*, Christophe Kolodziejczyk, Iben Bolvig and Kurt Houlberg

The Efficiency Potential in Social Programs: Accounting for Heterogeneity in Benefit-Cost Analysis[1]

Abstract: Traditionally, benefit-cost analyses focus on average benefits and average costs. However, heterogeneous treatment effects and/or costs are most often present, which means that there is an efficiency potential hidden in the implementation of public programs, if policies can be targeted at those who, net of costs, benefit the most. We introduce *efficiency potential* defined as the ratio between the net benefit achieved under perfect selection of the individuals with significant positive net benefit of program participation, and the actually realized benefits net of costs. Using data from a randomized control-trial experiment of a Danish return-to-work program combined with rich administrative records and survey data, we find that there is indeed a potential for increased efficiency. Results from the treatment literature indicate that, generally, it may be difficult to harvest the full potential. Our application corroborates this finding.

Keywords: benefit-cost analysis; heterogeneous treatment effects; Monte Carlo simulation; RCT; return-to-work coordination.

JEL classifications: I180; J48; H7.

1 We thank the editor and two anonymous referees for very helpful suggestions. We also thank The National Research Centre for the Working Environment (NFA) for access to data and Birgit Aust and Lars Skipper for helpful comments.
***Corresponding author: Nicolai Kristensen,** KORA, Danish Institute for Local and Regional Government Research, Denmark, Phone: +45 2511 7372, e-mail: nikr@kora.dk. Address: Koebmagergade 22, 1150 Copenhagen K, Denmark
and
Institute of the Study of Labor (IZA), Germany
Christophe Kolodziejczyk: KORA, Danish Institute for Local and Regional Government Research, Denmark
Iben Bolvig: KORA, Danish Institute for Local and Regional Government Research, Denmark
Kurt Houlberg: KORA, Danish Institute for Local and Regional Government Research, Denmark

https://doi.org/10.1017/bca.2017.1 Published online by Cambridge University Press

1 Introduction

A key purpose of Benefit-Cost Analysis (BCA) is to inform policy-makers about the value of publicly financed programs. Traditionally, benefits are based on parameter estimates of the average effect of a given program, which are then compared to the associated average costs. However, there will often be substantial heterogeneity in the effects and/or costs across individuals, and BCA based on the traditional approach may therefore hide valuable information about who benefits the most net of costs. Nevertheless, inclusion of heterogeneity in BCA is strikingly absent from both the research literature and consultants' applied work. The purpose of this paper is to analyze the difference between a traditional average BCA approach and an approach that takes heterogeneity into account.

Usually, the Kaldor–Hicks efficiency criterion is applied in BCA. According to this criterion, a public program can be defended if it, in total, has higher benefits than costs. A program can therefore have negative consequences for some individuals, as long as the overall benefits are higher than the overall costs.[2,3] We maintain the Kaldor–Hicks criterion, noting, though, that there is still a potential for increased efficiency. Making the most of program evaluations and social experiments involves taking into account heterogeneity in program impacts. This basic point was made in the seminal paper of Heckman, Smith and Clements (1997), who note that

> "*. . . many persons would judge programs to be successful if enough persons, or enough of the right kinds of persons, reaped benefits from them even if the average participant did not*.", Heckman et al. (1997, p. 488).

It is difficult not to concur with this point of view, and in a BCA setting it includes potential variations on the cost side. One important consequence of heterogeneous benefits and/or costs might be that public funding can potentially be targeted at those who gain the most net of costs. This will make society better off overall.

Taking account of heterogeneous effects in BCA is a rather basic idea. However, to our knowledge this has never been done. Why is it not standard? In recent years, there has been a surge in methodological developments and applied work related to identification and estimation of treatment effects, see e.g., Imbens and

2 As opposed to the Pareto criterion, which requires that no one will suffer from any change.

3 Harberger (1978) introduced weights into social benefit-cost analysis. Using individually different weights in BCA, it can be socially optimal to introduce a program if it benefits, say, "the poor," even though it comes at the expense of "the rich" (Harberger, 1978; Holtmann, 1991; Boardman, Greenberg, Vining & Weimer, 2014). In such a case, one would diverge from the Kaldor–Hicks criterion and in some sense also introduce heterogeneity. However, this type of heterogeneity is fundamentally different from the heterogeneity analyzed in this paper.

https://doi.org/10.1017/bca.2017.1 Published online by Cambridge University Press

Wooldridge (2009). This literature has focused mostly, but not entirely, on average treatment effects (ATEs).[4]

In this paper, we carry out a BCA of a Danish return-to-work (RTW) program, by which stakeholders within a municipality coordinate their efforts. To emphasize the potential gains of distributional effect analysis, we seek to understand who benefits the most from coordinated efforts and how. We study a large Danish municipality in which the RTW experiment was successfully implemented as a randomized controlled trial (RCT).[5,6]

The Danish National RTW program is based on the notion that long-term sickness absentees often face a series of challenges related to social, work and health status and, consequently, that employers, municipal case workers and health and rehabilitation professionals all play an important role in facilitating their RTW (Nielsen et al., 2013). Coordinated efforts among these stakeholders have often been suggested as an effective tool to reduce long-term sickness absence and have been found to generate small, but likely important, relative absolute benefits (Schandelmaier et al., 2012). The Danish National RTW program produced mixed results (Poulsen et al., 2012; Nielsen et al., 2013). For instance, Poulsen et al. (2012) found that the effects of the intervention in relation to duration of sickness absence differed markedly among the three municipalities examined. They argue that this indicates that contextual factors are of major importance for the success or failure of this complex intervention. Possibly, part of the reason for these mixed results can be related to mixed results within municipalities.

Understanding better how or why the program works is important, because RTW programs can be relatively costly. Hence, in order to be able to induce municipalities to make this additional effort and investment it is of great importance to ascertain whether or not the program is actually self-sustainable.

We find that benefits net of costs are positive, based on estimates of the average effect. However, quantile treatment-effect (QTE) analysis reveals that the program effect is heterogeneous. We define the "efficiency potential" (EP) as the increased benefits net of costs that would accrue to the municipality (or society), if it was possible to screen participants perfectly to those who benefit individually at least as

4 Another potential reason for inattention to heterogeneity in random controlled trials may be a concern about multiple comparisons and the consequent decrease in statistical power. Such an approach miss that the relevant emphasis for public policy purposes is net benefits, for which effect sizes and standard errors for use in Monte Carlos are needed.

5 Only 3 out of 21 participating municipalities introduced the RTW program as an RCT, and one of these had a problematic implementation process (Aust et al., 2015), while another could not credibly estimate the costs involved (Houlberg, Kolodziejczyk & Kristensen, 2012).

