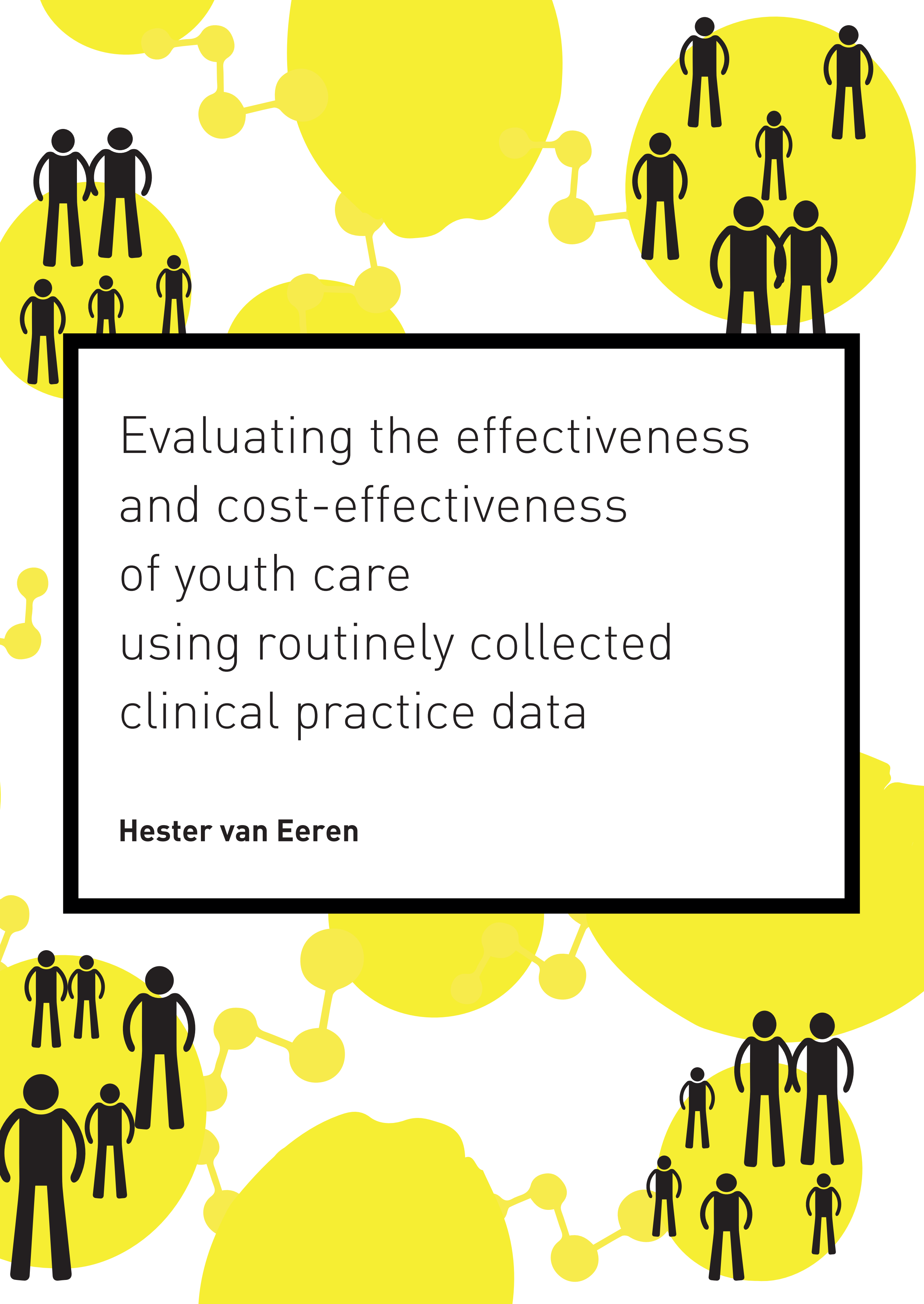


Evaluating the effectiveness and cost-effectiveness of youth care using routinely collected clinical practice data

Hester van Eeren

Evaluating the Effectiveness and Cost-Effectiveness of Youth Care using Routinely Collected Clinical Practice Data

Hester van Eeren

ISBN:	978-94-6299-709-7
Printed by:	Ridderprint BV – www.ridderprint.nl Iliana
Layout:	Boshoven-Gkini – www.AgileColor.nl
Cover design:	Nicolette Mulder

Evaluating the Effectiveness and Cost-Effectiveness of Youth Care using Routinely Collected Clinical Practice Data

De effectiviteit en kosteneffectiviteit van jeugdinterventies onderzocht
door gebruik te maken van Routine Outcome Monitoring data

Proefschrift

ter verkrijging van de graad van doctor aan de
Erasmus Universiteit Rotterdam
op gezag van de rector magnificus

Prof.dr. H.A.P. Pols

en volgens het besluit van het College van Promoties.

De openbare verdediging zal plaatsvinden op
woensdag 15 november 2017 om 11.30 uur

door

Hester van Eeren
geboren te Numansdorp

Erasmus University Rotterdam

Promotiecommissie

Promotoren:
Prof.dr. J.J. van Busschbach
Prof.dr. R.H.J. Scholte

Overige leden:
Prof.dr. C.D. Dirksen
Prof.dr. B. Orobio De Castro
Prof.dr. H. Raat

Copromotor:
Dr. R.E.A. van der Rijken

Contents

Chapter 1.

General introduction

Youth care is aimed at improving the well-being of children and adolescents and their parents. Moreover, youth care aims to prevent society from the burden of misbehavior and crime due to behavioral problems. Given these aims and the importance of them to society, substantial budgets are available for youth care in the Netherlands. To spend these budgets wisely, it would be worthwhile to be able to justify the reimbursement of youth care interventions based on cost-effectiveness data. Such information, however, is scarce because methods of cost-effectiveness analyses in youth care are not fully developed and data needed for such studies are difficult to obtain. As a result, little is known about the (cost-) effectiveness of most of the available youth care interventions. Therefore, this thesis aims at contributing to the development of research tools to determine the effectiveness and cost-effectiveness of interventions in youth care.

Dutch youth care

The Dutch youth care system comprises a broad range of interventions. To get an impression of the available interventions, the 'database effective youth interventions' can be consulted (DEI; 'databank effectieve jeugdinterventies' in Dutch; Netherlands Youth Institute, 2016a). The DEI provides insight into the quality, feasibility, and effectiveness of interventions and aims to connect academic research and theories with youth care practice (Netherlands Youth Institute, 2016b). The underlying thought is that practice-based evidence can lead to evidence-based practice to improve the services rendered to children, adolescents, and their parents (Veerman, van Yperen, Bijl, Ooms, & Roosma, 2008). According to APA standards, evidence-based practice is *"the integration of the best available research with clinical expertise in the context of patient characteristics, culture and preferences"* (APA Presidential Task Force on Evidence-based practice, 2006, p. 273). This definition underscores the aim of the DEI to be transparent about the available evidence for each youth care intervention. The effectiveness of interventions in the DEI is ranked according to the so-called 'effectladder' that comprises four categories of evidence: the intervention has potential (i.e., is well described), is promising (i.e., has a good underlying program theory), is effective in daily clinical practice (i.e., shows first signs of effectiveness based on, for example, routine outcome monitoring), or the intervention is effective based on a research design with an experimental and a control condition (Veerman & van Yperen, 2008). The higher the level of evidence, the more proof that the outcomes found after finishing an intervention indeed resulted from the intervention itself and were not caused by confounding factors (Veerman & van Yperen, 2008). Various committees consisting of clinicians and scientists with a specific area of expertise rank submitted interventions on their level of effectiveness (Netherlands Youth Institute, 2016c). Gathering available evidence and structuring it for all youth care interventions in the DEI supports clinicians, youths and families, but also policy makers in weighing the pros and cons of the interventions offered and, thus, making well-informed choices.

Not being randomized

A Randomized Controlled Trial (RCT) is considered to be the golden standard for comparing the effectiveness of interventions; its results are considered as the highest scientific evidence. In an RCT the internal validity is high, because participants are randomized over an experimental and control condition and the aim is to obtain unbiased treatment effect estimates: the treatment selection bias is assumed to be zero (APA, 2006; Imai, Kind, & Stuart, 2008; Stuart, Cole, Bradshaw, & Leaf, 2011). Proven effectiveness in an RCT is also the highest level of evidence for interventions to be marked as effective in the DEI.

However, conducting an RCT is not always feasible due to ethical or practical constraints (Black, 1996). Therefore, alternative options should be considered to obtain evidence for clinical practice. For example, a quasi-experimental study design using observational data could be used in evaluating the effectiveness of interventions in clinical practice. Using such a study design, regular clinical practice can be followed and participants are not randomly allocated to interventions. The external validity of the results found in such studies could be higher than in RCT's, since participants represent the actual target population and the sample selection bias is closer to zero (Imai et al., 2008; Stuart et al., 2011). However, to increase the internal validity of non-randomized samples, and decrease the treatment selection bias, one should minimize the effect of allocation bias in evaluating the effectiveness of interventions.

Without controlling for possible baseline differences due to allocation bias in non-randomized studies, the effect of the interventions could be confounded by these initial differences. To control for allocation bias in observational data, the propensity score (PS) method is a valid and frequently used method. The PS is defined as the conditional probability of assignment to an intervention given a set of observed pre-treatment differences (Rosenbaum & Rubin, 1983). When applying the PS, the strongly ignorability assumption should be considered carefully, meaning that the treatment assignment should be independent of the potential outcome, the treatment assignment should be independent given a set of measured covariates, and each adolescent should have a chance of being in either treatment arm (Shadish, 2013). Given the assumption of strongly ignorability, applying the PS to achieve balance in the treatment arms enables the achievement of results equivalent to randomized studies (Austin, 2011; Shadish, 2013; West, Cham, Thoemmes, Renneberg, Schulze, & Weiler, 2014).

This thesis explores the possibilities of using the PS in youth care evaluation studies, because using the PS enables comparative effectiveness studies in youth care and increases the external validity of the results. External validity is especially important in health policy and, thus, in cost-effectiveness research, since the results should be translated to spending public money to reimburse cost-effective interventions.

Common practice

In Dutch youth care, interventions are mostly compared to treatment as usual. As a consequence, though treatment as usual can be proven to be effective too, it is not known which intervention is more effective when one has to choose between two evidence-based interventions. It complicates the choice for the more effective option, especially when one should decide on what works best for whom. Furthermore, when effective interventions are already available for a certain target population, the question is what the standard treatment (i.e., control condition) should be when evaluating a new treatment. Therefore, it could be worthwhile to mutually compare evidence-based interventions on their effectiveness.

Reimbursement criteria

Establishing the effectiveness is the second criterion when deciding which health care interventions should be reimbursed in the Netherlands, according to the 'Trechter van Dunning' (Busschbach & Delwel, 2010; Roscam Abbin, 1991). The necessity of care, based on the burden of disease, is the first criterion, and the cost-effectiveness is the third criterion, after which it is decided if the patient can pay the costs of care or whether it should be reimbursed by society, which is the fourth criterion (Busschbach & Delwel, 2010). When we translate these criteria to youth care, the following questions should be answered subsequently: 1) Is it necessary to treat children and adolescents with a high burden of disease, that prevents them to participate at a societal level? 2) Have the treatment options been proven effective as presented in the DEI? 3) What is the cost-effectiveness of the various treatment options?, and 4) Can the youths and their families pay the costs of these interventions themselves or should they be reimbursed by society? Thus far, research in youth care has mainly focused on the first two questions. The third and fourth question have not been given as much attention.

These reimbursement decisions are, however, important in youth care, since the responsibility for the budgets available for youth care was transferred from the Dutch national government to local authorities in 2015 (Transitiebureau Jeugd, 2015). This was done to enable municipalities to develop integrated policies and to tailor youth care to local and individual situations and needs. The ultimate goal was to create more coherent, more effective, more transparent, and less expensive services for children and their families. Because of the shifting budgets, it became even more important to be able to show which interventions are available and which of them are proven to be effective and cost-effective.

Economic evaluation

Given this ‘current state of play’, municipalities have to choose between available interventions. However, though there is substantial budget available, the budgets were cut in recent years and the budgets are limited, since municipalities have to spend their money on various domains (Transitie Autoriteit Jeugd, 2017). As a result, municipalities only have a limited budget to help ‘their’ youth and their families by reimbursing these interventions. It would thus be useful to have information on both the effectiveness and the costs of the interventions offered. Costs and effects are jointly evaluated in a cost-effectiveness analysis, which is increasingly being used to inform decisions regarding the reimbursement of health care interventions. In these analyses, the costs are preferably considered from a societal perspective, in which all relevant costs are included, irrespective of the payer perspective (Drummond, Sculpher, Torrance, O’Brien, & Stoddart, 2005; Zorginstituut Nederland, 2015).

A cost-effectiveness analysis can reveal various outcomes. For example, if the new intervention is more effective and less costly than the alternative or current treatment, the new intervention is preferred over the alternative. Moreover, when the new intervention is more effective, and the new and alternative interventions cost the same, or when both interventions are equally effective while the new intervention is less costly, the new intervention is also preferred. However, the reimbursement decision becomes more complicated if the new intervention is more effective and more costly. It should then be decided whether the additional effects of the more effective intervention are worth the additional costs (Drummond et al., 2005).

In the situation where the new intervention is more effective and more costly, the additional costs of the new intervention can be divided by the additional effects, which represents the Incremental Cost-Effectiveness Ratio (ICER). This ICER value can be compared to the amount of money the decision-maker, or society, would be willing to pay for an additional unit of effect. If the ICER then is lower than the willingness-to-pay (WTP) value, the new interventions is cost-effective. However, if the ICER is higher than the WTP value, investing in the new intervention is not possible or rather questionable and one would stick with the alternative or current intervention (Briggs, Claxton, & Sculpher, 2006).

Although nowadays economic evaluations are often used when evaluating health care and especially when evaluating new medicines (Rutten, 2010; Zorginstituut Nederland, 2015), cost-effectiveness analyses in youth care are not yet broadly applied. In the Netherlands, the guidelines on cost-effectiveness analyses only shortly describe their application in mental and forensic health care (Zorginstituut Nederland, 2015). Nevertheless, the number of cost-effectiveness analyses in youth care, but also in mental health care and in crime prevention, has increased (e.g., Knapp, McDaid, Evers, Salvador-Carulla, Halsteinli, & MHEEN Group, 2008; Soeteman & Busschbach, 2008).

This thesis

This thesis outlines two issues that follow the current state of evidence in youth care. The first issue deals with the stringent health care budgets and the municipalities being responsible for the reimbursement of interventions in youth care from January 2015 onwards. In light of these recent developments, it could help municipalities if both costs and effects of an intervention are clear and when interventions are compared on both outcomes. Therefore, this thesis addresses the issue of whether cost-effectiveness research in the field of youth care is feasible. Chapter 2 addresses this issue by illustrating the use of a cost-effectiveness model in which Functional Family Therapy (FFT) is compared with treatment as usual (TAU). The results of a cost-effectiveness model are, however, subject to uncertainty in their cost and effects estimates. This parameter uncertainty can be reduced if we would have more information on the costs and effects. Therefore, further research could be helpful, but further research is not without costs. In a value of information analysis, as presented in Chapter 3, the value of conducting further research is estimated and the type of research that would be most useful is identified. As this type of analysis has not been applied in the field of youth care before, Chapter 3 presents an example of this analysis based on the cost-effectiveness model of Chapter 2.

The second issue is raised by using available, non-randomized, data in investigating the effectiveness of interventions in youth care practice. In such designs, statistical methods can help to control for initial, non-random, differences between adolescents assigned to different treatment groups. The PS method is such a method, and it is increasingly being used in psychological research. In Chapter 4, a Monte Carlo simulation study is conducted to find out how the PS can be used in subgroup analysis. In Chapter 5, everyday practice data is used to compare FFT and Multisystemic Therapy (MST) on their effectiveness, using the PS method. Chapter 6, the general discussion, summarizes the findings of this thesis and discusses future perspectives and implications for clinical practice, policy makers, and researchers in youth care.

References

APA Presidential Task Force on Evidence-based practice (2006). Evidence-based practice in psychology. *American Psychologist, 61*, 271-285.

Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research, 46*, 399-424.

Black, N. (1996). Why we need observational studies to evaluate the effectiveness of health care. *BMJ, 312*, 1215-1218.

Briggs, A., Claxton, K., & Sculpher, M. (2006). *Decision modelling for health economic evaluation*. New York: Oxford University Press.

Busschbach, J. J. V., & Delwel, G. O. (2010). *Het pakketprincipe kosteneffectiviteit: Achtergrondstudie ten behoeve van de 'appraisal' fase in pakketbeheer (publicatienummer 291) [A background study on the 'costeffectiveness' package principle for the benefit of the appraisal phase in package management]*. Diemen: College voor zorgverzekeringen.

Drummond, M. F., Sculpher, M. J., Torrance, G. W., O'Brien, B. J., & Stoddart, G. L. (2005). *Methods for the economic evaluation of health care programmes, Third edition*. Oxford: Oxford University Press.

Imai, K., Kind, G., & Stuart, E. A. (2008). Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society*: Series A, 171, 481-502.

Knapp, M., McDaid, D., Evers, S., Salvador-Carulla, L., Halsteinli, V., & MHEEN Group (2008). *Cost-effectiveness and mental health* (MHEEN Policy briefing). London: MHEEN network.

Netherlands Youth Institute (2016a). *Alfabetisch overzicht erkende interventies [Alfabetical overview of interventions being acknowledged as effective]*. Retrieved from http://www.nji.nl/nl/Databank/Databank-Effectieve-Jeugdinterventies/Alfabetisch-overzicht-erkende-interventies.

Netherlands Youth Institute (2016b). *Databank effectieve jeugdinterventies: Over de databank [Databank Effective youth interventions: About the databank]*. Retrieved from http://www.nji.nl/nl/Databank/Databank-Effectieve-Jeugdinterventies-Over-de-databank.html.

Netherlands Youth Institute (2016c). *Erkenningscommissie interventies [Committee of accreditation of youth interventions]*. Retrieved from http://www.nji.nl/nl/Databank/Databank-Effectieve-Jeugdinterventies/Erkenningscommissie-Interventies.

Roscam Abbin, H. D. C. (1991). Kiezen en delen; rapport van de commissie Keuzen in de zorg (Commissie-Dunning) [Choosing or sharing: A report of the committee on choices in health care]. *Nederlands Tijdschrift voor Geneeskunde, 135*, 2239-2241.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*, 41-55

Rutten, F. (2010). Historische ontwikkeling. In: M. Rutten-van Mölken, C. Uyl-de Groot, & F. Rutten (Eds.), *Van kosten tot effecten. Een handleiding voor economische evaluatiestudies in de gezondheidszorg (tweede druk) [Costs and effects: A guideline in economic evaluation studies in health care]* (pp. 17-21). Amsterdam: Elsevier Gezondheidszorg.

Shadish, W. R. (2013). Propensity score analysis: Promise, reality and irrational exuberance. *Journal of Experimental Criminology, 25*, 129-144.

Soeteman, D. I., & Busschbach, J. J. V. (2008). Cost-benefit and cost-effectiveness of prevention and treatment. In: R. Loeber, N. W. Slot, P. H. van der Laan, & M. Hoeve (Eds.). *Tomorrow's criminals. The development of child delinquency and effective interventions* (pp. 215–228). Hampshire: Ashgate Publishing Ltd.

Stuart, E. A., Cole, S. R., Bradshaw, C. P., & Leaf, P. J. (2011). The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 174*, 369-386.

Veerman, J. W., van Yperen, T., Bijl, B., Ooms, H., & Roosma, D. (2008). Praktijkgestuurd effectonderzoek maakt hulpverlening beter [Practice-driven effectiveness research improves youth care]. *Jeugd en Co Kennis, 2*, 8-18.

Veerman, J. W., & van Yperen, T. A. (2008). Wat is praktijkgestuurd effectonderzoek? In: T. A. van Yperen, & J. W. Veerman (Eds.), *Zicht op effectiviteit. Handboek voor praktijkgestuurd effectonderzoek in de jeugdzorg [Effectiveness in practice: Handbook to practice-driven effectiveness research in youth care]* (pp. 17-34). Delft: Eburon.

Transitie Autoriteit Jeugd (2017). *Zorgen voor de jeugd. Derde jaarrapportage [Caring for youth].* Den Haag: Transitie Autoriteit Jeugd.

Transitiebureau Jeugd (2015). *Spoorboekje implementatie transitie jeugdzorg [Implementing the transition of youth care].* Retrieved from https://voordejeugd.nl/.

West, S. G., Cham, H., Thoemmes, F. J., Renneberg, B., Schulze, J., & Weiler, M. (2014). Propensity scores as a basis for equating groups: Basic principles and application in clinical treatment outcome research. *Journal of Consulting and Clinical Psychology, 82*, 906-919.

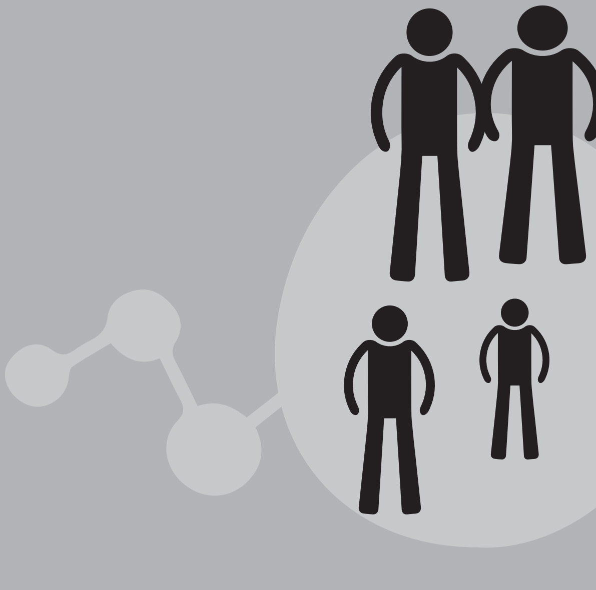
Zorginstituut Nederland (2015). *Richtlijn voor het uitvoeren van economische evaluaties in de gezondheidszorg [Guideline for economic evaluations in health care].* Diemen: Zorginstituut Nederland.

Chapter 2.

Framework for modelling the cost-effectiveness of systemic interventions aimed to reduce youth delinquency

Saskia J. Schawo, Hester V. Eeren, Djøra I. Soeteman, Marie-Christine van der Veldt, Marc J. Noom, Werner Brouwer, Jan J.V. Busschbach, & Leona Hakkaart

Journal of Mental Health Policy and Economics (2012), 15, 187-196



Abstract

Background: Many interventions initiated within and financed from the health care sector are not necessarily primarily aimed at improving health. This poses important questions regarding the operationalization of economic evaluations in such contexts.

Aims of the Study: We investigated whether assessing cost-effectiveness using state-of-the-art methods commonly applied in health care evaluations is feasible and meaningful when evaluating interventions aimed at reducing youth delinquency.

Methods: A probabilistic Markov model was constructed to create a framework for the assessment of the cost-effectiveness of systemic interventions in delinquent youth. For illustrative purposes, Functional Family Therapy (FFT), a systemic intervention aimed at improving family functioning and, primarily, reducing delinquent activity in youths, was compared to Treatment as Usual (TAU). "Criminal activity free years" (CAFYs) were introduced as central outcome measure. Criminal activity may e.g. be based on police contacts or committed crimes. In absence of extensive data and for illustrative purposes the current study based criminal activity on available literature on recidivism. Furthermore, a literature search was performed to deduce the model's structure and parameters.

Results: Common cost-effectiveness methodology could be applied to interventions for youth delinquency. Model characteristics and parameters were derived from literature and ongoing trial data. The model resulted in an estimate of incremental costs/CAFY and included long-term effects. Illustrative model results point towards dominance of FFT compared to TAU.

Discussion: Using a probabilistic model and the CAFY outcome measure to assess cost-effectiveness of systemic interventions aimed to reduce delinquency is feasible. However, the model structure is limited to three states and the CAFY measure was defined rather crude. Moreover, as the model parameters are retrieved from literature the model results are illustrative in the absence of empirical data.

Implications for Health Care Provision and Use: The current model provides a framework to assess the cost-effectiveness of systemic interventions, while taking into account parameter uncertainty and long-term effectiveness.

Implications for Health Policies: The framework of the model could be used to assess the cost-effectiveness of systemic interventions alongside (clinical) trial data. Consequently, it is suitable to inform reimbursement decisions, since the value for money of systemic interventions can be demonstrated using a decision analytic model.

Implications for Further Research: Future research could be focused on testing the current model based on extensive empirical data, improving the outcome measure and finding appropriate values for that outcome.

Introduction

Child delinquency poses a high economic burden on society (Welsh et al., 2008). Therefore, crime prevention and treatment of youth delinquents is of great importance to governments, in particular for Justice Departments. Systemic interventions, for instance Multisystemic Therapy (MST), Functional Family Therapy (FFT) or Parent Management Training Oregon (PTMO), are relatively costly interventions in youth health care aiming to reduce delinquent behavior (Aos, Lieb, Mayfield, Miller, & Pennucci, 2004). Cost-effectiveness studies are still limited in the field of youth health care. However, these costly systemic family interventions compete with medical treatments and other interventions for health care budgets, increasing the need for knowledge regarding the operationalization of economic evaluations in this context.

In the Netherlands, as part of an ongoing nationwide action plan of the Ministry of Justice, recently a selection was made of evidence-based treatments for delinquent youth (Ministry of Justice, 2008), among which MST, FFT and PMTO were implemented given their apparent effectiveness in reducing criminal activity in youths. The aim of these systemic interventions is not primarily to produce health in the sense of physical health and absence of disease, as measured in the Quality Adjusted Life Years (QALY) outcome. These interventions attempt to improve family functioning and may even intervene with the peers and school environment of the youth (i.e. Hendriks, van der Schee, & Blanken, 2011; Sexton & Alexander, 2000). Still, these treatments are reimbursed by the Dutch social health insurance system and, as such, part of the health care sector. Therefore, like other health care interventions, each intervention needs to demonstrate value for money since it competes for limited funds with other interventions. Efficiency considerations are deemed important in guiding decisions on which treatments to reimburse or initiate. However, given the atypical aim of these systemic interventions, i.e. reducing youth delinquency, an important question is how these types of interventions could demonstrate their efficiency or value for money. The conventional health economic approach of measuring improvements in terms of QALYs may fall short in this context.

Indeed, considering the literature on reducing youth delinquency, it becomes clear that important differences exist between economic evaluations performed in the health care sector and evaluations of crime prevention and treatment programs. It seems that both fields commonly perform sophisticated effect studies, including randomized controlled trials, meta-analyses and systematic reviews (Glisson et al., 2010; Henderson, Rowe, Dakof, Hawes, & Liddle, 2009; King et al., 2006; Teuffel et al., 2011; Ttofi, Farrington, Losel, & Loeber, 2011). Considering economic evaluations of crime prevention and treatment programs the classical cost-benefit analysis is conventionally used (Loeber & Farrington, 2000; Soeteman & Busschbach, 2008). An extensive cost-benefit evaluation of crime prevention and intervention programs has been performed by Aos and colleagues (2004) in the United States. That evaluation was based on a literature review, computation of average effects per treatment program, assignment of a monetary value to the effects and subsequently calculation of a net present value in a cost-benefit model structure. Furthermore, French and colleagues (2002a; 2002b), for example, conducted cost-benefit analyses on addiction treatment for substance abusers. These cost benefit analyses were deterministic models (Aos et al., 2004; French, McCollister, Cacciola, et al., 2002; French, McCollister, Sacks, et al., 2002). In addition, Aos and colleagues (2004) assessed costs and benefits from a taxpayer perspective. In health economic literature, cost-effectiveness analyses are preferably conducted from a societal perspective. Another difference between the two fields is, that in health economics sophisticated methodological guidelines for economic evaluations have been developed, while in the field of criminal justice such guidelines do not (yet) appear to exist. Furthermore, in health economic literature, cost-effectiveness or cost-utility analyses dominate (Soeteman & Busschbach, 2008). In the field of crime prevention and treatment, these analyses are limited. Nevertheless, McCollistar and colleagues (2003a; 2003b; 2004) and French and colleagues (2008) conducted various cost-effectiveness analyses related to substance abuse treatment, where the effectiveness is for example measured as days of re-incarceration (McCollister, French, Inciardi, et al., 2003; McCollister, French, Prendergast, et al., 2003; McCollister et al., 2004) or as a delinquency score (French et al., 2008). These studies show clearly the use of state of the art methods developed in the field of health care, applied in the field of crime prevention and treatment. On the other hand, these cost-effectiveness analyses were relatively conventional as parameter uncertainty was not captured in the model and long-term estimates were not taken into account. A common way to assess the cost-effectiveness in health care is the so-called decision analytic model (Briggs, Claxton, & Sculpher, 2006; Drummond, Sculpher, Torrance, O'Brien, & Stoddart, 2005). This approach provides a mathematical structure, synthesizing the evidence on costs and effects in a treated population under a variety of treatment options and makes the uncertainty around estimates visible. An additional advantage of this decision analytic modelling approach is that long-term effects can be modelled, even beyond the duration of the trial. Decision-analytic modelling and in particular inclusion of long-term effects may be especially relevant for interventions aiming to reduce criminal behavior. Several authors suggested that criminal behavior during adulthood tends to be preceded by behavioral disorders during childhood. Berger and Boendermaker (2003) stated that serious offenders often have a history of problematic behavior in their early years of life. Kim-Cohen and colleagues (2003) mentioned that most mental disorders in adults "...should be reframed as extensions of juvenile disorders". This suggests that systemic interventions for juvenile disorders may reduce future criminal activity later on in life. Estimates of long-term effects are therefore essential to the analysis of these interventions.

The current study aims to build a probabilistic decision analytic model like common models in health care for assessing interventions primarily aimed at crime prevention and treatment in youth care. In developing the model the following requirements had to be met:

i. The model should be applicable to assess costs and effects of systemic interventions primarily aimed in reducing delinquent behavior;
ii. The initial model should be fairly simple however easy to adjust to sophisticated details (i.e. severity of delinquency);
iii. The model should be probabilistic, taking uncertainty into account;
iv. The model should be suitable for long-term analysis;

As an illustration an initial assessment of the cost-effectiveness of Functional Family Therapy (FFT) compared to treatment as usual (TAU) is presented. As the aim of the study is the application of the probabilistic decision analytic modelling to interventions aimed at reducing delinquency, the interventions compared could be substituted by other systemic interventions mentioned.

The article is structured as follows. The methods section provides information on the health economic model type and general characteristics of the model. The results section elaborates on the applicability of the decision analytic model and outcome measure to the field of systemic interventions specifying necessary adaptations to the health economic approach based on an initial assessment of cost-effectiveness of FFT. The conclusion relates our findings to the general objective of applying health economic methods to systemic interventions not primarily aimed at improving health.

Methods

Model structure

We constructed a probabilistic Markov cohort model (Briggs et al., 2006). Disease progression in common Markov models is described using transitions between 'states', where a subject can move between states or remain in the current state. The transition rates between states are typically estimated based on short run data. Long-term predictions are made based on repetition of transition cycles and assumptions based on for example literature.

In order to keep the initial model as transparent as possible, a Markov model was constructed consisting of three states, i.e. A - criminal behavior, B - non criminal behavior and C - dead. The model structure is shown in Figure 1. All subjects in our study started in state A, moved to either state B or C or remained in state A and could then move between criminal and non-criminal states. Death acted as the absorbing state. Note that subjects could also remain in their present state (depicted by the u-turns).

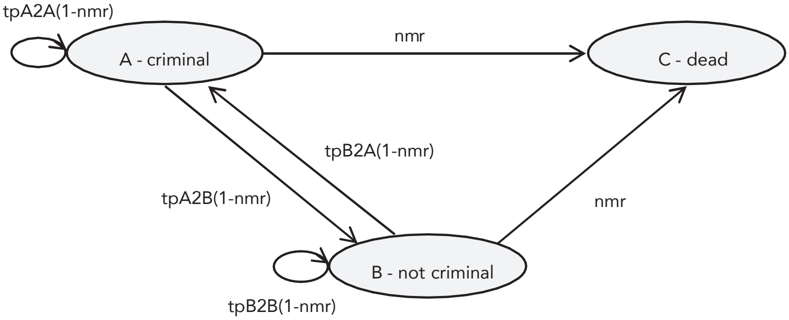


Figure 1. Markov model
nmr = natural mortality rate
tpA2A = transition probability of staying in state A
tpA2B = transition probability of moving from state A to state B
tpB2A = transition probability of moving from state B to state A
tpB2B = transition probability of staying in state B

Outcome measures and model parameters

In order to apply health economic methods meaningfully in the field of crime prevention and treatment, we introduce a new and neutral outcome measure of cost-effectiveness modified for this particular type of intervention: criminal activity free years (CAFYs). The CAFY was defined as a measure of time spent in a dichotomous criminal or non-criminal state. When extensive data is available, criminal activity can e.g. be defined as having had police contacts or committed crimes in the past half year. For the purpose of demonstrating the model functioning and in the absence of extensive clinical data, the criminal state in this study was based on adolescent recidivism derived from clinical trial findings reported by Sexton and Alexander (2000). Transition probabilities differed according to the treatments offered. Treatment costs also differed per treatment type whereas all other costs (Table 1) in the different states were assumed to be independent of the treatment arm but dependent on the state. The cycle length used in the model was six months. This corresponds to the period common for follow up intervals in clinical trials in the field of crime prevention (Glisson et al., 2010; Henderson et al., 2009; Hogue et al., 2008).

Table 1. Included types of costs

Cost categories	Direct	Indirect
Health care	Medical and mental health care child _(psychologist, psychiatrist, GP, specialist, ER, hospital (day) care, medication, youth welfare agency (bureau jeugdzorg)*, foster home*, residential institution, centre for addiction treatment, social worker)_ Medical and mental health care parent _(psychologist, psychiatrist, GP, specialist, foster care*, centre for addiction treatment, social worker)_	
Outside health care	Travel expenses (incl.parking) Time spent by child on exercises as part of therapy* Time spent by parent on exercises as part of therapy*	Productivity losses parent _(absence from work, inefficiency at work)_ Informal care/ support child _(community centre/ church/ moskee/ association, care/support by family or acquaintances)_ Criminal justice system child _(Council of child protection, Bureau Halt*, Police, Lawyer, Court, Incarceration costs)_ Informal care/ support parent _(community centre/ church/ moskee/ association)_

* Included until age 30

In the developed model two treatment alternatives were compared. To provide an example of a cost-effectiveness analysis of systemic interventions, a group receiving FFT therapy and a comparison group receiving TAU were evaluated. TAU refers to a comparable treatment, which delinquent youth would have received if they had not received FFT. As institutions offer diverse types of alternative therapies to FFT, TAU may differ between the different institutions. In one institution TAU may be MST, while another institution may offer Cognitive Behavioral Therapy (CBT) as an alternative to FFT. In our illustration subjects could not switch between FFT and TAU.

For an extensive comparison between two systemic interventions, the model should include several types of cost categories. Table 1 depicts the common cost categories in health economic evaluations; direct and indirect costs inside and outside the health care system adapted to the field of crime. The included types of costs are derived from a combination of the costs commonly included in health economic evaluations and literature on cost of crime (Cohen, 2005). These costs not only pertain to costs incurred by the delinquent juvenile, e.g. costs due to criminal activities or treatment, but also to costs falling on family, caregivers and the society as a whole. For reasons of comparability with other interventions in health care, the model included all relevant societal costs in accordance with the Dutch manual for costing in economic evaluations (Hakkaart, Tan, & Bouwmans, 2010).

Discount rates for future costs and effects were set consistent with guidelines for economic evaluations in the Netherlands (The Health Care Insurance Board, 2006). (Note that differential discounting is required in the Netherlands to account for the growth in the value of health over time. See for example Brouwer and colleagues (2005) for the rationale behind this. Therefore, by using these rates it was implicitly assumed here, that the value of a criminal activity free year (CAFY) will also increase over time, comparable to the rate of a QALY.)

Data analytic procedures: Cost-effectiveness and scenario analyses

In effect studies, uncertainty is generally represented as a confidence interval, i.e. the magnitude of uncertainty is expressed in standard deviations of the measurement error. This assumes that all relevant uncertainty is measurable in a single outcome measure, and that the distribution of the measurement error is reasonably normal. As both assumptions do not apply in typical health economic evaluations, normal t-tests and other parametric statistics are not particularly useful in health economic modelling. Instead, probabilistic analysis was conducted to take the uncertainty of the model parameters into account. In this analysis uncertainty was simulated by running the Markov model several times using a large cohort of subjects, each time with slightly different parameter values. These values were obtained by randomly sampling from each of the parameter distributions, i.e. gamma distributions for costs, and Dirichlet distributions for transition parameters (Briggs et al., 2006). One thousand Monte Carlo simulations were performed. In each simulation a random draw from the parameter distributions was taken, which creates a unique set of cost and effect parameters. The expected costs and effects were then calculated and could be plotted on a cost-effectiveness plane. Four additional scenarios were run to demonstrate model behavior under different assumptions. As the transition probabilities constitute important model parameters, a scenario was created in which probabilities for both interventions were equal. Subsequently, the intervention costs are important parameters, since systemic interventions are concerned to be relatively costly (Aos et al., 2004). From a societal perspective, family costs are assumed to be important, therefore it was investigated how exclusion of these costs would influence the results in the third scenario.

Results

The resulting health economic model for systemic interventions showed that modelling an intervention with a primary aim of decreasing delinquency was feasible. Based on the illustrative comparison of FFT versus TAU, costs and effects could be expressed in costs per CAFY. This section elaborates on the specific characteristics of the resulting decision analytic model. Obviously, the combination using different sources for the inputs of a model is certainly not without problems, but we stress that the emphasis here was on building an illustrative model and demonstrating the model functioning.

Model structure

Estimates of long-term effects were essential to the analysis and were taken into account in the current model. This required some (informed) assumption regarding the endurance of effects of treatment also taking into account the influence that reaching a certain age or experiencing certain life events may have on criminal behavior (Farrington, 2003). For the current model, information on these parameters was taken from the literature. Moffitt (1993) roughly suggested that after adolescence or at approximately age 30 subjects who are criminal during their entire life, life-course-persistent offenders, will remain criminal and subjects who only show criminal behavior during their adolescence, so-called adolescence-limited offenders, will have returned to non criminal behavior. This implies a stable state of criminal activity among individuals of age 30 and older. To illustrate the option of incorporating earlier theory and evidence on the development of offending and antisocial behavior we integrated parts of the long-term stabilizing effects described by Moffitt (1993) into the current model framework. This effect is implemented in the model by extending the effectiveness of the treatment till the age of 30 years. Consequently youth remain in their current state after that age. Thus after reaching the age of 30, youth reach a stable state in their criminal behavior, which means the transition probabilities in the model are from then on defined by mortality rates only. The time horizon of the model is 50 years.

To illustrate how long-term effects may influence model results, Figure 2 and Figure 3 present the percentage of youth in each model state over the time horizon of the model, for FFT and TAU respectively. Figure 2 and Figure 3 demonstrate that a stable state is already reached after about 1 year, which implies that the actual impact of the incorporation of a stabilizing effect, based on the theory of Moffit (1993) is minor in this model. However, as the model results are only as good as the available input used to fill the model, the current results only illustrate how long-term effects could be included in the present model, as it is mainly based on assumptions made and empirical data are lacking.

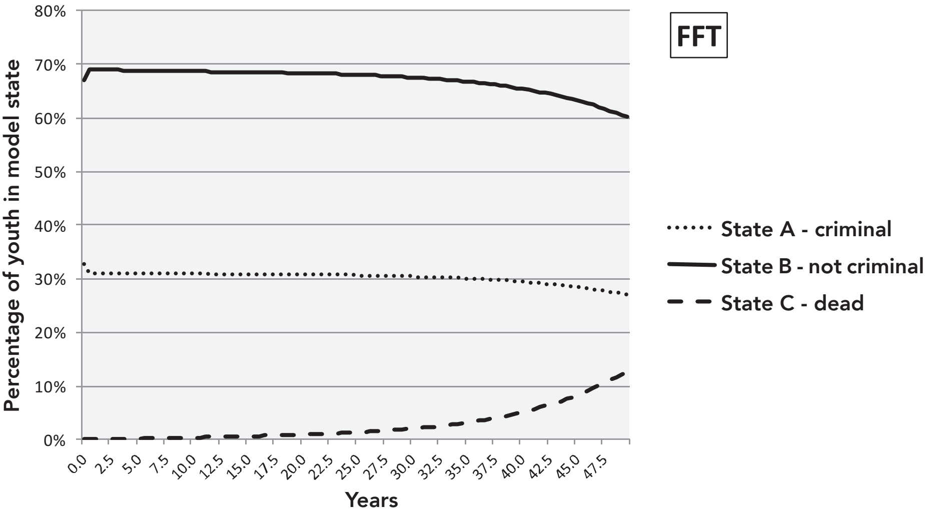


Figure 2. Percentage of youth in model states over time for FFT

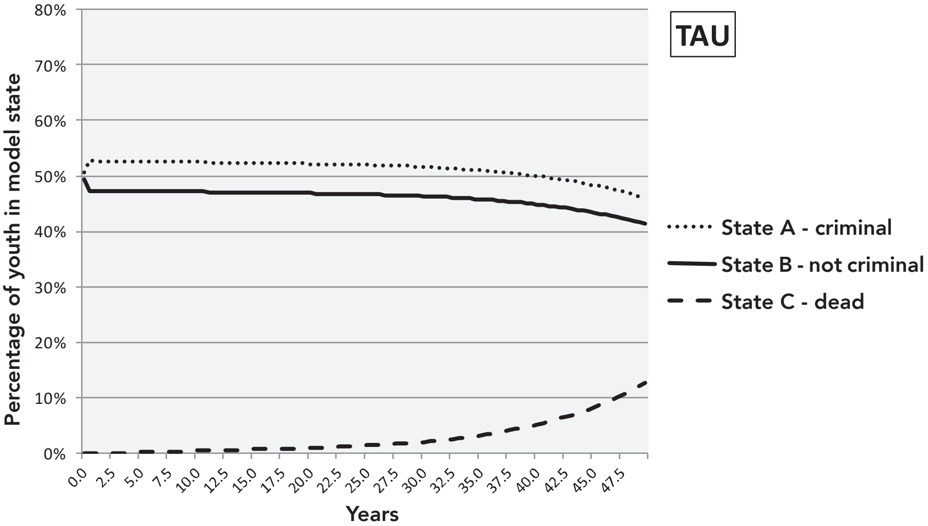


Figure 3. Percentage of youth in model states over time for TAU

Outcome measure: CAFY

In health economic evaluations, cost-effectiveness is most commonly estimated in cost per quality-adjusted life year (QALY). However, as the predominant effect of behavioral interventions for criminal youths is the reduction of criminal activity (Erkenningscommissie Gedragsinterventies Jeugd, 2011) and thus is not directly or exclusively linked to physical health and absence of disease, the effect measure QALY seems inadequate to capture the full benefit of interventions in adolescent mental health (Sindelar, Jofre-Bonet, French, & McLellan, 2004). Therefore, a different outcome measure that sufficiently captures the goals of crime prevention and treatment was required. Considering the societal perspective of the policymaker, a broad outcome measure, directly linked to the goal of a reduction in criminal activity, was chosen. As a first step in this context, we chose the outcome measure of criminal activity free years (CAFYs), which can be used to determine the (incremental) costs per CAFY, i.e. the costs per criminal activity free year. Using CAFYs as the effect measure enables decisions based on a non-monetary value that is comparable between interventions and that properly reflects the goals of the Ministry of Justice while fitting into the health economic modelling approach. Existing examples of an effectiveness measure that resembles the use of the CAFY measure, is the use of days re-incarcerated (McCollister et al., 2004).

As the model has two states defined as either being criminal or not being criminal, the transition from state A, criminal, to state B, not criminal, represents the rate of not being criminal after treatment. The transition of state B to state A on the other hand represents the rate becoming criminal after having been not criminal. It is assumed all youth enter the model as being criminal. The outcome of (incremental) costs per CAFY, was (as a first and rather simplified step) obtained by assigning different costs to individuals according to their current state, criminal state A or non criminal state B. Determining the net present value of the additional costs incurred in state A and state B over the full lifespan of subjects and dividing these by the amount of additional years the individual spends in the non criminal state B during his entire life (compared to TAU) yielded an estimate of incremental costs per CAFY. This process of calculating life-time costs and dividing these by life-time criminal-activity-free years was repeated 1000 times by means of simulation in order to reflect variability in input parameters.

Model parameters: Transition probabilities

Transition probabilities were dependent on the definition of the states reflecting the choice of outcome measure. In the current model, criminal behavior was chosen as most relevant outcome measure so that the states were defined as ‘criminal’ and ‘non criminal’ and transition probabilities between the states could be retrieved from literature.

Several studies showed the effectiveness of FFT compared to TAU (French et al., 2008; Gordon, Graves, & Arbuthnot, 1995; Sexton & Turner, 2010; Sexton & Alexander, 2000, 2002). Yet, there is no consistent outcome regarding the effectiveness of FFT in comparison to TAU. The results based on adolescent recidivism derived from clinical trial findings reported by Sexton and Alexander (2000) were most applicable and comparable to the formulation of our model parameters and definition of the comparison group.

So demonstrating the model, we used the effectiveness rates of that study (Sexton & Alexander, 2000). As the rate of recidivism based on the clinical trial reported in the study of Sexton and Alexander is 33 percent (Sexton & Alexander, 2000), we assumed this rate could be equal to the transition from state B to state A in the model and is therefore supposed to be equal to 33 percent. As the sum of all transition probabilities related to one state in the model sums up to 100 percent, the transition rate of state B to state B (individuals remaining in the non criminal state) is set at 67% (100% minus 33%). As for illustrative purposes we assumed here that the probability of individuals staying non-criminal (B to B) to be equal to the probability of becoming non-criminal (A to B),the transition from state A to state B, was fixed at 67 percent as well. Again subtracting this transition rate from 100% resulted in a probability of 33% for individuals remaining in the criminal state (A to A). Sexton and Alexander (2000) furthermore suggested that "FFT reduces recidivism and/or the onset of offending between 25 and 60 percent more effectively than other programs". As TAU refers to a comparable treatment, we took the average of this range as a reasonable and illustrative estimate of the effectiveness of TAU. The model therefore was constructed under the illustrative assumption that FFT reduces criminal activity 42.5 percent more effectively than TAU.

Transition probabilities were assumed to be fixed over the years, as no further long term effectiveness is known yet.

Model parameters: Costs

To fill in the cost parameters in the model the costs in the criminal state were retrieved from an ongoing trial of FFT (ZonMw, 2008). The volumes of costs in the non criminal state were derived from scaling volumes in the criminal state with a ratio of cost volumes of anti-social versus "normal" youths presented in a UK study on the financial costs of anti-social youths (Scott, Knapp, Henderson, & Maughan, 2001). Unit prices were taken from the Dutch manual for costing in economic evaluations (Hakkaart et al., 2010). In absence of Dutch unit costs, mean treatment costs of the interventions compared were derived from American costs presented in the study of Aos and colleagues (2004). These costs are not related to the states but depend on the intervention a youth received.

Cost-effectiveness

As the comparison of FFT with TAU in the current model is illustrative, the model results solely fulfil this objective. These illustrative cost-effectiveness results from the model point towards lower costs of FFT when compared to TAU. Taking the mean from the stochastic results, the number of CAFYs for FFT exceeds the number of CAFYs for TAU by 6.88 and the costs of FFT appear lower than TAU with incremental cost savings of 8,577EUR (Table 2), positioning the intervention in the South East quadrant of the cost-effectiveness plane (Figure 4). Incremental cost-effectiveness from the illustrative model data expressed in costs per CAFY amounts to cost savings of 1,246 EUR/CAFY. These exemplifying results suggest that FFT produces better effects at lower cost when compared to TAU.

Table 2. Scenario analyses

	CAFY's gained	Cost savings
Base case	6.88	8,577
Scenario 1: transition rate FFT=TAU	-0.02	-718
Scenario 2: TC FFT = TC TAU	6.85	9,112
Scenario 3: excl. family costs	6.88	6,307

FFT = Functional Family Therapy
TAU = Treatment As Usual
TC = Treatment Costs

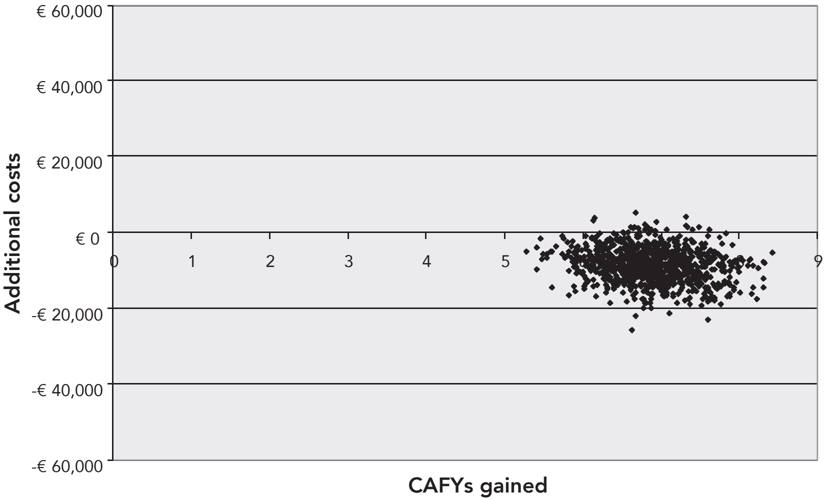


Figure 4. Cost-effectiveness results - Base case analysis

Scenario analyses

Scenario analysis can reveal how the results change if certain parameters are changed. The scenario analysis indicates that the model is particularly sensitive to changes in transition rates whereas the results appear rather robust to changes in other input parameters (Table 2). When transition rates of TAU and FFT are assumed equal (Table 2, Scenario 1), cost savings and CAFY gains entirely vanish. Simulation then results, on average, in an incremental effect of zero and negligible differences in costs between the interventions. The results of the model thus appear to strongly depend on accurate estimates of transition probabilities. Variation in intervention costs does not yield significant differences in costs or effects (Table 2, Scenario 2), whereas exclusion of family costs not only results in a decrease in cost savings but also decreases the variance of the incremental costs (Table 2, Scenario 3).

Discussion and Conclusions

This study created a framework for the evaluation of interventions aimed at reducing criminal activity in delinquent youth. A probabilistic Markov model approach was constructed allowing the assessment of the incremental cost-effectiveness of two systemic interventions. For illustrative purposes, the interventions considered were FFT and TAU. As the comparison of FFT with TAU in the current model is solely an example to demonstrate model functioning, the model results are illustrative in absence of empirical data. As a first step to come to suitable outcome measures in this field, we introduced the outcome measure of Criminal Activity Free Years (CAFY) in a probabilistic decision analytic model. The presented methodology may provide a basis for further development of the model and outcome measures and, ultimately, decision-making by both Ministries of Justice and, in particular, Health. Policymakers may compare cost and effects between different types of interventions aiming to reduce delinquency among youth.

An advantage of using decision analytic models is that this approach enables calculation of hypothetical scenarios. Hence, questions of policymakers, for example on differences in cost-effectiveness within subgroups of youth or on the optimal age for intervention may be answered. Moreover, the decision uncertainty is represented in the model results by taking into account the uncertainty surrounding the input parameters of the model. The current study showed that it was feasible to apply health economic methodology to assess interventions aimed at reducing delinquency rates. The approach was developed to be consistent with health economic guidelines. To our knowledge, this was the first economic evaluation using decision-analytic modelling in the evaluation of systemic interventions for crime prevention and treatment.

However, a number of important questions remain. First of all, the outcome measure presented here is clearly sector-specific. While this enables choosing between interventions with similar aims, it does not directly allow comparisons with other interventions. This problem is not unique for this context. For instance, interventions in elderly care or social care may not be primarily aimed at producing health as well. Outcome measures such as the OPUS and ICECAP have been proposed as better capturing the benefits of such care (Coast et al., 2008; Ryan, Netten, Skatun, & Smith, 2006). This does raise the question, however, of how to trade-off between interventions when their aim is not similar and when different outcome measures were used to assess cost-effectiveness. This seems to be an important area for future research.

Secondly, we proposed the measure of CAFY as a first step to demonstrate how interventions aimed to reduce delinquency could be evaluated within a probabilistic decision model. If such interventions were to be evaluated more systematically using methodology like the one presented here, clearly, the outcome measure deserves more attention. The outcome measure of the CAFY is a very simple and crude one. One could compare it to 'natural units' used in cost-effectiveness analysis like gained life years and event free life years. An important problem with these measures and the CAFY is that they do not reflect the seriousness of the events (e.g. living in a poor or good health state or, in this case, engaging in many and severe criminal activities or a few minor felonies). However, the definition of criminal activity free could be based on different measures, like the number of police contacts or youth self-report of committed crimes. Since not all committed crime, irrespective of the seriousness of the crime, is reported to the police, the difference in definition could give different effectiveness and cost-effectiveness results. Preference weighted measures (like the QALY) would be preferred in this context. Such measures could add a weight to different types of criminal activities and be more comprehensive in terms of the benefits they include (which could even entail a mix of health and crime-related outcomes).

Reducing delinquent behavior is an important outcome of systemic interventions, but multiple other outcomes may be relevant as well, among which for example the ability to live at home after treatment, school attendance or family functioning (Henggeler, 1999; Sindelar et al., 2004). As these multiple outcomes are not considered in the current model, it could be valuable to extend the model or broaden the outcome measure.

Before further use, the model would require improvement, since our analysis had a number of limitations. First, the model was limited to three states. Although a model is always a simplification of reality, and the current model even was an illustration, it should be investigated whether three states are sufficient to provide reasonable estimations of reality. Secondly, the states used now were dichotomous (criminal or non-criminal behavior). The severity of criminal offenses is likely to be important as well, also as a predictor of future criminal activity (Farrington, 2003). The frequency or the types of crime could be an important differentiating factor to discriminate more detailed states (Farrington, 2003). Using more differentiated states would therefore add validity to the model. However, a necessary condition for the formulation of a more complex model is the availability of more and detailed trial data. Third, an individual's history of offenses could be used to predict future behavior and, thus, it may be useful to relax the 'memoryless' feature of the Markov model (Briggs et al., 2006). This feature encompasses that once a subject has moved from one state to another, the Markov model will have 'no memory' regarding which state the subject has come from or the timing of that transition. Using the history of earlier offences in the model could also improve the resulting estimates. The incorporation of long-term effects in the model was based on the coarse assumption individuals reach a stable state of criminal behavior after an age of 30 (Moffitt, 1993). However, the impact of using this theory in the current model was minor. In future research one could consider incorporating other relevant theories like the one used here (Moffitt, 1993) to improve long-term effect modelling. Various other theories and studies about the development of offending and antisocial behavior exist (Farrington, 2003), that could be used to incorporate long-term effects into the model. For example, Sampson and Laub (1993) suggest that offending depends on the strength of bonding to society, like bonding to family, peers, school and social institutions (Moffitt, 1993). In addition, an early age of onset predicts a relatively long criminal career (Farrington, 2003; Loeber & Farrington, 2000) and several risk factors for the early onset of offending are acknowledged (Farrington, 2003). Besides using studies like those mentioned, a stabilizing effect could be modelled more smoothly over time or could be based on empirical, long-term follow-up data to add more detail to modelling long-term effects. Furthermore, Value of Information (VoI) analyses should explore the additional value of further research to characterize the uncertainty of the model inputs, including long-term effects (Briggs et al., 2006). Fourth, the cost parameters in the model are depicted from a combination of costs used in health economic evaluations and literature on cost of crime. However, victim costs and intangible costs, which include direct economic losses of the victims and indirect losses suffered by these victims, respectively, are not taken into account (McCollister, French, & Fang, 2010). Addition of these costs could be of value. Finally, model parameters were solely based on the limited evidence base of available literature and where retrieved out of different literature sources. Ideally, these parameters would be retrieved from more comprehensive empirical data. For example, the transition probabilities could be linked to the presence or absence of police contacts, contacts with judicial institutions or committed crimes. Availability of additional data can refine the input data of the model and increase the validity of the model structure and the accuracy of the results.

Concluding, we used the methods commonly employed in health economic evaluations to create a framework for determining the value for money of interventions targeted at reducing youth delinquency. The results are encouraging, but important further steps still need to be taken. A first next step may be the collection of empirical data to test the presented methodology. We further suggest the construction of a multidimensional outcome measure that enables researchers to capture the multiple dimensions of the treatment goals, in a preference-weighted manner. A final matter that deserves attention is the value we assign to outcomes such as reduced delinquency. Calculating cost-effectiveness is especially useful when the results can be judged against some 'threshold' value. What this should be in this context remains unclear as yet.

References

Aos, S., Lieb, R., Mayfield, J., Miller, M., & Pennucci, A. (2004). *Benefits and costs of prevention and early intervention programs for youth*. Olympia: Washington State Institute for Public Policy.

Berger, M., & Boendermaker, L. (2003). *Multisysteembehandeling in Nederland - Voorstel voor de introductie van MST [Multisystemic treatment in the Netherlands - Proposal for the introduction of MST]*. Utrecht, the Netherlands: Nederlands Instituut voor Zorg en Welzijn /NIZW Jeugd.

Briggs, A., Claxton, K., & Sculpher, M. (2006). *Decision Modelling for Health Economic Evaluation*. New York: Oxford University Press.

Brouwer, W. B., Niessen, L. W., Postma, M. J., & Rutten, F. F. (2005). Need for differential discounting of costs and health effects in cost effectiveness analyses. *BMJ, 331*, 446-448.

Coast, J., Flynn, T. N., Natarajan, L., Sproston, K., Lewis, J., Louviere, J. J., & Peters, T. J. (2008). Valuing the ICECAP capability index for older people. *Social science and Medicine, 67*, 874-882.

Cohen, M. A. (2005). *The costs of crime and justice*. New York: Routledge-Taylor Francis Group.

Drummond, M. F., Sculpher, M. J., Torrance, G. W., O'Brien, B. J., & Stoddart, G. L. (2005). *Methods for the economic evaluation of health care programmes (Third edition)*. Oxford, USA: Oxford University Press.

Erkenningscommissie Gedragsinterventies Jeugd (2011). Committee of approval in youth care. Retrieved from http://www.rijksoverheid.nl/onderwerpen/recidive/erkenningscommissie-gedragsinterventies.

Farrington, D. P. (2003). Developmental and life-course criminology: Key theoretical and empirical issues - The 2002 Sutherland Award Address. *Criminology, 41*, 221-255.

French, M. T., McCollister, K. E., Cacciola, J., Durell, J., & Stephens, R. L. (2002a). Benefit-cost analysis of addiction treatment in Arkansas: specialty and standard residential programs for pregnant and parenting women. *Substance Abuse, 23*, 31-51.

French, M. T., McCollister, K. E., Sacks, S., McKendrick, K., & De Leon, G. (2002b). Benefit-cost analysis of a modified therapeutic community for mentally ill chemical abusers. *Evaluation and Program Planning, 25*, 137-148.

French, M. T., Zavala, S. K., McCollister, K. E., Waldron, H. B., Turner, C. W., & Ozechowski, T. J. (2008). Cost-effectiveness analysis of four interventions for adolescents with a substance use disorder. *Journal of Substance Abuse Treatment, 34*, 272-281.

Glisson, C., Schoenwald, S. K., Hemmelgarn, A., Green, P., Dukes, D., Armstrong, K. S., & Chapman, J. E. (2010). Randomized trial of MST and ARC in a two-level evidence-based treatment implementation strategy. *Journal of Consulting and Clinical Psychology, 78*, 537-550.

Gordon, D. A., Graves, K., & Arbuthnot, J. (1995). The effect of functional family-therapy for delinquents on adult criminal behavior. *Criminal Justice and Behavior, 22*, 60-73.

Hakkaart, L., Tan, S. S., & Bouwmans, C. A. M. (2010). *Dutch costing manual for health care*. Amstelveen, the Netherlands: The Healthcare Insurance Board (CVZ).

Henderson, C. E., Rowe, C. L., Dakof, G. A., Hawes, S. W., & Liddle, H. A. (2009). Parenting practices as mediators of treatment effects in an early-intervention trial of multidimensional family therapy. *The American Journal of Drug and Alcohol Abuse, 35*, 220-226.

Hendriks, V., van der Schee, E., & Blanken, P. (2011). Treatment of adolescents with a cannabis use disorder: Main findings of a randomized controlled trial comparing multidimensional family therapy and cognitive behavioral therapy in the Netherlands. *Drug and Alcohol Dependence, 119*, 64-71.

Henggeler, S. W. (1999). Multisystemic therapy: An overview of clinical procedures, outcomes, and policy implications. *Child and Adolescent Mental Health, 4*, 2-10.

Hogue, A., Henderson, C. E., Dauber, S., Barajas, P. C., Fried, A., & Liddle, H. A. (2008). Treatment adherence, competence, and outcome in individual and family therapy for adolescent behavior problems. *Journal of Consulting and Clinical Psychology, 76*, 544-555.

Kim-Cohen, J., Caspi, A., Moffitt, T. E., Harrington, H., Milne, B. J., & Poulton, R. (2003). Prior juvenile diagnoses in adults with mental disorder: developmental follow-back of a prospective-longitudinal cohort. *Archives of General Psychiatry, 60*, 709-717.

King, S., Griffin, S., Hodges, Z., Weatherly, H., Asseburg, C., Richardson, G., . . . Riemsma, R. (2006). A systematic review and economic model of the effectiveness and cost-effectiveness of methylphenidate, dexamfetamine and atomoxetine for the treatment of attention deficit hyperactivity disorder in children and adolescents. *Health Technology Assessment, 10*, 1-146.

Loeber, R., & Farrington, D. P. (2000). Young children who commit crime: Epidemiology, developmental origins, risk factors, early interventions, and policy implications. *Development and Psychopathology, 12*, 737-762.

McCollister, K. E., French, M. T., & Fang, H. (2010). The cost of crime to society: New crime-specific estimates for policy and program evaluation. *Drug and Alcohol Dependence, 108*, 98-109.

McCollister, K. E., French, M. T., Inciardi, J. A., Butzin, C. A., Martin, S. S., & Hooper, R. M. (2003a). Post-release substance abuse treatment for criminal offenders: A cost-effectiveness analysis. *Journal of Quantitative Criminology, 19*, 389-407.

McCollister, K. E., French, M. T., Prendergast, M., Wexler, H., Sacks, S., & Hall, E. (2003b). Is in-prison treatment enough? A cost-effectiveness analysis of prison-based treatment and aftercare services for substance-abusing offenders. *Law and Policy, 25*, 63-82.

McCollister, K. E., French, M. T., Prendergast, M. L., Hall, E., & Sacks, S. (2004). Long-term cost effectiveness of addiction treatment for criminal offenders. *Justice Quarterly, 21*, 659-679.

Ministry of Justice (2008). *Afstemming van gedragsinterventies voor jeugdige delinquenten Programma Aanpak Jeugdcriminaliteit [Adjustment of behavioural interventions for youth delinquents: Program of procedure youth delinquency]*. The Hague, the Netherlands: Ministry of Justice.

Moffitt, T. E. (1993). Adolescence-limited and life-course-persistent antisocial behavior: a developmental taxonomy. *Psychological Review, 100*, 674-701.

Ryan, M., Netten, A., Skatun, D., & Smith, P. (2006). Using discrete choice experiments to estimate a preference-based measure of outcome--an application to social care for older people. *Journal of Health Economics, 25*, 927-944.

Sampson, R. J., & Laub, J. H. (1993). *Crime in the making: Pathways and turning points through life*. Cambridge, MA: Harvard University Press.

Scott, S., Knapp, M., Henderson, J., & Maughan, B. (2001). Financial cost of social exclusion: follow up study of antisocial children into adulthood. *BMJ, 323*, 191-194.

Sexton, T., & Turner, C. W. (2010). The effectiveness of functional family therapy for youth with behavioral problems in a community practice setting. *Journal of Family Psychology, 24*, 339-348.

Sexton, T. L., & Alexander, J. F. (2000). Functional Family Therapy. *Office of Juvenile Justice and Delinquency Prevention, Juvenile Justice Bulletin, 1*, 1-7.

Sexton, T. L., & Alexander, J. F. (2002). Functional family therapy for at-risk adolescents and their families (chapter 6). In F. W. Kaslow & T. Patterson (Eds.), *Comprehensive handbook of psychotherapy (Volume 2)* (pp. 117-140). New York: John Wiley & Sons, Inc..

Sindelar, J. L., Jofre-Bonet, M., French, M. T., & McLellan, A. T. (2004). Cost-effectiveness analysis of addiction treatment: Paradoxes of multiple outcomes. *Drug and Alcohol Dependence, 73*, 41-50.

Soeteman, D. I., & Busschbach, J. J. V. (2008). Cost-benefit and cost-effectiveness of prevention and treatment. In R. Loeber, W. Slot, P. van der Laan & M. Hoeve (Eds.), *Tomorrow's Criminals: The Development of Child Delinquency and Effective Interventions* (pp. 215-226). Surrey: Ashgate.

Teuffel, O., Kuster, S. P., Hunger, S. P., Conter, V., Hitzler, J., Ethier, M. C., . . . Sung, L. (2011). Dexamethasone versus prednisone for induction therapy in childhood acute lymphoblastic leukemia: A systematic review and meta-analysis. *Leukemia, 25*, 1232-1238.

The Health Care Insurance Board (2006). *Guidelines for Pharmacoeconomic Research*. Amstelveen, the Netherlands: The Healthcare Insurance Board (CVZ).

Ttofi, M. M., Farrington, D. P., Losel, F., & Loeber, R. (2011). The predictive efficiency of school bullying versus later offending: A systematic/meta-analytic review of longitudinal studies. *Criminal Behaviour and Mental Health, 21*, 80-89.

Welsh, B. C., Loeber, R., Stevens, B. R., Stouthamer-Loeber, M., Cohen, M. A., & Farrington, D. P. (2008). Cost of juvenile crime in urban areas: A longitudinal perspective. *Youth Violence Juvenile Justice, 6*, 3-27.

ZonMw (2008). *ZonMW programma zorg voor jeugd: weten wat werkt! [ZonMW program for youth]*. The Hague, the Netherlands: ZonMW.

Chapter 3.

Value of information analysis applied to the economic evaluation of interventions aimed at reducing juvenile delinquency: An illustration

Hester V. Eeren, Saskia J. Schawo, Ron H.J. Scholte, Jan J.V. Busschbach, & Leona Hakkaart

PLOS ONE (2015), 10, e0131255

Abstract

Objectives: To investigate whether a value of information analysis, commonly applied in health care evaluations, is feasible and meaningful in the field of crime prevention.

Methods: Interventions aimed at reducing juvenile delinquency are increasingly being evaluated according to their cost-effectiveness. Results of cost-effectiveness models are subject to uncertainty in their cost and effect estimates. Further research can reduce that parameter uncertainty. The value of such further research can be estimated using a value of information analysis, as illustrated in the current study. We built upon an earlier published cost-effectiveness model that demonstrated the comparison of two interventions aimed at reducing juvenile delinquency. Outcomes were presented as costs per criminal activity free year.

Results: At a societal willingness-to-pay of €71,700 per criminal activity free year, further research to eliminate parameter uncertainty was valued at €176 million. Therefore, in this illustrative analysis, the value of information analysis determined that society should be willing to spend a maximum of €176 million in reducing decision uncertainty in the cost-effectiveness of the two interventions. Moreover, the results suggest that reducing uncertainty in some specific model parameters might be more valuable than in others.

Conclusions: Using a value of information framework to assess the value of conducting further research in the field of crime prevention proved to be feasible. The results were meaningful and can be interpreted according to health care evaluation studies. This analysis can be helpful in justifying additional research funds to further inform the reimbursement decision in regard to interventions for juvenile delinquents.

Introduction

In order to guide policy decisions, it would be helpful to know the cost-effectiveness of interventions aimed at reducing juvenile delinquency. So far, cost-effectiveness analyses have informed an increasing number of reimbursement decisions in mental health-care (Evers, Salvador-Carulla, Halsteinli, McDaid, & MHEEN Group, 2007; Knapp et al., 2008). Accordingly, the number of cost-effectiveness analyses in the field of crime prevention is increasing (Barrett & Byford, 2012; Cary, Butler, Baruch, Hickey, & Byford, 2013; French et al., 2008; Knapp et al., 2008; McCollister et al., 2003a; McCollister et al., 2003b; McCollister, French, Prendergast, Hall, & Sacks, 2004; Romeo, Byford, & Knapp, 2005; Soeteman & Busschbach, 2008).

The inputs in a cost-effectiveness analysis can be uncertain, as available information about the costs and effects of interventions is rarely perfect. As a result, the decision whether or not to reimburse an intervention is marked by uncertainty. When a decision to reimburse an intervention turns out to be incorrect, it could lead to suboptimal interventions being approved. These interventions create costs in terms of foregone benefits and resources (Briggs, Claxton, & Sculpher, 2006; Claxton, 2008; Claxton, Neumann, Araki, & Weinstein, 2001; Claxton, Sculpher, & Drummond, 2002; Oostenbrink, Al, Oppe, & Rutten-van Mölken, 2008). Further research may eliminate this uncertainty and optimize the reimbursement decision.

This study aims to estimate the added value of future cost-effectiveness research. This type of analysis is referred to as a 'value of information' analysis and was introduced as part of statistical decision theory (Pratt, Raiffa, & Schlaifer, 1995; Raiffa, 1968). It has already been applied in other research areas, such as engineering and environmental risk analysis (Yokota & Thompson, 2004), before being introduced into health technology assessment (Briggs et al., 2006; Claxton, 1999, 2008; Claxton et al., 2001; Claxton et al., 2002; Oostenbrink et al., 2008), where the application of this analysis is now widely adopted, as well as in the field of mental health care (Mohseninejad, van Baal, van den Berg, Buskens, & Feenstra, 2013; Soeteman, Busschbach, Verheul, Hoomans, & Kim, 2011).

A value of information analysis reveals the value of conducting additional research and identifies the type of research that would be most useful. Its results can inform about further research on specific parameters, and more precisely inform the decision about which intervention should be reimbursed (Myers et al., 2011). Furthermore, a value of information analysis can be used to prioritize future research, for example by highlighting the merits of certain types of research which might add to the reduction of the parameter uncertainty in cost-effectiveness analysis (Carlson et al., 2013; Oostenbrink et al., 2008; Sculpher & Claxton, 2005). The potential value of further research could then be weighed against the costs of conducting this research in order to determine whether it is worthwhile (i.e. Briggs et al., 2006; Claxton, 2008).

Because a value of information analysis has not yet been applied in the field of crime prevention, we will present an example of this analysis based on an existing cost-effectiveness model in crime prevention and treatment (Schawo et al., 2012). We used two interventions aimed at reducing juvenile delinquency in the Netherlands, in adolescents aged 12-18 years. These interventions can be applied to prevent juvenile delinquency or used to prevent juveniles committing crimes in future, for example after an adolescent has been punished under the juvenile criminal laws. Juvenile law in the Netherlands applies to adolescents aged 12-17 years (van der Laan, 2006). Not only the criminal act itself is important, but there is a strong focus on for example the background and moral development of the adolescent (van der Laan, 2006).

As the present study was set up as an illustration, data was used solely to demonstrate the method. We did not aim to test the superiority of one of the interventions that were used to illustrate the method. Therefore, this article merely presents a demonstration of the relevance of a value of information analysis in the field of crime prevention and treatment. The presented input data and results should be interpreted in this context. We will start with a short summary of an earlier illustrative cost-effectiveness analysis (Schawo et al., 2012), and then introduce and illustrate the value of information analysis.

Methods

Interventions

We compared two interventions aimed at reducing juvenile delinquency. The 'Kursushuis' intervention (translated and referred to as the Course House) consists of a domestic foster home where several adolescents live for about 10 months and professional care is at close hand. The treatment costs and effects were described by Slot et al. (Slot, Jagers, & Beumer, 1992). The second intervention is a systemic intervention named Functional Family Therapy (FFT), which lasts about 4 to 6 months. The costs and effects of this intervention were obtained from a multicentre quasi-experimental study in the Netherlands (van der Veldt, Eenshuistra, & Campbell, 2011). The Medical Ethical Committee of the VU University Amsterdam approved this study (number 2008/152).

Cost-effectiveness model

The Markov model that was used for the value of information analysis consists of three mutually exclusive model states: A) criminal behavior, B) no criminal behavior, and C) dead (Schawo et al., 2012) (Figure 1). The time horizon of the model was 20 years, with a cycle length of six months (Schawo et al., 2012). A societal perspective was taken and results were expressed as costs per Criminal Activity Free Year (CAFY) (Schawo et al., 2012).

In line with health economic guidelines (The Health Care Insurance Board, 2006), the input parameters in the model were threefold. The first group of parameters were the transition probabilities. These reflect the probability that an adolescent transitions through the states. The measure of time an adolescent spends in a non-criminal state is used to estimate a CAFY. Criminal activity was based on the adolescents' self-reported contact with the police in connection with he/she having committed one or several crimes; having had no contacts was defined as criminal-activity free and having had one or more contacts as criminally active. Transition probabilities were extrapolated until the age of 30, as we integrated parts of the long-term stabilizing effects described by

Moffitt (Moffitt, 1993; Schawo et al., 2012). Dying because of committing crimes was not reflected in the CAFY. Adolescents were assumed to face a risk of death equivalent to the age specific mortality rates in the general population (Statistics Netherlands, 2015b). The second group consisted of costs of health-care use, productivity losses, and other societal costs such as costs of the criminal justice system. Both costs outside health care, and health care costs were included, such as the costs of visiting a psychiatrist or psychologist. As the family system is involved in the interventions provided, we included both the costs of the adolescent and those of one of the parents. The model state costs were fixed over time until the adolescent was 23 years. It was assumed that from that age onwards not all cost categories (such as a family guardian or foster care) would remain relevant. The third group comprised the intervention costs. The costs of one completed FFT treatment was calculated to be approximately €10,900 per adolescent, whereas the Course House was about €37,800 (retrieved from Slot et al. (Slot et al., 1992)). Both costs were extrapolated to 2013 Euro's accounting for inflation based on the consumer price index (Statistics Netherlands, 2015a). The cost and effects in the model were discounted (i.e. Brouwer, van Hout, & Rutten, 2000; van Hout, 1998), according to the guidelines for economic evaluations in the Netherlands (The Health Care Insurance Board, 2006).

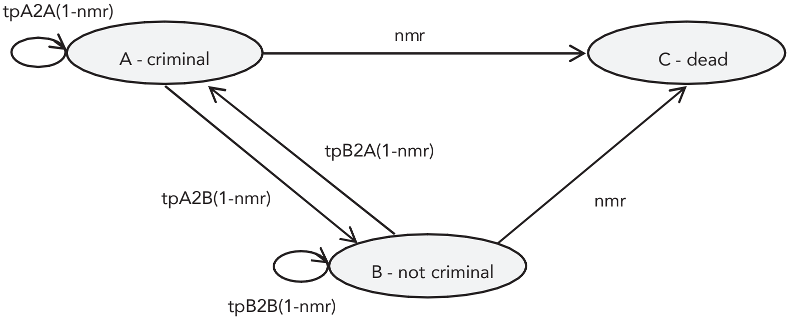


Figure 1. Markov model
nmr = natural mortality rate
tpA2A = transition probability of staying in state A
tpA2B = transition probability of moving from state A to state B
tpB2A = transition probability of moving from state B to state A
tpB2B = transition probability of staying in state B

To represent the uncertainty of each model parameter, we assigned parameter distributions (Table I in Supplemental Material). In a probabilistic analysis, uncertainty was simulated by running the model 10,000 times using a cohort of subjects and each time taking different parameter estimates from the parameter distributions (Briggs et al., 2006; Claxton, 2008). These 10,000 unique sets of parameter values were used to estimate the mean expected cost-effectiveness. For further details on the cost-effectiveness model, we refer to Schawo et al. (Schawo et al., 2012).

Cost-effectiveness analysis

The stochastic model resulted in the relative cost-effectiveness outcomes of the Course House intervention compared with FFT, represented as incremental costs/CAFY (Table 1; Figure 2). It showed that the Course House was more effective than FFT, but also produced higher costs. The cumulative number of CAFYs for the Course House exceeded the number of CAFYs for FFT by 0.7, while the incremental costs of the Course House exceeded those of FFT by €26,800, thereby positioning the intervention in the North East quadrant of the cost-effectiveness plane (Fenwick, Claxton, & Sculpher, 2008) (Figure 2). The incremental cost-effectiveness ratio (ICER) of the Course House compared with FFT was 39,000 €/CAFY.

Table 1. Cost-effectiveness results over 20 years[a]

Intervention	Cost	CAFY	ICER[b]	NMB[c]
Course House	€249,000	12.4	€39,000	€641,200
FFT	€222,200	11.7	-	€618,700

CAFY, criminal activity free year; ICER, incremental cost-effectiveness ratio; NMB, net monetary benefit; FFT, Functional Family Therapy.

[a] The results represented were averaged over the 10,000 simulations run.

[b] The ICER was calculated as the difference in cost divided by the difference in CAFYs between the Course House and FFT.

[c] The NMB was calculated by multiplying CAFYs by the WTPvalue of €71,700 per CAFY and subtracting cost. The Course House is cost-effective compared with FFT, because the NMB of the Course House is higher than the NMB of FFT. Due to decimals, the numbers in the table multiplied do not give the exact NMB values represented in this table.