6 The purpose of this paper is to understand *within* municipality variation, and for this purpose, analysis of one municipality is of interest, although the results cannot be carried over and generalized to other municipalities.

https://doi.org/10.1017/bca.2017.1 Published online by Cambridge University Press

much as the costs involved. Such a screening must necessarily be based on observable characteristics. In our RTW example, the EP is 0.22, which underscores the potential and the scope for future BCAs. However, we also find it difficult to realize the potential in our empirical example. In spite of the data being very rich, none of the observable characteristics interacted with the treatment were found to show significance in reducing the duration of sickness absence. Bitler, Gelbach and Hoynes (2006, 2014) point out that this type of finding will often occur due to variation not only between but also within subgroups.

The remainder of the paper is organized as follows. In the next section, we describe the RCT experiment and the detailed administrative data records, while Section 3 outlines the methodological framework. In Section 4, we first list the results based on traditional average benefits and costs and subsequently look into heterogeneity and the EP. Finally, we discuss implications and conclude with suggestions for future research and analysis.

2 Data and institutional background

As mentioned, the empirical application builds on experimental data from a randomized control trial of an RTW program in a large Danish municipality combined with detailed administrative panel data and survey data.

2.1 The institutional set-up related to sickness absence

In Denmark, employed, self-employed and insured unemployed people, are eligible for sickness benefits if they are unable to work due to sickness. The employer is obligated to pay either wages or sickness benefits during the first 30 days of the sick-leave period, and the municipality pays for the rest of the sick-leave period up to a maximum of 52 weeks. In some cases, dispensation can be granted up to another 52 weeks. All sickness benefit recipients are classified into three different categories: (1) if the sick person is expected to be able to return to work within 3 months, (2) if long-term sickness absence is assessed to last longer than three months, but the person on sick leave is able to attend labor market programs, and (3) if sickness absence is assessed to last longer than three months and the person on sick leave is unable to attend labor market programs.[7] For all persons on sick leave,

7 As of January 2014, the maximum number of weeks was reduced to 26 weeks, and the definitions of the categories were changed.

https://doi.org/10.1017/bca.2017.1 Published online by Cambridge University Press

the municipality must perform a sickness benefit management interview within the first eight weeks after the first day of sick leave. After this, the requirement for the frequency of meetings depends on the category. For category 2 individuals, the municipal case manager must carry out a sickness benefit management interview every four weeks. The goal is to develop an RTW plan that promotes the sick person's return to work, which can include participation in labor market programs.

2.2 Administrative registry data

A wide range of longitudinal administrative registry data from Statistics Denmark were used in the analysis. These records, all at the individual level, allow us to follow each person in both the treatment and the control group over time, both before and after the start of the RTW program. Specifically, the data include monthly hours worked and monthly income or welfare transfer, labor market status and purchases of prescription drugs, with information on date of purchase, price and type of drug. Finally, data include information on primary health care, i.e., all contacts with general practitioners, special practitioners and psychologists covered by the public insurance.[8]

2.3 A first look at the data

The experiment covered 2,310 sick-listed persons, of which 887 were being treated within the coordinated RTW program.[9] Ideally, we should find no difference in the characteristics between the treatment group and controls, see Table 1.

Generally, we only find few observed differences between the treatment (intervention) and control groups. One difference is that the rate of women among sick-listed persons under the RTW program is significantly higher than in the control group of sick-listed persons receiving ordinary sickness benefit management.

In addition, the treatment and control groups also differ in that individuals with self-assessed psychiatric illnesses are overrepresented in the treatment group vis-à-vis the control group. Moreover, prior job status and age differ significantly. These differences are not troublesome on their own, because we are able to condition on observed covariates in the treatment-effect analysis. However, the fact that several variables are significantly different raises the concern that, possibly, the random

8 The registers include e-income, DREAM, The Danish National Prescription Database and The National Health Insurance Service Registry.
9 See Appendix A for details of the RCT program.

https://doi.org/10.1017/bca.2017.1 Published online by Cambridge University Press

Table 1 Two sample t-tests for differences in control versus treated, average values.

	Control	RTW	Difference	P-value
% out of sick leave	62.1	71.2	−9.1***	0.000
Age	42.6	41.8	0.8*	0.084
Age < 30	13.6	15.5	−1.9	0.208
30 ⩽ Age < 40	24.7	26.5	−1.8	0.333
40 ⩽ Age < 50	32.5	32.5	0.0	0.999
50 ⩽ Age	29.2	25.5	3.7*	0.054
Female	53.1	59.8	−6.7***	0.002
Compulsory education	35.6	34.7	0.9	0.656
Vocational education	39.8	39.1	0.7	0.729
Further education	24.6	26.2	−1.6	0.379
Further education short/medium	17.0	16.9	0.1	0.957
Further education long	7.6	9.3	−1.7	0.143
% Cause: musculoskeletal	36.2	33.6	2.6	0.204
% Cause: depression	23.6	21.1	2.5	0.166
% Cause: stress	12.1	12.6	−0.5	0.713
% Cause: psychiatric illness	5.0	8.1	−3.1***	0.002
% With previous sickness leave	49.2	49.2	−0.0	0.986
% With previous unemployment	55.4	56.9	−1.6	0.461
% With previous medical expenses	90.5	89.6	0.9	0.465
% Admitted to hospital or psych.	43.6	40.8	2.8	0.186
Mean medical expenses the previous year (DKK)	3250.9	2808.0	443.0	0.124
Unemployment level past 3 years	0.2	0.2	−0.0	0.319
No. weeks on sick leave 1 year before	5.2	5.1	0.1	0.805
No. weeks on sick leave 2 years before	6.6	7.0	−0.4	0.450
No. weeks on sick leave 3 years before	6.4	6.9	−0.5	0.399
Prior job status: fully unemployed	34.0	34.4	−0.3	0.876
Prior job status: self-sufficient	57.9	55.2	2.8	0.189
Prior job status: other income support	8.0	10.5	−2.5**	0.043
Prior job status: employed	36.8	36.0	0.8	0.708
Observations (2,310 in total)	887	1,423		

Note: *$p < 0.10$, **$p < 0.05$, ***$p < 0.01$.

assignment was flawed or there was sample selection after random assignment. Concerning the differences in age the differences found are merely marginal since only the oldest age category differs and the mean age does not differ. As for prior

https://doi.org/10.1017/bca.2017.1 Published online by Cambridge University Press

job status and psychiatric illness they both are highly gender correlated.[10] That is, these differences should not be seen as three independent factors being different, rather they are likely to be a consequence of the share of women being higher in the treatment group vis-à-vis the control group. However, we have no explanation for this gender difference. Both Aust et al. (2015) and Poulsen et al. (2012) have documented the design and implementation of the RCT study and they conclude that this municipality has performed very well. Poulsen et al. (2012) are also puzzled by the gender difference, but conclude that it must be a simple coincidence. We balance the two groups by including covariates in the treatment-effect analysis.

3 Methodological approach

3.1 Average and quantile treatment effects

The methodological approach was chosen in accordance with the available experimental data described above. The availability of experimental data makes it possible to obtain the ATE without the standard type of selection problem that often hampers treatment analyses. The RCT lends itself to a simple nonparametric analysis of the treatment effect. Using common notation, we may write

$$\text{ATT} = \text{ATE} = \bar{Y}_1 - \bar{Y}_0 \tag{1}$$

where $\bar{Y}_1$ is the average number of weeks on sickness absence for the treatment group, while $\bar{Y}_0$ is the average number of weeks on sickness absence for the control group. The nonparametric average effect of the RTW program is simply the difference between the two averages.

As usual, ATT denotes average treatment of the treated, and ATE denotes average treatment effect (for the population). Given the RCT design, these two types of average effects coincide here. The interpretation, therefore, is that for the entire population we estimate what an extension of the program would yield. The results do not carry over to other municipalities, but do carry over to extended versions of the program within this municipality. As shown in Table 1, though, control variables are needed, in order to balance the treatment and control groups on a few of the observable characteristics. We do this by estimating a weighted OLS model.

One important purpose of this paper is to shed light on the mechanisms that make RTW coordination (economically) successful. One avenue is simply to intro-

10 If we take out maternity leave from the category "other income support," there is no difference between the two groups.

https://doi.org/10.1017/bca.2017.1 Published online by Cambridge University Press

duce interactions between RTW participation and observable characteristics, for instance, interaction with type of diagnosis.[11] Here, we make additional use of the tremendously good data available. In order to make the most of the available information, we use confirmatory factor analysis (CFA) to create a combined measure that may capture the severity of the condition of the sick-listed person. In particular, the detailed individual historical records of employment, purchases of prescription drugs and hospital diagnoses. Moreover, the self-assessed diagnoses recorded by the person who did the initial screening of the long-term sick-listed person are combined into a single factor that can be thought of as a measure of complexity combined with sickness history. Such latent factors are computed overall as are within types of reported diagnoses and medicine used in relation to that condition.

It is intrinsically difficult to measure the program impact for each individual, since, by definition, the counterfactual for each person is unobserved. However, we may compute the entire distribution of outcomes for treated persons and compare this with the distribution of outcomes for controls (see Heckman et al. (1997)). If the distance between the two distributions is not constant over the population, this will be sign of heterogeneity. Hence, although we are unable to estimate the impacts individually, we might nevertheless gain knowledge about whether there is heterogeneity in the distribution of effects. It is important to note the limitations of QTE. With this approach, we can ascertain that there is some degree of heterogeneity if the two distributions, treated vs controlled, are not alike, even though QTE does not reveal the individual impact. However, if we invoke an *order assumption*, whereby individuals with the longest sickness duration without treatment also are the ones with longest duration among the treated (and so forth following the order of all participants), we obtain a measure of individual net benefits. While surely false to some degree in almost any application, this assumption may be approximately true for many cases including our application.[12]

In practice, we do this by estimating QTEs. Following Bitler et al. (2006), consider two distributions of predicted duration of sick leave for treated and nontreated persons, respectively, F_1 and F_0, and define QTE as $\Delta_q = y_q(1) - y_q(0)$, where $y_q(t)$ is the qth quantile of distribution F_t. Graphically, this is equivalent to the horizontal distance between two cumulative distributions.

11 Based on a more narrow data window, it has previously proven hard to find any significant differences in the impact of RTW across observable characteristics, Winzor et al. (2013).

12 Hence, QTE can reveal the existence of heterogeneity but the converse is not true: QTE estimates may be identical even though heterogeneity exists.

https://doi.org/10.1017/bca.2017.1 Published online by Cambridge University Press

Table 2 Elements included in the BCA.

Costs	Benefits
1. Municipal extra costs related to RTW	1. Wages (= production)
2. Marginal costs of public funds of additional public cost for public transfers.	2. Reduced costs for medical treatment
	3. Reduced costs for prescription drugs (incl. share paid by user).

3.2 Benefit-cost analysis

We link the estimated effects to the costs of achieving them through a BCA for society as a whole. The RTW program is implemented at the municipal level, and given the economic strain put on municipalities, costs and benefits related to such a decision are likely very important. However, there are returns beyond the municipality level, and we therefore assess the economic effects of the program for society at large. The elements included in the BCA are listed in Table 2.

The cost side includes extra costs to the coordinator and marginal costs of public funds. Taxes and transfers between the municipality and citizens do not impact the production in society in general and are therefore not to be included in the BCA at society level, (Mishan and Quah (2007)). Still, distortionary effects from collecting taxes, the marginal costs of public funds (MCF), are included at the society level. The additional or reduced distortion expected from this program stems from the additional or reduced use of public transfers due to the program. We use the treatment-effect estimate from the weighted OLS model to calculate the reduced public cost for sickness benefits and the increased public cost for other public transfers. The remaining costs and benefits are measured directly in data.[13] The range of MCF is discussed further below. Among benefits, wage is the primary component, but reduced (public) costs for medical treatment and medicine are also accounted for.

All BCAs are uncertain, and ours is no exception. For instance, the estimated benefits are essentially random variables, and the municipal costs are partly based on subjective assessments by civil servants and are intrinsically difficult to determine, (see Appendix B). The appropriate approach for a BCA, therefore, is to treat the multiple uncertainties jointly through the use of Monte Carlo simulations. This is also strongly recommended by Vining and Weimer (2010); previous examples

13 We estimate competing risk hazards to define the distribution of those leaving sickness benefit for other public transfers. Shadow prices for the different types of transfers are presented in Table 4 below.

https://doi.org/10.1017/bca.2017.1 Published online by Cambridge University Press

include Weimer and Sager (2009), Nichol (2001) and Briggs, Mooney and Wonderling (1999). We follow the recommendation and perform a Monte Carlo analysis of the uncertainty related to the BCA.

In practice, we assume that MCF is uniformly distributed across the interval [20%; 120%]. This covers several of the elasticity scenarios in Kleven and Kreiner (2006).[14] The use of a uniform distribution in place of a normal or log-normal distribution increases the variance, and this underscores the uncertainty involved for this parameter. The estimated treatment-effect parameter is assumed to follow a normal distribution with standard deviation equal to the estimated standard deviation. It appears reasonable to assume that the uncertainty linked to the municipal assessment of costs is bell-shaped around the "best guess" given. For this reason, we assume a normal distribution for these costs as well, but lacking any prior about the uncertainty we assume that the 95% confidence interval for the "best guess" assessment varies by a factor of plus/minus 100%. In this manner, we combine a continuum of municipal cost scenarios with the Monte Carlo analysis. This approach is useful in situations where one (important) variable is uncertain, and where priors about the uncertainty are arbitrary to some degree and has previously mostly been used for dealing with uncertainty of the social discount rate.