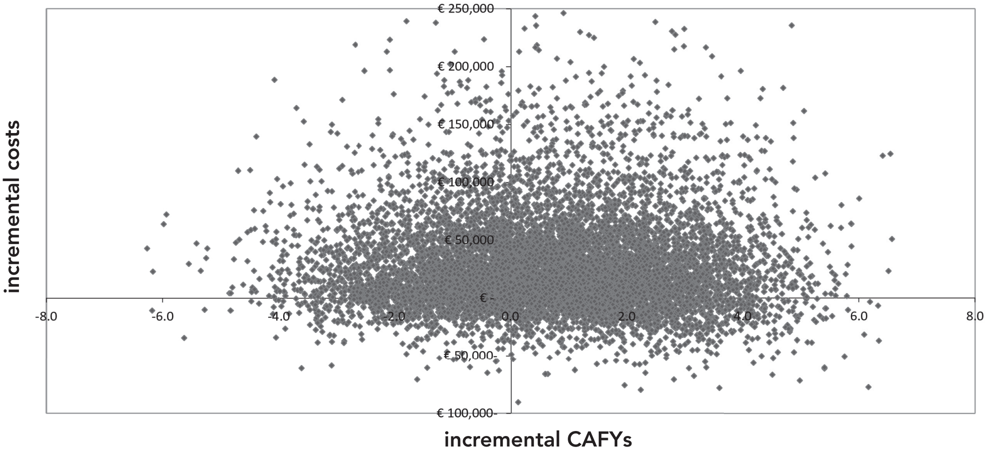


Figure 2. Incremental cost-effectiveness plane for Course House compared with FFT (10,000 simulations)
FFT = Functional Family Therapy
CAFY = Criminal Activity Free Year

Parameter uncertainty

The influence of parameter uncertainty on the model outcomes was shown in the cost-effectiveness acceptability frontier (CEAF). In the CEAF, the probability of being cost-effective compared to the other intervention is shown for the intervention with the highest expected net monetary benefit (NMB) for a range of societal willingness-to-pay (WTP) values per CAFY, and is therefore cost-effective compared with the alternative, given a certain WTP (Barton, Briggs, & Fenwick, 2008; Fenwick, Claxton, & Sculpher, 2001). Here, the overall maximum expected net benefit guides the decision on which intervention is cost-effective compared with the alternative intervention (Barton et al., 2008; Fenwick et al., 2001). The NMB was calculated by multiplying CAFYs by the WTP value per CAFY and subtracting cost (Briggs et al., 2006; Claxton, 2008). The CEAF is illustrated in the results' section.

Value of information analysis

In the value of information analysis, the parameter uncertainty in the model is monetarized. More precisely, we estimated the value of 'knowing everything': the 'expected value of having perfect information' (EVPI) (Claxton, 2008; Claxton et al., 2002). Having perfect information would eliminate parameter uncertainty and optimize the reimbursement decision. In estimating the 'value of knowing everything', the EVPI places an upper boundary on the value of performing further research (Briggs et al., 2006; Claxton, 2008). It can be interpreted as the maximum value society 'should' be willing to pay for additional evidence to reduce decision uncertainty around which intervention is preferred and, therefore, inform the reimbursement decision in the future (Briggs et al., 2006; Claxton, 2008). The EVPI is computed by first taking the difference between the expected NMB with perfect information and the expected NMB with current information per simulation. This difference is equal to the expected benefits foregone when making the decision based on current evidence (Briggs et al., 2006; Claxton, 2008). Comparing the EVPI estimates with the costs of this future research reveals whether further research is worthwhile.

As the value of further information is related to the size of the eligible population of adolescents to be treated, the EVPI was multiplied with the eligible population of adolescents in the population EVPI (pEVPI). About 825 adolescents annually were assumed to be eligible for FFT in the Netherlands. When we discount this number over five years, which is the assumed lifetime of the intervention for which additional research would be useful (Briggs et al., 2006; The Health Care Insurance Board, 2006), it resulted in an eligible population of 3,820 adolescents. We assumed that the eligible number of adolescents for the Course House was equal to that for FFT.

In a value of information analysis one could also focus on specific groups of model parameters. To identify the model parameters that contribute to most of the uncertainty and for which future research is the most promising, we estimated the expected value of partial perfect information (EVPPI) (Briggs et al., 2006; Claxton, 2008). The EVPPI was estimated using the Sheffield Accelerated Value of Information application of Strong et al. (Strong, Oakley, & Brennan, 2014). Multiplying the EVPPI values with the eligible population results in the population EVPPI (pEVPPI).

The EVPI and EVPPI not only depend on the uncertainty of the model parameters, but also on the WTP per CAFY. In the absence of a WTP per CAFY in the Netherlands, we used WTP estimates to reduce crime of Cohen et al. (Cohen & Piquero, 2009; Cohen, Piquero, & Jennings, 2010). These WTP values per crime indicate the value society wants to pay to prevent one crime, for example €32,200 per burglary (Table 2). Table 2 provides an overview of these estimates, adjusted for inflation and purchasing power parities (OECD Library, 2015). Although WTP to prevent one crime is definitely not equal to WTP per CAFY, we used it to illustrate what is meant by WTP in crime prevention and how the concept can be used in a value of information analysis. We hereby implicitly assumed that one crime is committed per year, and thus exactly one crime per year is avoided in a CAFY. We estimated the EVPI and EVPPI for various WTP values, and we chose an average WTP value to illustrate the result in the results section, which was €71,700 (Table 2).

Table 2. Willingness-to-pay values for crimes (Cohen & Piquero, 2009)

Crime	WTP in 2007 dollars	WTP in 2013 euro's
Murder	$140,000	€128,700
Rape	$290,000	€266,600
Armed robbery	$280,000	€257,400
Robbery	$39,000	€35,900
Aggravated assaults	$85,000	€78,100
Simple assaults	$19,000	€17,4500
Burglary	$35,000	€32,200
Moter vehicle theft	$17,000	€15,600
Larceny	$4,000	€3,700
Druk driving crash	$60,000	€55,200
Arson	$115,000	€105,700
Vandalism	$2,000	€1,800
Fraud	$5,500	€5,100
Other offenses	$1,000	€900
Average	**$140,000**	**€71,700**

WTP, willingness-to-pay

The model parameters were grouped into the following ten subsets to indicate the direction of research as a result of the EVPPI analysis: research on 1) transition probabilities for FFT; 2) transition probabilities for the Course House; 3) direct health-care costs of the criminal state; 4) direct health-care costs of the non criminal state; 5) direct non health-care costs related to the criminal state; 6) direct non health-care costs related to the non criminal state; 7) indirect non health-care costs related to the criminal state; 8) indirect non health-care costs related to the non criminal state; 9) intervention costs of FFT; and 10) intervention costs of the Course House.

Results

Model uncertainty

The CEAF shows that FFT had the highest NMB for a WTP ranging from €0 - €39,000 (Figure 3). At a WTP of €39,000, FFT was cost-effective in 49% of the 10,000 model simulations, or a probability of 0.49, whereas the Course House was cost-effective in 51% of the simulations. Above the €39,000 WTP, the Course House had the highest NMB and thus was the optimal intervention. This switching point in the CEAF is where the NMB for FFT is equal to the NMB of the Course House. At this point, the WTP was exactly equal to the ICER value (€39,000 per CAFY).

The CEAF (Figure 3) shows a large error probability. At the WTP of €71,700 the Course House was cost-effective in 57% of the 10,000 model simulations, which suggests that there is an error probability of 0.43 that could be reduced by collecting additional evidence.

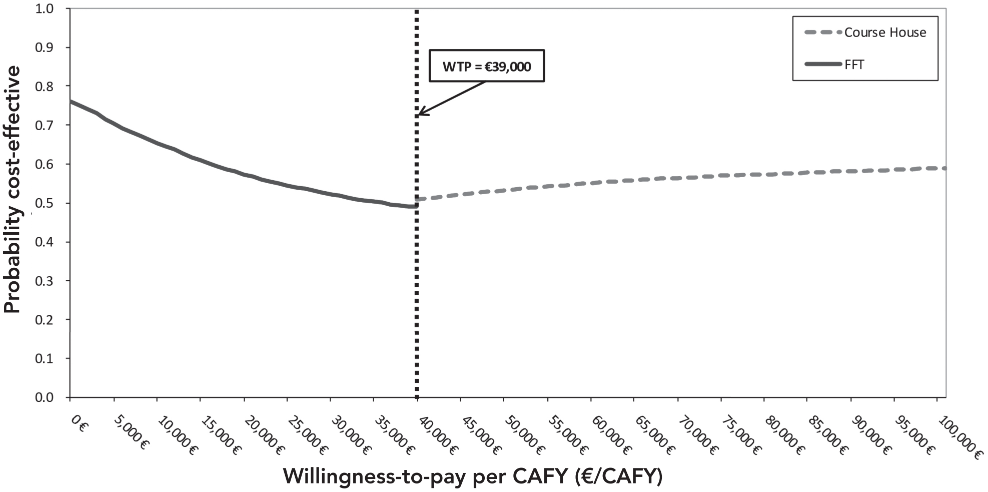


Figure 3. Cost-effectiveness Acceptability Frontier (CEAF)
FFT = Functional Family Therapy
CAFY = Ciminal Activity Free Year
WTP = Willingness-to-pay

Value of information analysis

In order to know the value of reducing the error probability and to assign a value to additional research, we estimated the EVPI. Table 3 illustrates the EVPI estimation (based on Soeteman et al., 2011). The table shows the generated NMB for each intervention for 6 of the 10,000 simulations, given a WTP value of €71,700 per CAFY. The EVPI was determined as follows: First, we assumed that decision makers have perfect information for each simulation instead of making one single choice over all simulations. For example, for simulation 1 and 2, this would result in the choice for FFT (see Table 3). Second, we determined the choice based on current information. In this case, the Course House had the highest expected NMB (€641,200) over all simulations and hence was the preferred intervention. Finally, we took the difference between the decision based on perfect information per simulation and the optimal choice over all simulations. This difference resulted in the EVPI value or the benefits foregone per simulation. The expectation of all benefits foregone over the 10,000 simulations is the EVPI per adolescent, which is €46,000 at a WTP value of €71,700 per CAFY. Perfect information for an individual adolescent was thus valued at €46,000. Multiplying this EVPI value by 3,820 eligible adolescents resulted in a pEVPI of €176 million. This pEVPI value suggests that, at a societal WTP value of €71,700 per CAFY, there is room to reduce the parameter uncertainty in the model by a maximum of €176 million.

Table 3. Calculation of expected value of perfect information (EVPI) for individual adolescent

Simulation		**Net monetary benefits**[a]		**Maximum net benefit**	**Benefits foregone**
		Course House	FFT		
	Expectation	***€641,200***	*€618,700*	*€687,200*	*€46,000*
1		€481,000	**€650,000**	€650,000	€169,000
2		€553,800	**€710,300**	€710,300	€156,500
3		€513,800	**€768,000**	€768,000	€254,200
4		**€717,500**	€562,700	€717,500	€0
5		€516,200	**€671,000**	€671,000	€154,800
…	…	…	…	…	…
10,000		**€602,300**	€587,200	€602,300	€0

[a] Net monetary benefit (NMB) was calculated by multiplying CAFYs by the WTP value of €71,700 per CAFY and subtracting cost.

<u>Explanation:</u>
Decision based on current information: Course House.
Decision based on perfect information: bold.
Expected net benefit with current information: €641,200.
Expected net benefit with perfect information: €687,200.
Expected value of perfect information (EVPI): €687,200 - €641,200 = €46,000.

Perfect information can be valued at different WTP values. The extent of the monetarized uncertainty surrounding the decision for a range of WTP values is represented in the pEVPI curve. Figure 4 presents the pEVPI curve for an eligible population of 3,820 adolescents. As an example we consider the point where research costs society €50 million (i.e. the pEVPI value at the y-axis in Figure 4). At this point further research would potentially be cost-effective if society were willing to pay more than €17,600 per CAFY (i.e. the value at the x-axis, if the pEVPI is €50 million). At lower values of the WTP per CAFY, the benefits of further research cannot offset the costs (Briggs et al., 2006; Fenwick, O'Brien, & Briggs, 2004). At a WTP of €39,000 per CAFY, the pEVPI shows a local maximum of €127 million. At this point, the parameter uncertainty in the model is the highest and thus decision uncertainty is highest, as already shown in the CEAF curve (Figure 3).

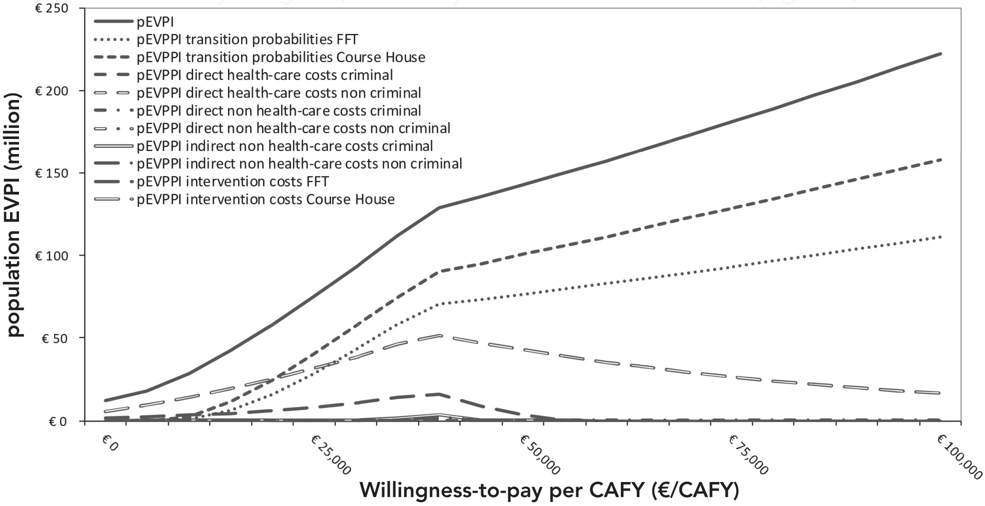


Figure 4: population Expected Value of Perfect Information (pEVPI) and population Expected Value of Partial Perfect Information (pEVPPI) curve
FFT = Functional Family Therapy
CAFY = Criminal Activity Free Year
pEVPI = population Expected Value of Perfect Information
pEVPPI = population Expected Value of Partial Perfect Information

Perfect information of subsets of parameters was valued in the pEVPPI. This pEVPPI was estimated for a range of WTP values (Figure 4). At the illustrative WTP value of €71,700 per CAFY, future research would be most valuable for three subsets of parameters: the transition probabilities and the intervention costs of the Course House and the transition probabilities of FFT (see Figure 4). The pEVPPI of the transition probabilities of the Course House was €125 million (€32,700 per adolescent), and the pEVPPI for the transition probabilities of FFT was €91 million (€23,800 per adolescent). The pEVPPI for the intervention costs of the Course House was €28 million (€7,400 per adolescent). The pEVPPIs for the direct non health-care costs of the criminal state and the non criminal state were respectively €8,400 and €43,300 (respectively €2 and €11per adolescent). The pEVPPIs for the other parameter groups were all estimated to be zero (Figure 4), meaning there was no potential value of further research into these parameters. Given a WTP of €71,700 per CAFY further research for these parameters would not reduce decision uncertainty. The EVPI and EVPPI values depend highly on the WTP value per CAFY, as can be seen in Figure 4 and Table II in Supplemental Material. At a WTP of for example €40,000, there was indeed potential value of further research into all model states costs. Note that due to the interactions within the model structure, the pEVPPI for the groups of parameters do not sum up to the overall pEVPI for the model (see Figure 4) (Briggs et al., 2006; Fenwick et al., 2004).

Discussion

While cost-effectiveness analyses are increasingly being used in the field of crime prevention, the value of further research has not yet been estimated for comparison between interventions aimed at reducing juvenile delinquency. An earlier developed cost-effectiveness model was used to estimate this value of further research. This study demonstrated that it was feasible to estimate the value of conducting further research in this context, using a value of information framework common in health economic evaluations. The results can be interpreted as similar to cost/QALY (Quality-Adjusted Life Year) studies in health care evaluation.

In this value of information analysis, the results indicated the parameters for which further research was valuable. Our findings show particular uncertainty in three groups of parameters: the transition probabilities of the Course House and of FFT, and to a lesser extent, the intervention costs of the Course House and the direct non health-care costs in both model states. Performing additional research in the suggested fields can reduce parameter uncertainty, and hence, can reduce decision uncertainty.

Therefore, the results of a value of information analysis can prioritize further research to optimize the final reimbursement decision, thereby increasing the probability that adolescents will be assigned to the intervention that is cost-effective, compared with the alternative. Given this information, future interventions could be reimbursed (or not), and they could also be approved 'only in research' (OIR) (i.e. further research is required before the intervention can be approved) or 'approved with research' (AWR) (i.e. research can be conducted while the intervention is approved) (Claxton et al., 2012; McKenna & Claxton, 2011). For example, from this study we can conclude that given a WTP of €40,000 per CAFY, the Course House could be 'approved with research'. The Course House would then be reimbursed while further research would be required, for example on the effectiveness of the Course House. Current practice in adolescent care in the Netherlands illustrates this approval condition: the Netherlands Youth Institute identifies effective youth interventions, while still conducting research on the effectiveness of some of these interventions (Netherlands Youth Institute, 2014). However, approval might lead to irrecoverable costs when the approval is revised due to subsequent research revealing that the Course House was not as effective as expected. Then, approval 'only in research' might be preferred, because commitment to future costs is avoided until the results of further research are known. Approval might even be dependent on any change in the effective price of an intervention (Claxton et al., 2012; Walker, Sculpher, Claxton, & Palmer, 2012).

This study was a first attempt to apply a value of information framework to the field of crime prevention and treatment of juvenile delinquents. Therefore, some considerations should be kept in mind. The value of information analysis estimates the monetary value of eliminating all or part of the parameter uncertainty of the presented model. However, two other sources of uncertainty can influence the results: structural and methodological uncertainty (Bojke, Claxton, Sculpher, & Palmer, 2009; Briggs, 2000). Structural uncertainty relates to structural aspects of the model (Bilcke, Beutels, Brisson, & Jit, 2011; Bojke et al., 2009; Haij Ali Afzali & Karnon, 2015), such as the conceptual framework or the transitions between the model states (Haij Ali Afzali &

Karnon, 2015), and it can lead to different estimated model outcomes (i.e. Frederix et al., 2014). This structural uncertainty is likely to be present in our model. For example, we did not account for the severity of crimes in the model states or the elevated risk of death for adolescents in the criminal state (i.e. Chassin, Piquero, Losoya, Mansion, & Schubert, 2013; Teplin, McClelland, Abram, & Mileusnic, 2005). The uncertainty of these aspects was not represented in the current value of information analysis. Future cost-effectiveness models in the field of crime prevention should therefore carefully characterize the structural uncertainty (Haij Ali Afzali & Karnon, 2015), and account for it when possible, for example by parameterization (Bilcke et al., 2011; Haij Ali Afzali & Karnon, 2015) or model averaging (Bojke et al., 2009; Haij Ali Afzali & Karnon, 2015; Jackson, Bojke, Thompson, Claxton, & Sharples, 2011). The second additional source of uncertainty is methodological uncertainty, which relates to the analytical method chosen (Bilcke et al., 2011; Briggs, 2000). Our model also represents some methodological uncertainties, such as whether or not to include the costs of crime in the model (i.e. McCollister, French, & Fang, 2010). Here, three methodological uncertainties in our model are discussed in more detail. These uncertainties could be resolved through, for example, formulating guidelines (i.e. Bilcke et al., 2011; Bojke et al., 2009) to model cost-effectiveness research in the field of crime prevention.

The first methodological uncertainty concerns the societal perspective used in the model, which means that we included the costs and effects relevant to society. When considering this perspective in health care, the focus is merely on the patient, whereas this will be different in the area of crime prevention and treatment (i.e. van Zutphen, Goderie, & Janssen, 2014). In this study, we already included the direct and indirect costs of one parent, as well as direct non medical costs of the adolescent, such as the costs of contact with the police. Other costs that we did not account for, but are nevertheless relevant in the field of crime prevention are: the effect of the intervention reflected in both costs and effects, such as increased wellbeing and reduced productivity losses (i.e. van Zutphen et al., 2014), in regard to family members (e.g. parents or siblings of the adolescents). Further additional categories are reduced victim costs and increased victim wellbeing, the reduction of the number of out-of-home placements, the reduction of the costs of committed crimes to society, reduced costs of avoided crimes to society and the value of reduced fear of crime (i.e. Cohen & Piquero, 2009; van Zutphen et al., 2014).

The second methodological uncertainty deals with the WTP value for a CAFY. Although we used the WTP values of Cohen et al. (Cohen & Piquero, 2009; Cohen et al., 2010) to illustrate the use of WTP in crime prevention, WTP to prevent a crime like burglary is definitely not equal to WTP per CAFY. Therefore, it is important to carefully estimate the WTP value per CAFY. In this study, for example, we could have weighted the WTP values of Cohen et al. (Cohen & Piquero, 2009; Cohen et al., 2010) by the frequency of the crimes as yearly committed by the adolescents in this study, or by the number of yearly registered crimes in the Netherlands. For clarity reasons and due to a lack of more detailed data regarding the crimes committed, we chose not to use a weighted WTP value. Furthermore, for the WTP values used, it is not exactly known which components of crime, such as investigation, prosecution, witnesses, legal aid, prevention programs, the costs to victims, and the valuation of fear (Cohen & Piquero, 2009) are included in this valuation (Cohen & Piquero, 2009; Cohen, Rust, Steen, & Tidd, 2004). Therefore, further research is needed into which categories of costs of crime are included in a WTP value, before determining what society is willing to pay for one CAFY. The cost categories included in the cost-effectiveness model should be reflected in the WTP value and vice versa. Also, other estimations of the societal WTP might be considered. These could, for example, be based on the cost of crime using a bottom-up approach or a breaking-down approach (Moolenaar, 2009). These methods take into account only the costs of crime, not the willingness to reduce crime levels.

Third, to estimate a WTP per CAFY, it should be known what type of criminal activity is avoided in a CAFY. The seriousness of the crime, the number of times the crime is committed and the types of criminal activity can also be taken into account in defining criminal activity. Furthermore, it is important to decide on how to measure criminal activity. The CAFY used in our study was based on the adolescents' self-reported contact with the police. However, criminal activity may as well be determined on the basis of police registries (Slot et al., 1992), contacts with other judicial institutions (McCollister et al., 2003a; McCollister et al., 2003b; McCollister et al., 2004), rates of reconviction (Barrett & Byford, 2012), or a delinquency score (French et al., 2008). Different definitions of criminal activity can influence the model results. For example not all committed crimes are recorded in police registrations, while self-reported measures could yield socially desirable answers. Using the CAFY in further research thus requires a clear definition of criminal activity.

A final remark on this analysis concerns the interventions chosen. FFT and the Course House were chosen to illustrate the analysis in the field of crime prevention. The interventions under study, however, could be replaced by other interventions aimed at reducing juvenile delinquency, such as Multisystemic Therapy, Multidimensional Foster Treatment Care or Multidimensional Family Therapy (Netherlands Youth Institute, 2014). Contrary to a broader range of cost-effectiveness studies in the UK and US (i.e. Aos, Lieb, Mayfield, Miller, & Pennucci, 2004; Cary et al., 2013), in the Netherlands, to the best of our knowledge, the cost-effectiveness of such interventions has not yet been investigated, except for a cost-benefit analyses of 'Maatregel Inrichting Stelstelmatige Daders' or a case study into 'Strafrechtelijke Opvang Verslaafden' (van Zutphen et al., 2014; Versantvoort et al., 2005), which are both aimed at adults. Furthermore, in studying interventions in this field, the context of the interventions under study is highly important. In our illustration, we assumed that in practice, the interventions would be applied completely equivalently. However, preferences for an intervention may influence the choice for a certain intervention in reality, such as earlier experience with an intervention, specific characteristics of an adolescent, or the availability of the intervention itself. In our illustration, FFT may be, for example, used more often to avoid committing crimes, whereas the Course House could be used as an addition to a punishment under juvenile justice law, where the adolescent has already committed a crime. There may then be a higher probability of recidivism if the adolescents already have a history of committing crimes (Donker & de Bakker, 2012). These non-equivalent baseline situations may influence the measured effectiveness of the intervention. Moreover, the situation after treatment may also be different. This may affect the acceptance of possible or required further care (i.e. Donker & de Bakker, 2012), and therefore may influence the final degree of committing crimes in the future. In modelling the cost-effectiveness of interventions in the field of crime prevention, the application of interventions in practice should therefore be taken into account in a cost-effectiveness model, or at least, this should be clarified when modelling the cost-effectiveness of such interventions.

In conclusion, an analysis to estimate the value of performing further research had not yet been conducted in the field of crime prevention. The findings of the current study illustrate how such an analysis might be estimated and interpreted in this field. Future investment in cost-effectiveness research on interventions aimed at reducing juvenile delinquency could use this value of information framework to efficiently conduct further cost-effectiveness research.

References

Aos, S., Lieb, R., Mayfield, J., Miller, M., & Pennucci, A. (2004). *Benefits and costs of prevention and early intervention programs for youth*. Olympia: Washington State Institute for Public Policy.

Barrett, B., & Byford, S. (2012). Costs and outcomes of an intervention programme for offenders with personality disorders. *The British Journal of Psychiatry, 200*, 336-341.

Barton, G. R., Briggs, A. H., & Fenwick, E. A. L. (2008). Optimal cost-effectiveness decisions: The rule of the cost-effectiveness acceptability curve (CEAC), the cost-effectiveness acceptability frontier (CEAF), and the expected value of perfection information (EVPI). *Value in Health, 11*, 886-897.

Bilcke, J., Beutels, P., Brisson, M., & Jit, M. (2011). Accounting for methodological, structural, and parameter uncertainty in decision-analytic models: A practical guide. *Medical Decision Making, 31*, 675-692.

Bojke, L., Claxton, K., Sculpher, M., & Palmer, S. (2009). Characterizing structural uncertainty in decision analytic models: A review and application of methods. *Value in Health, 12*, 739-749.

Briggs, A., Claxton, K., & Sculpher, M. (2006). *Decision Modelling for Health Economic Evaluation*. New York: Oxford University Press.

Briggs, A. H. (2000). Handling uncertainty in cost-effectiveness models. *Pharmacoeconomics, 17*, 479-500.

Brouwer, W. B., van Hout, B. A., & Rutten, F. (2000). A fair approach to discounting future effects: Taking a societal perspective. *Journal of Health Services Research and Policy, 5*, 114-118.

Carlson, J. J., Thariani, R., Roth, J., Gralow, J., Henry, N. L., Esmail, L., . . . Veenstra, D. L. (2013). Value-of-information analysis within stakeholder-driven research prioritization process in a US setting: An application in cancer genomics. *Medical Decision Making, 33*, 463-471.

Cary, M., Butler, S., Baruch, G., Hickey, N., & Byford, S. (2013). Economic evaluation of Multisystemic Therapy for young people at risk for continuing criminal activity in the UK. *PloS ONE, 8*, e61070.

Chassin, L., Piquero, A. R., Losoya, S. H., Mansion, A. D., & Schubert, C. A. (2013). Joint consideration of distal and proximal predictors of premature mortality among serious juvenile offenders. *Journal of Adolescent Health, 52*, 689-696.

Claxton, K. (1999). The irrelevance of inference: a decision-making approach to the stochastic evaluation of health care technologies. *Journal of Health Economics, 18*, 341-364.

Claxton, K. (2008). Exploring uncertainty in cost-effectiveness analysis. *Pharmacoeconomics, 26*, 781-798.

Claxton, K., Neumann, P. J., Araki, S., & Weinstein, M. C. (2001). Bayesian value-of-information analysis. An application to a policy model of Alzheimer's disease. *International Journal of Technology Assessment in Health Care, 17*, 38-55.

Claxton, K., Palmer, S., Longworth, L., Bojke, L., Griffin, S., McKenna, C., . . . Youn, J. (2012). Informing a decision framework for when NICE should recommend the use of health technologies only in the context of an appropriately designed programme of evidence development. *Health Technology Assessment, 16*.

Claxton, K., Sculpher, M. J., & Drummond, M. F. (2002). A rational framework for decision making by the National Institute For Clinical Excellence (NICE). *The Lancet, 360*, 711-715.

Cohen, M. A., & Piquero, A. R. (2009). New evidence on the monetary value of saving a high risk youth. *Journal of Quantitative Criminology, 25*, 25-49.

Cohen, M. A., Piquero, A. R., & Jennings, W. G. (2010). Studying the costs of crime across offender trajectories. *Criminology and Public Policy, 9*, 279-305.

Cohen, M. A., Rust, R. T., Steen, S., & Tidd, S. T. (2004). Willingness-to-pay for crime control programs. *Criminology, 42*, 89-109.

Donker, A., & de Bakker, W. (2012). *Vrij na een PIJ. Voorspellende factoren van acceptatie vrijwillige nazorg en recidive na een PIJ-maatregel [In freedom after PIJ. The predictive factors of accepting voluntary after care and recidivism after a PIJ-order]*. Leiden, the Netherlands: WODC, Ministerie van Veiligheid en Justitie, Hogeschool Leiden.

Evers, S., Salvador-Carulla, L., Halsteinli, V., McDaid, D., & Group, M. (2007). Implementing mental health economic evaluation evidence: Building a bridge between theory and practice. *Journal of Mental Health 16*, 223-241.

Fenwick, E., Claxton, K., & Sculpher, M. J. (2001). Representing uncertainty: The role of cost-effectiveness acceptability curves. *Health Economics, 10*, 779-787.

Fenwick, E., Claxton, K., & Sculpher, M. J. (2008). The value of implementation and the value of information: Combined and uneven development. *Medical Decision Making, 28*, 21-32.

Fenwick, E., O'Brien, B. J., & Briggs, A. (2004). Cost-effectiveness acceptability curves – facts, fallacies and frequently asked questions. *Health Economics, 13*, 405-415.

Frederix, G. W. J., van Hasselt, J. G. C., Schellens, J. H. M., Hövels, A. M., Raaijmakers, J. A. M., Huitema, A. D. R., & Severens, J. L. (2014). The impact of structural uncertainty on cost-effectiveness models for adjuvant endocrine breast cancer treatments: The need for disease-specific model standardization and improved guidance. *Pharmacoeconomics 32*, 47-61.

French, M. T., Zavala, S. K., McCollister, K. E., Waldron, H. B., Turner, C. W., & Ozechowski, T. J. (2008). Cost-effectiveness analysis of four interventions for adolescents with a substance use disorder. *Journal of Substance Abuse Treatment, 34*, 272-281.

Haij Ali Afzali, H., & Karnon, J. (2015). Exploring structural uncertainty in model-based economic evaluations. *Pharmacoeconomics, 20 January 2015.*

Netherlands Youth Institute (2014). *Database of effective youth interventions*. Retrieved from http://www.nji.nl/nl/Kennis/Databanken/Databank-Effectieve-Jeugdinterventies/Erkende-interventies.

Jackson, C. H., Bojke, L., Thompson, S. G., Claxton, K., & Sharples, L. D. (2011). A framework for addressing structural uncertainty in decision models. *Medical Decision Making, 31*, 662-674.

Knapp, M., McDaid, D., Evers, S., Salvador-Carulla, L., Halsteinli, V., & MHEEN Group (2008). *Cost-effectiveness and mental health (MHEEN Policy briefing)*. London: MHEEN network.

McCollister, K. E., French, M. T., & Fang, H. (2010). The cost of crime to society: New crime-specific estimates for policy and program evaluation. *Drug and Alcohol Dependence, 108*, 98-109.

McCollister, K. E., French, M. T., Inciardi, J. A., Butzin, C. A., Martin, S. S., & Hooper, R. M. (2003a). Post-release substance abuse treatment for criminal offenders: A cost-effectiveness analysis. *Journal of Quantitative Criminology, 19*, 389-407.

McCollister, K. E., French, M. T., Prendergast, M., Wexler, H., Sacks, S., & Hall, E. (2003b). Is in-prison treatment enough? A cost-effectiveness analysis of prison-based treatment and aftercare services for substance-abusing offenders. *Law and Policy, 25*, 63-82.

McCollister, K. E., French, M. T., Prendergast, M. L., Hall, E., & Sacks, S. (2004). Long-term cost effectiveness of addiction treatment for criminal offenders. *Justice Quarterly, 21*, 659-679.

McKenna, C., & Claxton, K. (2011). Addressing adoption and research design decisions simultaneously: The role of value of sample information analysis. *Medical Decision Making, 31*, 853-865.

Moffitt, T. E. (1993). Adolescence-limited and life-course-persistent antisocial behavior: A developmental taxonomy. *Psychological Review, 100*, 674-701.

Mohseninejad, L., van Baal, P., van den Berg, M., Buskens, E., & Feenstra, T. L. (2013). Value of information analysis from a societal perspective: A case study in prevention of major depression. *Value in Health, 16*, 490-497.

Moolenaar, D. E. G. (2009). Modelling Criminal Justice System Costs by Offence. *European Journal on Criminal Policy and Research, 15*, 309-326.

Myers, E., Sanders, G. D., Ravi, D., Matchar, D., Havrilesky, L., Samsa, G., . . . Gray, R. (2011). *Evaluating the potential use of modeling and value of information analysis for future research prioritization within the evidence-based practice centre program (AHRQ Publication No. 11-EHC030-EF)*. Rockville, MD: Agency for Healthcare Research and Quality.

OECD Library (2015). *Purchasing power parities for GDP*. Retrieved from http://www.oecd-ilibrary.org/economics/purchasing-power-parities-for-gdp-2014-5_ppp-gdp-table-2014-5-en;jsessionid=j1wryvqunpvk.x-oecd-live-03.

Oostenbrink, J. B., Al, M. J., Oppe, M., & Rutten-van Mölken, M. P. M. H. (2008). Expected value of perfect information: An empirical example of reducing decision uncertainty by conducting additional research. *Value in Health, 11*, 1070-1080.

Pratt, J., Raiffa, H., & Schlaifer, R. (1995). *Statistical decison theory*. Cambridge, MA: MIT Press.

Raiffa, H. (1968). *Decision analysis: Introductory lectures on choices under uncertainty*. New York: Random House.

Romeo, R., Byford, S., & Knapp, M. (2005). Annotation: Economic evaluations of child and adolescent mental health interventions: A systematic review. *Journal of Child Psychology and Psychiatry, 46*, 919-930.

Schawo, S. J., van Eeren, H., Soeteman, D. I., van der Veldt, M. C. A. E., Noom, M. J., Brouwer, W., . . . Hakkaart, L. (2012). Framework for modelling the cost-effectiveness of systemic interventions aimed to reduce youth delinquency. *Journal of Mental Health Policy and Economics, 15*, 187-196.

Sculpher, M. J., & Claxton, K. (2005). Establishing the cost-effectiveness of new pharmaceuticals under conditions of uncertainty - When is there sufficient evidence? *Value in Health, 8*, 433-446.

Slot, N. W., Jagers, J. D., & Beumer, M. H. (1992). Tien jaar Kursushuis: Ervaringen en follow-up-gegevens [Ten years Course House: Experiences and follow-up data]. *Kind en adolescent, 13*, 62-71.

Soeteman, D. I., Busschbach, J. J., Verheul, R., Hoomans, T., & Kim, J. J. (2011). Cost-effective psychotherapy for personality disorders in the Netherlands: The value of further research and active implementation. *Value in Health, 14*, 229-239.

Soeteman, D. I., & Busschbach, J. J. V. (2008). Cost-benefit and cost-effectiveness of prevention and treatment. In R. Loeber, N. W. Slot, P. H. van der Laan & M. Hoeve (Eds.), *Tomorrow's criminals: The development of child delinquency and effective interventions* (pp. 215-228). Hampshire: Ashgate Publishing Ltd.

Statistics Netherlands, S. (2015a). *Statistics Netherlands: Consumer price index*. Retrieved from http://statline.cbs.nl.

Statistics Netherlands, S. (2015b). *Statistics Netherlands: Life tables*. Retrieved from http://statline.cbs.nl.

Strong, M., Oakley, J. E., & Brennan, A. (2014). Estimating multiparameter partial expected value of perfect information from a probabilistic sensitivity analysis sample: A nonparametric regression approach. *Medical Decision Making, 34*, 311-326.

Teplin, L. A., McClelland, G. M., Abram, K. M., & Mileusnic, D. (2005). Early violent death among delinquent youth: A prospective longitudinal study. *Pediatrics, 115*, 1586-1593.

The Health Care Insurance Board (2006). *Guidelines for Pharmacoeconomic Research*. Amstelveen, the Netherlands: CVZ.

van der Laan, P. H. (2006). Just Desert and Welfare: Juvenile Justice in the Netherlands. In J. Junger-Tas & S. H. Decker (Eds.), *International handbook of juvenile justice* (pp. 145-172). New York: Springer.

van der Veldt, M. C. A. E., Eenshuistra, R. M., & Campbell, E. E. (2011). *FFT versterkt: Een evaluatiestudie naar de implementatie en de effecten van Functional Family Therapy in Nederland [FFT strenghtens: An evaluationstudy into the implementation and effectiveness of FFT in the Netherlands]*. Amsterdam, the Netherlands: PI Research.

van Hout, B. A. (1998). Discounting costs and effects: A reconsideration. *Health Economics, 7*, 581-594.

van Zutphen, F., Goderie, M., & Janssen, J. (2014). *De maatregel Inrichting Stelselmatige Daders (ISD). Maatschappelijke kosten-batenanalyse van een eventuele verlenging [A societal cost-benefit analysis into the possible extension of 'ISD']*. Rotterdam, the Netherlands: van Zutphen Economisch Advies.

Versantvoort, M. C., Verster, A. C. M., Jannink, J., van den Broek, L. G. J. M., van Zutphen, F., & Donker van Heel, P. A. (2005). *Kosten en baten van justitiele interventies. Ontwikkeling van een analyse- en rekenmodel [Costs and benefits of judicial interventions. Development of a model to analyse these costs and benefits]*. Rotterdam, the Netherlands: ECORYS.

Walker, S., Sculpher, M., Claxton, K., & Palmer, S. (2012). Coverage with evidence development, only in research, risk sharing, or patient access scheme? A framework for coverage decisions. *Value in Health*, *15*, 570-579.

Yokota, F., & Thompson, K. M. (2004). Value of information analysis in environmental health risk management decisions: Past, present, and future. *Risk Analysis*, *24*, 635-650.