4 Results

4.1 Average treatment and BCA

We first present results from a standard BCA based on the ATEs. We find that, on average, sick-listed individuals under the RTW program have 24 days shorter periods of sickness leave compared to sick-listed persons under the traditional RTW program; see Table 3. The comparison of the treatment and control groups presented in Table 1 showed that the treatment group differed from the control group in various aspects (age, gender, self-assessed cause of sickness absence and prior job status). Hence, we estimate the treatment effect both nonparametrically and in a parametrically weighted OLS model including the covariates that differ significantly between the treatment and control groups. Even though three out of the four

14 This range of MCF may appear high, and by international comparison, it is indeed high. Kleven and Kreiner (2006) extend previous measures of MCF to account for labor force participation responses including a distinction between extensive (participation) and intensive (hours of work) responses. Accounting for discrete participating decisions generates first-order effects of government revenue.

https://doi.org/10.1017/bca.2017.1 Published online by Cambridge University Press

Table 3 Average treatment effect of RTW on duration of sickness leave, nonparametric and weighted OLS regression.

	Nonparametric		Weighted OLS	
	Coeff.	**S.E.**	**Coeff.**	**S.E.**
RTW	−24.00***	(5.027)	−24.64***	(5.147)
Age			0.664***	(0.239)
Female			10.54**	(5.220)
Cause: psychiatric illness			34.95***	(9.834)
Prior job status: other income support			−9.305	(9.097)
Observations	2,310		2,310	

Note: Standard errors in parentheses. Duration is measured in days. *$p < 0.10$, **$p < 0.05$, ***$p < 0.01$. Covariates centered on their mean.

covariates significantly affect the duration of sickness leave, including them in the model does not alter the ATE significantly. The new point estimate is 24.6 days.[15]

Economically, the overwhelming benefits from RTW stem from lower public transfers, and consequently reduced marginal costs for public funds, as well as increased production, see Table 4.

A part of the reduction in sickness benefits is transferred to increased costs for unemployment insurance. This is because being reported fit for duty means that the individual is either back in employment or has been transferred to unemployment insurance or, if the person is not eligible for UI benefits, transferred to social welfare. The major costs are operating costs, but the net effect, the BCA result, is a large positive economic benefit from the RTW program in this municipality. The point estimate is 5,568 DKK per person in net benefit for society. The standard errors, calculated using 50,000 Monte Carlo iterations, indicate a relatively high uncertainty, even though the BCA result is positive at the 95% confidence level.

The most uncertain of all inputs in the BCA is the partly subjective assessment of the municipal operating costs. As noted above, we therefore introduce a transparent manner by which one can elicit to what extent the BCA results respond to this uncertainty. A common manner for dealing with uncertainty in the social discount rate is to vary the assessed costs from plus/minus 100% (the 100% level means we use the reported best assessment from the municipality, −100% means we apply zero cost and +100% means we apply double the assessed cost); see

15 The treatment effect found here is consistent with the findings in Nielsen et al. (2013). Based on the same data, but with censoring taking place earlier than in our data, they estimate a parametric duration model and find an estimated reduction due to RTW of 5.4 weeks.

https://doi.org/10.1017/bca.2017.1 Published online by Cambridge University Press

Table 4 Benefit-cost analysis, by item (DKK).

Item	Mean	S.E.	CI_{lower}	CI_{upper}
1. Change in health costs during sickness	−36.5	0.7	−38	−35
2. Reduced health costs after reported fit for duty	50.9	10	31	72
3. Increased production (wage)	6,578	1,351	3,946	9,246
4. Change in marginal cost of public funds				
– Sickness benefits	9,104	4,268	2,638	17,990
– Social welfare	−478	224	−944	−138
– Unemployment insurance	−2,768	1,297	−5,469	−802
– Other (e.g., rehabilitation)	−600	281	−1,185	−174
5. Fixed set-up costs	−714	312	−1,307	−222
6. Operating costs	−5,569	2,431	−10,208	−1,729
Economic effect (benefits-costs) per person in RTW program	5,568	2,639	111	10,577

Note: Standard errors are bootstrapped based on 50,000 iterations. CI = 95% Confidence Interval. A minus denotes a cost. 5.6 DKK equaled approximately 1 USD, in the years covered by our sampling period. Further details given in Appendix C.

Moore, Boardman, Vining, Weimer and Greenberg (2004).[16] The normal distribution is truncated to zero in order to avoid that the costs become negative. The results of this exercise are shown in Figure 1.

As the gray area indicates, the horizontal confidence interval broadens as the assessed program cost increases.

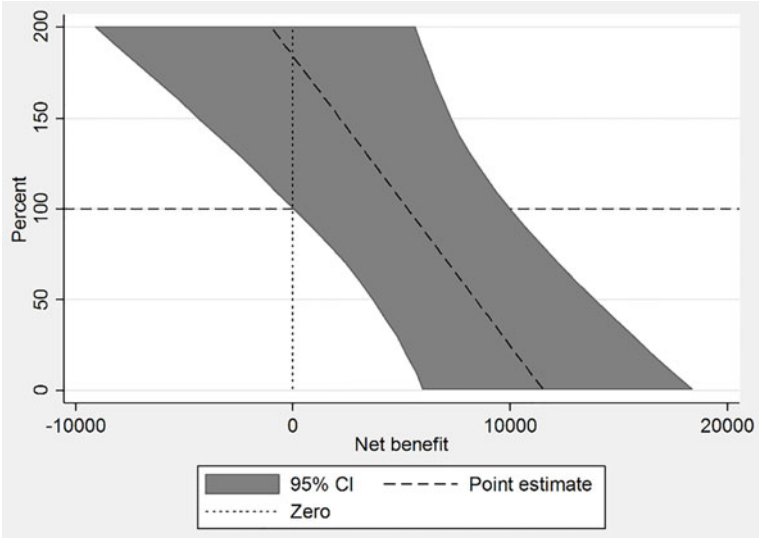
4.2 Accounting for heterogeneity

To explore the possibility of a heterogeneous treatment effect, we estimate QTEs. The difference in the distribution of length of absence is informative with regard to how the effect of the RTW program may vary over the distribution. Treatment-effect differences within the distribution of sickness leave duration will indicate whether, for instance, the treatment effects are stronger for the long-term sick-listed than the short-term sick-listed.