Supplemental Material

Table I. Model parameters and parameter distributions

EVPPI group of parameters	Parameters
	Transition probabilities
transition probabilities FFT	***FFT***
	tpA2A
	tpA2B
	tpB2A
	tpB2B
transition probabilities Course House	***Course House***
	tpA2A
	tpA2B
	tpB2A
	tpB2B
	Intervention costs
intervention costs FFT	FFT
intervention costs Course House	Course House
	Model state costs
	Criminal state (A)
direct health-care costs criminal	**Direct health-care - adolescent**
	Psychiatrist
	Psychologist
	Psychiatric nurse
	Social worker
	GP
	GP at school
	Pediatrician
	Medical specialist
	Alternative healer
	Family guardian
	Youth welfare agency
	ER
	Foster care
	Residential institution
	Day hospitalization
	Hospitalization
	Centre for addiction treatment
	Direct health-care - parent
	Psychiatrist
	Psychologist
	Psychiatric nurse

Probability	Events	Complements		Distribution	Reference
0.23	6	20		Dirichlet	Multicentre trial
0.77	20	6		Dirichlet	Multicentre trial
0.37	7	12		Dirichlet	Multicentre trial
0.63	12	7		Dirichlet	Multicentre trial
0.39	13	20		Dirichlet	Slot et al. (1992)
0.61	20	13		Dirichlet	Slot et al. (1992)
0.24	4	13		Dirichlet	Slot et al. (1992)
0.76	13	4		Dirichlet	Slot et al. (1992)
Mean	**SE mean***	**alpha**	**beta**	**Distribution**	**Reference**
€10,900	€10,900	1	10900	Gamma	Costdata mental health institutions
€37,800	€37,800	1	37800	Gamma	Slot et al. (1992)
** SE was not known, therefore taken conservative and set equal to mean*					
Mean	**SE mean**	**alpha**	**beta**	**Distribution**	**Reference**
€112	€67	2.8	39.9	Gamma	Multicentre trial
€183	€103	3.2	57.5	Gamma	Multicentre trial
€47	€47	1.0	47.2	Gamma	Multicentre trial
€113	€78	2.1	54.2	Gamma	Multicentre trial
€43	€26	2.7	15.6	Gamma	Multicentre trial
€3	€3	1.0	3.0	Gamma	Multicentre trial
€0	€0	0.0	0.0	Gamma	Multicentre trial
€8	€1	100.0	0.1	Gamma	Multicentre trial
€0	€0	0.0	0.0	Gamma	Multicentre trial
€0	€0	0.0	0.0	Gamma	Multicentre trial
€142	€75	3.6	39.5	Gamma	Multicentre trial
€66	€50	1.7	38.4	Gamma	Multicentre trial
€0	€0	0.0	0.0	Gamma	Multicentre trial
€25	€25	1.0	25.3	Gamma	Multicentre trial
€0	€0	0.0	0.0	Gamma	Multicentre trial
€1,438	€1,438	1.0	1437.5	Gamma	Multicentre trial
€0	€0	0.0	0.0	Gamma	Multicentre trial
€61	€61	1.0	61.1	Gamma	Multicentre trial
€55	€47	1.4	40.5	Gamma	Multicentre trial
€0	€0	0.0	0.0	Gamma	Multicentre trial

	Model state costs
	Criminal state (A)
direct health-care costs criminal	Social worker
	GP
	Medical officer
	Medical specialist
	Alternative healer
	Family guardian
	Centre for addiction treatment
direct non health-care costs criminal	**Direct non health-care - adolescent**
	Council of child protection
	Bureau Halt
	Police
	Lawyer
	Court
	Social rehabilitation
	Incarceration costs
indirect non health-care costs criminal	**Indirect non health-care - adolescent**
	Time spent on exercises as part of intervention
	Indirect non health-care - parent
	Absence from work
	Inefficiency at work
	Productivity losses due to unpaid support
	Productivity losses due to paid support
	Time spent on exercises as part of intervention
	Non Criminal state (B)
direct health-care costs non criminal	**Direct health-care - adolescent**
	Psychiatrist
	Psychologist
	Psychiatric nurse
	Social worker
	GP
	GP at school
	Pediatrician
	Medical specialist
	Alternative healer
	Family guardian
	Youth welfare agency
	ER
	Foster care
	Residential institution
	Day hospitalization
	Hospitalization

Mean	SE mean	alpha	beta	Distribution	Reference
€6	€6	1.0	5.9	Gamma	Multicentre trial
€94	€71	1.7	53.9	Gamma	Multicentre trial
€12	€7	3.3	3.7	Gamma	Multicentre trial
€21	€15	2.0	10.9	Gamma	Multicentre trial
€9	€9	1.0	9.1	Gamma	Multicentre trial
€0	€0	0.0	0.0	Gamma	Multicentre trial
€80	€80	1.0	79.7	Gamma	Multicentre trial
€25	€16	2.3	10.7	Gamma	Multicentre trial
€44	€27	2.8	15.9	Gamma	Multicentre trial
€90	€37	6.0	15.0	Gamma	Multicentre trial
€57	€57	1.0	57.1	Gamma	Multicentre trial
€29	€29	1.0	28.6	Gamma	Multicentre trial
€171	€111	2.4	71.4	Gamma	Multicentre trial
€0	€0	0.0	0.0	Gamma	Multicentre trial
€3	€2	1.9	1.6	Gamma	Multicentre trial
€1,317	€835	2.5	528.8	Gamma	Multicentre trial
€4,450	€2,775	2.6	1730.4	Gamma	Multicentre trial
€0	€0	0.0	0.0	Gamma	Multicentre trial
€0	€0	0.0	0.0	Gamma	Multicentre trial
€2	€2	1.0	2.1	Gamma	Multicentre trial
€31	€17	3.3	9.2	Gamma	Multicentre trial
€78	€45	3.1	25.4	Gamma	Multicentre trial
€0	€0	0.0	0.0	Gamma	Multicentre trial
€60	€45	1.8	33.9	Gamma	Multicentre trial
€42	€15	8.3	5.1	Gamma	Multicentre trial
€0	€0	0.0	0.0	Gamma	Multicentre trial
€0	€0	0.0	0.0	Gamma	Multicentre trial
€12	€8	2.0	5.6	Gamma	Multicentre trial
€0	€0	0.0	0.0	Gamma	Multicentre trial
€16	€12	1.9	8.4	Gamma	Multicentre trial
€47	€27	3.0	15.7	Gamma	Multicentre trial
€30	€2	289.6	0.1	Gamma	Multicentre trial
€0	€0	0.0	0.0	Gamma	Multicentre trial
€247	€247	1.0	246.6	Gamma	Multicentre trial
€0	€0	0.0	0.0	Gamma	Multicentre trial
€9	€9	1.0	9.4	Gamma	Multicentre trial

	Model state costs
	Non Criminal state (B)
direct health-care costs non criminal	Centre for addiction treatment
	Direct health-care - parent
	Psychiatrist
	Psychologist
	Psychiatric nurse
	Social worker
	GP
	Medical officer
	Medical specialist
	Alternative healer
	Family guardian
	Centre for addiction treatment
direct non health-care costs non criminal	**Direct non health-care - adolescent**
	Council of child protection
	Bureau Halt
	Police
	Lawyer
	Court
	Social rehabilitation
	Incarceration costs
indirect non health-care costs non criminal	**Indirect non health-care - adolescent**
	Time spent on exercises as part of intervention
	Indirect non health-care – parent
	Absence from work
	Inefficiency at work
	Productivity losses due to unpaid support
	Productivity losses due to paid support
	Time spent on exercises as part of intervention

Italics - costs not relevant from 23 years onwards

EVPPI, Expected Value of Partial Perfect Information; FFT, Functional Family Therapy; SE, Standard error

Mean	SE mean	alpha	beta	Distribution	Reference
€0	€0	0.0	0.0	Gamma	Multicentre trial
€0	€ 0	0.0	0.0	Gamma	Multicentre trial
€80	€ 40	3.9	20.3	Gamma	Multicentre trial
€0	€ 0	0.0	0.0	Gamma	Multicentre trial
€17	€ 17	1.0	16.8	Gamma	Multicentre trial
€18	€ 7	5.9	3.0	Gamma	Multicentre trial
€6	€ 6	1.0	6.1	Gamma	Multicentre trial
€39	€ 2	292.4	0.1	Gamma	Multicentre trial
€0	€ 0	0.0	0.0	Gamma	Multicentre trial
€27	€ 19	1.9	14.3	Gamma	Multicentre trial
€0	€ 0	0.0	0.0	Gamma	Multicentre trial
€0	€0	0.0	0.0	Gamma	Multicentre trial
€12	€9	1.6	7.2	Gamma	Multicentre trial
€25	€15	3.0	8.4	Gamma	Multicentre trial
€0	€0	0.0	0.0	Gamma	Multicentre trial
€0	€0	0.0	0.0	Gamma	Multicentre trial
€36	€29	1.6	23.0	Gamma	Multicentre trial
€0	€0	0.0	0.0	Gamma	Multicentre trial
€4	€4	1.2	3.4	Gamma	Multicentre trial
€1,020	€558	3.3	305.9	Gamma	Multicentre trial
€3,866	€1,503	6.6	584.5	Gamma	Multicentre trial
€0	€0	0.0	0.0	Gamma	Multicentre trial
€0	€0	0.0	0.0	Gamma	Multicentre trial
€8	€5	2.7	2.8	Gamma	Multicentre trial

Table II. pEVPI and pEVPPI for a range of WTP values

WTP	transition probabilities FFT	transition probabilities Course House	direct health-care costs criminal	direct health-care costs non criminal	direct non health-care costs criminal
€0	€0	€0	€0	€0	€0
€5,000	€23,000	€0	€0	€0	€1,700
€10,000	€919,600	€1,990,500	€0	€0	€12,400
€15,000	€6,402,400	€11,155,100	€0	€0	€21,200
€20,000	€16,263,300	€24,720,800	€600	€0	€29,800
€25,000	€28,832,200	€40,322,400	€8,900	€0	€34,300
€30,000	€42,919,000	€57,149,700	€16,900	€0	€60,400
€35,000	€57,863,500	€74,630,700	€41,300	€0	€264,800
€40,000	€70,749,600	€89,965,700	€590,500	€357,800	€1,546,500
€45,000	€73,470,900	€95,078,000	€0	€0	€185,800
€50,000	€76,407,100	€100,394,600	€0	€0	€75,900
€55,000	€79,537,400	€105,903,400	€0	€0	€42,900
€60,000	€82,789,400	€111,483,800	€0	€0	€25,100
€65,000	€86,144,500	€117,144,600	€0	€0	€14,200
€70,000	€89,582,300	€122,808,500	€0	€0	€9,400
€75,000	€93,083,900	€128,578,700	€0	€0	€7,300
€80,000	€96,638,800	€134,392,600	€0	€0	€6,700
€85,000	€100,232,100	€140,405,200	€0	€0	€6,100
€90,000	€103,864,300	€146,115,500	€0	€0	€5,400
€95,000	€107,525,900	€152,098,000	€0	€0	€4,800
€100,000	€111,211,000	€157,921,800	€0	€0	€4,100

WTP, willingness-to-pay; FFT, Functional Family Therapy; pEVPI, population Expected Value of Perfect Information; pEVPPI, population Expected Value of Partial Perfect Information

direct non health-care costs non criminal	indirect non health-care costs criminal	indirect non health-care costs non criminal	intervention costs FFT	intervention costs Course House	pEVPI
€0	€0	€0	€1,798,900	€5,646,700	€11,911,500
€0	€2,300	€0	€2,462,300	€9,568,600	€18,197,300
€0	€13,300	€0	€3,351,400	€14,051,900	€28,674,800
€900	€41,300	€0	€4,519,500	€19,242,800	€42,195,800
€0	€107,000	€0	€6,029,400	€25,103,700	€57,847,600
€0	€255,400	€0	€7,980,200	€31,601,900	€74,913,300
€0	€609,000	€10,700	€10,567,700	€38,703,200	€92,970,400
€10,700	€1,761,100	€559,100	€14,004,000	€46,337,600	€111,815,800
€1,089,900	€3,732,300	€2,326,300	€15,926,700	€51,773,800	€128,621,700
€24,000	€374,800	€21,300	€8,788,000	€47,140,300	€135,339,200
€32,500	€20,300	€0	€3,432,200	€42,879,500	€142,393,000
€35,000	€800	€0	€427,800	€38,985,800	€149,713,600
€37,500	€0	€0	€0	€35,421,100	€157,238,100
€40,000	€0	€0	€0	€32,177,300	€164,940,500
€42,500	€0	€0	€0	€29,245,900	€172,780,800
€44,900	€0	€0	€0	€26,603,300	€180,738,000
€47,400	€0	€0	€0	€24,164,400	€188,804,300
€49,900	€0	€0	€0	€21,956,200	€196,955,500
€52,300	€0	€0	€0	€19,947,500	€205,186,900
€54,700	€0	€0	€0	€18,102,800	€213,485,700
€57,200	€0	€0	€0	€16,420,600	€221,852,400

Chapter 4.

Estimating subgroup effects using the propensity score method: A practical application in outcomes research

Hester V. Eeren, Marieke D. Spreeuwenberg, Anna Bartak, Mark de Rooij, &
Jan J.V. Busschbach

Medical Care (2015), 53, 366-373

Abstract

Objectives: Our aim was to demonstrate the feasibility of the univariate and generalized propensity score (PS) method in subgroup analysis of outcomes research.

Methods: First, to estimate subgroup effects, we tested the performance of 2 different PS methods, using Monte Carlo simulations: 1) the univariate PS with additional adjustment on the subgroup; and 2) the generalized PS, estimated by crossing the treatment options with a subgroup variable. The subgroup effects were estimated in a linear regression model using the 2 PS adjustments. We further explored whether the subgroup variable should be included in the univariate PS. Second, the 2 methods were compared using data from a large effectiveness study on psychotherapy in personality disorders. Using these data we tested the differences between short and long-term treatment , with the severity of patients' problems defining the subgroups of interest.

Results: The Monte Carlo simulations showed minor differences between both PS methods, with the bias and mean squared error overall marginally lower for the generalized PS. When considering the univariate PS, the subgroup variable can be excluded from the PS estimation and only adjusted for in the outcome equation. When applied to the psychotherapy data, the univariate and generalized PS estimations gave similar results.

Conclusions: The results support the use of the generalized PS as a feasible method, compared to the univariate PS, to find certain subgroup effects in non-randomized outcomes research.

Introduction

In non-randomized studies, the propensity score (further denoted as PS) method has gained popularity as a statistical method to overcome selection bias due to differences in observed pre-treatment variables of patient groups (Winship & Mare, 1992) and the "dimensionality" problem of alternative methods such as stratification and matching (D'Agostino, 1998). The univariate propensity score (Rosenbaum & Rubin, 1983) is a valid solution to compare 2 treatment categories (Bartak et al., 2009; Rubin, 1974), whereas the generalized PS can be used if >2 treatment categories are compared (Feng, Zhou, Zou, Fan, & Li, 2012; Imbens, 2000; Spreeuwenberg et al., 2010). An equal distribution on the covariates is assumed after adjustment on the PS (Austin, 2009; Rubin, 1997; Spreeuwenberg et al., 2010). Although the PS can control for overt bias due to (many) observed pre-treatment variables (Rosenbaum, 1991; Rubin, 1997), hidden bias could still be present (Rosenbaum, 1991).

The PS is predominantly used to estimate treatment effects (Austin, 2011; Hirano, Imbens, & Ridder, 2003). However, it may also be important to define which treatment is specifically effective for a (sub)group of patients (Norcross & Wampold, 2011). Treatment options can then be applied more efficiently by directly allocating a group of patients to a relevant treatment. For instance, one could argue that long-term psychotherapy is more effective than short-term psychotherapy for patients having severe problems. Because the patients are not randomly assigned to the treatment options, patients having either mild or severe problems can differ on the observed pre-treatment variables. Therefore, there is a need to apply PS modelling methods when studying subgroup effects.

Several authors have described methods to estimate treatment effects for particular subgroups while using the univariate PS. Rosenbaum and Rubin already recommended sub-classifying or matching on additional covariates to identify differences in treatment effect between subgroups. To reduce bias in the estimated treatment effect, Rubin and Thomas (2000) and Stürmer and colleagues (2006) advised that the covariate together with the PS be included when estimating the treatment effect. Such an additional covariate could define subgroups. The treatment effect can also vary according to quantiles of the PS estimations. Effect modification (e.g. interaction effects) can result in different estimated treatment effects for different PS quantiles (Glynn, Schneeweiss, & Sturmer, 2006; Kurth et al., 2006; Lunt et al., 2009; Sturmer, Rothman, & Glynn, 2006; Ye, Bond, Schmidt, Mulia, & Tam, 2012). However, it is not possible to relate a specific subgroup to the PS quantiles. More recently, Liem and colleagues (2010) determined the treatment effect for subgroups by adding interaction terms in a multivariable adjusted model in which the PS was also included. Another method was described by Radice and colleagues (2012) and Kreif and colleagues (2012) who estimated the univariate PS within each subgroup separately. Yet, only the univariate PS was used in subgroup analyses and the PS was not made multiple, as a generalized PS, by crossing the treatment options with a subgroup variable.

Because the univariate PS is mainly used in subgroup analyses, this study investigated whether a generalized PS could be used to estimate subgroup effects in outcomes research, and compared it to using a univariate PS. We first used Monte Carlo simulations to investigate whether and how the generalized or univariate PS could be used to estimate subgroup effects. These 2 PS estimations were subsequently compared using data from a Dutch research project on psychotherapy effectiveness: SCEPTRE (Study on Cost-Effectiveness of Personality Disorder TREatment) (Bartak et al., 2010).

Methods

First, we describe the univariate propensity score (PS), the generalized PS and the simulation study in which we tested these methods. Then, we describe the case study where we compare the 2 PS methods. To estimate the treatment effects for subgroups of patients, we used the 2 PS estimations in covariate adjustment, as this method is the most frequently used PS method in the medical literature (Austin, 2009; Shah, Laupacis, Hux, & Austin, 2005; Stürmer et al., 2006; Weitzen, Lapane, Toledano, Hume, & Mor, 2004) because it leaves the sample size intact.

Univariate PS method

The univariate PS is defined according to Rosenbaum and Rubin (Rosenbaum & Rubin, 1983) as:

$$\mathrm{PS}(x)=\mathrm{pr}\left\langle D=1|X=x\right\rangle \tag{1}$$

where if D = 1 the PS defines the conditional probability of assignment to the treatment of interest, given a set of observed covariates (X) (Rosenbaum & Rubin, 1983). The ignorability assumption defines that the potential outcomes and the treatment assignment are independent given the observed covariates (X) (Rosenbaum & Rubin, 1983; Rubin, 1997). The PS was estimated in a univariate logistic regression function (Hirano & Imbens, 2001). To estimate the treatment effect for subgroups of patients, this PS estimation was used as an extra predictor in a linear regression model with treatment outcome (Y) as the dependent variable:

$$OUTCOME = \beta_0 + \beta_1 D + \beta_3 Z + \beta_4 DZ \tag{2}$$

The treatment groups (D), subgroups (Z) and the interaction between these (DZ) were the independent variables, and the effects of interest (Liem et al., 2010).

Generalized PS method

The generalized PS is an extension of the univariate PS defined by Rosenbaum and Rubin (Rosenbaum & Rubin, 1983) and is further defined by Imbens (2000). To calculate the generalized PS used in this study, we estimated the joint conditional probability of the treatment assignment (D) and subgroup (Z) given all covariates (X):

$$\mathrm{PS}(d,z,x)=\mathrm{pr}\left\langle D=d,Z=z|X=x\right\rangle \tag{3}$$

Using 2 treatment options and 2 subgroups, the PS was estimated for 4 groups. The assumption on the strongly ignorable treatment assignment is crucial (Imbens, 2000; Rosenbaum & Rubin, 1983; Rubin, 1997). When this assumption was adjusted to the combined categories on which the generalized PS was estimated, the joint distributions of the potential outcomes and the covariates X should be equal between the 4 groups (Fujii, Henmi, & Fujita, 2012). The generalized PS was estimated in a multinomial regression model. To estimate the treatment effects, 3 estimated generalized PSs and 3 dummy variables indicating group membership (G) were adjusted for in a regression model with treatment outcome (Y) as the dependent variable:

$$OUTCOME = \beta_0 + \beta_1 PS_1 + \beta_2 PS_2 + \beta_3 PS_3 + \beta_4 G_1 + \beta_5 G_2 + \beta_6 G_3 \tag{4}$$

The coefficients related to the 3 dummy variables were the effects of interest.

Monte Carlo simulation study

A Monte Carlo simulation study was designed to test the 2 PS estimations. Therefore, we simulated 2 treatment categories, a subgroup variable with 2 categories and 3 additional variables that served as covariates. These 3 covariates were continuous variables (such as age, length, or body weight) related to (a) only the treatment assignment, (b) both treatment assignment and outcome, such that it is a true confounder (Brookhart et al., 2006), or (c) outcome alone (Table 1).

Table 1. Variables and characteristics of Monte Carlo simulation*

Variables	**Type**	**Function**
X1	Covariate	Multivariate normal distribution (0, 1)
X2	Covariate	Multivariate normal distribution (0, 1)
X3	Covariate	Multivariate normal distribution (0, 1)
Z	Covariate – forms subgroups	Bernoulli distribution (1, 0.4)
D	Treatment assignment	Defined in treatment assignment, values 0 or 1
Y	Outcome	Defined in outcome
$\varepsilon_1, \varepsilon_2$	Error terms	Multivariate normal distribution (0, 1)
Treatment assignment		
Scenario 1: $f \sim 0.5X1+0.5X2+\varepsilon_1$; *if f<0, D=0; otherwise D=1*		
Scenario 2: $f \sim 0.5X1+0.5X2+0,3Z+\varepsilon_1$ *; if f<0, D=0; otherwise D=1*		
Outcome		
$Y \sim 0.5X2+0.5X3+\alpha_1 D+\alpha_2 Z+\alpha_3 DZ+\varepsilon_2$; *linear regression model* where $\alpha_1 = 0.7, \alpha_2 = 0.4, \alpha_3 = 0.2$		
Characteristics that define simulated datasets		**Categories**
Correlation between covariates (*X*1-*X*3)		0; 0.3; 0.7
Correlation Z – covariates (*X*1-*X*3)		yes; no
Sample size		250; 500; 1000
Variables selected in PS		
Univariate PS		X2,X3; X1,X2,X3; X2,X3,Z; X1,X2,X3,Z; X1,X2; X1,X2,Z; X2; X2,Z
Generalized PS		X2,X3; X1,X2,X3; X1,X2; X2

* The parameter values related to the different variables in the description of the treatment assignment and outcome are arbitrary.

PS indicates propensity score

The covariates were multivariate normally distributed with a mean of zero and variance of 1, except for the subgroup, that followed a Bernoulli distribution with, for example, a probability of having severe problems of 0.4 (Table 1). The outcome was simulated from a linear regression model and its error term was multivariate normally distributed, just as for the error terms of the treatment assignment (Table 1). The correlation between the error terms was set to zero as only overt bias was simulated. We simulated 2 scenarios: in scenario 1 the subgroup was not related to the treatment assignment, whereas in scenario 2, this relationship was simulated. In both the scenarios, the subgroup was related to the outcome (Table 1). Within each scenario we varied the simulated data on 3 levels of the correlation between the covariates, on the presence or absence of a correlation with the subgroup, and on 3 different sample sizes (Table 1). Under each combination of characteristics of the simulated data, 1,000 datasets were created, which resulted in 18,000 datasets per scenario.

In the literature there is no consensus on how to select the variables in PS estimation (Austin, Grootendorst, & Anderson, 2007; Brookhart et al., 2006). Therefore, we varied the inclusion of variables in the PS estimations (Table 1). For the univariate PS, we investigated whether the subgroup variable should be in- or excluded in the PS estimation. The subgroup variable cannot be selected for the generalized PS, as it is part of its definition (Table 1).

To evaluate the performance of the 2 PS methods in the simulation study, we estimated the bias, mean squared error (MSE) and SE of the relevant effects over the total number of simulations. In Eq. (2) the relevant effects were the treatment effect, the effect of the subgroups and the interaction term. In Eq. (4), these were the coefficients related to each dummy variable. Because we were interested in 3 coefficients per regression model [Eqs (2) or (4)] and these coefficients were not comparable one-to-one, we averaged the bias, MSE and SE over the 3 relevant coefficients per regression model [Eqs (2) or (4)]. We then used this value for the bias, MSE and SE per PS method to compare the PS methods.

Case study

Our sample consisted of a total of 841 patients with personality disorders(Association, 2000) who had enrolled for different types of psychotherapy in 6 mental health care institutes in The Netherlands. The patients were selected for either short-term (up to 6 mo) or long-term (> 6 mo) psychotherapy in various settings (Bartak, Andrea, Spreeuwenberg, Thunnissen, et al., 2011; Bartak, Andrea, Spreeuwenberg, Ziegler, et al., 2011; Bartak et al., 2010; Soeteman et al., 2011). The mean age was 34.12 (SD 9.83, range 17 – 62y) and 68.6 % were female. To compare the PS methods, we investigated whether the treatment effect was modified by the severity of problems, that is, having mild or severe problems. Although we were aware of more recent possible classifications of severity (Crawford, Koldobsky, Mulder, & Tyrer, 2011), for comparison purposes we differentiated between the patients having personality difficulties or a simple personality disorder versus patients having complex or more severe personality disorders based on a classification of personality disorders by Tyrer and colleagues (2004; 1996).

The primary outcome measure was psychiatric symptomatology and was measured with the Global Severity Index (GSI), which is the mean score of the 53 items of the Brief Symptom Inventory (Arrindell & Ettema, 2003; Derogatis, 1986). The GSI ranges from 0 to 4, with higher scores indicating more problems. Three treatment institutes conducted their follow-up measures on the GSI at 12, 24, 36 and 60 months after baseline. The 3 remaining treatment institutions conducted their follow-up measures at the end of treatment, 6 and 12 months after end of treatment, and again at 36 and 60 months after baseline. As in an earlier study by Spreeuwenberg and colleagues (2010) we used the mean GSI score of all follow-up measures as a primary outcome measure (range 0.01 - 3.17) (American Psychiatric Association, 2000). We excluded 114 cases that had missing values on one of the potential confounders, leaving 727 patients in the final sample. The excluded cases were not significantly different on the outcome GSI.

The potential confounders were assessed at baseline, that is, age, gender, civil status, living situation, care of children, employment, level of education, duration of psychological complaints, treatment history, alcohol and drug abuse, motivation, treatment preferences, level of psychiatric symptomatology, level of personality pathology, interpersonal functioning, social role functioning, quality of life, number of Diagnostic and Statistical Manual (DSM)-IV Axis II cluster A personality disorders, number of DSM-IV Axis II cluster B personality disorders, number of DSM-IV Axis II cluster C personality disorders, and psychological capacities. For specific details of this study, we refer the reader to the literature (Bartak, Andrea, Spreeuwenberg, Thunnissen, et al., 2011; Bartak, Andrea, Spreeuwenberg, Ziegler, et al., 2011; Bartak et al., 2010).

Computation

The analyses were performed with IBM SPSS for Windows, version 20 (SPSS Inc., Chicago, IL). All simulations were performed in R programming language, version 2.13.0 (R Development Core Team, 2010).

Results

Monte Carlo simulation results

We evaluated the bias, MSE and standard error of the relevant effects in the simulation study (see Table I, Supplemental Material for the bias, MSE and SE in scenario 1; for scenario 2, see Table II, Supplemental Material). Because taking an average over 3 estimated bias values related to the 3 relevant coefficients per PS method can average out positive and negative bias values, the MSE was used to find which PS estimations was most efficient (Table 2). In almost all simulated datasets within scenario 1, when the subgroup was not related to the treatment assignment, the MSE was closest to zero if the variables related to the outcome only were included in the univariate PS and the generalized PS (Table 2). If the subgroup was related to the treatment assignment, as in scenario 2, the MSE was closest to zero when the variables related to the outcome were included in the PS model in all simulated datasets (Table 2). In both the scenarios, including the subgroup variable in the univariate, PS estimation gave larger MSE values.

Table 2. MSE of simulations

Sample size	Correlation covariates X1, X2, X3	Correlation Z - X1,X2,X3		Variables in propensity score (PS) model			
				Univariate PS			
				X2, X3	X1, X2, X3	X2, X3, Z	X1, X2, X3, Z
Scenario 1*							
N = 250	0	Absent	MSE	*0.0523*	0.0579	0.0642	0.0605
		Present	MSE	*0.0524*	0.0561	0.0641	0.0583
	0.3	Absent	MSE	*0.0491*	0.0562	0.0618	0.0608
		Present	MSE	*0.0556*	0.0683	0.0687	0.0724
	0.7	Absent	MSE	*0.0481*	0.0518	0.0593	0.0592
		Present	MSE	*0.0657*	0.0792	0.0926	0.0907
N = 500	0	Absent	MSE	*0.0261*	0.0294	0.0330	0.0311
		Present	MSE	0.0292	0.0264	0.1095	0.0306
	0.3	Absent	MSE	*0.0245*	0.0285	0.0318	0.0311
		Present	MSE	*0.0294*	0.0381	0.0328	0.0394
	0.7	Absent	MSE	0.0270	0.0248	0.0313	0.0315
		Present	MSE	*0.0331*	0.0432	0.0489	0.0494
N = 1000	0	Absent	MSE	*0.0132*	0.0148	0.0160	0.0154
		Present	MSE	*0.0130*	0.0146	0.0160	0.0152
	0.3	Absent	MSE	*0.0124*	0.0141	0.0158	0.0153
		Present	MSE	*0.0162*	0.0214	0.0174	0.0227
	0.7	Absent	MSE	*0.0121*	0.0131	0.0153	0.0964
		Present	MSE	*0.0174*	0.0256	0.0270	0.0290
Scenario 2*							
N = 250	0	Absent	MSE	*0.0537*	0.0575	0.0994	0.0678
		Present	MSE	*0.0576*	0.0622	0.1116	0.0770
	0.3	Absent	MSE	*0.0525*	0.0604	0.1117	0.0850
		Present	MSE	*0.0605*	0.0717	0.1382	0.0763
	0.7	Absent	MSE	*0.0512*	0.0558	0.0944	0.0872
		Present	MSE	*0.0684*	0.0809	0.1901	0.1042
N = 500	0	Absent	MSE	*0.0289*	0.0330	0.0682	0.0427
		Present	MSE	*0.0269*	0.0295	0.0660	0.0383
	0.3	Absent	MSE	*0.0245*	0.0279	0.0702	0.0467
		Present	MSE	*0.0317*	0.0394	0.0943	0.0390
	0.7	Absent	MSE	*0.0259*	0.0283	0.0634	0.0560
		Present	MSE	*0.0370*	0.0443	0.1483	0.0633
N = 1000	0	Absent	MSE	*0.0128*	0.0145	0.0453	0.0219
		Present	MSE	*0.0145*	0.0161	0.0502	0.0249
	0.3	Absent	MSE	*0.0127*	0.0146	0.0523	0.0308
		Present	MSE	*0.0173*	0.0220	0.0713	0.0203
	0.7	Absent	MSE	*0.0121*	0.0129	0.0468	0.0384
		Present	MSE	*0.0189*	0.0247	0.1193	0.0395

*Results in *italic* indicate MSE was closest to zero.

MSE indicates mean squared error; PS, propensity score.

				Generalized PS			
X1, X2	X1, X2, Z	X2	X2, Z	X2, X3	X1, X2, X3	X1, X2	X2
0.0600	0.0624	0.0541	0.0664	*0.0329*	0.0351	0.0428	0.0411
0.0608	0.0630	0.0551	0.0674	*0.0342*	0.0367	0.0451	0.0424
0.0585	0.1322	0.0569	0.0675	*0.0348*	0.0375	0.0449	0.0472
0.0682	0.0722	0.0673	0.0709	*0.0356*	0.0381	0.0578	0.0725
0.0560	0.0635	0.0561	0.0664	*0.0365*	0.0391	0.0422	0.0456
0.0741	0.0884	0.0805	0.0892	*0.0476*	0.0514	0.0719	0.0993
0.0312	0.0329	0.0278	0.0350	*0.0162*	0.0183	0.0218	0.0209
0.0297	0.0310	0.0277	0.0334	*0.0163*	0.0186	0.0227	0.0213
0.0285	0.0311	0.0285	0.0344	*0.0174*	0.0185	0.0217	0.0248
0.0385	0.0400	0.0403	0.0350	*0.0172*	0.0186	0.0361	0.0519
0.0254	0.0299	0.0268	0.0326	*0.0183*	0.0198	0.0215	0.0250
0.0396	0.0469	0.0493	0.0450	*0.0238*	0.0249	0.0455	0.0776
0.0149	0.0155	0.0136	0.0164	*0.0081*	0.0090	0.0110	0.0105
0.0154	0.0161	0.0138	0.0168	*0.0085*	*0.0081*	0.0099	0.0094
0.0145	0.0157	0.0156	0.0184	*0.0087*	0.0090	0.0108	0.0144
0.0232	0.0245	0.0262	0.0195	*0.0084*	0.0093	0.0235	0.0394
0.0123	0.0145	0.0146	0.0174	*0.0093*	0.0095	0.0103	0.0149
0.0237	0.0277	0.0354	0.0238	*0.0116*	0.0122	0.0312	0.0643
0.0591	0.0698	0.0545	0.1030	*0.0329*	0.0344	0.0436	0.0414
0.0643	0.0787	0.0595	0.1152	*0.0348*	0.0362	0.0445	0.0430
0.0618	0.0876	0.0589	0.1106	*0.0343*	0.0374	0.0458	0.0453
0.0745	0.0778	0.0739	0.1252	*0.0375*	0.0404	0.0627	0.0750
0.0562	0.0883	0.0577	0.0977	*0.0373*	0.0388	0.0428	0.0465
0.0798	0.1045	0.0865	0.1732	*0.0490*	0.0531	0.0765	0.1042
0.0335	0.0433	0.0294	0.0696	*0.0177*	0.0186	0.0220	0.0210
0.0302	0.0390	0.0275	0.0666	*0.0163*	0.0171	0.0208	0.0199
0.0287	0.0478	0.0287	0.0675	*0.0172*	0.0183	0.0214	0.0245
0.0404	0.0392	0.0413	0.0784	*0.0183*	0.0196	0.0376	0.0517
0.0287	0.0565	0.0303	0.0652	*0.0187*	0.0194	0.0209	0.0249
0.0445	0.0639	0.0528	0.1207	*0.0254*	0.0268	0.0471	0.0777
0.0149	0.0224	0.0133	0.0463	*0.0083*	0.0089	0.0105	0.0101
0.0166	0.0254	0.0148	0.0511	*0.0088*	0.0093	0.0114	0.0107
0.0150	0.0315	0.0160	0.0491	*0.0087*	0.0092	0.0108	0.0141
0.0230	0.0205	0.0260	0.0555	*0.0090*	0.0098	0.0256	0.0404
0.0131	0.0387	0.0153	0.0470	*0.0090*	0.0094	0.0102	0.0141
0.0260	0.0399	0.0356	0.0893	*0.0124*	0.0134	0.0342	0.0651

When comparing the univariate and generalized PS methods, in which only the covariates related to the outcome were included, the MSE was smaller for the generalized PS in all simulated datasets in scenario 1 and scenario 2 (Table 2).

In addition, if the sample size increased, the MSE decreased in all simulations. If the correlation increased when there was a correlation between the subgroup and covariates, the MSE increased. However, if the correlation between the subgroup and covariates was absent, the MSE showed a rather inconsistent pattern. Comparing the simulations when the correlation between the subgroup and covariates was either present or absent, gave overall lower MSE values if this correlation was absent.

Univariate PS: case study

The univariate PS was applied according to the protocol described by Bartak and colleagues (Bartak et al., 2009). In total, 28 covariates related to outcome (*P* = 0.10) were selected in the PS estimation. We added four sociodemographic variables as these are considered highly relevant in psychotherapy research (Bartak et al., 2009). The subgroup of interest, that is, the severity of problems, was related to treatment assignment ($\chi^2(1)$= 10.80, *P* =.001) and to the outcome (B = .318, *P* = .000; in a linear regression on the outcome). Thus, in applying the PS methods in the case study, we followed the results of scenario 2: for the univariate PS, we excluded the variable that reflected the severity of problems from the PS estimation.

The PS was estimated in a logistic regression analysis on the treatment assignment and no interaction terms between the covariates were added. The distributions of the estimated PS scores showed considerable overlap (Figure 1). A lack of overlap would yield imprecise estimates of the treatment effect (Spreeuwenberg et al., 2010).

The PS was then added to a regression model on the outcome GSI, in which treatment duration, severity of problems, and an interaction term were the independent variables:

$$OUTCOME = \beta_0 + \beta_1 PS + \beta_2 Treatment + \beta_3 Severity + \beta_4 Treatment \cdot Severity \quad (5)$$

If patients had mild problems, long-term treatment yielded more favorable results than short-term treatment (standardized coefficient of 0.092; Table 3). For patients having severe problems, both treatment options were equally effective: for patients having severe problems in the short-term treatment the standardized coefficient was 0.240, whereas for the long-term treatment, the final standardized coefficient was 0.248. The interaction effect was, however, not significant. Excluding this coefficient indicated that long-term treatment was preferred for patients with severe problems (Table 3).

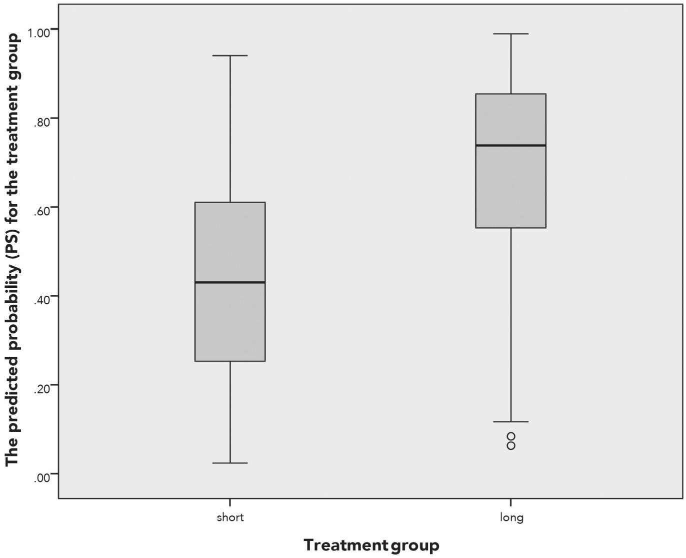


Figure 1. Boxplots of the overlap of the univariate propensity score (PS) distributions
Circle is a mild outlier (defined when the value > 3rd quartile + 1.5 • interquartile range or < 1st quartile - 1.5 • interquartile range).
Aterisk is an extreme outlier (defined when the value > 3rd quartile + 3 • interquartile range or < 1st quartile - 3 • interquartile range).