Figure 2 shows the cumulative distributions for the treatment groups and the control groups of the RTW program. The horizontal difference in the distributions

16 It might appear far-fetched to allow for such a wide range of operating costs, but our experience from retrieving such assessments from municipalities is that it is very difficult for them to estimate the true value. Further, by receiving only one value, we have no information on the distribution of the assessed operational cost, and a uniform distribution therefore arguably is a good choice.

https://doi.org/10.1017/bca.2017.1 Published online by Cambridge University Press



Note: Uncertainty of costs is handled jointly using 10,000 Monte Carlo simulations and a 95% confidence interval (horizontal) for the assessed operating costs, which vary along the vertical axis. At 100%, we use the point estimate, and at 0% the mean costs are set to zero.

Figure 1 Net benefits of RTW program.

indicates the impact of the program, and this difference corresponds to the QTE effect, which is presented in Figure 3. A horizontal line for the ATE is presented in Figure 3.

Figure 3 clearly indicates that the RTW program has varying impact across the distribution. Given the order assumption holds, the largest effects are found approximately between the 40th and 80th percentiles of sickness absence duration. Also, note that the duration of sickness absence is lower for the treatment group at almost all points in the distribution, but not significantly lower at the two extremes.

The comparison between ATE and QTE, shown in Figure 3, reveals substantial heterogeneity in the treatment effect. If we were able to perfectly screen participants and only allocate treatment to individuals with a treatment effect above the average cost level, (in the absence of individualized cost data), i.e., individuals above the break-even line shown in Figure 4, then, obviously, we would be able to improve the bottom line.

The potential room for improving the bottom line in the BCA is what we define as an EP.

Definition. The "Efficiency Potential" (EP) is the increased benefits net of costs that would accrue to the society, if it was possible to perfectly screen participants

https://doi.org/10.1017/bca.2017.1 Published online by Cambridge University Press

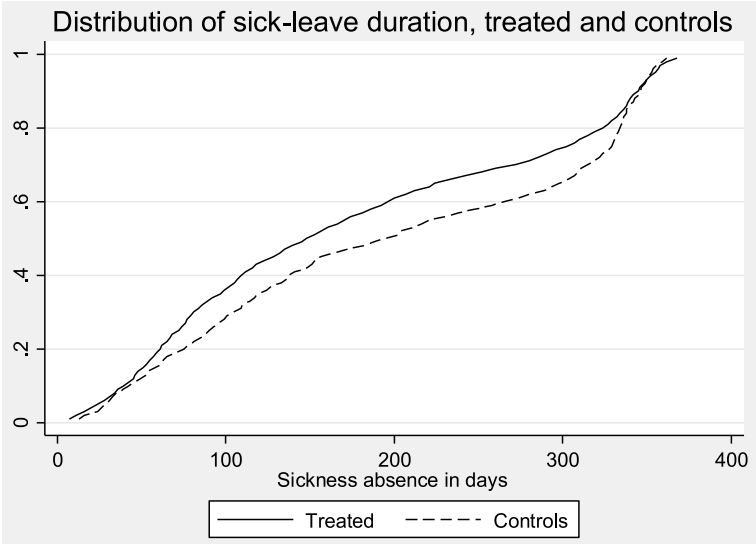


Figure 2 Cumulative distribution of sick-leave duration.

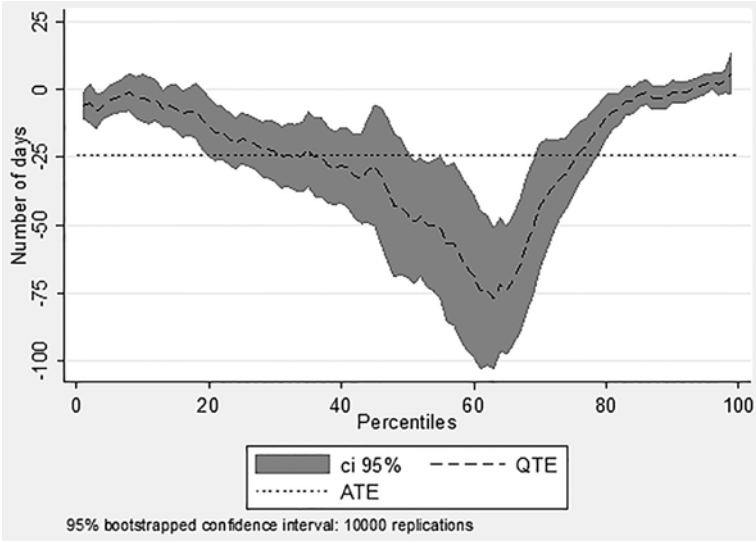
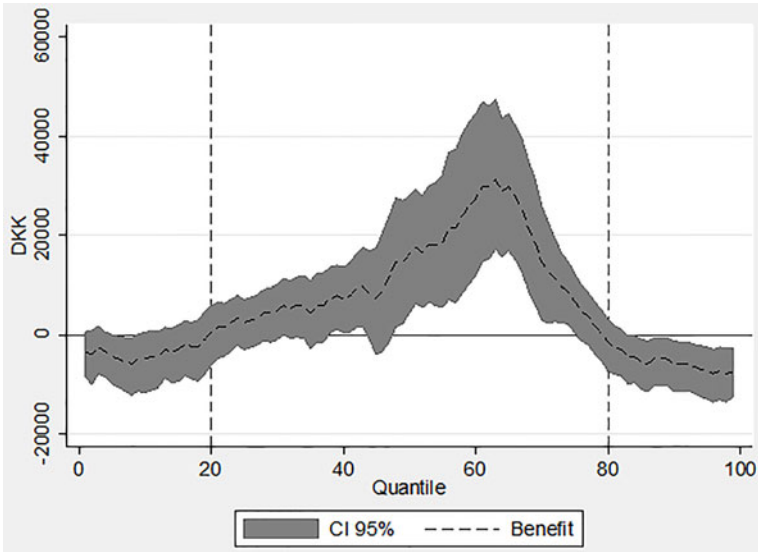


Figure 3 Average and quantile treatment effects.

https://doi.org/10.1017/bca.2017.1 Published online by Cambridge University Press



Note: Break-even at the horizontal zero-line

Figure 4 Quantile treatment and break-even analysis.

to those who benefit individually at least as much as the costs involved. Formally, the EP is given by

$$\text{EP} = \frac{(\text{BCA Perfect} - \text{BCA realized})}{\text{BCA realized}}.$$

In our application, the BCA results in an economic net effect of 7,923,264 DKK (Table 4). The corresponding sum of BCA among the share of individuals with a significantly positive treatment effect is 9,630,864 DKK.[17] Hence, the implied EP in this application is 0.22.[18]

The heterogeneity can be split between observed and unobserved heterogeneity and, clearly, the EP can only be harvested if we are able to screen individuals based on observable characteristics. One may expect heterogeneity in the treatment effect to arise from specific observable characteristics of the treated. For instance, coordination could very well be of greater use for sick-listed persons with more complicated diseases. Next, we therefore seek to identify which observable covariates drive the heterogeneous treatment effect.