Table 3. Linear regression on GSI outcome

Variables (N=727)	B	95% CI	Standardized coefficient
Using the univariate PS			
Intercept	0.512**	0.408 – 0.614	-
Long duration treatment group	0.106*	0.006 – 0.206	0.092
Severe problems	0.366**	0.168 – 0.564	0.240
Interaction	-0.147	-0.382 – 0.088	-0.084
Using the generalized PS			
Intercept	0.536**	0.426 - 0.646	-
Short duration – mild problems	Reference	-	-
Short duration – severe problems	0.348**	0.091 – 0.605	0.129
Long duration – mild problems	0.113*	0.011 – 0.215	0.099
Long duration – severe problems	0.244**	0.064 – 0.424	0.139

*p<0.05.
**p<0.01.
CI indicate confidence interval; GSI, Global Severity Index; PS, propensity score.

Generalized PS: case study

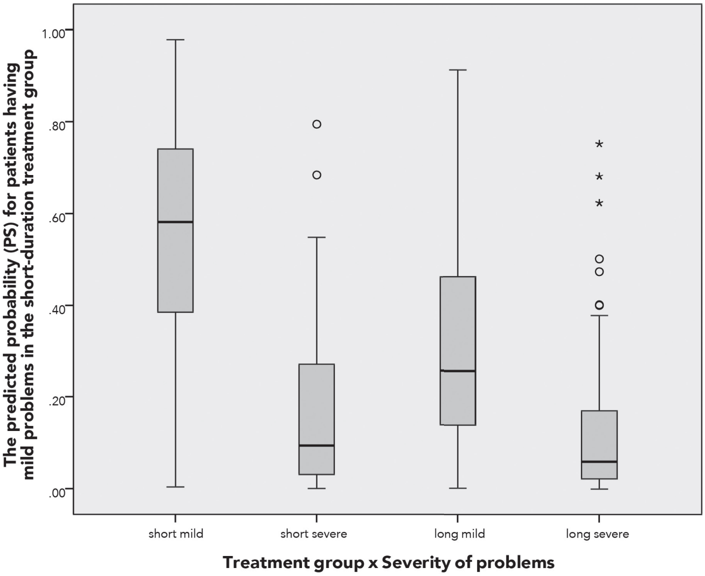
The generalized PS was applied according to the protocol described by Spreeuwenberg and colleagues (2010) The generalized PS was estimated as a combination variable of treatment duration and severity of problems: short-term treatment for patients having mild problems (reference category, N=268), short-term treatment for patients having severe problems (N=34), long-term treatment for patients having mild problems (N=338) and long-term treatment for patients having severe problems (N=87). The same list of covariates selected in the univariate PS was included in the generalized PS estimation. Here, we followed the simulation results of scenario 2. The PS was estimated by multinomial regression analysis, with the combination variable of treatment duration and severity of problems as dependent variable and not including interaction terms between the covariates. However, the number of cases in 2 groups was small and validity of the model fit was therefore uncertain (Austin, 2009; Shah et al., 2005; Stürmer et al., 2006; Weitzen et al., 2004). Because the four estimated generalized PSs add up to 1 and are complementary, only 3 of 4 were used in further analyses. As required when using the PS, the ranges of the estimated PS scores showed overlap (Figure 2).

In the final regression model on the outcome GSI, 3 generalized PSs and 3 dummies indicating group membership were included [Eq. (6)]:

$$OUTCOME = \beta_0 + \beta_1 PS_1 + \beta_2 PS_2 + \beta_3 PS_3 + \beta_4 LongMild_1 + \beta_5 ShortSevere_2 + \beta_6 LongSevere_3$$

These results indicated that long- term treatment was more favorable for patients having mild problems. For patients having severe problems, both treatment options were almost equally effective, just as was presented when the univariate PS was applied while taking the interaction effect into account (Table 3).

To compare the relative effects of this model to the results of using the univariate PS, we used the standardized coefficients. These coefficients can be interpreted independent of the intercept and PS scores added in each model. The coefficient for patients having severe problems in short-term treatment was 0.240 using the univariate versus 0.129 using the generalized PS. The coefficient of patients having mild problems in long-term treatment was almost equal: 0.092 using the univariate PS versus 0.099 using the generalized PS. For patients having severe problems in long-term treatment we combined the standardized coefficients of the model in which the univariate PS was applied and compared it to the corresponding coefficient in the model of the generalized PS. For these patients, the combined coefficient was 0.248 using the univariate versus 0.139 using the generalized PS (Table 3).



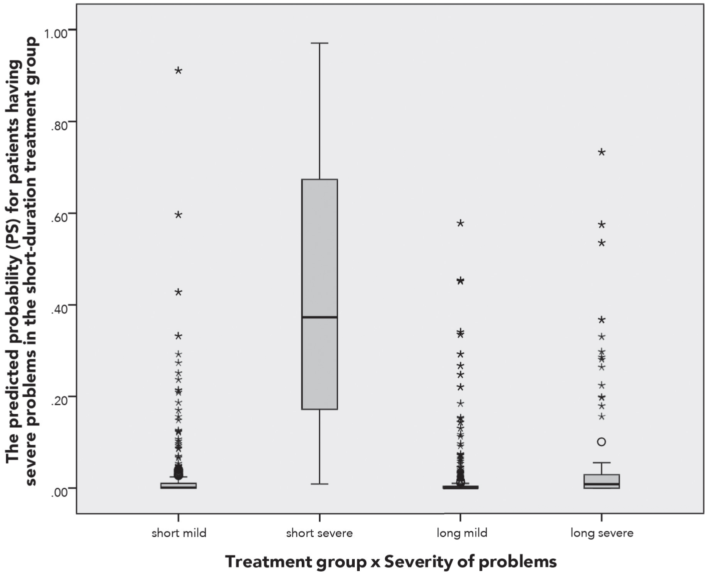
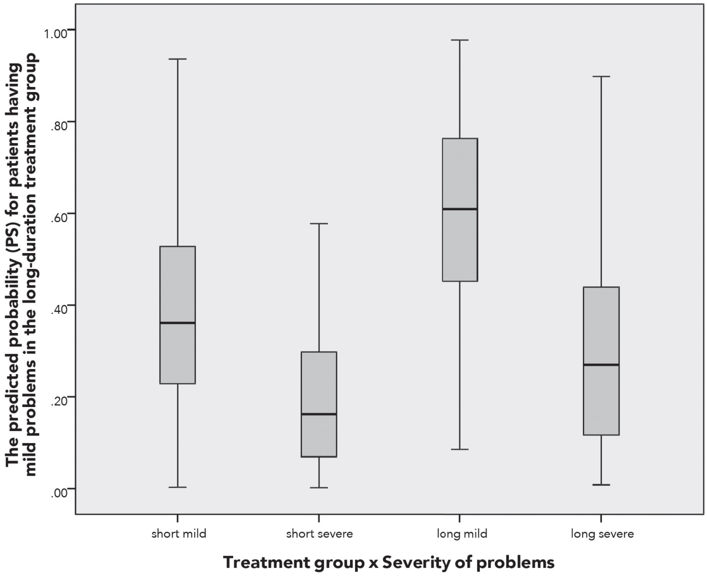


Figure 2. Boxplots of the overlap of the generalized propensity score (PS) distributions
Circle is a mild outlier (defined when the value > 3rd quartile + 1.5 • interquartile range or < 1st quartile - 1.5 • interquartile range).
Aterisk is an extreme outlier (defined when the value > 3rd quartile + 3 • interquartile range or < 1st quartile - 3 • interquartile range).



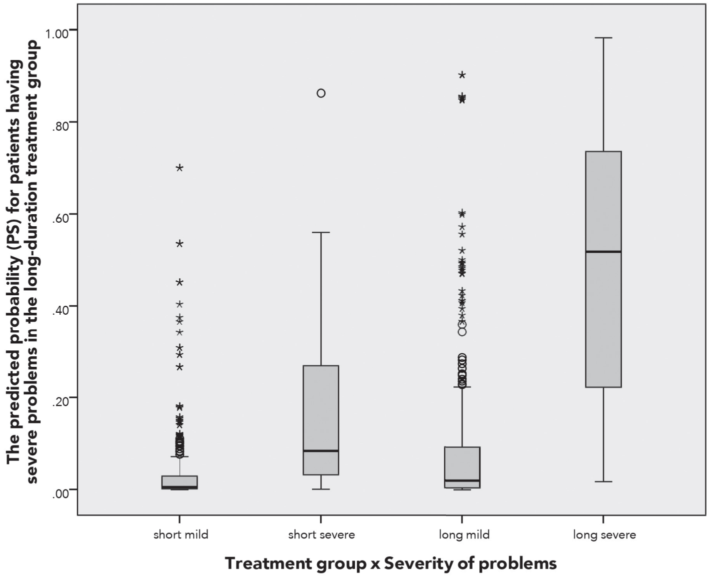


Figure 2. Boxplots of the overlap of the generalized propensity score (PS) distributions
Circle is a mild outlier (defined when the value > 3rd quartile + 1.5 • interquartile range or < 1st quartile - 1.5 • interquartile range).
Aterisk is an extreme outlier (defined when the value > 3rd quartile + 3 • interquartile range or < 1st quartile - 3 • interquartile range).

Discussion

The present study illustrates the use of the univariate and generalized PS in subgroup analyses in non-randomized outcomes research, and describes how the generalized PS could be used in subgroup analysis. The results indicate that the generalized PS – estimated by crossing the treatment options with a subgroup variable – could be a feasible option and should be seriously considered when assessing subgroup effects while correcting for observed pre-treatment differences. In the Monte Carlo simulation study, the generalized PS gave more efficient results overall than the univariate PS, regardless of whether there was a relationship between the subgroup and treatment assignment. In both PS methods, the variables related to the outcome should be included in the PS estimation. These results follow earlier studies of Brookhart and colleagues (2006) and Austin and colleagues (2007) in selecting only the covariates related to the outcome. Furthermore, when the univariate PS was used, the subgroup of interest should be excluded from the PS estimation. Applying the 2 PSs estimations on real-world data produced almost equal model results, illustrating the modifying effect of the severity of problems on the differential effectiveness of 2 psychotherapy treatment arms.

In applying the generalized PS when analyzing subgroups effects, a researcher should take into account additional characteristics of their datasets. Firstly, the characteristics of the subgroup variable should be taken into account. For example, the independence of irrelevant alternatives (IIA) assumption can be violated. This assumption will be violated if, for example, short-term psychotherapy for patients having mild problems is no longer available and this influences the relative risks of the remaining categories. We tested this assumption in our study and it was not violated. However, when it is violated, a nested structure can overcome this violation by first defining the probability that a patient belongs to a particular subgroup and is subsequently assigned to a treatment option. Fuji and colleagues (2012) focused on a 2-stage structure (i.e. a nested structure), in which patients could be assigned sequentially to 2 treatment options, each consisting of 2 sub-options. Yet, if for example a patient characteristic evolves after treatment, it could mediate the relationship between the independent and dependent variable (i.e. a mediator) and should be analyzed differently from the proposed methods (VanderWeele & Vansteelandt, 2009).

Secondly, for various reasons, other application methods when using the PS could be more advantageous (Austin, 2011; Rubin, 1979). For example, the PS can be estimated in each subgroup separately (Kreif et al., 2012; Radice et al., 2012), requiring a large study population to have sufficient power. If the sample size is large enough, a multivariable adjusted model including interaction terms for the subgroup effects can also be a valid alternative (Liem et al., 2010). In this study we applied the PS using covariate adjustment, as sample sizes in clinical practice can be small and this method uses the complete sample size. However, covariate adjustment inherently assumes a correctly specified outcome regression model (Rubin, 2004), whereas for matching this is not required. Inverse probability weighting on the PS is a third and efficient method to control for selection bias (Hirano et al., 2003). The latter 2 methods can indeed eliminate most systematic differences between treated and untreated subjects (Austin, 2009), but matching for example can result in very small comparison groups (Spreeuwenberg et al., 2010). To improve precision of the effect estimates, those methods can also be used in combination with regression analysis (Hirano & Imbens, 2001; Imbens, 2004).

Thirdly, when using the PS, researchers should assess carefully which method is most appropriate for their specific research question. For example, when the treatment effect itself is more important, using the univariate PS and adjusting for extra covariates additionally reduces the bias of the treatment effect estimation (Rubin & Thomas, 2000; Stürmer et al., 2006). However, incorporating effect modification reduces the direct interpretability of the main treatment effect (Liem et al., 2010; Sturmer et al., 2006). Furthermore, if the distributions of the effect modifying variables vary highly, the overall estimates may differ across PS quantiles (Lunt et al., 2009). Different adjustment methods can thus result in divergent results, which all may be correct, but strongly depends on the research question and the population in which the estimation is most suitable (Kurth et al., 2006; Liem et al., 2010).

Our study has several limitations. First, only a basic simulation study was designed. The characteristics of the simulated data were known in advance to the analyzer, which could have influenced the analysis and method chosen. Testing the methods using new simulated data that could be based on a real dataset, is recommended to further investigate the performance of the methods. Furthermore, the number of simulated datasets was rather small, which could have caused the small differences and inconsistencies in the simulation results, due to Monte Carlo error. Secondly, the overlap for the generalized PSs in the case study appeared to be less than optimal (Figure 2). This could have caused the difference in the estimated coefficients when the PS methods were compared. A distance score defined by Cochran and Rubin (1973) can be used to precisely test and define the overlap. A third limitation deals with the selection of variables into the PS. We left out the subgroup variable in the univariate PS, while Rubin and Thomas (2000) state that no prognostic variable should be left out. Although the results of our simulations and the case study only slightly changed when we added the subgroup variable to the univariate PS, we recommend investigating its influence in more detail. Fourth, although we controlled for observed pre-treatment variables, hidden bias due to unobserved confounders could not be controlled for in the case study. As we did not include hidden bias in the simulated datasets either, we do not know the effect of hidden bias in using the PSs in subgroup analysis.

This study supports the idea that the generalized PS can be used in estimating the treatment effect when this is modified by a subgroup variable. As patient-tailored treatment becomes more and more important in outcomes research (Norcross & Wampold, 2011), this study contributes to the literature on how to handle effect estimation in non-randomized outcomes studies of patient subgroups using the PS.

References

American Psychiatric Association (2000). *Diagnostic and statistical manual of mental disorders (4th ed., Text Revision)*. Washington, DC: American Psychiatric Association.

Arrindell, W. A., & Ettema, J. H. M. (2003). *Herziene Handleiding bij de Multidimensionele Psychopathologie-Indicator SCL-90-r [Revised Manual for a Multidimensional Indicator of Psychopathology]*. Lisse, the Netherlands: Swets & Zeitlinger.

Austin, P. C. (2009). The relative ability of different propensity score methods to balance measured covariates between treated and untreated subjects in observational studies. *Medical Decision Making, 29*, 661-677.

Austin, P. C. (2011). An Introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioural Research, 46*, 399-424.

Austin, P. C., Grootendorst, P., & Anderson, G. M. (2007). A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: A Monte Carlo study. *Statistics in Medicine, 26*, 734-753.

Bartak, A., Andrea, H., Spreeuwenberg, M. D., Thunnissen, M., Ziegler, U. M., Dekker, J., . . . Emmelkamp, P. M. (2011). Patients with cluster A personality disorders in psychotherapy: An effectiveness study. *Psychotherapy and Psychosomatics, 80*, 88-99.

Bartak, A., Andrea, H., Spreeuwenberg, M. D., Ziegler, U. M., Dekker, J., Rossum, B. V., . . . Emmelkamp, P. M. (2011). Effectiveness of outpatient, day hospital, and inpatient psychotherapeutic treatment for patients with Cluster B personality disorders. *Psychotherapy and Psychosomatics, 80*, 23-38.

Bartak, A., Spreeuwenberg, M. D., Andrea, H., Busschbach, J. J., Croon, M. A., Verheul, R., . . . Stijnen, T. (2009). The use of propensity score methods in psychotherapy research: A practical application. *Psychotherapy and Psychosomatics, 78*, 26-34.

Bartak, A., Spreeuwenberg, M. D., Andrea, H., Holleman, L., Rijnierse, P., Rossum, B. V., . . . Emmelkamp, P. M. (2010). Effectiveness of different modalities of psychotherapeutic treatment for patients with cluster C personality disorders: Results of a large prospective multicentre study. *Psychotherapy and Psychosomatics, 79*, 20-30.

Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J., & Sturmer, T. (2006). Variable selection for propensity score models. *American Journal of Epidemiology, 163*, 1149-1156.

Cochran, W. G., & Rubin, D. B. (1973). Controlling bias in observational studies: A review. *Sankhya, 35*, 417-446.

Crawford, M. J., Koldobsky, N., Mulder, R., & Tyrer, P. (2011). Classifying personality disorder according to severity. *Journal of Personality Disorders, 25*, 321-330.

D'Agostino, R. B., Jr. (1998). Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in Medicine, 17*, 2265-2281.

Derogatis, L. R. (1986). *SCL-90 (R): Administration, scoring and procedure. Manual-ii for the revised version*. Townson, MD: Clinical Psychometric Research.

Feng, P., Zhou, X. H., Zou, Q. M., Fan, M. Y., & Li, X. S. (2012). Generalized propensity score for estimating the average treatment effect of multiple treatments. *Statistics in Medicine, 31*, 681-697.

Fujii, Y., Henmi, M., & Fujita, T. (2012). Evaluating the interaction between the therapy and the treatment in clinical trials by the propensity score weighting method. *Statistics in Medicine, 31*, 235-252.

Glynn, R. J., Schneeweiss, S., & Sturmer, T. (2006). Indications for propensity scores and review of their use in pharmacoepidemiology. *Basic & Clinical Pharmacology & Toxicology, 98*, 253-259.

Hirano, K., & Imbens, G. W. (2001). Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. *Health Services and Outcomes Research Methodology, 2*, 259-278.

Hirano, K., Imbens, G. W., & Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica, 71*, 1161-1189.

Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika, 87*, 706-710.

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *The Review of Economics and Statistics, 86*, 4-29.

Kreif, N., Grieve, R., Radice, R., Sadique, Z., Ramsahai, R., & Sekhon, J. S. (2012). Methods for estimating subgroup effects in cost-effectiveness analyses that use observational data. *Medical Decision Making, 32*, 750-763.

Kurth, T., Walker, A. M., Glynn, R. J., Chan, K. A., Gaziano, J. M., Berger, K., & Robins, J. M. (2006). Results of multivariable logistic regression, propensity matching, propensity adjustment, and propensity-based weighting under conditions of nonuniform effect. *American Journal of Epidemiology, 163*, 262-270.

Liem, Y. S., Wong, J. B., Hunink, M. M., de Charro, F. T., & Winkelmayer, W. C. (2010). Propensity scores in the presence of effect modification: A case study using the comparison of mortality on hemodialysis versus peritoneal dialysis. *Emerging Themes in Epidemiology, 7*, 1-8.

Lunt, M., Solomon, D., Rothman, K., Glynn, R., Hyrich, K., Symmons, D. P., . . . British Society for Rheumatology Biologics Register Control Centre, C. (2009). Different methods of balancing covariates leading to different effect estimates in the presence of effect modification. *American Journal of Epidemiology, 169*, 909-917.

Norcross, J. C., & Wampold, B. E. (2011). What works for whom: Tailoring psychotherapy to the person. *Journal of Clinical Psychology, 67*, 127-132.

Radice, R., Ramsahai, R., Grieve, R., Kreif, N., Sadique, Z., & Sekhon, J. S. (2012). Evaluating treatment effectiveness in patient subgroups: A comparison of propensity score methods with an automated matching approach. *International Journal of Biostatistics, 8*, 25.

R Development Core Team (2010). R: A language and environment for statistical computing (Version 2.13.0). Vienna, Austria: R Foundation for Statistical Computing.

Rosenbaum, P. R. (1991). Discussing hidden bias in observational studies. *Annals of Internal Medicine, 115*, 901-905.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*, 41-55.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology, 66*, 688-701.

Rubin, D. B. (1979). Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association, 74*, 318-324.

Rubin, D. B. (1997). Estimating causal effects from large data sets using propensity scores. *Annals of Internal Medicine, 127*, 757-763.

Rubin, D. B. (2004). On principles for modeling propensity scores in medical research. *Pharmacoepidemiology and Drug Safety, 13*, 855-857.

Rubin, D. B., & Thomas, N. (2000). Combining propensity score matching with additional adjustments for prognostic covariates. *Journal of the American Statistical Association, 95*, 573-585.

Shah, B. R., Laupacis, A., Hux, J. E., & Austin, P. C. (2005). Propensity score methods gave similar results to traditional regression modeling in observational studies: A systematic review. *Journal of Clinical Epidemiology, 58*, 550-559.

Soeteman, D. I., Verheul, R., Meerman, A. M., Ziegler, U., Rossum, B. V., Delimon, J., . . . Kim, J. J. (2011). Cost-effectiveness of psychotherapy for cluster C personality disorders: A decision-analytic model in the Netherlands. *Journal of Clinical Psychiatry, 72*, 51-59.

Spreeuwenberg, M. D., Bartak, A., Croon, M. A., Hagenaars, J. A., Busschbach, J. J., Andrea, H., . . . Stijnen, T. (2010). The multiple propensity score as control for bias in the comparison of more than two treatment arms: An introduction from a case study in mental health. *Medical Care, 48*, 166-174.

Stürmer, T., Joshi, M., Glynn, R. J., Avorn, J., Rothman, K. J., & Schneeweiss, S. (2006). A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. *Journal of Clinical Epidemiology, 59*, 437-447.

Sturmer, T., Rothman, K. J., & Glynn, R. J. (2006). Insights into different results from different causal contrasts in the presence of effect-measure modification. *Pharmacoepidemioly and Drug Safety, 15*, 698-709.

Tyrer, P. (2004). New approaches to the diagnosis of psychopathy and personality disorder. *Journal of the Royal Society of Medicine, 97*, 371-374.

Tyrer, P., & Johnson, T. (1996). Establishing the severity of personality disorder. *American Journal of Psychiatry, 153*, 1593-1597.

VanderWeele, T. J., & Vansteelandt, S. (2009). Conceptual issues concerning mediation, interventions and composition. *Statistics and Its Interface, 2*, 457-468.

Weitzen, S., Lapane, K. L., Toledano, A. Y., Hume, A. L., & Mor, V. (2004). Principles for modeling propensity scores in medical research: A systematic literature review. *Pharmacoepidemiology and Drug Safety, 13*, 841-853.

Winship, C., & Mare, R. D. (1992). Models for sample selection bias. *Annual Review of Sociology, 18*, 327-350.

Ye, Y., Bond, J. C., Schmidt, L. A., Mulia, N., & Tam, T. W. (2012). Toward a better understanding of when to apply propensity scoring: A comparison with conventional regression in ethnic disparities research. *Annals of Epidemiology, 22*, 691-697.

Supplemental Material

Table I. Bias, standard error (SE) and mean squared error (MSE) of simulations– scenario 1*

Sample size	Correlation covariates X1, X2, X3	Correlation Z - X1, X2,X3		**Variables in propensity score (PS) model**				
				Univariate PS				
				X2, X3	X1, X2, X3	X2, X3, Z	X1,X2,X3,Z	X1, X2
N = 250	0	absent	Bias	***0.0005***	0.0015	0.0014	0.0018	0.0046
			SE	***0.0071***	0.0075	0.0079	0.0076	0.0076
			MSE	***0.0523***	0.0579	0.0642	0.0605	0.0600
		present	Bias	*0.0033*	0.0048	0.0059	0.0057	0.0046
			SE	*0.0071*	0.0074	0.0079	0.0075	0.0077
			MSE	*0.0524*	0.0561	0.0641	0.0583	0.0608
	0.3	absent	Bias	*-0.0041*	-0.0032	-0.0019	-0.0024	**0.0023**
			SE	*0.0069*	0.0074	0.0077	0.0077	**0.0075**
			MSE	*0.0491*	0.0562	0.0618	0.0608	**0.0585**
		present	Bias	*0.0313*	0.0491	-0.0186	0.0478	0.0477
			SE	*0.0071*	0.0076	0.0080	0.0079	0.0077
			MSE	*0.0556*	0.0683	0.0687	0.0724	0.0682
	0.7	absent	Bias	***0.0019***	0.0038	0.0029	0.0042	0.0037
			SE	***0.0068***	0.0071	0.0076	0.0076	0.0073
			MSE	***0.0481***	0.0518	0.0593	0.0592	0.0560
		present	Bias	*0.0319*	0.0611	-0.0255	0.0589	0.0560
			SE	*0.0078*	0.0082	0.0091	0.0088	0.0080
			MSE	*0.0657*	0.0792	0.0926	0.0907	0.0741
N = 500	0	absent	Bias	*0.0015*	0.0005	-0.0014	-0.0010	**0.0001**
			SE	*0.0050*	0.0053	0.0056	0.0055	**0.0055**
			MSE	*0.0261*	0.0294	0.0330	0.0311	**0.0312**
		present	Bias	0.0006	*0.0010*	0.1738	**0.0005**	0.0036
			SE	0.0053	*0.0050*	0.0047	**0.0054**	0.0053
			MSE	0.0292	*0.0264*	0.1095	**0.0306**	0.0297
	0.3	absent	Bias	***0.0003***	-0.0004	-0.0006	-0.0005	0.0008
			SE	***0.0049***	0.0052	0.0055	0.0055	0.0053
			MSE	***0.0245***	0.0285	0.0318	0.0311	0.0285
		present	Bias	*0.0370*	0.0539	**-0.0123**	0.0526	0.0540
			SE	*0.0049*	0.0053	**0.0055**	0.0055	0.0054
			MSE	*0.0294*	0.0381	**0.0328**	0.0394	0.0385
	0.7	absent	Bias	**0.0002**	*0.0005*	-0.0008	-0.0010	-0.0003
			SE	**0.0051**	*0.0049*	0.0055	0.0055	0.0050
			MSE	**0.0270**	*0.0248*	0.0313	0.0315	0.0254
		present	Bias	*0.0264*	0.0546	-0.0297	0.0548	0.0555
			SE	*0.0055*	0.0058	0.0065	0.0063	0.0056
			MSE	*0.0331*	0.0432	0.0489	0.0494	0.0396

			Generalized PS			
X1, X2, Z	X2	X2, Z	X2, X3	X1, X2, X3	X1, X2	X2
0.0051	0.0041	0.0055	***-0.0022***	-0.0035	-0.0066	-0.0088
0.0078	0.0072	0.0080	***0.0057***	0.0059	0.0065	0.0064
0.0624	0.0541	0.0664	***0.0329***	0.0351	0.0428	0.0411
0.0054	**0.0021**	0.0047	***0.0027***	0.0045	0.0009	0.0003
0.0078	**0.0073**	0.0081	***0.0058***	0.0061	0.0067	0.0065
0.0630	**0.0551**	0.0674	***0.0342***	0.0367	0.0451	0.0424
0.1683	0.0266	0.0287	*-0.0063*	0.0081	**0.0062**	0.0595
0.0069	0.0073	0.0080	*0.0059*	0.0061	**0.0067**	0.0065
0.1322	0.0569	0.0675	*0.0348*	0.0375	**0.0449**	0.0472
0.0457	0.0765	**0.0143**	***0.0007***	-0.0029	0.0926	0.1599
0.0080	0.0074	**0.0082**	***0.0060***	0.0062	0.0066	0.0064
0.0722	0.0673	**0.0709**	***0.0356***	0.0381	0.0578	0.0725
0.0043	0.0288	0.0297	***0.0015***	0.0051	0.0049	0.0579
0.0079	0.0072	0.0079	***0.0060***	0.0062	0.0065	0.0064
0.0635	0.0561	0.0664	***0.0365***	0.0391	0.0422	0.0456
0.0520	0.1007	**0.0088**	*0.0087*	**0.0055**	0.1065	0.1999
0.0088	0.0077	**0.0090**	*0.0069*	**0.0071**	0.0073	0.0069
0.0884	0.0805	**0.0892**	*0.0476*	**0.0514**	0.0719	0.0993
-0.0014	-0.0004	-0.0034	*0.0028*	**0.0017**	0.0044	0.0019
0.0057	0.0052	0.0058	*0.0040*	**0.0043**	0.0047	0.0046
0.0329	0.0278	0.0350	*0.0162*	**0.0183**	0.0218	0.0209
0.0033	0.0036	0.0032	***0.0050***	-0.0070	-0.0060	-0.0062
0.0055	0.0052	0.0057	***0.0040***	0.0043	0.0048	0.0046
0.0310	0.0277	0.0334	***0.0163***	0.0186	0.0227	0.0213
0.0006	0.0277	0.0267	***-0.0015***	0.0011	0.0012	0.0515
0.0055	0.0051	0.0056	***0.0042***	0.0043	0.0047	0.0045
0.0311	0.0285	0.0344	***0.0174***	0.0185	0.0217	0.0248
0.0525	0.0839	0.0233	***0.0023***	-0.0002	0.0961	0.1627
0.0056	0.0051	0.0057	***0.0041***	0.0043	0.0047	0.0045
0.0400	0.0403	0.0350	***0.0172***	0.0186	0.0361	0.0519
-0.0014	0.0270	0.0256	***0.0048***	0.0041	0.0043	0.0558
0.0054	0.0049	0.0055	***0.0043***	0.0044	0.0046	0.0045
0.0299	0.0268	0.0326	***0.0183***	0.0198	0.0215	0.0250
0.0550	0.1025	0.0137	***0.0079***	0.0082	0.1103	0.2053
0.0061	0.0054	0.0064	***0.0048***	0.0049	0.0052	0.0049
0.0469	0.0493	0.0450	***0.0238***	0.0249	0.0455	0.0776

Sample size	Correlation covariates X1, X2, X3	Correlation Z - X1, X2,X3		Variables in propensity score (PS) model				
				Univariate PS				
				X2, X3	X1, X2, X3	X2, X3, Z	X1,X2,X3,Z	X1, X2
N = 1000	0	absent	Bias	*-0.0003*	**-0.0002**	0.0009	0.0005	0.0016
			SE	*0.0035*	**0.0038**	0.0039	0.0038	0.0038
			MSE	*0.0132*	**0.0148**	0.0160	0.0154	0.0149
		present	Bias	*0.0009*	0.0003	0.0007	**0.0000**	0.0037
			SE	*0.0035*	0.0037	0.0039	**0.0038**	0.0038
			MSE	*0.0130*	0.0146	0.0160	**0.0152**	0.0154
	0.3	absent	Bias	*0.0001*	0.0008	-0.0011	**0.0000**	-0.0033
			SE	*0.0035*	0.0037	0.0039	**0.0038**	0.0038
			MSE	*0.0124*	0.0141	0.0158	**0.0153**	0.0145
		present	Bias	*0.0345*	0.0502	**-0.0149**	0.0498	0.0501
			SE	*0.0034*	0.0037	**0.0040**	0.0039	0.0039
			MSE	*0.0162*	0.0214	**0.0174**	0.0227	0.0232
	0.7	absent	Bias	*0.0012*	0.0012	0.0027	0.1706	**0.0003**
			SE	*0.0034*	0.0036	0.0039	0.0034	**0.0035**
			MSE	*0.0121*	0.0131	0.0153	0.0964	**0.0123**
		present	Bias	*0.0304*	0.0578	-0.0267	0.0580	0.0521
			SE	*0.0039*	0.0041	0.0046	0.0045	0.0040
			MSE	*0.0174*	0.0256	0.0270	0.0290	0.0237

*Results in **bold** indicate bias was closest to zero, results in i*talic* indicate MSE was closest to zero

			Generalized PS			
X1, X2, Z	X2	X2, Z	X2, X3	X1, X2, X3	X1, X2	X2
0.0023	0.0010	0.0022	***-0.0011***	0.0021	0.0026	0.0014
0.0039	0.0036	0.0040	***0.0028***	0.0030	0.0033	0.0032
0.0155	0.0136	0.0164	***0.0081***	0.0090	0.0110	0.0105
0.0034	0.0028	0.0024	0.0032	***0.0012***	0.0015	0.0013
0.0039	0.0036	0.0040	0.0029	***0.0028***	0.0031	0.0031
0.0161	0.0138	0.0168	0.0085	***0.0081***	0.0099	0.0094
-0.0042	0.0226	0.0216	***-0.0015***	0.0023	0.0018	0.0522
0.0039	0.0036	0.0040	***0.0030***	0.0030	0.0033	0.0032
0.0157	0.0156	0.0184	***0.0087***	0.0090	0.0108	0.0144
0.0496	0.0814	0.0216	***0.0019***	-0.0050	0.0878	0.1543
0.0041	0.0037	0.0042	***0.0029***	0.0030	0.0033	0.0032
0.0245	0.0262	0.0195	***0.0084***	0.0093	0.0235	0.0394
0.0018	0.0277	0.0292	***0.0030***	0.0039	0.0032	0.0563
0.0038	0.0035	0.0038	***0.0030***	0.0031	0.0032	0.0032
0.0145	0.0146	0.0174	***0.0093***	0.0095	0.0103	0.0149
0.0522	0.1013	**0.0129**	*0.0090*	**0.0052**	0.1087	0.2039
0.0044	0.0038	**0.0045**	*0.0034*	**0.0035**	0.0036	0.0034
0.0277	0.0354	**0.0238**	*0.0116*	**0.0122**	0.0312	0.0643

Table II. Bias, standard error (SE) and mean squared error (MSE) of simulations – scenario 2*

Sample size	**Correlation covariates** X1,X2,X3	**Correlation** Z - X1, X2,X3		**Variables in propensity score (PS) model**				
				Univariate PS				
				X2, X3	X1, X2, X3	X2, X3, Z	X1, X2, X3, Z	X1, X2
N = 250	0	absent	Bias	*0.0017*	0.0032	-0.0964	-0.0438	0.0021
			SE	*0.0072*	0.0074	0.0082	0.0077	0.0076
			MSE	*0.0537*	0.0575	0.0994	0.0678	0.0591
		present	Bias	*-0.0038*	-0.0051	-0.1062	-0.0555	-0.0041
			SE	*0.0074*	0.0077	0.0084	0.0080	0.0079
			MSE	*0.0576*	0.0622	0.1116	0.0770	0.0643
	0.3	absent	Bias	***-0.0025***	-0.0032	-0.1120	-0.0736	-0.0039
			SE	***0.0071***	0.0076	0.0082	0.0081	0.0078
			MSE	***0.0525***	0.0604	0.1117	0.0850	0.0618
		present	Bias	*0.0299*	0.0446	-0.1255	**-0.0163**	0.0528
			SE	*0.0074*	0.0079	0.0089	**0.0085**	0.0080
			MSE	*0.0605*	0.0717	0.1382	**0.0763**	0.0745
	0.7	absent	Bias	*0.0037*	**0.0031**	-0.0927	-0.0791	0.0041
			SE	*0.0070*	**0.0074**	0.0080	0.0080	0.0074
			MSE	*0.0512*	**0.0558**	0.0944	0.0872	0.0562
		present	Bias	*0.0285*	0.0540	-0.1354	**-0.0226**	0.0622
			SE	*0.0080*	0.0084	0.0100	**0.0095**	0.0083
			MSE	*0.0684*	0.0809	0.1901	**0.1042**	0.0798
N = 500	0	absent	Bias	*-0.0006*	-0.0001	-0.0998	-0.0482	**-0.0001**
			SE	*0.0053*	0.0056	0.0060	0.0058	**0.0057**
			MSE	*0.0289*	0.0330	0.0682	0.0427	**0.0335**
		present	Bias	*-0.0008*	-0.0014	-0.1006	-0.0497	-0.0012
			SE	*0.0051*	0.0053	0.0058	0.0055	0.0054
			MSE	*0.0269*	0.0295	0.0660	0.0383	0.0302
	0.3	absent	Bias	*0.0001*	-0.0001	-0.1091	-0.0713	**0.0000**
			SE	*0.0049*	0.0052	0.0056	0.0055	**0.0053**
			MSE	*0.0245*	0.0279	0.0702	0.0467	**0.0287**
		present	Bias	*0.0297*	0.0444	-0.1238	**-0.0168**	0.0509
			SE	*0.0052*	0.0056	0.0062	**0.0060**	0.0056
			MSE	*0.0317*	0.0394	0.0943	**0.0390**	0.0404
	0.7	absent	Bias	***0.0013***	0.0015	-0.0976	-0.0838	0.0019
			SE	***0.0050***	0.0052	0.0056	0.0057	0.0053
			MSE	***0.0259***	0.0283	0.0634	0.0560	0.0287
		present	Bias	***0.0248***	0.0488	-0.1445	-0.0339	0.0591
			SE	***0.0058***	0.0060	0.0072	0.0068	0.0059
			MSE	***0.0370***	0.0443	0.1483	0.0633	0.0445

			Generalized PS			
X1, X2, Z	X2	X2, Z	X2, X3	X1, X2, X3	X1, X2	X2
-0.0466	**0.0009**	-0.1033	*0.0046*	0.0067	**0.0041**	0.0030
0.0078	**0.0073**	0.0083	*0.0057*	0.0059	**0.0066**	0.0064
0.0698	**0.0545**	0.1030	*0.0329*	0.0344	**0.0436**	0.0414
-0.0548	**-0.0023**	-0.1084	***-0.0084***	-0.0093	-0.0119	-0.0100
0.0081	**0.0076**	0.0085	***0.0059***	0.0060	0.0067	0.0065
0.0787	**0.0595**	0.1152	***0.0348***	0.0362	0.0445	0.0430
-0.0760	0.0208	-0.0798	*-0.0033*	**-0.0024**	-0.0073	0.0389
0.0082	0.0074	0.0084	*0.0059*	**0.0061**	0.0068	0.0065
0.0876	0.0589	0.1106	*0.0343*	**0.0374**	0.0458	0.0453
-0.0185	0.0809	-0.0847	***0.0049***	0.0058	0.1016	0.1618
0.0086	0.0078	0.0089	***0.0061***	0.0063	0.0069	0.0066
0.0778	0.0739	0.1252	***0.0375***	0.0404	0.0627	0.0750
-0.0796	0.0295	-0.0623	***0.0117***	0.0126	0.0132	0.0584
0.0081	0.0073	0.0082	***0.0061***	0.0062	0.0065	0.0064
0.0883	0.0577	0.0977	***0.0373***	0.0388	0.0428	0.0465
-0.0303	0.1043	-0.0968	*0.0134*	**0.0133**	0.1244	0.2120
0.0094	0.0081	0.0101	*0.0069*	**0.0072**	0.0074	0.0070
0.1045	0.0865	0.1732	*0.0490*	**0.0531**	0.0765	0.1042
-0.0490	-0.0007	-0.1024	***0.0015***	0.0035	0.0031	0.0008
0.0059	0.0053	0.0060	***0.0042***	0.0043	0.0047	0.0046
0.0433	0.0294	0.0696	***0.0177***	0.0186	0.0220	0.0210
-0.0497	**-0.0003**	-0.1015	*0.0017*	0.0008	**0.0007**	0.0014
0.0055	**0.0051**	0.0058	*0.0040*	0.0041	**0.0046**	0.0044
0.0390	**0.0275**	0.0666	*0.0163*	0.0171	**0.0208**	0.0199
-0.0719	0.0240	-0.0749	***0.0019***	0.0024	0.0026	0.0469
0.0056	0.0051	0.0058	***0.0041***	0.0043	0.0046	0.0046
0.0478	0.0287	0.0675	***0.0172***	0.0183	0.0214	0.0245
-0.0198	0.0799	-0.0814	*0.0007*	**0.0006**	0.0962	0.1582
0.0060	0.0054	0.0062	*0.0043*	**0.0044**	0.0049	0.0047
0.0392	0.0413	0.0784	*0.0183*	**0.0196**	0.0376	0.0517
-0.0841	0.0268	-0.0668	***0.0089***	0.0113	0.0093	0.0528
0.0057	0.0052	0.0057	***0.0043***	0.0044	0.0046	0.0045
0.0565	0.0303	0.0652	***0.0187***	0.0194	0.0209	0.0249
-0.0377	0.1033	-0.0979	*0.0077*	**0.0073**	0.1139	0.2033
0.0068	0.0058	0.0072	*0.0050*	**0.0051**	0.0052	0.0050
0.0639	0.0528	0.1207	*0.0254*	**0.0268**	0.0471	0.0777

Sample size	Correlation covariates X1,X2,X3	Correlation Z - X1, X2,X3		Variables in propensity score (PS) model				
				Univariate PS				
				X2, X3	X1, X2, X3	X2, X3, Z	X1,X2,X3,Z	X1, X2
N = 1000	0	absent	Bias	***0.0009***	0.0013	-0.0963	-0.0461	0.0014
			SE	***0.0035***	0.0037	0.0040	0.0038	0.0038
			MSE	***0.0128***	0.0145	0.0453	0.0219	0.0149
		present	Bias	*-0.0011*	**-0.0008**	-0.1020	-0.0499	-0.0013
			SE	*0.0037*	**0.0039**	0.0042	0.0040	0.0040
			MSE	*0.0145*	**0.0161**	0.0502	0.0249	0.0166
	0.3	absent	Bias	*0.0002*	**-0.0001**	-0.1074	-0.0695	-0.0003
			SE	*0.0035*	**0.0037**	0.0040	0.0039	0.0038
			MSE	*0.0127*	**0.0146**	0.0523	0.0308	0.0150
		present	Bias	*0.0296*	0.0437	-0.1210	-0.0167	0.0507
			SE	*0.0037*	0.0039	0.0044	0.0042	0.0039
			MSE	*0.0173*	0.0220	0.0713	0.0203	0.0230
	0.7	absent	Bias	***0.0000***	0.0004	-0.1002	-0.0862	0.0003
			SE	***0.0034***	0.0035	0.0038	0.0038	0.0036
			MSE	***0.0121***	0.0129	0.0468	0.0384	0.0131
		present	Bias	***0.0239***	0.0462	-0.1445	-0.0375	0.0571
			SE	***0.0040***	0.0043	0.0052	0.0049	0.0043
			MSE	***0.0189***	0.0247	0.1193	0.0395	0.0260

*Results in **bold** indicate bias was closest to zero, results in i*talic* indicate MSE was closest to zero

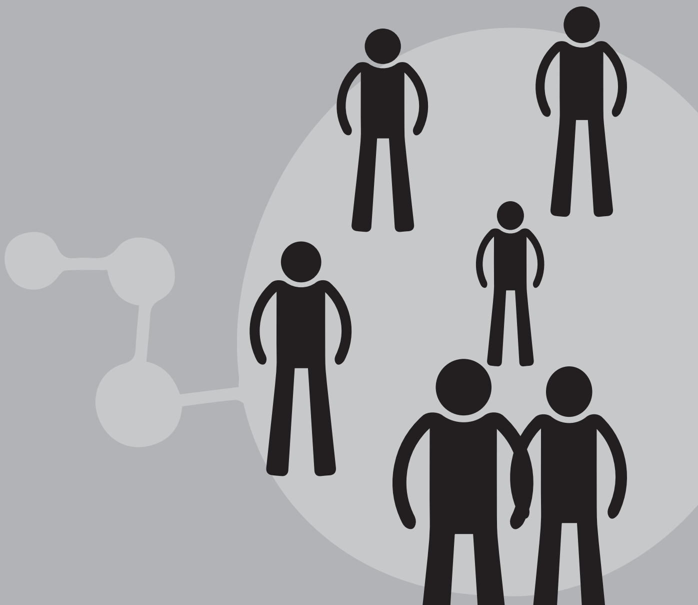
			Generalized PS			
X1, X2, Z	X2	X2, Z	X2, X3	X1, X2, X3	X1, X2	X2
-0.0463	0.0011	-0.0972	***0.0030***	0.0046	0.0036	0.0020
0.0039	0.0036	0.0040	***0.0029***	0.0030	0.0032	0.0032
0.0224	0.0133	0.0463	***0.0083***	0.0089	0.0105	0.0101
-0.0508	-0.0017	-0.1042	***0.0005***	0.0019	-0.0012	-0.0031
0.0041	0.0038	0.0042	***0.0030***	0.0031	0.0034	0.0033
0.0254	0.0148	0.0511	***0.0088***	0.0093	0.0114	0.0107
-0.0702	0.0239	-0.0730	*0.0032*	0.0038	**0.0030**	0.0488
0.0040	0.0037	0.0040	*0.0029*	0.0030	**0.0033**	0.0032
0.0315	0.0160	0.0491	*0.0087*	0.0092	**0.0108**	0.0141
-0.0184	0.0798	-0.0771	***-0.0005***	-0.0006	0.0955	0.1566
0.0042	0.0038	0.0044	***0.0030***	0.0031	0.0034	0.0033
0.0205	0.0260	0.0555	***0.0090***	0.0098	0.0256	0.0404
-0.0867	0.0254	-0.0690	***0.0047***	0.0066	0.0061	0.0516
0.0038	0.0036	0.0039	***0.0030***	0.0031	0.0032	0.0031
0.0387	0.0153	0.0470	***0.0090***	0.0094	0.0102	0.0141
-0.0394	0.1025	-0.0960	***0.0112***	0.0128	0.1174	0.2056
0.0049	0.0041	0.0052	***0.0035***	0.0036	0.0038	0.0036
0.0399	0.0356	0.0893	***0.0124***	0.0134	0.0342	0.0651

Chapter 5.

Multisystemic Therapy and Functional Family Therapy compared on their effectiveness using the propensity score method

Hester V. Eeren, Lucas M.A. Goossens, Ron H.J. Scholte, Jan J.V. Busschbach, &
Rachel E.A. van der Rijken

Submitted



Abstract

Objective: To compare the effectiveness of Multisystemic Therapy (MST) and Functional Family Therapy (FFT) using a quasi-experimental design in the Netherlands.

Method: Between October, 2009 and June, 2014, outcome data were collected from 697 adolescents assigned to either MST or FFT (422 MST; 275 FFT). Data were gathered during Routine Outcome Monitoring. The primary outcome was externalizing problem behavior (Child Behavior Checklist and Youth Self Report). Secondary outcomes were the proportion of adolescents living at home, engaged in school or work, and who lacked police contact during treatment. Because of the non-random assignment, a propensity score method was used to control for observed pre-treatment differences. Because the risk-need-responsivity (RNR) model guided treatment assignment, effectiveness was also estimated in youth with and without a court order as an indicator of their risk level.

Results: In the study sample, no difference was found with regard to externalizing problems. For adolescents without a court order, effects on externalizing problems were larger from MST. Because many more adolescents with a court order were assigned to MST compared to FFT, the propensity score method could not balance the treatment groups in this subsample.

Conclusions: In accordance with previous results, few differences between MST and FFT can be found in the Netherlands. Though treatment assignment was based on the RNR model, results in the group without a court order were not in accordance with this model, while higher-risk adolescents with a court order were indeed more often assigned to the more intensive treatment, namely MST.

Introduction

Multisystemic Therapy (MST) and Functional Family Therapy (FFT) both originated in the US. Their proven effectiveness in reducing adolescents' antisocial behavior and delinquency has led to the worldwide dissemination of these interventions. Both MST and FFT are aimed at reducing the behavioral problems of 12–18 year old adolescents by intervening in the youth's family and environmental system. Functional Family Therapy has an integrated theoretical base in which behavioral techniques, system perspectives, and cognitive theory are combined while remaining informed by intrapsychic perspectives (Breuk et al., 2006; Sexton & Alexander, 2003). Antisocial behavior is thought to be mediated and embedded in a complex sequence of relations between the adolescent and his or her family members (Sexton & Alexander, 2003). Therefore, FFT is specifically aimed at improving family communication and supportiveness while decreasing negativity and dysfunctional behavioral patterns (Blueprints for healthy youth development, 2015). The therapy mainly consists of direct contact with family members, but may be coupled with support system services, such as school or work. Research has shown that FFT is effective in reducing (delinquent) behavioral problems, recidivism, and substance abuse, and that it guides family members in improving their family situation (Alexander & Sexton, 2002; Sexton & Turner, 2010; Sexton & Alexander, 2000)).

Caregivers are also seen as the most important link in the treatment process of MST, but MST also actively involves all other systems surrounding the youth, such as friends, schools, and neighborhoods (Henggeler, Schoenwald, Borduin, Rowland, & Cunningham, 2009). This approach is founded in the social-ecological theory of Bronfenbrenner (Bronfenbrenner, 1979), in which it is thought that antisocial behavior is multi-determined by the different social systems in which an individual acts. By intervening in and with these social systems, risk factors are reduced and a youth's social environment is changed such that it stimulates prosocial activities instead of antisocial behavior (Henggeler et al., 2009). Multisystemic Therapy is more intensive than FFT because a therapist visits the family at home and is available to the family round-the-clock. Research has shown that MST effectively reduces behavioral problems and delinquency, recidivism, substance abuse, out-of-home placement, family problems, and involvement with deviant peers (Henggeler, 2011; van der Stouwe, Asscher, Stams, Deković, & van der Laan, 2014).

The effectiveness of both MST and FFT is well-established compared to regular treatment, such as individual treatment or family-based interventions, such as parenting counseling (Asscher, Dekovic, Manders, van der Laan, & Prins, 2013; Sundell et al., 2008). Multisystemic Therapy and FFT clearly show overlap in their target populations and treatment goals (e.g., Chorpita et al., 2011; Henggeler, 2011; Sexton & Turner, 2010). However, little is known about their relative effectiveness (i.e., whether one intervention outperforms the other). A recent study by Baglivio and colleagues (2014) compared the effectiveness of MST and FFT in juvenile practice in the US. In this study, youth receiving MST or FFT had been referred by probation officers from the juvenile justice department. Results showed little significant difference in the effectiveness of the two interventions. However, low-risk youth receiving FFT committed fewer offenses during treatment than low-risk youth receiving MST.

Because referral practices and treatment populations differ between countries (Asscher et al., 2013; Sundell et al., 2008), the relative effectiveness of MST and FFT is unknown in other countries. In the Netherlands, youth are referred to MST or FFT by various referral agencies, including the Child Protection Council, juvenile judges, local referral institutions, and primary health care providers. To allocate adolescents and their families to either one of the treatments, a well-known model, the Risk-Need-Responsivity (RNR) model, is often used. Following this model, the intensity of the treatment is matched to risks and characteristics of the adolescent. The higher the risk of delinquent behavior, the more intensive treatment should be (Andrews & Bonta, 2010; Andrews, Bonta, & Wormith, 2006). The model implies that adolescents should be assigned to FFT unless there are indications that MST would be more suitable, such as serious delinquent behavior, a high risk that the family cannot provide a safe environment, and an increased risk of recidivism (Oudhof, Ten Berge, & Berger, 2009). In practice, this assignment procedure is followed by clinicians, assigning youth to either FFT or MST. A previous Dutch study comparing both treatment populations found that more youth receiving MST had a court order than youth receiving FFT, and that youth receiving MST had more risk factors than those receiving FFT (Hendriks, Lange, Boonstoppel-Boender, & van der Rijken, 2014). This finding is in accordance with the results of a Swedish study which demonstrated that youth receiving MST had more behavioral problems than youth receiving FFT (Gustle, Hansson, Sundell, Lundh, & Lofhölm, 2006). However, although both European studies showed that the most at risk youth received the most intensive treatment (i.e., MST), the model leaves room for interpretation and may be subject to chance. In fact, the target populations of MST and FFT show substantial overlap (Hendriks et al., 2014). Therefore, it appears that criteria used to allocate adolescents and their families to either one of the treatments are not fully mutually exclusive. Because these studies only looked into treatment populations and did not consider treatment effects, it remains unknown which intervention is the most effective for these overlapping target populations.

Therefore, the current study aimed to investigate the relative effectiveness of MST and FFT in the Netherlands. Because interventions are compared in their everyday practice settings, a quasi-experimental design was used, meaning that youth were not randomly allocated to one of the interventions. Without controlling for pre-treatment differences, a difference in outcomes is either caused by the intervention itself, or by pre-treatment characteristics of adolescents and their families. Therefore, a propensity score (PS) was estimated and used to control for this 'allocation bias'. Using the PS, the treatment arms can be balanced from a large set of observed, pre-treatment characteristics (Austin, 2011; Rosenbaum & Rubin, 1983; Rubin, 2001). When all important covariates are measured, applying the PS to achieve the balance of the treatment arms enables controlling for allocation bias and may even yield results equivalent to randomized studies (Austin, 2011; Shadish, 2013; West et al., 2014). If randomization is not feasible in clinical practice, the use of a PS is a valid solution (Shadish, 2013; West et al., 2014). It should, however, be noted that, in contrast to randomized studies, a PS can only control for overt bias (i.e., bias due to observed pre-treatment differences) and not for hidden bias (i.e., bias due to unmeasured or unobserved differences) (Rosenbaum, 1991). Furthermore, for each adolescent, there must be a chance of being in either treatment group (Shadish, 2013).

The use of a PS in psychological research has increased in recent years (e.g., Austin, 2011; Green & Stuart, 2014; Thoemmes & Kim, 2011; West et al., 2014). The current study used these tutorials and literature as a starting point in comparing MST and FFT. Because previous research has shown that youth receiving MST were more at risk than youth receiving FFT (Gustle et al., 2006; Hendriks et al., 2014), and because the only study to directly compare the effectiveness of FFT and MST thus far takes risk level into account as well (Baglivio et al., 2014), the current study compared the treatment effects not only for the whole sample, but also in two subsamples of youth: with and without a court order. Having a court order can be interpreted as a risk factor and indicate the risk level of an adolescent before treatment. Based on this model, FFT could be sufficiently effective in the group of adolescents without a court order since FFT is expected to be less intensive. Multisystemic Therapy, on the other hand, could be more effective in adolescents with a court order. Since MST and FFT are both aimed at reducing behavioral problems, the primary outcome measure was externalizing problem behavior. Secondary outcomes were the proportion of youth living at home (i.e., the adolescent had not been placed out of home), engaged in school or work at the end of treatment, and without new police contact during the treatment period.

With a growing body of research examining evidence-based treatment, and given today's stringent health care budgets, it seems only logical to allocate youth to a more intensive and likely more expensive treatment only when there is no effective alternative. By comparing evidence-based interventions, budget allocation and the assignment of youth to proper interventions can be optimized.

Methods

Participants and Procedure

As part of the treatment procedure, adolescents and their families filled in questionnaires for Routine Outcome Monitoring (ROM) at the beginning of and after completing treatment. Routine Outcome Monitoring is a measurement system to routinely collect data on the outcome of treatment, evaluate individual treatment progress, and provide transparency regarding the effectiveness of treatment (Buwalda, Nugter, Swinkels, & Mulder, 2011). Within ROM, adolescents and their families provide consent concerning the collection of data and its use for quality control and research. The Medical Ethical Committee of the Erasmus Medical Centre approved this study (METC-2015-124).

Between October, 2009 and June, 2014, 1,714 adolescents and their families began either FFT (N=640) or MST (N=1074) at De Viersprong, institute for personality disorders and behavioral problems in the Netherlands. After finishing treatment, 697 (40.7%) participants completed the primary outcome measure on the Child Behavior Checklist (CBCL) (275 [43%] adolescents who had received FFT and 422 [39.3%] adolescents who had received MST). Such a low percentage of completed questionnaires after treatment is not uncommon within ROM because data is not gathered for specific research purposes (Stichting Benchmark GGZ (SBG), 2016). To reduce uncertainty in the statistical analyses and results, these 697 families formed the study sample for the statistical analyses. Adolescents who had received FFT and completed the primary outcome measure differed significantly from those who did not with regard to their country of birth, living situation, and whether or not they had a court order before treatment (see Table I in Supplemental Material). Adolescents who received MST and completed the assessment after finishing treatment differed from those who did not with regard to their country of birth, living situation, engagement in school or work, whether or not they had a court order before treatment, as well as the country of birth, level of education, and employment status of their primary caregiver, and whether or not this primary caregiver had a partner (see Table II in Supplemental Material).

In addition to the study sample of 697 adolescents, the effectiveness of the treatments was compared between the two subsamples of youth with and without a court order. Of the 422 adolescents who received MST, 246 had a court order and 168 did not (for 10 adolescents [2 FFT; 8 MST], the judicial status was unknown). For FFT, 71 adolescents had a court order, while 202 did not.

Because the assignment procedure following the RNR model implies that adolescents should be assigned to FFT unless there are indicators that MST would be more suitable (Oudhof et al., 2009), FFT was considered the reference treatment and MST the 'new' treatment. Both interventions are continuously monitored on their fidelity and implementation research has shown that both MST and FFT are provided with fidelity in the Netherlands (Manders, Deković, Asscher, van der Laan, & Prins, 2011; van der Rijken, 2015).

Instruments

Baseline measurements

To correct for initial differences between treatment groups, an extensive set of questionnaires were completed at the beginning of treatment. The therapist recorded several demographics of the adolescents and their primary caregiver. Age, gender, country of birth, living situation, level of education, previous treatment, engagement in school or work, previous court orders, police contacts, and the relation with their father, mother, siblings, and peers were reported for each adolescent. Furthermore, the country of birth, level of education, employment status, and presence of a partner were reported for the primary caregiver (Praktikon/MST-NL, 2012). Table 1 shows all demographic characteristics at baseline separately for both treatment groups.

Furthermore, parents completed the CBCL (Achenbach & Rescorla, 2001; Dutch version by Verhulst & van der Ende, 2001a) and the youths themselves completed the Youth Self Report (YSR) (Achenbach & Rescorla, 2001; Dutch version by Verhulst & van der Ende, 2001b). A youth's internalizing problem behavior, externalizing problem behavior, and the total score of the problem behavior were used for analyses. On both questionnaires, items were completed on a 3-point scale (ranging from 0 = never to 2

= often). T-scores were computed and used for analyses. A higher T-score indicates that an adolescent has more problems. Both CBCL and YSR scales were used to measure problem behavior from different perspectives. The Cronbach's alpha coefficients of the study sample for internalizing, externalizing, and total problem behavior measured with the CBCL were .88, .93, and .96 respectively. For the YSR these coefficients were .92, .90, and .95 respectively. The Cronbach's alpha coefficients found in the study sample were similar to those reported in the CBCL and YSR manual (i.e., CBCL:.90, .94, and .97, YSR: .90, .90, and .95) (Achenbach & Rescorla, 2001).

Finally, until September, 2012, parenting stress was measured with the 'Nijmeegse Ouderlijke Stress Index' (NOSI-R) (De Brock, Vermulst, Gerris, Veerman, & Abidin, 2004) in which the primary caregiver completes 42 items on a 4-point scale (ranging from 1 = fully disagree to 4 = fully agree). These items are used to estimate a score for parenting stress wherein a higher score indicates more stress. The reliability coefficient was .95. From October, 2012 onwards, the 'Opvoedingsbelasting Vragenlijst' (OBVL) (Vermulst, Kroes, De Meyer, Nguyen, & Veerman, 2012) was used to measure parenting stress. For this measure, the primary caregiver completes 34 items on a 4-point scale (ranging from 1 = not true, to 4 = true). The scores of all items are summed for a total score regarding parenting stress. The alpha coefficient for this measure was .94. Because parenting stress was measured with two different questionnaires, the deviance score of the scales was used to express the level of parenting stress for both questionnaires in one score concerning parenting stress. This was estimated by subtracting the normscore from the score of the adolescent and dividing this by the standard deviation of the norm group.

Treatment variables, such as length of treatment and dosage of treatment, were not controlled for in the propensity score since these treatment characteristics are part of the treatment itself and the treatment is adapted to the specific situation of the adolescent and his or her family.

Outcome measures

Because both FFT and MST are primarily aimed at reducing externalizing problem behavior, this was defined as the primary outcome measure and was measured with the CBCL and with the YSR (Achenbach & Rescorla, 2001). The primary caregiver reported the externalizing problems of the adolescent with the CBCL, while the youth reported this behavior with the YSR. Both measures were completed at the start of and the end of treatment by completing 35 items on a 3-point scale (ranging from 0 = never to 2 = often). T-scores were computed and used for the analyses. A higher T-score indicates that an adolescent has more problems. The alpha reliability coefficient for the current sample at the end of the treatment with the CBCL is .94. For externalizing problems with the YSR, it is .88.

Table 1. Baseline differences between adolescents assigned to FFT and MST and standardized bias in full sample (N=697)

Variable		**FFT**	***(N = 275)***	
Continuous variables		Mean	SD	N
Age		15.9	1.59	275
CBCL	Internalizing problems	62.51	9.26	263
Primary outcome	Externalizing problems	67.08	9.57	263
	Total behavioral problems †	66.04	8.61	263
YSR	Internalizing problems	54.79	11.31	246
	Externalizing problems	59.27	9.73	246
	Total behavioral problems	57.35	9.78	246
Parenting stress		1.97	1.78	258
Categorical variables		%		N
Gender	Male	53.6		141
	Female	46.4		122
Country of birth	Netherlands	95.8		253
	Western country	1.1		3
	Non-Western country	3.0		8
Living situation adolescent	Together with one parent	36.1		97
	Together with multiple parents	60.6		163
	Other	3.3		9
Living situation adolescent	Lived not at home	0.8		2
Secondary outcome	Lived at home	99.2		260
Level of education	None	7.1		19
	Primary education	3.7		10
	Lower secondary education	54.5		146
	Higher secondary education	34.7		93
Previous treatment	Absent	9.8		26
	Present	90.2		240
Engagement in school or work	Absent	14.5		37
Secondary outcome	Present	85.5		219
Court order	No	74.0		202
	Civil	10.6		29
	Criminal	15.4		42
Police contacts during treatment	Absent	66.9		176
Secondary outcome	Present	33.1		87
Relation father	Absent	6.8		17
	Present	93.2		234
Relation mother	Absent	0.4		1
	Present	99.6		249
Relation siblings	Absent	7.6		18
	Present	92.4		218

MST	*(N = 422)*		Test statistic	Standardized bias	
Mean	SD	N	T-test	Before PS application	After PS application
15.67	1.35	422	1.96	0.17	0.12
61.04	9.68	409	1.95	0.15	0.01
68.29	10.06	409	-1.56	0.12	0.07
65.32	9.76	409	1.00	0.07	0.01
50.78	11.5	356	4.24***	0.35	0.11
57.54	10.87	356	2.04*	0.16	0.04
53.59	11.02	356	4.40***	0.34	0.07
2.06	2.07	397	-0.61	0.05	0.06
%		N	Chi-Square statistic	Before PS application	After PS application
67.2		275	12.60***	0.29	0.23
32.8		134		0.29	0.23
83.4		341	24.04***	0.19	0.06
4.6		19		0.05	0.04
12.0		49		0.13	0.02
42.9		179	6.93*	0.12	0.06
51.1		213		0.16	0.04
6.0		25		0.05	0.02
2.9		12	3.64	0.13	0.01
97.1		400		0.13	0.01
13.7		56	32.55***	0.08	0.03
2.7		11		0.01	0.01
66.8		274		0.15	0.11
16.8		69		0.21	0.07
5.5		23	4.46*	0.19	0.01
94.5		395		0.19	0.01
22.6		91	6.61**	0.19	0.16
77.4		312		0.19	0.16
40.6		168	75.91***	0.41	0.05
30.9		128		0.25	0.18
28.5		118		0.16	0.13
50.8		198	16.74***	0.32	0.01
49.2		192		0.32	0.01
9.2		37	1.22	0.09	0.15
90.8		364		0.09	0.15
0.7		3	0.30	0.04	0.06
99.3		402		0.04	0.06
6.0		23	0.64	0.07	0.02
94.0		361		0.07	0.02

Variable		**FFT** *(N = 275)*	
Categorical variables		%	N
Relation peers	Absent	0.0	0
	Present	100.0	249
Country of birth primary caregiver	the Netherlands	88.5	232
	Western country	4.2	11
	Non-Western country	7.3	19
Level of education primary caregiver	None	1.2	3
	Primary education	4.1	10
	Lower secondary education	27.9	68
	Higher secondary education	45.5	111
	Higher education	21.3	52
Employment primary caregiver	Employed	71.8	186
	Unemployed	28.2	73
Partner primary caregiver	Absent	21.7	55
	Present	78.3	198

** p < .05, ** p < .01, *** p < .001*

† Not selected for PS estimation.

NOTE:

Values depict the mean values and standard deviations. Except for age and parenting stress all other scores are standardized T-scores, having a mean of 50 and a standard deviation of 10. For NOSI-R and parenting stress, normed z-scores are displayed.

MST	(N = 422)	Test statistic	Standardized bias	
%	N	Chi-Square statistic	Before PS application	After PS application
1.3	5	3.15	0.11	0.00
98.7	393		0.11	0.00
79.3	325	11.28**	0.13	0.02
4.9	20		0.01	0.02
15.9	65		0.12	0.01
3.0	12	8.17	0.02	0.05
8.0	32		0.04	0.00
31.3	126		0.04	0.01
39.1	157		0.07	0.04
18.7	75		0.03	0.02
61.9	253	6.98**	0.21	0.17
38.1	156		0.21	0.17
23.9	94	0.41	0.05	0.01
76.1	299		0.05	0.01

Three secondary outcome measures were assessed at the end of the treatment: 1) whether or not the youth was living at home (i.e., the adolescent had not been placed out of home); 2) whether or not the adolescent was engaged in school or work for at least 20 hrs/week at the end of the treatment; and 3) whether or not the adolescent had new police contact due to inappropriate or illegal behavior during the treatment period. The therapist registered these treatment outcomes after treatment and in consultation with the primary caregiver. These three outcomes have been operationalized and standardized by MST Services to ensure that these outcomes are scored identically by all therapists (Institute, 2016). This scoring procedure was also followed by FFT. The quality assurance systems of both treatments ensure that their ultimate outcomes are monitored by the therapist, the team supervisor, and the team consultant.

Statistical analysis

Development of the propensity score

The PS is defined as the conditional probability of assignment to an intervention given a set of observed, pre-treatment variables (Rosenbaum & Rubin, 1983). Moreover, the PS is a balancing score which can be used to achieve a balanced distribution for the observed covariates of the treated and control group (Austin, 2011; Rubin, 2001). The PS was estimated in a univariate logistic regression function for the intervention groups. Here, MST is considered the treated group (coded as 1), and FFT the comparison group (coded as 0). This is because, according to the RNR model, adolescents should be assigned to FFT unless there are serious indications to assign an adolescent to MST (Oudhof et al., 2009).

The observed pre-treatment variables of adolescents are the independent variables added to the model (Austin, 2011; D'Agostino R.B., 1998; Thoemmes & Kim, 2011). These variables, the potential confounders, were selected for the PS model based on clinical knowledge and their expected relation to at least the outcome, and possibly to the treatment itself (Ali et al., 2015; Austin, 2011; Brookhart et al., 2006; Stuart, 2010). Variables solely related to treatment assignment or influenced by treatment should not be included in a PS model (Ali et al., 2015; Austin, 2011; Brookhart et al., 2006).

Weighting by the propensity score

The PS was applied by weighting groups by the odds of their estimated PS score (Stuart, 2010). Weighting by their odds was preferred because there were more treated MST cases than control FFT cases and the interest lies in the average treatment effect in the treated (ATT) rather than the average treatment effect (ATE) (Stuart, 2010). The ATT is the average effect that would be found if all adolescents treated with MST had been treated with FFT. The ATE, however, estimates the average effect if all adolescents (MST and FFT) had received MST compared to all of them received FFT (Harder, Stuart, & Anthony, 2010). In other words, because the ATT is estimated, treatment effects for adolescents who received MST are compared with treatment effects that would have been found had they received FFT (Harder et al., 2010; Stuart, 2010). The MST group was therefore weighted with 1, while the FFT group was weighted with the odds of the PS, that is, the

PS score divided by one subtracted by the PS score (Harder et al., 2010). The PS scores that showed no overlap in the treatment groups were removed. The treatment effect was then estimated in the sample containing overlap with the estimated PSs. Though this restricts the generalizab7ility of the results to cases for which overlap is present, removing cases without overlap allows for more precisely balancing the treatment arms (Harder et al., 2010).

Missing indicator approach

The baseline covariates in the dataset of 697 adolescents who completed either FFT or MST had missing values. To manage these missing values, a missing indicator approach was used while estimating the PS (Cham & West, 2016; D'Agostino, Lang, Walkup, Morgan, & Karter, 2001; Harder et al., 2010; Rosenbaum & Rubin, 1984; West et al., 2014). This method can be theoretically justified and works well to balance observed and missing value patterns across treatment groups without removing cases from the analysis (Cham & West, 2016; Harder et al., 2010; Rosenbaum & Rubin, 1984). In applying this method, the covariate and a missing indicator for this covariate were included in the PS estimation, coded 1 if there was a missing value for the covariate and 0 if not (D'Agostino et al., 2001; Haviland, Nagin, & Rosenbaum, 2007; Rosenbaum, 2010). The missing values of the covariates included in the PS were replaced with an arbitrary value in the range of the values of the covariate itself (Haviland et al., 2007; Rosenbaum, 2010). Using a missing indicator and the covariate with substitution of missing values in the PS estimation enables the use of all cases and balances observed values in the covariates, as well as the missing patterns of these covariates. After PS estimation, balance was assessed for the missing indicators and covariates without missing value substitution. In estimating treatment effects, the missing value substitution was also removed. Thus, this substitution does not affect the evaluated balance, nor does it affect the treatment effect estimation. Furthermore, missing indicators were not taken into account in estimating treatment effects (Haviland et al., 2007; Rosenbaum, 2010).

Balance assessment

An important step in applying the PS is to assess the balance of the observed covariates between the two treatment arms instead of assessing the parameter estimates of the PS model itself (Stuart, 2010). Balance was evaluated for the covariate without missing value substitution and for the missing indicators of the covariates (Harder et al., 2010; Haviland et al., 2007). Balance is achieved when the distribution of the baseline covariates is similar for the two interventions. Balance was assessed with the standardized bias which is independent of the sample size of the study. It was calculated by dividing the difference of the means of the covariates between the treated (MST) and comparison (FFT) group by the standard deviation of the treated group (Ali et al., 2015; Austin, 2009; Austin, 2011; Harder et al., 2010; Stuart, 2010; West et al., 2014). As such, the difference in means was divided by the standard deviation of the MST group (Harder et al., 2010; McCaffrey et al., 2013). For the categorical covariates, the standardized bias was estimated per level. For instance, if the covariate had three levels, the standardized bias was calculated for all three (Harder et al., 2010).

The standardized bias was assessed before and after applying the PS to determine whether balance was achieved. The balance of the baseline covariates and missing indicators was assessed in the weighted sample. As a rule of thumb, it was assumed that balance was achieved when the standardized bias was less than .25 (Harder et al., 2010; Ho, Imai, King, & Stuart, 2007; West et al., 2014). The standardized bias of all covariates was carefully evaluated in addition to the balance of important, prognostic covariates (Ho et al., 2007).

In addition to the standardized bias, the variance ratio and the five-number-summary of the continuous covariates were assessed to determine whether these distributions were similar in higher order moments (Austin, 2009). The distributions of the estimated variances are assumed to follow an F-distribution (Austin, 2009). The 2.5th and 97.5th percentiles can serve as a guide as to which variance ratios are tested to be equal between the treatment groups (Austin, 2009). The five-number summaries should also be used as a qualitative assessment because there is no method to test the similarity of these summaries between treatment groups (Austin, 2009).

Analysis of treatment effect

Regression analysis was used to estimate treatment effect estimates in the weighted sample. The treatment effect on the primary outcome measure — externalizing problem behavior measured with the CBCL — was estimated with an OLS regression on the outcome and the treatment indicator as an independent variable. The effect of interventions on the secondary outcome measures—living at home, being in school or having a job, new contact with the police—was analyzed with logistic regression analyses. The results were used to calculate average risk differences and risk ratios, as these measures are collapsible among subgroups. Odds ratios are not collapsible, meaning they are not comparable when they result from analyses with different sets of covariates or over different subgroups (Goossens, Redekop, & van Gils, 2015). These measures were estimated using ordinary cross tabs of the outcomes and treatment indicators in the weighted sample. For example, for the outcome 'living at home after treatment', the risk ratio was estimated as the probability of living at home after MST divided by the probability of living at home after FFT. The risk difference is the difference between these probabilities, estimated as the probability of living at home after MST minus the probability of living at home after FFT. For 'engaged in school or work' and 'new police contacts', the probability of being engaged in school or work and of having had police contact during treatment were looked at. The 95% confidence intervals of the final treatment effects were estimated using simple bootstrapping, as advised by Austin and Small (2014). In total, 5,000 bootstrap samples were drawn from the weighted sample and in each bootstrapped sample, treatment effects were estimated as described. A nonparametric percentile-based approach was used to define the 95% interval (Austin & Small, 2014).

Regression analyses were done with and without adjustment for the covariates used to calculate the PS. Analyses with the treatment indicator as the only covariate in the study sample (N=697) were followed by analyses in the complete case sample (N=361; 132 FFT and 229 MST)—with and without all covariates—to overcome possible misspecification of this model and to assess whether results were robust (Harder et al., 2010; Rubin & Thomas, 2000). A case was determined 'complete' when all baseline data were available and there were no missing values for the baseline variables. [1] When adding covariates to the regression analyses in the complete case sample, backward selection of the covariates on the treatment effect was used to find a parsimonious model for the outcome. For the secondary outcomes, it was not possible to also control for covariates, as the number of events needed per covariate in a logistic regression was not met (Peduzzi, Concato, Kemper, Holford, & Feinstein, 1996). Therefore, treatment effects on the secondary outcomes in the complete case sample were estimated without controlling for additional covariates.

Subgroup effects

Finally, within the study sample, analyses were repeated for the subsamples of youth who had a court order (246 adolescents assigned to MST; 71 FFT) and youth who did not have a court order (168 MST; 202 FFT). Within the complete case sample, analyses were also repeated for the subsamples of youth with (125 MST; 26 FFT) and without (104 MST; 106 FFT) a court order. Here, the PS within each subsample was estimated separately, as recommended by Green and Stuart (2014). Within each subsample, again the balance between the treatment arms was assessed and then the PS was applied by weighting groups by the odds of the estimated PS score (Green & Stuart, 2014).

The analyses were performed with IBM SPSS for Windows, version 22 (IBM Corp, 2013) and Microsoft Excel (2013). The 95% confidence intervals were bootstrapped in Stata 12 (StataCorp, 2011).

Results

This section first describes the sample characteristics, then the balance in the covariates, and finally the treatment effect for respectively all adolescents in the study sample (N = 697), the subsample of adolescents without a court order (N = 370), and the subsample of adolescents with a court order (N = 317). Within each sample, the results of the complete cases analyses are also described.

Study sample: All adolescents

Within the study sample of 697 adolescents, 422 completed MST and 275 completed FFT. Of the adolescents who completed MST, 67.2% were male and 83.4% were born in the Netherlands. For FFT, 53.6% of the adolescents were male and 95.8% were born in the Netherlands (see Table 1 for an extensive comparison of the treatment arms). Comparing the treatment groups on baseline characteristics showed substantial differences in internalizing, externalizing, and total behavioral problems reported by adolescents. Furthermore, treatment groups differed in gender, country of birth, the adolescent's living situation, level of education, previous treatment, engagement in school or work, previous court order, previous police contact, and country of birth and employment status of the primary caregiver (Table 1).

[1] Because the complete case sample had no missing values on any of the covariates, no missing indicators were needed in estimating the PS and assessing balance of the covariates between the treatment arms

Balance assessment

Before the PS application, balance was assessed in all measured baseline characteristics. Table 1 represents the standardized biases. The largest imbalances were found for internalizing problems reported on the YSR (mean T-score: 54.79 for FFT and 50.78 for MST), total behavioral problems measured with the YSR (mean T-score: 57.35 for FFT and 53.59 for MST), gender (53.6% male for FFT and 67.2% male for MST), previous court order (26% had a court order for FFT and 59.4% had a court order for MST), and having police contact before treatment (33.1% had police contact for FFT and 49.2% had police contact for MST) (Table 1). The standardized bias of these baseline variables was higher than the accepted .25 (Table 1).

After weighting, balance for all of the covariates was found when the PS model contained all covariates except for the total score of behavioral problems measured by the CBCL (Table 1). Balance was inspected in the sample with overlapping PS scores. As a result, 8 MST and 12 FFT cases were removed from the resulting sample. As Table 1 shows, values for the standardized bias after PS application are all lower than .25. The values of the standardized bias for the missing indicator variables were also lower than .25 (Table III in Supplemental Material shows standardized bias for missing indicators before and after applying the PS).

Table 2 shows the variance ratio and five-number summaries of the continuous variables as additional measures for inspecting balance. In the weighted sample, the 2.5th and 97.5th percentiles are .78 and 1.22. The estimated variance ratios are within these boundaries, and thus equality between treatment groups using this measure can be assumed. Moreover, the five-number summaries are also roughly equal in the PS weighted sample (Table 2).

In the complete case sample of 361 adolescents and their families, balance was achieved when the variables of age, internalizing and externalizing problems measured with the CBCL, parenting stress, gender, country of birth, previous treatment, engagement in school or work, court order, police contacts, and employment status of the primary caregiver were included in the PS estimation. Because balance was inspected in the sample with overlapping PS scores, 49 MST and 3 FFT cases were removed from this sample.

Analysis of treatment effect

After assessing the balance, the effectiveness of MST and FFT was compared in the outcome model. Table 3 shows no difference in externalizing problem behavior (CBCL: 0.14; 95% CI -3.23 – 3.49, YSR: -0.29; 95% CI -2.45 – 1.90), with a small effect size of *d* = 0.01 and *d* = 0.03, respectively. The risk ratios (RR) and risk differences (RD) of the secondary outcomes showed no differences between MST and FFT for the proportion of youth living at home and having had police contact (Table 3). However, a significantly higher proportion of adolescents who had completed MST were engaged in school or work after treatment compared with FFT (RR 1.27; 95% CI 1.06 – 1.57, RD 19.2%; 95% CI 5.2% - 32.9%) (Table 3).