17 This maximum benefit is calculated as the sum of BCA among the Monte Carlo draws that came out with a positive BCA.

18 There may be heterogeneity in benefits across individuals not captured by the quantile methods employed in this paper. Hence, the EP is not a maximum benefit that might be obtained.

https://doi.org/10.1017/bca.2017.1 Published online by Cambridge University Press

Table 5 Duration of sickness spell, weighted OLS regression with different interactions.

	Model (1)		Model (2)	
RTW	−16.89**	(8.040)	−27.98***	(8.729)
RTW interaction (ref = public transfer recipient)				
RTW*employed	−12.17	(14.07)		
RTW*self-supporting	−5.020	(13.86)		
RTW interaction (ref = compulsory education)				
RTW*vocational education			0.927	(11.95)
RTW*college			15.29	(15.06)
RTW*long education			10.10	(21.24)
Covariates as controls	Yes		Yes	
Observations	2,310		2,310	

Note: Standard error in parentheses. *$p < 0.10$, **$p < 0.05$, ***$p < 0.01$. Balanced with covariates, as in Table 3.

We apply three different strategies to investigate the hypothesis regarding heterogeneous treatment effects.

First, we interact the RTW treatment with prior employment status and level of education. Part of the coordination involves maintaining the individuals' connection with the labor market (for those who were in employment prior to their illness). One may hypothesize that such a connection would mean that the sick-listed person could benefit from being able to return to his/her former job once ready, and that this could even shorten the sickness spell. Along similar lines, highly educated persons might have a greater degree of empowerment, (Kristensen, Andersen and Pedersen (2012)), and this could perhaps result in greater benefits from coordination.[19] However, these interaction terms are insignificant, and so the heterogeneity observed in Figure 2 is not linked to these covariates, see Table 5.

The second approach is to investigate the hypothesis that individuals with complicated sickness spells would gain more from a coordinated RTW program. In order to capture the level of complication for each sickness spell, we make use of factor analysis. This allows us to exploit the wide range of good quality data available by summarizing them according to different measures of sickness severity. Using CFA, we estimate factors that are based on a predefined structure.

Hence, three separate measures, one each for severe psychiatric diseases, severe somatic diseases and severe pain diseases, are estimated. In addition, a joint mea-

19 For instance, it could be the case that coordination would be better for people who are able to speak up for themselves. This type of ability is captured by "empowerment" and has been shown to be closely linked to education, (Kristensen et al. (2012)).

https://doi.org/10.1017/bca.2017.1 Published online by Cambridge University Press

Table 6 Average treatment effects and interactions with different factors for severity, weighted OLS regression.

	(1) No factor	(2) Psych	(3) Pain	(4) Somatic	(5) General
RTW = 1	−22.47***	−22.53***	−22.82***	−22.86***	−22.27***
	(−4.28)	(−4.34)	(−4.36)	(−4.37)	(−4.27)
Factor		5.114***	1.794	1.866	4.641***
		(4.05)	(1.40)	(1.14)	(3.22)
RTW* factor		1.246	0.632	1.948	0.230
		(0.73)	(0.39)	(0.93)	(0.13)
Constant	175.5***	182.1***	182.6***	180.6***	190.1***
	(13.76)	(14.29)	(13.81)	(13.90)	(14.41)

Note: t statistics in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$. Based on 2,310 observations. Balanced with covariates as in Table 3.

sure for general severe sickness spells is estimated and applied to the treatment-effect estimation. The estimated factors draw upon information on previous hospitalizations, previous medical costs, diagnosis, previous sickness leave periods and previous psychiatric costs.

The estimated factor-specific treatment effects are found to be insignificant for all four factors, see Table 6.

Hence, although the QTEs clearly predict the existence of heterogeneity, our measure for sickness severity is apparently not able to capture this heterogeneity. According to the findings in Table 6, having a history of either psychiatric illness, somatic illness or pain-affected illness does not increase the effect gained from being part of the RTW program.

Third, using individual characteristics (including measures of illness) we fit a model of duration, including a dummy for treatment but not including interactions, and then use the predicted values of this model to interact with the treatment. Once again (results now shown) we find very insignificant estimation results of the RTW interactions.

5 Discussion

Based on observable covariates, nothing seems to indicate that the RTW program has different effects on sickness beneficiaries across observable differences. Following Bitler et al. (2006, 2014), this finding is not surprising. They argue that mean

https://doi.org/10.1017/bca.2017.1 Published online by Cambridge University Press

subgroup impacts, which is what we estimate using subgroup interactions, may be poorly suited to capturing treatment-effect heterogeneity when both within-group and between-group treatment-effect heterogeneity are present. Our results indicate that within-subgroup heterogeneity may drive the QTE results. Bitler et al. (2014) also build on RCT data in their case related to labor supply. They find that adjusting subgroups' mean impacts for extensive margin impacts of reforms leads to a substantial improvement in the ability of cross-subgroup mean impact treatment effects to explain distributional effects of welfare reform. In our data, labor force participation might also be an important issue, and hence the extensive margin may also be relevant here. Our outcome, however, is limited to observing whether or not the sick-listed persons return to work, with no further analysis of whether they return to work, say, 12 or 37 hours a week.

Our findings corroborate the general message found in Bitler et al. (2006, 2014) and may consequently draw a rather pessimistic picture of the true potential of accounting for heterogeneity in BCAs. However, while it may be impossible to harvest close to the maximum EP subgroup analyses have nevertheless often been argued to be highly important, (e.g., in the vast literature on ALMP, see the handbook chapter in Heckman, LaLonde and Smith (1999) and the meta-analysis in Card, Kluve and Weber (2010)). Furthermore, our sample size of 1,423 treated individuals and 887 individuals in the control group is relatively modest and consequently provides relatively little statistical power.

Making the most of BCA by including heterogeneity in the analysis ought to include not only heterogeneous treatment effects but preferably also heterogeneous costs. In our application, the municipality had a relatively difficult task in coming up with a credible estimate of the average costs, and individualized cost data may generally be hard to collect. This aspect of heterogeneity in BCA might in fact be just as important as the heterogeneity in treatment effects, though the potential seems to be even more difficult to realize.

6 Conclusion

Heterogeneity in BCA is strikingly absent, even though it is well known that one should expect treatment-effect heterogeneity, and hence expect a scope for targeting publicly funded programs better. Our example here – a large scale RTW program in a large Danish municipality achieving significant average net benefits – shows that there might indeed be a large potential. However, it also shows that, even with the extremely rich data available, we have not been able to gain knowledge about how any screening system could tap into the EP. In part, this lack of significant

https://doi.org/10.1017/bca.2017.1 Published online by Cambridge University Press

subgroup effects is a result of a relatively small sample size and the consequent low test power.

The fundamental idea to take away from this analysis remains: There is a large potential for improved efficiency and increased knowledge from including heterogeneous treatment effects in BCAs. Future research, as well as applied consultancy work, should seek to collect the type of data needed for individualized (or subgroup-dependent) costs, and treatment effects should allow for more than a simple average.

Appendix A. The Danish National RTW program

The Danish National RTW program was targeted at sickness benefit receivers in category 2. With a view to improving RTW, it consisted of three core elements that aimed at changing the procedures of sickness benefit management. The core elements were: (i) establishment of a multidisciplinary RTW team consisting of two RTW coordinators (municipal case workers), a psychologist, a physiotherapist or ergotherapist, a part-time psychiatrist and a part-time physician of occupational, social or general medicine, (ii) implementation of standardized workability assessments tools, and (iii) a comprehensive training course for all multidisciplinary R teams prior to the program (for details see Aust et al., 2012, 2015). Both the assessment tools and the training course had a special focus on supporting better detection and handling of psychological health problems. In contrast to ordinary sickness benefit management, RTW coordinators in the RTW program had the option of referring sick-listed persons to the RTW team for further assessment. However, it was expected that only half of the category 2 sickness beneficiaries would need to be referred to the interdisciplinary RTW team. Hence, part of the task of the RTW coordinator was to assess whether a referral was needed. If referred to the RTW team, the sick-listed person was asked to attend a consultation with a psychologist *and* with a physiotherapist or ergotherapist. In this way, the RTW team members were able to identify sick-listed persons with both somatic and psychological symptoms, which generally serves as a predictor of risk of long-term sickness.

If further assessment was needed, sick-listed persons could be referred to the psychiatrist and/or the physician of occupational, social or general medicine. The RTW program was implemented in 21 Danish municipalities (Aust et al., 2015). Only in three of these municipalities was the evaluation design a randomized controlled trial (RCT). This paper builds on the larger of the three municipalities, which gives a highly homogeneous sample with only within municipality variation. The experimental design in this municipality was especially successful, as the random-

https://doi.org/10.1017/bca.2017.1 Published online by Cambridge University Press

ization was conducted centrally before the cases were divided into two physical divided units, one carrying out the coordinated RTW program, the other offering the ordinary sickness benefit management. Although the caseworkers were placed in geographically separate units, the sick-listed persons were spread geographically independently of being in the treatment group or not.

Appendix B. Municipal economic data

The economic program costs are provided by the municipalities. The costs are assessed both for the treatment and the control group, and the difference thus lies in measuring the extra cost per sickness benefit case for providing the RTW program compared to the ordinary sickness benefit management. The total costs are separated into initial start-up costs and operating costs. Initial start-up costs consist of start-up expenses, on education, office arrangement etc., and operating costs in the form of administrative expenses for case administration and program expenses as labor market training and activation. The cost of the sickness benefit itself is not included.

The costs associated with the program are very difficult to trace out for the municipalities, as these are not registered on specific accounts, and as a set of locally estimated algorithms for distributing personnel costs between RTW and other activities have to be applied. Consequently, the reported costs for the individual municipalities are fraught with uncertainty. In order to reduce sensitivity, and as a conservative approach, we apply the average costs across the 18 municipalities that took part in the program. The cost estimate is conservative in the sense that the costs here are higher than the cost assessment made by the municipality in question.[20]

Municipal cost sheet

Average costs per case	Treatment	Control	Difference
Initial start-up costs	1,020	0	1,020
Operating costs	27,839	19,884	7,955

20 Municipal level costs were applied in Houlberg et al. (2012), in which the cost estimates were less sensitive to local algorithms, as the study included all 18 municipalities in the program. One could argue that the reason why costs differ so much across municipalities is, in part, structural and organizational differences, though the municipalities themselves stated that they were very uncertain about the costs levels. There does not seem to be a first best choice, but the costs applied here can be considered a conservative estimate.

https://doi.org/10.1017/bca.2017.1 Published online by Cambridge University Press

Appendix C. Shadow prices and BCA

Weekly shadow prices for public transfers are:

- Sickness benefit: 3,830 DKK
- UI benefit: 3,830 DKK
- Social welfare: 3,069 DKK
- Other transfers: 3,657 DKK

When program participation reduces the duration of sickness absence it enters the BCA as the difference between these weekly shadow prices. If a person makes a transition from sickness benefits to social welfare, the net impact on public funds is $(3{,}830 - 3{,}069) = 761$ DKK. If a person makes a transition from sickness benefits to work the net impact becomes both the reduced sickness benefits (3,830 DKK) and the increased production (= the wage). In both cases, change in the need for public funds is also included as a change in the marginal cost of public funds.

References

Aust, Birgit, Helverskov, Trine, Nielsen, Maj Britt D., Bjorner, Jakob Bue, Rugulies, Reiner, Nielsen, Karina, Sørensen, Ole H, Grundtvig, Gry, Andersen, Malene F., Hansen, Jørgen V., Buchardt, Helle L., Nielsen, Lisbeth, Lund, Trine L., Andersen, Irene, Andersen, Mogens H., Clausen, Aksel S., Heinesen, Eskil, Mortensen, Ole S., Ektor-Andersen, John, Ørbæk, Palle, Winzor, Glen, Bültmann, Ute & Poulsen, Otto M. (2012). The Danish National Return-to-work program – Aims, Content, and Design of the Process and Effect Evaluation. *Scandinavian Journal of Work, Environment & Health*, *38*(2), 120–133.

Aust, Birgit, Nielsen, Maj Britt D., Grundtvig, Gry, Buchardt, Helle L., Ferm, Linnea, Andersen, Irene, Lund, Trine L., Jelle, Martin O. C., Andersen, Malene F., Hansen, Jørgen V., Tverborgvik, T., Helverskov, Trine, Bjorner, Jakob Bue, Rugulies, Reiner, Ørbæk, Palle, Winzor, Glen, Bültmann, Ute & Poulsen, Otto M. (2015). Implementation of the Danish Return-to-work Program: Process Evaluation of a Trial in 21 Danish Municipalities. *Scandinavian Journal of Work, Environment & Health*, *41*(6), 529–541.

Bitler, Marianne P., Gelbach, Jonah. B. & Hoynes, Hilary W. (2006). What Mean Impacts Miss: Distributional Effects of Welfare Reform Experiments. *American Economic Review*, *96*(4), 988–1012.