Table 2. Variance ratio and 5-number summary of continuous covariates after PS application in full sample (N = 697)

			Variance ratio‡	Minimum	25th percentile	Median	75th percentile	Maximum
Age		FFT	0.79	12.10	14.76	15.95	16.79	20.39
		MST		11.07	14.80	15.83	16.72	18.34
CBCL	Internalizing problems	FFT	0.96	33.00	55.00	61.00	68.00	88.00
		MST		33.00	55.00	62.00	69.00	82.00
	Externalizing problems	FFT	0.92	34.00	61.00	70.00	74.00	92.00
		MST		34.00	63.00	69.00	75.00	88.00
	Total behavioral problems	FFT	0.87	24.00	60.00	68.28	71.00	85.00
		MST		27.00	60.00	67.00	72.00	83.00
YSR	Internalizing problems	FFT	0.97	30.00	44.00	54.00	61.00	83.00
		MST		27.00	44.00	50.00	58.00	85.00
	Externalizing problems	FFT	1.02	29.00	52.00	59.00	66.00	80.00
		MST		29.00	51.00	58.00	66.00	93.00
	Total behavioral problems	FFT	1.06	28.00	47.00	56.00	62.00	77.00
		MST		26.00	46.00	54.00	62.00	82.00
Parenting stress		FFT	1.21	-1.40	0.61	1.98	3.34	7.78
		MST		-1.52	0.45	1.92	3.42	8.95

‡ In the weighted sample the 2.5th and 97.5th percentiles of the F-distribution are 0.78 and 1.22 respectively.

Repeating the analyses in the complete case sample of 361 adolescents and their families also showed no difference in externalizing problem behavior (CBCL: -0.16; 95% CI -4.32 – 4.14, YSR: -0.56; 95% CI -3.79 – 2.68), with, again, a small effect size of d = 0.01 and d = 0.05, respectively. While there was no difference in this sample between the treatment groups concerning the proportion of youth living at home and the proportion of youth engaged in school or work after treatment, it was significantly more likely that adolescents assigned to MST had had police contact during treatment than those assigned to FFT (RR 2.40; 95% CI 1.26 – 5.94, RD 15.3 %; 95% CI 4.7% - 24.5%) (Table 3).

After this, covariates selected for the PS in the outcome regression model were added in the complete case sample to overcome possible misspecifications of the model. All covariates were selected except for level of education of the adolescent, level of education and employment of the primary caregiver, and internalizing behavioral problems reported by adolescents in a parsimonious model on the outcome. This model can be interpreted as an additional check on the results found earlier. Again, it showed no difference between the effect of MST and FFT on externalizing problem behavior (CBCL: -0.56; 95% CI -3.06 – 1.49, YSR: 0.35; 95% CI -1.58 – 2.31) (Table 3).

Table 3. Comparing MST with FFT average treatment effects of the treated

	All adolescents	
	Study sample (N=697)	
	B	95% CI
Externalizing problem behavior CBCL	0.14	-3.23 - 3.49
Externalizing problem behavior YSR	-0.29	-2.45 - 1.90
	RR	95% CI
Police contact during treatment	1.61	0.98 - 3.08
Living at home after treatment	0.98	0.96 - 1.01
Engaged in school or work after treatment	1.27**	1.06 - 1.57
	Youth without a court order	
	Study sample (N=370)	
	B	95% CI
Externalizing problem behavior CBCL	-3.24*	-5.97 - -0.39
Externalizing problem behavior YSR	-3.33*	-5.81 - -0.86
	RR	95% CI
Police contact during treatment	1.20	0.72 - 2.77
Living at home after treatment	0.97	0.94 - 1.01
Engaged in school or work after treatment	1.09	0.94 - 1.31
	Youth with a court order	
	Study sample (N=317)	
	B	95% CI
Externalizing problem behavior CBCL	*Balance not achieved* ‖	
Externalizing problem behavior YSR		
	RR	95% CI
Police contact during treatment	*Balance not achieved* ‖	
Living at home after treatment		
Engaged in school or work after treatment		

** Confidence interval does not contain 0*

*** Confidence interval does not contain 1*

‖ *Balance was not achieved, therefore the differential effectiveness of FFT and MST could not be estimated*

Study sample (N=697)		Complete cases (N = 361)			
		B	95% CI	B adjusted	95% CI
		-0.16	-4.32 - 4.14	-0.56	-3.06 - 1.49
		-0.56	-3.79 - 2.68	0.35	-1.58 - 2.31
RD	95% CI	RR	95% CI	RD	95% CI
0.10	-0.01 - 0.19	2.41**	1.26 - 5.94	0.15*	0.05 - 0.25
-0.02	-0.04 - 0.01	0.98	0.95 - 1.01	-0.02	-0.06 - 0.01
0.19*	0.05 - 0.33	1.09	0.96 - 1.29	0.08	-0.03 - 0.21
Study sample (N=370)		Complete cases (N = 210)			
		B	95% CI	B adjusted	95% CI
		-4.55*	-8.30 - -0.41	-3.21*	-5.89 - -0.76
		-3.21*	-6.03 - -0.14	-1.55	-4.26 - 0.99
RD	95% CI	RR	95% CI	RD	95% CI
0.05	-0.10 - 0.20	0.90	0.50 - 2.53	-0.03	-0.24 - 0.19
-0.03	-0.06 - 0.01	1.00	0.95 - 1.07	0.00	-0.05 - 0.06
0.07	-0.05 - 0.21	1.14	0.92 - 1.54	0.11	-0.08 - 0.32
Study sample (N=317)		Complete cases (N = 151)			
		B	95% CI	B adjusted	95% CI
	Balance not achieved ‖				*Balance not achieved* ‖
RD	95% CI	RR	95% CI	RD	95% CI
	Balance not achieved ‖				*Balance not achieved* ‖

Subsample: Youth without a court order

Of the 697 adolescents in the study sample, 370 (168 MST; 202 FFT) had no court order before beginning the intervention. Of adolescents who had completed MST, 61.5% were male and 90.3% born in the Netherlands. For FFT, 52.3% of the adolescents were male and 97.4% born in the Netherlands (for an extensive comparison of the treatment arms, see Table IV in Supplemental Material). Comparing the treatment groups within this subsample on baseline characteristics showed significant differences in age, externalizing and total behavioral problems measured with the CBCL, parenting stress, country of birth, level of education, previous treatment, engagement in school or work, and previous police contact (Table IV in Supplemental Material).

Balance assessment

Before the PS application, the largest imbalances—standardized bias higher than the accepted .25—were found for age, externalizing problems on the CBCL, level of education, previous treatment, and having had police contact before treatment (Table IV in Supplemental Material). After PS application, balance was found when all covariates except for the total score of behavioral problems measured by the CBCL were selected for the PS estimation. Before inspecting balance, 11 MST and 29 FFT cases were removed for which there was no overlap on the PS scores. Except for the standardized bias of the level of education of the adolescent, values of the standardized bias after PS application were lower than .25 (Table IV in Supplemental Material). Values for the standardized bias for the missing indicator variables were also lower than .25 (Table V in Supplemental Material). The variance ratios of the continuous variables, except for parenting stress, were within the boundaries defined by the 2.5th and 97.5th percentiles of the F-distribution in the weighted sample. Thus, except for parenting stress, balance can be assumed given these values (Table VI in Supplemental Material). The five-number summaries show roughly equally distributed continuous variables between the treatment groups (Table VI in Supplemental Material).

In the complete case sample of 210 adolescents without court orders (104 MST; 106 FFT), balance was achieved when the variables of age, internalizing and externalizing problems measured with the CBCL, parenting stress, gender, country of birth, previous treatment, engagement in school or work, court order, police contacts, and employment status of the primary caregiver were included in the PS estimation. Except for the level of education of the adolescent, all standardized bias values were lower than .25 after PS application. Because balance was inspected in the sample with overlapping PS scores, 6 MST and 5 FFT cases were removed from this sample.

Analysis of treatment effect

In the subsample of adolescents without a court order, MST and FFT differed significantly in terms of externalizing problem behavior. Multisystemic Therapy resulted in lower scores on externalizing problem behavior than FFT (CBCL: -3.24; 95% CI -5.97 – -.39, YSR: -3.33; 95% CI -5.81 – -.86), with a medium effect size of *d* = 0.32 and *d* = 0.34, respectively. The differences (RR and RD) between MST and FFT on the three secondary outcomes were insignificant (Table 3).

Repeating the analyses within the complete case sample of 210 adolescents without a court order yielded the same results. Again, MST showed lower scores than FFT on externalizing problems (CBCL: -4.55; 95% CI -8.30 – -.41, YSR: -3.21; 95% CI -6.03 – -.14), with a medium effect size of *d* = 0.43 and *d* = 0.33, respectively. As before, no differences were found for the secondary outcomes within this sample.

To overcome possible misspecifications in the outcome model for this subgroup in the complete case sample, all covariates were selected except for the level of education of the adolescent, level of education of the primary caregiver, previous treatment, engagement in school or work, relation with mother, internalizing behavioral problems measured with the CBCL and YSR, and total problem behavior measured with the CBCL in a parsimonious model on the outcome. Again, a significant difference between MST and FFT on externalizing problem behavior was found for the CBCL (-3.21; 95% CI -5.89 – -.76), while no difference was found for the YSR (-1.55; 95% CI -4.26 – -.99) (Table 3).

Subsample: Youth with a court order

In total, 317 (246 MST; 71 FFT) of the 697 adolescents in the study sample had a court order before starting treatment. Of the adolescents who had completed MST, 70.4% were male and 78.2% were born in the Netherlands, while for FFT, 56.1% of the adolescents were male and 91% were born in the Netherlands (for an extensive comparison of the treatment arms, see Table VII in Supplemental Material). Multisystemic Therapy and FFT showed significant differences in terms of age, externalizing behavioral problems measured with the CBCL, internalizing problems measured with the YSR, gender, relation with father, and employment status of the primary caregiver at the baseline (Table VII in Supplemental Material).

Balance assessment

Before the PS application, the standardized bias was higher than the accepted .25 for age, externalizing problem behavior on the CBCL, internalizing problems on the YSR, gender, relation with father, and employment status of the primary caregiver (Table VII in Supplemental Material). After PS application, balance was not achieved using different PS estimations. Either there were some variables with a standardized bias higher than .25, or there were numerous variables with a standardized bias just below.25. Furthermore, if balance was assessed in the sample with overlapping scores on the PS, roughly 60–80 MST cases had to be removed each time when testing various PS estimations. This indicates that the sample of adolescents assigned to MST could not be balanced to the sample of adolescents assigned to FFT (West et al., 2014).

In the complete case sample, 151 adolescents had a court order before treatment (125 MST; 26 FFT). Again, no balance was found between the treatment groups when testing and applying different PS estimations.

Analysis of treatment effect

Because there was not confidence in assuming balance was achieved in this subsample of youth with a court order, the effectiveness could not be estimated without ensuring the control of allocation bias. The same holds for the complete case sample of 151 adolescents.

Discussion

Using the PS method to control for the non-random assignment of adolescents to either MST or FFT, this study compared these two interventions on their effectiveness in the Netherlands. In the study sample, target populations were balanced and no differences between the interventions were found regarding externalizing problem behavior. This result proved to be robust in the complete case sample. Some additional results were found, but these were not robust: adolescents assigned to MST were more often engaged in school or work after treatment. This treatment objective likely receives greater emphasis during MST than FFT. Moreover, in the complete case sample, adolescents who had received FFT had less police contact during treatment than adolescents who had received MST. Because the MST sample included a higher percentage of adolescents who had a court order *before* treatment, they were probably more likely to have additional police contact during treatment.

In the present study, the average treatment effect of the treated was estimated and the finding suggests that adolescents who receive MST may display the same treatment effects if they would have received FFT. This treatment effect, however, is only applicable for adolescents and their sample characteristics for whom there were outcome measurements after treatment. Finding only a few robust differences when comparing the effectiveness of MST and FFT in the overall study sample is in accordance with previous findings by Baglivio and colleagues (2014).

As the present study demonstrates that adolescents with a court order — interpreted as a possible risk factor following the RNR-model (Andrews et al., 2006; Laan, Slotboom, & Stams, 2010) — were more often assigned to MST (246 MST; 71 FFT), MST could also be expected to be more effective in this subsample. However, due to the incomparability of the FFT and MST subsamples of youth with a court order, the present study cannot confirm this. On the other hand, following the RNR model, FFT could be expected to be sufficiently effective in the subsample of adolescents without a court order, as these adolescents would be expected to have lower risks, and therefore, less intensive treatment would be adequate (Andrews et al., 2006; Laan et al., 2010). However, MST was more effective in reducing externalizing problems in the subsample without a court order. This may be explained by the fact that, although some risk factors were less present in the group without a court order, such as engagement in school or police contact (Table I and IV in Supplemental Material), this group nevertheless reported more problem behavior measured with the CBCL and the YSR (Tables I and IV in Supplemental Material). Another explanation may be that having or not having a court order only provides a rough indication of the risk level of an adolescent, while clinicians assign adolescents to either MST or FFT based on other risk factors as well. The RNR model thus leaves room for interpretation, or a single characteristic cannot fully represent the risk level of an adolescent. Even more, it could be possible that more intensive treatment in a less severe target population is always likely to be more effective, but the question remains as to whether it is appropriate and proportional treatment. For the secondary outcomes, however, no differences were found between the interventions, though these outcomes may be highly relevant to society. This should be taken into account when interpreting the overall effectiveness of the interventions in this subgroup. Furthermore, future research could focus on the applicability and validity of a checklist based on the RNR model, for example, to support stepped care when applicable, and assigning adolescents directly to more intensive interventions when needed (Krugten et al., 2016).

In addition to the effectiveness and assignment procedures of the interventions, and with stringent health care budgets, the costs of an intervention should be taken into account. If costs of a more effective intervention are higher than the costs of its alternative, it can be worthwhile to compare the interventions and their cost-effectiveness. Earlier studies in the US and UK have shown MST to be cost-effective compared with alternatives like individual therapy (Cary, Butler, Baruch, Hickey, & Byford, 2013; Klietz, Borduin, & Schaeffer, 2010). The cost-benefit ratio of FFT compared to MST in the US has been shown to be in favor of FFT (Lee et al., 2012). In the Netherlands, Vermeulen and colleagues (2016) compared MST to treatment as usual, including FFT, and found MST to be more cost-effective. Thus, cost-effectiveness depends on the context of the study, e.g. sample or country. With regard to the current study, it would for example be beneficial to implement a cost-effectiveness analysis in the subsample of adolescents without a court order. In this subsample, MST was more effective at reducing externalizing problems than FFT. Although it is unknown what the precise costs of MST and FFT are in the Netherlands, it is expected that MST is more expensive due to the intensity of the intervention. Cost-effectiveness analysis could reveal whether additional costs for MST are worth the higher effects. Future research must focus on estimating the exact costs of MST and FFT in the Netherlands and estimating health services use of this population to indeed estimate the cost-effectiveness. Moreover, it is of additional interest to determine the cost-effectiveness of intervention options when following a stepped care procedure, i.e. should youth with a lower risk be assigned to MST directly, or should a less intensive option be the first choice.

Comparing evidence-based interventions within overlapping target populations could eventually result in greater knowledge about which interventions work best for whom (Yirmiya, 2010). Therefore, it is important to examine treatment through client interactions and understand and study the assignment procedure based on the RNR model in greater detail. However, it is likely even more necessary—given the broad range of interventions currently available—to study practice elements or program elements of interventions to determine overlapping, effective elements (Chorpita & Daleiden, 2009; Evenboer, Huyghen, Tuinstra, Knorth, & Reijneveld, 2012; Lee et al., 2014). Furthermore, it would be of interest to compare the long-term effects of MST and FFT to find out whether their comparative effectiveness changes over time.

This study also shows that using clinical practice data, like ROM data, is worthwhile for evaluating treatments. It increases both the external validity of the study and the clinical utility because data was gathered in regular clinical practice and sample selection bias is less present (Hodgson, Bushe, & Hunter, 2007). The current study shows that the PS method is a useful and important method for using these data (West et al., 2014). It is, however, relevant to evaluate the chosen treatment outcomes in light of the selected dataset. The current study selected data from the Viersprong and not from other youth care institutions. Moreover, of the data selected, a sample was selected for which there was an outcome measure after treatment. The study sample — within which the comparative effectiveness was studied — consisted of adolescents with overall less risk factors (i.e., less reported court orders, see Table I and II in Supplemental Material) compared to the group for which no data was available after treatment, which could in turn result in less differences between interventions because this group might have shown better results overall. And thus, though clinical practice data were used, the findings can only be generalized to the selected group of adolescents and the findings should be interpreted in light of this sample selection. On the one hand, this study sample is likely larger and has less sample selection bias compared to data from randomized clinical trials (RCTs). But using observational data still merits reflection on the generalizability of the findings and evaluation given the selections, regardless of the study design (Stuart, Cole, Bradshaw, & Leaf, 2011). Furthermore, partial replication of a previous study (Baglivio et al., 2014) supports prior evidence and shows that the results are robust across different clinical settings and study designs (Duncan, Engel, Claessens, & Dowsett, 2014).

Despite the clinical relevance and use of this study, some limitations merit reflection. First, although a wide range of initial differences between adolescents in the treatment arms were controlled for, there could still be differences that were unmeasured and thus not controlled for. For example, the quality of life of the adolescent was not measured. This could have led to hidden biases in the presented results (Rosenbaum, 1991; Shadish, 2013). Second, though a response rate of ~40% is common when using clinical practice data from ROM in the Netherlands and not gathered for specific research purposes, there were a number of families who did not complete the CBCL at the end of the treatment. When comparing adolescents who did and did not complete this primary outcome measure, there were differences within the MST and FFT group. As a result, the external validity of this study is not optimal because the effect of the treatments in the group with missing data could not be measured. Third, the interventions are monitored in a quality system, follow detailed protocols, and require therapists to have completed higher education in a relevant domain. Differences between interventions, however, could be related to the duration of the treatment, the dosage and intensity of the interventions, and adherence of therapists to treatment protocol. Because the duration and intensity of treatment depend on the particular situation of an adolescent assigned to MST and FFT which could be related to specific background characteristics of the adolescent and the family, controlling for these factors would not fully represent the services as provided. Even more, it is yet unclear how the intensity of treatment can be defined. It could, for example, be related to the number of sessions, the amount of time, directly and indirectly, given to an adolescent and his or her family, and the length of treatment. Fourth, we had not data on adolescents assigned to treatment as usual or a control group consisting of adolescents not receiving treatment. However, when decision makers decided on the use of these interventions, it would have been helpful to include a reference treatment option. Fifth, though the chosen method was thoroughly considered, and all assumptions checked, and although results were robust over different samples (the study sample and the complete case sample), the choice of methods could influence the outcomes. There could, for example, be different estimation methods, e.g., matching with the PS or stratification using the PS, which arrive even closer to the true effect (Cham & West, 2016; Harder et al., 2010). Even more, using different approaches can help reducing uncertainty surrounding outcomes. Finally, the subgroup was chosen to indicate risk level according to the RNR model, but other demographic characteristics (in combination) could have also been used, such as living situation or education level.

In conclusion, the current study found few differences in the relative effectiveness of MST and FFT. This paper also stresses the necessity of investigating effects within subgroups of adolescents, as conclusions can change in looking at specific subgroups. Though RCTs are considered to be most effective for evaluating treatment options, using clinical practice data is certainly a viable alternative when carefully applied. By thoroughly controlling for treatment selection, the approach even enhances external validity because sample selection is less present than in RCTs (Stuart et al., 2011).

References

Achenbach, T. M., & Rescorla, L. A. (2001). *Manual for the ASEBA school-age forms and profiles*. Burlington, VT: University of Vermont, Research Centre for Children, Youth & Families.

Alexander, J. F., & Sexton, T. L. (2002). Functional Family Therapy: A model for treating high-risk, acting-out youth. In F. W. Kaslow & J. Lebow (Eds.), *Comprehensive handbook of psychotherapy: Integrative/eclectic* (pp. 111-132). New York, NY: John Wiley.

Ali, M. S., Groenwold R.H., Belitser, S. V., Pestman, W. R., Hoes, A. W., Roes, K. C., . . . Klungel, O. H. (2015). Reporting of covariate selection and balance assessment in propensity score analysis is suboptimal: A systematic review. *Journal of Clinical Epidemiology, 16*, 112-121.

Andrews, D. A., & Bonta, J. (2010). Rehabilitating criminal justice policy and practice. *Psychology, Public Policy and Law, 16*, 39-55.

Andrews, D. A., Bonta, J., & Wormith, J. S. (2006). The recent past and near future of risk and/or need assessment. *Crime & Delinquency, 52*, 7-27.

Asscher, J. J., Dekovic, M., Manders, W. A., van der Laan, P. H., & Prins, P. J. M. (2013). A randomized controlled trial of the effectiveness of Multisystemic therapy in the Netherlands: post-treatment changes and moderator effects. *Journal of Experimental Criminology, 9*, 169-187.

Austin, P. C. (2009). Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity score matched samples. *Statistics in Medicine, 28*, 3083-3107.

Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research, 46*, 399-424.

Austin, P. C., & Small, D. S. (2014). The use of bootstrapping when using propensity score matching without replacement: A simulation study. *Statistics in Medicine, 33*, 4306-4319.

Baglivio, M. T., Jackowski, K., Greenwald, M. A., & Wolff, K. T. (2014). Comparison of Multisystemic Therapy and Functional Family Therapy effectiveness: A multiyear statewide propensity score matching analysis of juvenile offenders. *Criminal Justice and Behavior, 41*, 1033-1056.

Blueprints for healthy youth development (2015). Factsheet Functional Family Therapy (FFT). Retrieved from http://www.blueprintsprograms.com/factSheet.php?pid=0a57cb53ba59c46fc4b692527a38a87c78d84028

Breuk, R. E., Sexton, T. L., van Dam, A., Disse, C., Doreleijers, T. A. H., Slot, W. N., & Rowland, M. K. (2006). The implementation and the cultural adjustment of functional family therapy in a Dutch psychiatric day-treatment centre. *Journal of Marital and Family Therapy, 32*, 515-529.

Bronfenbrenner, U. (1979). *The ecology of human development*. Cambridge, MA: Harvard University Press.

Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J., & Stürmer, T. (2006). Variable selection for propensity score models. *American Journal of Epidemiology, 163*, 1149-1156.

Buwalda, V. J. A., Nugter, M. A., Swinkels, J. A., & Mulder, C. L. (2011). *Praktijkboek ROM in de ggz: Een leidraad voor gebruik en implementatie van meetinstrumenten [Manual ROM in mental health care: Guidance for use and implementation of measurement instruments]*. Utrecht, the Netherlands: De Tijdstroom uitgeverij B.V..

Cary, M., Butler, S., Baruch, G., Hickey, N., & Byford, S. (2013). Economic evaluation of multisystemic therapy for young people at risk for continuing criminal activity in the UK. *PLoSONE, 8*, e61070.

Cham, H., & West, S. G. (2016). Propensity score analysis with missing data. *Psychological Methods*.

Chorpita, B. F., & Daleiden, E. L. (2009). Mapping evidence-based treatments for children and adolescents: Application of the distillation and matching model to 615 treatments from 322 randomized trials. *Journal of Consulting and Clinical Psychology, 77*, 566-579.

Chorpita, B. F., Daleiden, E. L., Ebesutani, C., Young, J., Becker, K. D., Nakamura, B. J., . . . Starace, N. (2011). Evidence-based treatments for children and adolescents: An updated review of indicators of efficacy and effectiveness. *Clinical Psychology: Science and Practice, 18*, 154-172.

D'Agostino, R.B.Jr. (1998). Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in Medicine, 17*, 2265-2281.

D'Agostino, R. B., Jr., Lang, W., Walkup, M., Morgan, T., & Karter, A. (2001). Examining the impact of missing data on propensity score estimation in determining the effectiveness of self-monitoring of blood blucose (SMBG). *Health Services and Outcomes Research Methodology, 2*, 291-315.

De Brock, A., Vermulst, A., Gerris, J., Veerman, J. W., & Abidin, R. (2004). *Nijmeegse Ouderlijke Stress Index-R. Voorlopige handleiding [Nijmeegse Parentins Stress Index Revisited]*. Nijmegen, the Netherlands: Behavioural Science Institute.

Duncan, G. J., Engel, M., Claessens, A., & Dowsett, C. J. (2014). Replication and robustness in developmental research. *Developmental Psychology, 50*, 2417-2425.

Evenboer, K. E., Huyghen, A. N., Tuinstra, J., Knorth, E. J., & Reijneveld, S. A. (2012). A taxonomy of care for youth: Results of an empirical development procedure. *Research on Social Work Practice, 22*, 637-646.

Goossens, L., Redekop, K., & van Gils, C. (2015). Noncollapsibility and censoring: What's the bias in estimating effects on survival? *Epidemiology, 26*, e1.

Green, K. M., & Stuart, E. A. (2014). Examining moderation analyses in propensity score methods: Application to depression and substance use. *Journal of Consulting and Clinical Psychology, 82*, 773-783.

Gustle, L.-H., Hansson, K., Sundell, K., Lundh, L.-G., & Lofhölm, C. A. (2006). Blueprints in Sweden. Symptom load in Swedish adolescents in studies of Functional Family Therapy (FFT), Multisystemic Therapy (MST) and Multidimensional Treatment Foster Care (MTFC). *Nordic Journal of Psychiatry, 61*, 443-451.

Harder, V. S., Stuart, E. A., & Anthony, J. C. (2010). Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. *Psychological Methods, 15*, 234-249.

Haviland, A., Nagin, D. S., & Rosenbaum, P. R. (2007). Combining propensity score matching and group-based trajectory analysis in an observational study. *Psychological Methods, 12*, 247-267.

Hendriks, M. E. D., Lange, A. M. C., Boonstoppel-Boender, M., & van der Rijken, R. E. A. (2014). Functional Family Therapy en Multi Systeem Therapie: Een vergelijking van doelgroepen [Functional Family Therapy and Multisystemic Therapy: A comparison of target populations]. *Orthopedagogiek: Onderzoek en Praktijk, 53*, 355-366

Henggeler, S. W. (2011). Efficacy studies to large-scale transport: The development and validation of Multisystemic therapy programs. *Annual Review of Clinical Psychology, 7*, 351-381.

Henggeler, S. W., Schoenwald, S. K., Borduin, C. M., Rowland, M. D., & Cunningham, P. B. (2009). *Multisystemic therapy for antisocial behavior in children and adolescents (2nd ed.)*. New York, NY: The Guilford Press

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis, 15*, 199-236.

Hodgson, R., Bushe, C., & Hunter, R. (2007). Measurement of long-term outcomes in observational and randomised controlled trials. *Britisch Journal of Psychiatry, 191*, 78-84.

IBM Corp. Released 2013. IBM SPSS Statistics for Windows, Version 22.0. Armonk, NY: IBM Corp.

Klietz, S. J., Borduin, C. M., & Schaeffer, C. M. (2010). Cost-benefit analysis of multisystemic therapy with serious and violent juvenile offenders. *Journal of Family Psychology, 24*, 657-666.

Krugten, F. C., Kaddouri, M., Goorden, M., van Balkom, A. J., Ruhé, H. G., van Schaik, D. J., . . . Hakkaart-van Roijen, L. (2016). Feasibility, reliability and validity of the decision too unipolar depression (DTUD) in identifying patients with major depressive disorder in need of highly specialized care. *Value in Health, 19*.

Lee, B. R., Ebesutani, C., Kolivoski, K. M., Becker, K. D., Lindsey, M. A., Brandt, N. E., . . . Barth, R. P. (2014). Program and practice elements for placement prevention. A review of interventions and their effectiveness in promoting home-based care. *American Journal of Orthopsychiatry, 84*, 244-256.

Lee, S., Aos, S., Drake, E., Pennucci, A., Miller, M., & Anderson, L. (2012). *Return on investment: Evidence-based options to improve statewide outcomes*. Olympia: Washington State Institute for Public Policy.

Manders, W. A., Deković, M., Asscher, J. J., van der Laan, P. H., & Prins, P. J. M. (2011). De implementatie van Multisysteem Therapie in Nederland: de invloed van behandelintegriteit en nonspecifieke factoren op behandeluitkomsten [The implementation of Multisystemic Therapy in the Netherlands: the influence of treatment integrity and nonspecific factors on treatment outcomes]. *Gedragstherapie, 44*, 327-340

McCaffrey, D. F., Griffin, B. A., Almirall, D., Slaughter, M. E., Ramchand, R., & Burgette, L. F. (2013). A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Statistics in Medicine, 32*, 3388-3414.

MST Institute (2016). *Frequently Asked Questions. MST Institute Enhanced Website*. Retrieved from https://www.msti.org/documents/EW_FAQs.pdf

Oudhof, M., Ten Berge, I., & Berger, M. (2009). *Checklist MST/FFT. De ontwikkeling van een indicatie-instrument voor MST en FFT in de vorm van een checklist [Checklist MST/FFT. The development of a checklist-instrument to indicate assignment to MST and FFT]*. Utrecht, the Netherlands: Nederlands Jeugdinstituut.

Peduzzi, P., Concato, J., Kemper, E., Holford, T. R., & Feinstein, A. R. (1996). A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology, 49*, 1373-1379.

Praktikon/MST-NL. (2012). *Sociaal Demografische Informatie. Ongepubliceerde vragenlijst [Demographic information. Unpublished questionnaire].*

Rosenbaum, P. R. (1991). Discussing hidden bias in observational studies. *Annals of Internal Medicine, 115*, 901-905.

Rosenbaum, P. R. (2010). *Design of observational studies*. New York, NY: Springer-Verlag.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*, 41-55.

Rosenbaum, P. R., & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association, 79*, 516-524.

Rubin, D. B. (2001). Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Services & Outcomes Research Methodology, 2*, 169-188.

Rubin, D. B., & Thomas, N. (2000). Combining propensity score matching with additional covariate adjustments for prognostic covariates. *Journal of the American Statistical Association, 450*, 573-585.

Sexton, T., & Turner, C. W. (2010). The effectiveness of Functional Family Therapy for youth with behavioral problems in a community practice setting. *Journal of Family Psychology, 24*, 339-348.

Sexton, T. L., & Alexander, J. F. (2000). Functional Family Therapy. Office of Juvenile Justice and Delinquency Prevention. *Juvenile Justice Bulletin, 1*, 1-7

Sexton, T. L., & Alexander, J. F. (2003). Functional Family Therapy for at-risk adolescent and their families (chapter 6). In F. W. Kaslow & T. Patterson (Eds.), *Comprehensive handbook of psychotherapy: Cognitive-Behavioral approaches (volume 2)*. New York: John Wiley & Sons, Inc.

Shadish, W. R. (2013). Propensity score analysis: Promise, reality and irrational exuberance. *Journal of Experimental Criminology, 25*, 129-144.

StataCorp. (2011). Stata Statistical Software: Release 12. College Station, TX: StataCorp LP.

Stichting Benchmark GGZ (SBG) (2016). *Position Paper: Benchmarken is beter worden door te vergelijken [Position paper: Benchmarking is becoming better by comparing].* Retrieved from https://www.sbggz.nl/Over-SBG?contentitem=b49041b8-bb96-4d2a-abc6-56ac55f3ec25&paragraph=14cc6e96-474e-4a3d-a365-b69077113168#Position-Paper.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science, 25*, 1-21.

Stuart, E. A., Cole, S. R., Bradshaw, C. P., & Leaf, P. J. (2011). The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 174*, 369-386.

Sundell, K., Hansson, K., Löfholm, C. A., Olsson, T., Gustle, L. H., & Kadesjö, C. (2008). The transportability of multisystemic therapy to Sweden: short-term results from a randomized trial of conduct-disordered youths. *Journal of Family Psychology, 22*, 550-560.

Thoemmes, F. J., & Kim, E. S. (2011). A systematic review of propensity score methods in the social sciences. *Multivariate Behavioral Research, 46*, 90-118.

van der Laan, A. M., Slotboom, A.-M., & Stams, G. J. (2010). Wat werkt? Bijdrage aan het terugdringen van recidive [What works? Contribution to reducing recidivism]. In H. M. P.J. Koppen, M. Jelicic, & J.W. Keijser (Ed.), *Reizen met mijn rechter. Psychologie van het recht* (pp. 987-1001). Deventer, the Netherlands: Kluwer.

van der Rijken, R. E. A. (2015). *Treatment fidelity of MST and FFT*: Unpublished report.

van der Stouwe, T., Asscher, J. J., Stams, G. J. J. M., Deković, M., & van der Laan, P. H. (2014). The effectiveness of Multisystemic Therapy (MST): A meta-analysis. *Clinical Psychology Review, 34*, 468-481.

Verhulst, F. C. & van der Ende, J. (2001a). *Gedragsvragenlijst voor kinderen van 6 tot 18 jaar [CBCL 6-18] [Child Behavior Checklist for children aged 6 to 18]*. Rotterdam, the Netherlands: Erasmus MC-Sophia Kinderziekenhuis.

Verhulst, F. C., & van der Ende, J. (2001b). *Zelf in te vullen vragenlijst voor 11-18 jarigen [YSR 11-18] [Youth Self Report for adolescents aged 11 to 18]*. Rotterdam, the Netherlands: Erasmus MC-Sophia Kinderziekenhuis.

Vermeulen, K. M., Jansen, D. E. M. C., Knorth, E. J., Buskens, E., & Reijneveld, S. A. (2016). Cost-effectiveness of multisystemic therapy versus usual treatment for young people with anitosocial problems. *Criminal Behaviour and Mental Health*.

Vermulst, A., Kroes, G., De Meyer, R., Nguyen, L., & Veerman, J. W. (2012). *Opvoedingsbelastingvragenlijst (OBVL). Handleiding [Questionnaire on parenting stress. Manual]*. Nijmegen, the Netherlands: Praktikon.

West, S. G., Cham, H., Thoemmes, F. J., Renneberg, B., Schulze, J., & Weiler, M. (2014). Propensity scores as a basis for equating groups: Basic principles and application in clinical treatment outcome research. *Journal of Consulting & Clinical Psychology, 82*, 906-919.

Yirmiya, N. (2010). Editorial: Early prevention and intervention – the Five W (and one H) questions. *Journal of Child Psychology and Psychiatry, 51*, 1297-1299.

Supplemental Material

Table I: Excluded versus included adolescents due to missing outcome measure after treatment (FFT)

Variable		**Excluded**	***(N = 365)***
Continuous variables		Mean	SD
Age		15.95	1.65
CBCL	Internalizing problems	62.5	10.45
Primary outcome	Externalizing problems	65.88	10.49
	Total behavioral problems	65.17	9.93
YSR	Internalizing problems	53.54	10.78
	Externalizing problems	58.01	10.28
	Total behavioral problems	65.05	9.53
Parenting stress		2	1.92
Categorical variables		%	
Gender	Male	50.1	
	Female	49.9	
Country of birth	Netherlands	90.1	
	Western country	3.3	
	Non-Western country	6.6	
Living situation adolescent	Together with one parent	45.0	
	Together with multiple parents	48.8	
	Other	6.1	
Living situation adolescent	Lived not at home	2.1	
Secondary outcome	Lived at home	97.9	
Level of education	None	9.2	
	Primary education	4.0	
	Lower secondary education	52.5	
	Higher secondary education	34.0	
	Higher education	0.3	
Previous treatment	Absent	10.4	
Engagement in school or work	Absent	20.3	
Secondary outcome	Present	79.7	
Court order	No	67.0	
	Civil	18.7	
	Criminal	14.3	
Police contacts during treatment	Absent	73.4	
Secondary outcome	Present	26.6	
Relation father	Absent	10.2	
	Present	89.8	
Relation mother	Absent	1.2	
	Present	98.8	
Relation siblings	Absent	9.2	
	Present	90.8	
Relation peers	Absent	0.6	
	Present	99.4	

	Included	***(N = 275)***		**Test statistic**
N	Mean	SD	N	T-test
365	15.9	1.59	275	0.70
268	62.51	9.26	263	-0.02
268	67.08	9.57	263	-1.37
268	66.04	8.61	263	-1.08
228	54.79	11.31	246	-1.24
228	59.27	9.73	246	-1.37
228	57.35	9.78	246	-1.46
248	1.97	1.78	258	0.21
N	%		N	Chi-Square statistic
183	53.6		141	0.74
182	46.4		122	
301	95.8		253	7.17*
11	1.1		3	
22	3.0		8	
154	36.1		97	9.20**
167	60.6		163	
21	3.3		9	
7	0.8		2	1.76
326	99.2		260	
30	7.1		19	1.77
13	3.7		10	
171	54.5		146	
111	34.7		93	
1	0.0		0	
35	9.8		26	0.06
64	14.5		37	3.34
251	85.5		219	
229	74.0		202	7.76*
64	10.6		29	
49	15.4		42	
240	66.9		176	2.94
87	33.1		87	
33	6.8		17	2.04
292	93.2		234	
4	0.4		1	1.07
330	99.6		249	
29	7.6		18	0.42
287	92.4		218	
2	0.0		0	1.51
330	100.0		249	

Variable		Excluded	(N = 365)
Categorical variables		%	
Country of birth primary caregiver	the Netherlands	85.6	
	Western country	3.1	
	Non-Western country	11.3	
Level of education primary caregiver	None	3.8	
	Primary education	5.2	
	Lower secondary education	31.4	
	Higher secondary education	41.5	
	Higher education	18.1	
Employment primary caregiver	Employed	66.0	
	Unemployed	34.0	
Partner primary caregiver	Absent	25.3	
	Present	74.7	

** p < .05, ** p < .01, *** p < .001*

NOTE: Values depict the mean values and standard deviations. Except for age and parenting stress all other scores are standardized T-scores, having a mean of 50 and a standard deviation of 10. For NOSI-R and parenting stress, normed z-scores are displayed.