Bitler, Marianne P., Gelbach, Jonah B. & Hoynes, Hilary W. (2014). Can Variation in Subgroups' Average Treatment Effects Explain Treatment Effect Heterogeneity? Evidence from a Social Experiment. *Review of Economics and Statistics*, forthcoming.

Boardman, Anthony, Greenberg, David, Vining, Aidan & Weimer, David (2014). Cost-Benefit Analysis. In *Concepts and Practice*. (4th ed.). Upper Saddle River, NJ: Prentice Hall.

Briggs, Andrew H., Mooney, Chrisptoher Z. & Wonderling, David E. (1999). Constructing Confidence Intervals for Cost-Effectiveness Ratios: An Evaluation of Parametric and

https://doi.org/10.1017/bca.2017.1 Published online by Cambridge University Press

Non-parametric Techniques Using Monte Carlo Simulation. *Statististics in Medicine*, *18*, 3245–3262.

Card, David, Kluve, Jochen & Weber, Andrea (2010). Active Labour Market Policy Evaluations: A Meta-Analysis. *The Economic Journal*, *120*(November), F452–F477.

Harberger, Arnold C. (1978). On the Use of Distributional Weights in Social Cost-Benefit Analysis. *Journal of Political Economy*, *86*(2), 87–120; pt. 2.

Heckman, James, J., LaLonde, Robert J. & Smith, J. Jeffrey (1999). The Economics and Econometrics of Active Labor Market Programs. In Orley Ashenfelter & David Card (Eds.), *Handbook of Labor Economics* (Vol. III, pp. 1865–2097). Amsterdam: North-Holland.

Heckman, James, J., Smith, Jeffrey & Clements, Nancy (1997). Making The Most Out Of Program Evaluations and Social Experiments: Accounting For Heterogeneity in Program Impacts. *Review of Economic Studies*, *64*, 487–535.

Holtmann, Alphonse G. (1991). Beyond Efficiency: Economics and Distributional Analysis. In David L. Weimer (Ed.), *Policy Analysis and Economics*. Developments, Tensions, Prospects. Norwell, MA: Kluwer Academic Publishers.

Houlberg, Kurt, Kolodziejczyk, Christophe & Kristensen, Nicolai (2012). Evaluering af Det Store TTA-projekt. Delrapport om økonomisk Evaluering. Report. København: KORA (In Danish).

Imbens, Guido W. & Wooldridge, Jeffrey M. (2009). Recent Developments in the Econometrics of Program Evaluation. *Journal of Economic Literature*, *47*(1), 5–86.

Kleven, Henrik J. & Kreiner, Claus T. (2006). The Marginal Cost of Public Funds: Hours of Work Versus Labor Force Participation. *Journal of Public Economics*, *90*, 1955–1973.

Kristensen, Nicolai, Andersen, Lotte B. & Pedersen, Lene H. (2012). Public Service Efficacy. *International Journal of Public Administration*, *35*(14), 947–958.

Le Grand, Julian (2003). *Motivation, Agency and Public Policy: Of Knights and Knaves, Pawns and Queens*. Oxford: Oxford University Press.

Le Grand, Julian (2010). Knights and Knaves Return: Public Service Motivation and the Delivery of Public Services. *International Public Management Journal*, *13*(1), 56–71.

Mishan, Ezra J. & Quah, Euston (2007). *Cost-Benefit-Analysis*. (5th ed.). London and New York: Routledge.

Moore, Mark A., Boardman, Anthony E., Vining, Aidan R., Weimer, David L. & Greenberg, David H. (2004). Just Give Me a Number! Practical Values for the Social Discount Rate. *Journal of Policy Analysis and Management*, *23*(4), 789–812.

Nichol, Kristin L. (2001). Cost-Benefit Analysis of a Strategy to Vaccinate Healthy Working Adults Against Influenza. *Archived of Internal medicine*, *161*(5), 749–759.

Nielsen, Maj Britt D., Hansen, Jorgen V., Aust, Birgit, Tverborgvik, Torill, Thomsen, Birthe L., Bjorner, Jakob Bue, Mortensen, Ole Steen, Rugulies, Reiner, Winzor, Glen, Ørbæk, Palle, Helverskov, Trine, Kristensen, Nicolai & Poulsen, Otto Melchior (2013). A Multisite Randomized Controlled Trial on Time to Self-support Among Sickness Absence Beneficiaries. The Danish National Return-to-work Program. *European Journal of Public Health*, *25*(1), 96–102.

Poulsen, Otto Melchior, Aust, Birgit, Bjorner, Jakob Bue, Rugulies, Reiner, Hansen, Jorgen V., Tverborgvik, Torill, Winzor, Glen, Mortensen, Ole Steen, Ørbæk, Palle & Nielsen, Maj Britt D. (2012). Effect of the Danish Return-to-work Program on Long-term Sickness Absence: Results From A Randomized Controlled Trial in Three Municipalities. *Scandinavian Journal of Work, Environment & Health*, *38*(2), 120–133.

https://doi.org/10.1017/bca.2017.1 Published online by Cambridge University Press

Schandelmaier, Stefan, Ebrahim, Shanil, Burkhardt, Susan C. A., de Boer, Wout E. L., Zumbrunn, Thomas, Guyatt, Gordon H., Busse, Jason W. & Kunz, Regina (2012). Return to Work Coordination Programs for Work Disability: A Meta-Analysis of Randomised Controlled Trials. *PLoS One*, *7*(11), e49760.

Vining, Aidan & Weimer, David L. (2010). An Assessment of Important Issues Concerning the Application of Benefit-Cost Analysis to Social Policy. *Journal of Benefit-Cost Analysis*, *1*(1), 1–38.

Weimer, David L. & Sager, Mark A. (2009). Early Identification and Treatment of Alzheimer's Disease: Social and Fiscal Outcomes. *Alzheimer's & Dementia*, *5*, 215–226.

Winzor, Glen, Poulsen, Otto M., Andersen, Irene, Andersen, Malene Friis, Andersen, Mogens H., Aust, Birgit, Bjørner, Jakob, Buchardt, Helle Løvdal, Clausen, Aksel S., Ferm, Linnea, Grundtvig, Gry, Hansen, Christoffer R., Hansen, Jørgen Vinsløv, Helverskov, Trine, Hviid, Kirsten, Jelle, Martin O., Juhl, Martin, Lund, Trine Lehmann, Mortensen, Ole Steen, Nielsen, Lisbeth, Nielsen, Maj Britt Dahl, Lykke, Birthe, Helle Tornemand, Thomsen & Tverborgvik, Torill 2013. *Det store TTA-projekt. Proces-, effekt- og økonomisk evaluering*. København: Det Nationale Forskningscenter for Arbejdsmiljø (NFA) (In Danish).

https://doi.org/10.1017/bca.2017.1 Published online by Cambridge University Press