	Included	*(N = 275)*		Test statistic
N	%		N	Chi-Square statistic
274	88.5		232	3.04
10	4.2		11	
36	7.3		19	
11	1.2		3	5.47
15	4.1		10	
90	27.9		68	
119	45.5		111	
52	21.3		52	
214	71.8		186	2.22
110	28.2		73	
80	21.7		55	0.99
236	78.3		198	

Table II. Excluded versus included adolescents due to missing outcome measure after treatment (MST)

Variable		**Excluded**	***(N = 652)***
Continuous variables		Mean	SD
Age		15.65	1.43
CBCL	Internalizing problems	61.42	9.71
Primary outcome	Externalizing problems	68.09	10.13
	Total behavioral problems	65.45	9.29
YSR	Internalizing problems	51.58	11.19
	Externalizing problems	57.2	10.59
	Total behavioral problems	53.4	10.91
Parenting stress		2.09	2.01
Categorical variables		%	
Gender	Male	68.9	
	Female	31.1	
Country of birth	Netherlands	73.7	
	Western country	3.3	
	Non-Western country	23.0	
Living situation adolescent	Together with one parent	53.5	
	Together with multiple parents	43.1	
	Other	3.4	
Living situation adolescent	Lived not at home	0.9	
Secondary outcome	Lived at home	99.1	
Level of education	None	14.8	
	Primary education	3.8	
	Lower secondary education	64.3	
	Higher secondary education	16.9	
Previous treatment	Absent	7.6	
	Present	92.4	
Engagement in school or work	Absent	31.2	
Secondary outcome	Present	68.8	
Court order	No	30.6	
	Civil	40.1	
	Criminal	29.2	
Police contacts during treatment	Absent	46.3	
Secondary outcome	Present	53.7	
Relation father	Absent	13.1	
	Present	86.9	
Relation mother	Absent	1.0	
	Present	99.0	
Relation siblings	Absent	7.4	
	Present	92.6	
Relation peers	Absent	1.1	
	Present	98.9	

	Included	*(N = 422)*		Test statistic
N	Mean	SD	N	T-test
652	15.67	1.35	422	-0.22
485	61.04	9.68	409	0.58
485	68.29	10.06	409	-0.31
485	65.32	9.76	409	0.19
431	50.78	11.5	356	0.99
431	57.54	10.87	356	-0.45
431	53.59	11.02	356	-0.24
425	2.06	2.07	397	0.21
N	%		N	Chi-Square statistic
447	67.2		275	0.31
202	32.8		134	
468	83.4		341	20.32***
21	4.6		19	
146	12.0		49	
345	42.9		179	13.09***
278	51.1		213	
22	6.0		25	
6	2.9		12	5.76*
631	97.1		400	
93	13.7		56	2.07
24	2.7		11	
404	66.8		274	
106	16.8		69	
49	5.5		23	1.75
597	94.5		395	
193	22.6		91	8.99**
426	77.4		312	
197	40.6		168	13.10**
258	30.9		128	
188	28.5		118	
285	50.8		198	1.87
330	49.2		192	
81	9.2		37	3.51
539	90.8		364	
6	0.7		3	0.13
623	99.3		402	
45	6.0		23	0.74
563	94.0		361	
7	1.3		5	0.04
615	98.7		393	

Variable		Excluded	(N = 652)
Categorical variables		%	
Country of birth primary caregiver	the Netherlands	59.3	
	Western country	4.2	
	Non-Western country	36.5	
Level of education primarycaregiver	None	8.7	
	Primary education	18.1	
	Lower secondary education	29.7	
	Higher secondary education	31.0	
	Higher education	12.5	
Employment primary caregiver	Employed	47.8	
	Unemployed	52.2	
Partner primary caregiver	Absent	32.7	
	Present	67.3	

** p < .05, ** p < .01, *** p < .001*
NOTE:
Values depict the mean values and standard deviations. Except for age and parenting stress all other scores are standardized T-scores, having a mean of 50 and a standard deviation of 10. For NOSI-R and parenting stress, normed z-scores are displayed.

	Included	***(N = 422)***	**Test statistic**
N	%	N	Chi-Square statistic
369	79.3	325	52.10***
26	4.9	20	
227	15.9	65	
53	3.0	12	40.86***
110	8.0	32	
181	31.3	126	
189	39.1	157	
76	18.7	75	
301	61.9	253	19.75***
329	38.1	156	
196	23.9	94	8.89**
403	76.1	299	

Table III. Standardized bias of missing indicators in full sample (N = 697)

Missing indicators§			**Before PS application**	**After PS application**
CBCL	Internalizing problems	Missing	0.07	0.07
		Not missing	0.07	0.07
	Externalizing problems	Missing	0.07	0.07
		Not missing	0.07	0.07
	Total behavioral problems	Missing	0.07	0.07
		Not missing	0.07	0.07
YSR	Internalizing problems	Missing	0.14	0.11
		Not missing	0.14	0.11
	Externalizing problems	Missing	0.14	0.11
		Not missing	0.14	0.11
	Total behavioral problems	Missing	0.14	0.11
		Not missing	0.14	0.11
Parenting stress		Missing	0.01	0.06
		Not missing	0.01	0.06
Gender		Missing	0.07	0.07
		Not missing	0.07	0.07
Country of birth		Missing	0.05	0.04
		Not missing	0.05	0.04
Living situation adolescent		Missing	0.09	0.02
		Not missing	0.09	0.02
Living situation adolescent		Missing	0.16	0.04
Secondary outcome		Not missing	0.16	0.04
Level of education		Missing	0.02	0.07
		Not missing	0.02	0.07
Previous treatment		Missing	0.24	0.05
		Not missing	0.24	0.05
Engagement in school or work		Missing	0.12	0.09
Secondary outcome		Not missing	0.12	0.09
Court order		Missing	0.09	0.01
		Not missing	0.09	0.01
Police contacts during treatment		Missing	0.12	0.03
Secondary outcome		Not missing	0.12	0.03
Relation father		Missing	0.17	0.03
		Not missing	0.17	0.03
Relation mother		Missing	0.26	0.03
		Not missing	0.26	0.03
Relation siblings		Missing	0.18	0.09
		Not missing	0.18	0.09
Relation peers		Missing	0.16	0.05
		Not missing	0.16	0.05
Country of birth primary caregiver		Missing	0.11	0.20
		Not missing	0.11	0.20
Level of education primary caregiver		Missing	0.31	0.01
		Not missing	0.31	0.01
Employment primary caregiver		Missing	0.16	0.05
		Not missing	0.16	0.05
Partner primary caregiver		Missing	0.05	0.11
		Not missing	0.05	0.11

§ *No missing values were present for the variable 'Age', thus the missing indicator was not needed.*

Table IV. Baseline differences between adolescents assigned to FFT and MST and standardized bias for youth without a court order (N=370)

Variable		**FFT**	***(N = 202)***	
Continuous variables		Mean	SD	N
Age		15.72	1.61	202
CBCL	Internalizing problems	62.99	8.72	195
Primary outcome	Externalizing problems	68.54	8.48	195
	Total behavioral problems †	67.31	7.39	195
YSR	Internalizing problems	55.21	11.15	182
	Externalizing problems	59.96	8.84	182
	Total behavioral problems	58.34	8.96	182
Parenting stress		2.23	1.71	188
Categorical variables		%		N
Gender	Male	52.3		102
	Female	47.7		93
Country of birth	Netherlands	97.4		190
	Western country	1.0		2
	Non-Western country	1.5		3
Living situation adolescent	Together with one parent	31.3		62
	Together with multiple parents	66.7		132
	Other	2.0		4
Living situation adolescent	Lived not at home	0.0		0
Secondary outcome	Lived at home	100.0		194
Level of education	None	5.1		10
	Primary education	3.6		7
	Lower secondary education	53.8		106
	Higher secondary education	37.6		74
Previous treatment	Absent	21.5		20
	Present	78.5		73
Engagement in school or work	Absent	12.1		23
Secondary outcome	Present	87.9		167
Court order	No	100.0		202
	Civil	0.0		0
	Criminal	0.0		0
Police contacts during treatment	Absent	72.0		139
Secondary outcome	Present	28.0		54
Relation father	Absent	8.2		15
	Present	91.8		169
Relation mother	Absent	0.6		1
	Present	99.4		180
Relation siblings	Absent	38.5		10
	Present	61.5		16

MST	*(N = 168)*		Test statistic		Standardized bias
Mean	SD	N	T-test	Before PS application	After PS application
15.28	1.39	168	2.76**	0.32	0.05
62.99	8.58	161	0.00	0.00	0.01
72.2	7.87	161	-4.18***	0.47	0.04
68.84	6.8	161	-2.02*	0.23	0.10
52.76	11.42	139	1.93	0.22	0.04
60.73	9.91	139	-0.73	0.08	0.02
56.76	9.8	139	1.50	0.16	0.02
2.67	2.11	161	-2.12*	0.21	0.09
%		N	Chi-Square statistic	Before PS application	After PS application
61.5		99	3.03	0.19	0.10
38.5		62		0.19	0.10
90.3		149	8.97*	0.13	0.06
1.8		3		0.01	0.00
7.9		13		0.12	0.06
35.5		59	1.23	0.08	0.22
61.4		102		0.10	0.25
3.0		5		0.02	0.03
0.0		0	NA	0.00	0.00
100.0		165		0.00	0.00
8.4		14	23.81***	0.05	0.12
3.6		6		0.00	0.06
73.1		122		0.27	0.27
15.0		25		0.31	0.09
4.8		8	3.94*	0.26	0.06
95.2		160		0.26	0.06
20.0		33	4.14*	0.20	0.06
80.0		132		0.20	0.06
100.0		168	NA	0.00	0.00
0.0		0		0.00	0.00
0.0		0		0.00	0.00
55.1		87	10.90***	0.34	0.04
44.9		71		0.34	0.04
6.1		10	0.55	0.09	0.03
93.9		154		0.09	0.03
0.6		1	0.00	0.01	0.00
99.4		165		0.01	0.00
5.6		9	0.00	0.00	0.03
94.4		151		0.00	0.03

Variable		FFT	*(N = 202)*
Categorical variables		%	N
Relation peers	Absent	0.0	0
	Present	100.0	180
Country of birth primary caregiver	the Netherlands	89.5	171
	Western country	4.2	8
	Non-Western country	6.3	12
Level of education primary caregiver	None	1.1	2
	Primary education	2.2	4
	Lower secondary education	23.3	42
	Higher secondary education	50.6	91
	Higher education	22.8	41
Employment primary caregiver	Employed	69.8	132
	Unemployed	30.2	57
Partner primary caregiver	Absent	21.1	39
	Present	78.9	146

** p < .05, ** p < .01, *** p < .001*

† Not selected for PS estimation.

NOTE:

Values depict the mean values and standard deviations. Except for age and parenting stress all other scores are standardized T-scores, having a mean of 50 and a standard deviation of 10. For NOSI-R and parenting stress, normed z-scores are displayed.

MST	*(N = 168)*	**Test statistic**		**Standardized bias**
%	N	Chi-Square statistic	Before PS application	After PS application
0.6	1	1.11	0.08	0.00
99.4	161		0.08	0.00
88.3	144	2.96	0.02	0.09
1.8	3		0.04	0.00
9.8	16		0.06	0.09
0.6	1	1.14	0.01	0.00
3.1	5		0.01	0.03
22.7	37		0.01	0.02
47.2	77		0.04	0.04
26.4	43		0.04	0.02
69.1	112	0.02	0.02	0.15
30.9	50		0.02	0.15
16.1	26	1.37	0.13	0.09
83.9	135		0.13	0.09

Table V. Standardized bias of missing indicators in sample of youth without a court order (N = 370)

Missing indicators§			**Before PS application**	**After PS application**
CBCL	Internalizing problems	Missing	0.04	0.06
		Not missing	0.04	0.06
	Externalizing problems	Missing	0.04	0.06
		Not missing	0.04	0.06
	Total behavioral problems	Missing	0.04	0.06
		Not missing	0.04	0.06
YSR	Internalizing problems	Missing	0.19	0.03
		Not missing	0.19	0.03
	Externalizing problems	Missing	0.19	0.03
		Not missing	0.19	0.03
	Total behavioral problems	Missing	0.19	0.03
		Not missing	0.19	0.03
Parenting stress		Missing	0.14	0.06
		Not missing	0.14	0.06
Gender		Missing	0.04	0.06
		Not missing	0.04	0.06
Country of birth		Missing	0.13	0.00
		Not missing	0.13	0.00
Living situation adolescent		Missing	0.07	0.06
		Not missing	0.07	0.06
Living situation adolescent		Missing	0.16	0.11
Secondary outcome		Not missing	0.16	0.11
Level of education		Missing	0.24	0.08
		Not missing	0.24	0.08
Previous treatment		Missing	0.00	0.00
		Not missing	0.00	0.00
Engagement in school or work		Missing	0.31	0.11
Secondary outcome		Not missing	0.31	0.11
Court order		Missing	0.00	0.00
		Not missing	0.00	0.00
Police contacts during treatment		Missing	0.06	0.06
Secondary outcome		Not missing	0.06	0.06
Relation father		Missing	0.43	0.00
		Not missing	0.43	0.00
Relation mother		Missing	0.85	0.06
		Not missing	0.85	0.06
Relation siblings		Missing	0.36	0.00
		Not missing	0.36	0.00
Relation peers		Missing	0.39	0.00
		Not missing	0.39	0.00
Country of birth primary caregiver		Missing	0.15	0.00
		Not missing	0.15	0.00
Level of education primary caregiver		Missing	0.46	0.00
		Not missing	0.46	0.00
Employment primary caregiver		Missing	0.15	0.10
		Not missing	0.15	0.10
Partner primary caregiver		Missing	0.21	0.03
		Not missing	0.21	0.03

§ *No missing values were present for the variable 'Age', thus the missing indicator was not needed.*

Table VI: Variance ratio and 5-number summary of continuous covariates after PS application in sample of youth without court order (N = 370)

			Variance ratio‡	Minimum	25th percentile	Median	75th percentile	Maximum
Age		FFT	0.83	12.10	14.25	15.18	16.08	20.39
		MST		11.07	14.54	15.29	16.39	17.88
CBCL	Internalizing problems	FFT	0.96	39.00	57.00	65.00	69.00	83.00
		MST		34.00	57.75	64.00	70.00	81.00
	Externalizing problems	FFT	1.09	43.00	69.00	72.00	76.00	92.00
		MST		46.00	68.75	73.50	77.00	88.00
	Total behavioral problems	FFT	1.03	44.00	65.00	70.00	74.00	84.00
		MST		50.00	65.00	70.00	74.00	82.00
YSR	Internalizing problems	FFT	1.17	30.00	45.00	54.00	60.54	83.00
		MST		30.00	46.00	53.00	62.00	81.00
	Externalizing problems	FFT	1.25	37.00	55.00	60.96	68.01	80.00
		MST		29.00	54.00	61.00	68.00	88.00
	Total behavioral problems	FFT	1.18	34.00	51.00	56.06	63.20	77.00
		MST		26.00	50.00	56.00	64.00	82.00
Parenting stress		FFT	1.40	-1.16	1.23	2.62	3.54	7.78
		MST		-1.38	1.20	2.59	4.13	8.95

‡ In the weighted sample the 2.5th and 97.5th percentiles of the F-distribution are 0.63 and 1.37 respectively.

Table VII: Baseline differences between adolescents assigned to FFT and MST and standardized bias for youth with court order (N=317)

Variable		FFT	*(N = 71)*	
Continuous variables		Mean	SD	N
Age		16.4	1.42	71
CBCL	Internalizing problems	60.77	10.25	66
Primary outcome	Externalizing problems	62.85	11.25	66
	Total behavioral problems	62.18	10.73	66
YSR	Internalizing problems	53.16	11.64	62
	Externalizing problems	57.21	11.85	62
	Total behavioral problems	54.29	11.52	62
Parentingl stress		1.24	1.82	68
Categorical variables		%		N
Gender	Male	56.1		37
	Female	43.9		29
Country of birth	Netherlands	91.0		61
	Western country	1.5		1
	Non-Western country	7.5		5
Living situation adolescent	Together with one parent	47.8		33
	Together with multiple parents	44.9		31
	Other	7.2		5
Living situation adolescent	Lived not at home	3.0		2
Secondary outcome	Lived at home	97.0		64
Level of education	None	13.0		9
	Primary education	4.3		3
	Lower secondary education	58.0		40
	Higher secondary education	24.6		17
Previous treatment	Absent	8.5		6
	Present	91.5		65
Engagement in school or work	Absent	21.9		14
Secondary outcome	Present	78.1		50
Court order	No	0.0		0
	Civil	40.8		29
	Criminal	59.2		42
Police contacts during treatment	Absent	51.5		35
Secondary outcome	Present	48.5		33
Relation father	Absent	3.0		2
	Present	97.0		64
Relation mother	Absent	0.0		0
	Present	100.0		68
Relation siblings	Absent	12.1		7
	Present	87.9		51
Relation peers	Absent	0.0		0
	Present	100.0		68

MST	*(N = 246)*		Test statistic	Standardized bias	
Mean	SD	N	T-test	Before PS application	After PS application
15.94	1.26	246	2.58**	0.36	NA
59.84	10.14	240	0.66	0.09	NA
65.78	10.52	240	-1.97*	0.28	NA
63.09	10.7	240	-0.61	0.09	NA
49.59	11.52	209	2.14*	0.31	NA
55.55	11.01	209	1.02	0.15	NA
51.27	11.38	209	1.65	0.24	NA
1.68	1.93	228	-1.66	0.23	NA
%		N	Chi-Square statistic	Before PS application	After PS application
70.4		169	4.85*	0.31	NA
29.6		71		0.31	NA
78.2		186	5.92	0.18	NA
6.7		16		0.07	NA
15.1		36		0.10	NA
48.6		119	0.10	0.01	NA
43.3		106		0.03	NA
8.2		20		0.01	NA
5.0		12	0.45	0.09	NA
95.0		229		0.09	NA
17.6		42	2.90	0.05	NA
2.1		5		0.02	NA
61.9		148		0.04	NA
18.4		44		0.07	NA
5.7		14	0.70	0.12	NA
94.3		231		0.12	NA
24.8		58	0.23	0.07	NA
75.2		176		0.07	NA
0.0		0	2.76	0.00	NA
52.0		128		0.22	NA
48.0		118		0.22	NA
47.6		108	0.32	0.08	NA
52.4		119		0.08	NA
11.6		27	4.30*	0.27	NA
88.4		206		0.27	NA
0.9		2	0.58	0.09	NA
99.1		233		0.09	NA
6.4		14	2.14	0.23	NA
93.6		206		0.23	NA
1.7		4	1.19	0.13	NA
98.3		228		0.13	NA

Variable		FFT	*(N = 71)*
Categorical variables		%	N
Country of birth primary caregiver	the Netherlands	85.5	59
	Western country	4.3	3
	Non-Western country	10.1	7
Level of education primary caregiver	None	1.6	1
	Primary education	9.7	6
	Lower secondary education	41.9	26
	Higher secondary education	30.6	19
	Higher education	16.1	10
Employment primary caregiver	Employed	76.5	52
	Unemployed	23.5	16
Partner primary caregiver	Absent	24.2	16
	Present	75.8	50

** p < .05, ** p < .01, *** p < .001*

NOTE:

Values depict the mean values and standard deviations. Except for age and parenting stress all other scores are standardized T-scores, having a mean of 50 and a standard deviation of 10. For NOSI-R and parenting stress, normed z-scores are displayed.

MST	*(N = 246)*	Test statistic	Standardized bias	
%	N	Chi-Square statistic	Before PS application	After PS application
72.6	175	4.87	0.16	NA
7.1	17		0.03	NA
20.3	49		0.13	NA
4.7	11	1.88	0.03	NA
11.5	27		0.02	NA
37.2	87		0.05	NA
32.9	77		0.02	NA
13.7	32		0.02	NA
57.3	138	8.26**	0.39	NA
42.7	103		0.39	NA
30.1	68	0.85	0.13	NA
69.9	158		0.13	NA

Table VIII: Standardized bias of missing indicators in sample of youth with court order (N = 317)

Missing indicators§			**Before PS application**	**After PS application**
CBCL	Internalizing problems	Missing	0.30	NA
		Not missing	0.30	NA
	Externalizing problems	Missing	0.30	NA
		Not missing	0.30	NA
	Total behavioral problems	Missing	0.30	NA
		Not missing	0.30	NA
YSR	Internalizing problems	Missing	0.07	NA
		Not missing	0.07	NA
	Externalizing problems	Missing	0.07	NA
		Not missing	0.07	NA
	Total behavioral problems	Missing	0.07	NA
		Not missing	0.07	NA
Parenting stress		Missing	0.12	NA
		Not missing	0.12	NA
Gender		Missing	0.30	NA
		Not missing	0.30	NA
Country of birth		Missing	0.13	NA
		Not missing	0.13	NA
Living situation adolescent		Missing	0.38	NA
		Not missing	0.38	NA
Living situation adolescent		Missing	0.35	NA
Secondary outcome		Not missing	0.35	NA
Level of education		Missing	0.00	NA
		Not missing	0.00	NA
Previous treatment		Missing	0.06	NA
		Not missing	0.06	NA
Engagement in school or work		Missing	0.23	NA
Secondary outcome		Not missing	0.23	NA
Court order		Missing	0.00	NA
		Not missing	0.00	NA
Police contacts during treatment		Missing	0.13	NA
Secondary outcome		Not missing	0.13	NA
Relation father		Missing	0.08	NA
		Not missing	0.08	NA
Relation mother		Missing	0.01	NA
		Not missing	0.01	NA
Relation siblings		Missing	0.25	NA
		Not missing	0.25	NA
Relation peers		Missing	0.06	NA
		Not missing	0.06	NA
Country of birth primary caregiver		Missing	0.06	NA
		Not missing	0.06	NA
Level of education primary caregiver		Missing	0.36	NA
		Not missing	0.36	NA
Employment primary caregiver		Missing	0.16	NA
		Not missing	0.16	NA
Partner primary caregiver		Missing	0.04	NA
		Not missing	0.04	NA

§ No missing values were present for the variable 'Age', *thus the missing indicator was not needed.*

Chapter 6.

General discussion

This dissertation addresses two issues of contemporary interest in the field of youth care: 1) the feasibility of cost-effectiveness research in this field and 2) the use of available, non-randomized, data in investigating the effectiveness of interventions in youth care practice. These issues will be discussed in light of the results in this dissertation and, subsequently, several recommendations will be provided for clinical practice and further research.

Cost-effectiveness research in youth care

We investigated whether state of the art methods commonly applied in health care evaluation studies can be applied to systemic interventions in youth care. In Chapter 2, Functional Family Therapy (FFT) was compared with Treatment as usual (TAU) in an illustrative probabilistic Markov model, in which parameter uncertainty and long-term cost-effectiveness were taken into account. Estimating long-term cost-effectiveness is essential when interventions are applied to youth aged 12 to 18 years and when treatment effects are expected to last into adulthood. By expressing the cost-effectiveness ratio in costs per Criminal Activity Free Year (CAFY), we used an outcome measure that addressed not only the clinical, but also the societal effect of the interventions. By doing so, clinically relevant issues and health economic questions were brought together. Chapter 2 thus showed that commonly applied economic evaluation methods are applicable in evaluating youth care.

However, when modelling the cost-effectiveness of interventions, the results in both cost and effect estimates can be subject to uncertainty. This parameter uncertainty can be reduced by information obtained from further research. A value of information analysis can reveal the value and justify the direction of further research. In Chapter 3, an illustrative value of information analysis was applied using the cost-effectiveness model developed in Chapter 2. Thereby, it was needed to assume a 'willingness-to-pay (WTP) value' for one CAFY (i.e., what is society willing to pay for one year without criminal activity of one adolescent?) since the expected value of further information depends on this WTP value. Findings from a value of information analysis may lead to reimbursing the intervention studied or not under certain conditions. Chapter 3 showed that a value of information analysis in the field of youth care can also be interpreted as similar to cost/ Quality Adjusted Life Year (QALY) studies in health care evaluations and is particularly meaningful in this field, because the interest in cost-effectiveness research is increasing, and its analyses could use a wider range of input parameters than cost-effectiveness research in health care, for instance.

Both the cost-effectiveness analysis in Chapter 2 and the value of information analysis in Chapter 3 revealed important issues that should be considered in future cost-effectiveness research in youth care. First, in contrast to health care evaluation studies, in youth care not only the referred client (i.e., the adolescent) should be the focus of research, but also the systems surrounding him. For example, the societal perspective chosen in the analysis should include the effect of the intervention in terms of reduced costs and increased well-being of family members, reduced victim costs, and reduced costs because of avoided crimes, all taken into account over the defined time-horizon of the analysis. Second, when measuring CAFYs, the type of criminal activity that is avoided, the seriousness of the crime, and the number of times it is committed should be defined. Even more, the outcome measure should be clinically relevant and should allow making comparisons with other interventions. A preference weighted measure, like the QALY may be preferred as it adds weights to different outcomes of the interventions. Third, we should be able to assign a value to the defined outcome, the so-called WTP value, in order to judge the cost-effectiveness results against a threshold value. This WTP value depends on the outcome defined and the cost categories included in the analysis, which should be reflected in this WTP value and vice versa.

In conclusion, when evaluating the cost-effectiveness of youth care interventions, commonly applied economic evaluation methods are feasible and their results can be interpreted in the same manner as in health care evaluation studies.

Use of observational data in treatment evaluation

Evaluating the effectiveness of interventions in youth care becomes increasingly important, but due to practical and ethical constraints it is not always possible to randomly allocate adolescents and their families to treatment. As an alternative, research could follow clinical practice in gathering data. In that case, the propensity score (PS) method can be used to control for initial differences between treatment groups. It is thereby of interest to study subgroup effects when using the PS to tailor youth care to adolescents' situations and needs. One can adjust for subgroup effects in several ways, for example by additionally adjusting for the subgroup or by splitting the dataset into the relevant subgroups. The first manner was studied in a Monte Carlo simulation study in Chapter 4. The research question was whether the subgroups should be added to the outcome model, together with an interaction term between treatment and subgroups, or whether the PS should be made multiple to estimate the specific treatment effects within subgroups. Both methods were found to be feasible, while the latter option (i.e., making the PS multiple on the subgroup and treatment options) gave less biased results compared to the first option (i.e., adding the subgroups to the outcome model).

In Chapter 5, two youth care interventions, FFT and Multisystemic Therapy (MST) were compared on their effectiveness using Routine Outcome Monitoring (ROM) data and subgroups were investigated by splitting the dataset. The outcomes were externalizing problems of the adolescent, whether the adolescent was living at home after treatment, was engaged in school or work after treatment, and had had police contacts during treatment. Minor differences were found between the interventions. However, when splitting the dataset into subgroups of adolescents who had a court order before treatment and those who had not, different results were obtained: MST was more effective than FFT in reducing externalizing problems when adolescents had no court order. Because many more adolescents with a court order were assigned to MST than to FFT, the PS could not balance the intervention groups in this subsample. Therefore, no comparative treatment effect could be estimated in this subsample.

In general, both Chapter 4 and Chapter 5 showed the applicability of the PS when evaluating youth care, and more importantly, the use of clinical practice data to answer questions about the effectiveness of interventions. When using available, non-randomized clinical practice data, however, the following considerations should be kept in mind. The treatments evaluated and the results of the analyses should be considered in light of the selected dataset. For example, using data from one institution can complicate the applicability and generalizability of the findings to other institutions, as referral practices and outcome measures may differ between institutions. In addition, characteristics of the selected dataset influence the choice for the analyses, which may influence the results and the conclusions drawn. For instance, when one is interested in a subgroup effect, the definition of this subgroup determines whether it could moderate or mediate the relations between treatment and outcome. Even more, the findings should be interpreted in light of daily clinical practice because clinicians and patients should be able to use these findings in clinical practice.

The two issues addressed in this thesis, cost-effectiveness analyses in youth care and treatment evaluation using observational data are related, since effectiveness studies are needed to decide on the necessity and relevance of a cost-effectiveness study. On the other hand, when a cost-effectiveness study is needed, data on the costs and effects of the interventions studied are needed. When data of a randomized trial are not available, alternatives such as clinical practice data gathered within ROM could be useful.

Implications for youth care

Interventions seem to be most efficiently studied and compared on effectiveness using clinical practice data. Setting up randomized controlled trials (RCT's) and asking clinicians and adolescents and their parents to fill in questionnaires in addition to the questionnaires that are used to routinely monitor the treatment process can be too expensive and too burdensome for patients and clinicians (Borah, Moriarty, Crown, & Doshi, 2014). Within mental health care, and within youth care, various ROM systems already systematically and repeatedly collect data on patients' mental health and function as an indicator of the treatment outcome. Not only can these data be used to monitor individual treatment progress, but they can also be used to evaluate interventions on their outcomes. Using such research findings in clinical practice can help youth and their families receive an intervention that is proven to be effective and evidence-based (APA, 2006). To accomplish this, research findings should effectively be communicated to clinical practice and should be translated into clinically relevant actions and policy considerations. Only then, practice-based evidence can lead to evidence-based practice (Veerman, van Yperen, Bijl, Ooms, & Roosma, 2008).

Because Dutch budgets available for youth care were reduced and were transferred to local authorities in 2015, these authorities should have insight in which interventions are available and which of them are evidence-based: research findings from clinical practice are needed to gain these insights. When, for example, two interventions have the same target population and the intervention that is more effective is also more costly, it is relevant to evaluate these youth care interventions on their cost-effectiveness as an addition to the effectiveness study. This cost-effectiveness analysis is mostly seen as an 'additional' part of the evaluation of interventions, but should become part of the evaluation process itself. Collecting relevant cost data on health care expenditures of patients in for example the ROM system, is a first step towards an economic evaluation. A second step is creating more awareness about the relevance of these types of analyses among clinicians and youth care institutions. For them, evidence based practice should also mean that one knows about the effectiveness of the intervention, its costs, and its cost-effectiveness and that they should be able to explain these results when needed. Moreover, policymakers who decide on the expenditure of youth care budgets (i.e., the municipalities in the Netherlands) should be informed about these types of analyses and the interpretation of the results too, since they should distribute youth care budgets wisely. Additionally, municipalities can play an important role in supporting institutions, scientists, and clinicians translating the data and findings to practically relevant outcomes. When the policymakers, youth care institutions, and clinicians understand each other's language, results can more easily be applied and translated into reimbursement decisions in order to avoid wasting money on less effective and costly youth care interventions.

Recommendations for future research and policy

The importance of evaluating youth care interventions on effectiveness and cost-effectiveness should not only be emphasized within clinical practice, but also within research and policymaking. The following recommendations are important for future research and policy in the field of youth care.

First, evaluating interventions using 'real world data' is becoming more widely accepted (Berger, Dreyer, Anderson, Towse, Sedrakyan, & Normand, 2012). Findings and conclusions from such datasets should be interpreted in light of the selected data and differences with the broader population should be emphasized. Furthermore, the statistical method that is chosen to analyse the data is likely to depend on the content of the data available and should take into account uncertainties within the dataset, such as missing values or having data only for a subgroup of adolescents. Using clinical practice data, or 'real world data', is not a substitution of conducting RCTs, since these study designs can address the comparative efficacy of two interventions without allocation bias (Borah et al., 2014). It is, however, a valid alternative to be able to use clinical practice data and apply correction methods such as the PS. When using methods like the PS, one should consider existing guidelines for observational research (i.e., Berger et al., 2012) so that using non randomized clinical practice data does not become an excuse for not gathering crucial data on patient characteristics and assignment to treatment (Borah et al., 2014). Even more, it could be necessary to develop standards in reporting results of PS methods to be transparent about the assessed balance and treatment outcome (Borah et al., 2014) and to register these observational studies (Berger et al., 2012; Williams, Tse, Harlan, & Zarin, 2010). In addition, the context of the field of youth care at the moment the data was gathered is also relevant, since for example reimbursement decisions or referral policies can change yearly within each municipality. If intervention A, for example, is reimbursed within year 'X', while it is not in year 'Y', the number and 'type' of adolescents assigned to intervention A can differ between these years, which complicates the comparison of intervention A and intervention B over year X and Y. Thus, the context of the selected dataset and of the evaluation study should be considered and discussed when interpreting its findings.

Second, not only research into the effectiveness and cost-effectiveness of youth care should be brought together, but also the criteria used to evaluate this research. The criteria used to rank the level of effectiveness of interventions could for instance be combined with the criteria used to decide on the reimbursement of interventions in youth care. The level of effectiveness of youth interventions can be found in the Database Effective Interventions (Netherlands Youth Institute, 2016). As described in the introduction of this thesis, the DEI is based on the so-called 'effectladder' in which an intervention is marked according to four ranked categories (Veerman & van Yperen, 2008). The criteria to decide upon reimbursement, on the other hand, are formulated from a health economic perspective that is originally based on the 'Trechter van Dunning' (Busschbach & Delwel, 2010; Roscam Abbin, 1991). The decision to reimburse health care interventions is based on four criteria, of which proven effectiveness is the second criterion, after having evaluated the need for intervening, the necessity, in the first place (Busschbach & Delwel, 2010; Roscam Abbin, 1991; Zwaap, Knies, van der Meijden, Staal, & van der Heijden, 2015). The third criterion considers the cost-effectiveness, or efficiency, of the intervention. Combining the criteria could be accomplished by incorporating the youth care perspective, for example given in the criteria of the 'effectladder', in the second criterion of the 'Trechter van Dunning', which is also described in a report on how to decide upon the level of evidence in research and practice concerning health care (Zwaap et al., 2015). Another way to combine the effectiveness and the cost-effectiveness criteria could be to add the criteria for reimbursement of interventions to the level of evidence included in the DEI. This could be a first step in bringing evidence and decisions about reimbursements together in youth care, or at least to make the criteria to decide upon the level of evidence more explicit. A future step might even be to introduce and apply more detailed and explicit decision criteria, which could be brought together in a multi-criteria decision analysis (Baltussen & Niessen, 2006; Thokala & Duenas, 2012). A multi-criteria decision analysis is an approach to explicitly incorporate and weigh several criteria in a systematic and transparent way (Baltussen & Niessen, 2006; Thokala & Duenas, 2012). These criteria are scored and weighted to a sum score per treatment alternative. Policy makers can use this score as a tool to inform their reimbursement decision. As this approach is more widely applied in health technology assessment nowadays, it could be a tool to indeed bringing criteria to decide upon the reimbursement of youth care together and make them explicit and transparent.

Third, cost-effectiveness research in youth care needs guidelines that can improve the implementation of health economic evaluations in this field, because this type of research is not yet widely applied. For example, youth care could incorporate the already existing guidelines and best practices in health economics (Zorginstituut Nederland, 2015), by which it would also follow international health technology assessment standards because these are included in these guidelines. In addition, these guidelines could be adjusted on advises and best practices that specifically fit this type of research in youth care, to support this field in evaluating interventions on cost-effectiveness. The Netherlands organization for health research and development (ZonMw) recently funded research into the cost-effectiveness of youth care interventions (ZonMw, 2016). Part of this project was finding out whether economic evaluations in the youth sector could be standardized and which methodological issues and practical challenges would appear when applying economic evaluations in youth care (Dirksen & Evers, 2016). Some of the issues found were concerned with the perspective of the economic evaluation: If a societal perspective is taken, how are costs and effects distributed over different stakeholders? What time horizon should be used? This would preferably be a long-term estimation to model cost-effectiveness into adulthood, especially because childhood risks can predict economic burden into adulthood (Caspi, Houts, Belsky, Harrington, Hogan, Ramrakha, et al., 2016). However, long-term follow-up data is often lacking and if modelling over long-term, what effects should be taken into account? Another issue of concern is the identification, measurement, and valuation of costs and outcomes: Should we measure costs of the child or adolescent alone, or also costs of the parents, and how are all service types valued if these are not mentioned in the Dutch costing manual (Hakkaart van Roijen, van der Linden, Bouwmans, Kanters, & Tan, 2015)? Another question is whether we should use a generic outcome measure such as Quality Adjusted Life Years (QALY) in health economics, or not? Some of these issues also became apparent while conducting the research in this thesis, since we applied economic evaluation methods in youth care. For example, we modelled long-term cost-effectiveness estimates because we made assumptions on how these effects would last over time. Measuring these effects over time would be preferred, but highly depends on the study design and needs including adolescents and parents for a longer period. Although we measured the effect of the interventions with criminal activity free years, this measure cannot broadly be applied to compare interventions not aimed at reducing criminal activity in youth care. Therefore, a generic outcome which is clinically relevant as well is highly recommended. Though this thesis showed that it is possible to conduct economic evaluation studies in youth care, in light of the issues raised by Dirksen and Evers (2016) it is important to develop standards to conduct or at least report on these issues in an economic evaluation study to be able to compare studies and the results of the interventions evaluated. The methodological and practical challenges raised by Dirksen and Evers (2016) thus provide a starting point to conduct further research and develop future guidelines. Moreover, there are prominent and important steps made in conducting economic evaluation studies in youth care in the Netherlands, of which Kremer and colleagues (2016) gave a first overview. They recommend to validate the model used, to use a generic outcome measure such as the QALY, to include all relevant costs and to report carefully and in detail about the analyses to be able to compare the results with other studies in youth care.

Fourth, an additional recommendation for future studies pertains to the judicial context of some interventions. Adolescents referred to FFT and MST, for instance, often have a court order (Baglivio, Jackowski, Greenwald, & Wolff, 2014). When interventions are used in a judicial context, criminological theories and studies should be considered in evaluating their effectiveness and cost-effectiveness (i.e., Velthoven, 2008). For example, in estimating willingness-to-pay values, one should wonder 'Do we want to pay only for avoiding crimes, or do we want to pay for avoiding crimes in which the costs of investigations, prosecution, witnesses, legal aid, prevention programs, and the valuation of fear are also incorporated?' (Cohen & Piquero, 2009; Cohen, Rust, Steen, & Tidd, 2004). Though the most recent Dutch guidelines on health economic studies mentioned this shortly, if youth care overlaps with the field of criminology there could be an additional set of issues that should be given attention in research and guidelines.

In sum, evaluation studies and economic evaluations studies are an important tool in further developing the content, accessibility, and affordability of youth care interventions in the Netherlands. The developments mentioned above can further direct these types of studies and can bring together clinical practice and research in youth care.

Limitations

Despite the strengths of the studies that are described in this thesis, such as the use of clinical practice data to evaluate the effectiveness of interventions, meaning that interventions and questionnaires were not adjusted for research purposes, and the use of an economic model to illustrate the applicability of these analyses in youth care, a number of limitations merit reflection.

Firstly, the cost-effectiveness model that was introduced in Chapter 2 and further used in the value of information analysis in Chapter 3 was illustrative in the absence of underlying trial data. The analyses and interpretations illustrated how such a method could be applied when evaluating youth care interventions, specifically aimed at reducing juvenile delinquency. In applying these methods, health economic guidelines, nationally and internationally, prescribe the steps that should be taken in conducting a valid and reliable health economic evaluation study (Husereau, Drummond, Petrou, Carswell, Moher, Greenberg, et al., 2013; Zorginstituut Nederland, 2015. We were not able to follow these steps in much detail, because we only had limited data. However, future economic evaluations in youth care should take into account these guidelines and use them as a starting point to apply these methods in youth care, while different recommendations should be made for youth care specifically. Furthermore, the cost-effectiveness study in this thesis was based on methods in health technology assessment and specifically focused on cost-effectiveness analyses. However, there are alternative methods such as a cost-utility analysis, in which the effect is measured and expressed in a generic, preference weighted outcome measure like the QALY. Another alternative is a cost-benefit analysis in which the benefit is expressed in costs as well. In the US, these cost-benefit analyses are widely adopted in the evaluation of youth interventions (Aos, Lieb, Mayfield, Miller, & Pennucci, 2004). In the Netherlands, guidelines for societal cost-benefit analysis have been developed as well (Pomp, Schoemaker, & Polder, 2014). This approach, however, differs notably from the cost-effectiveness analysis presented in this thesis. If economic evaluation methods will be further applied in youth care, a uniform approach that is well documented and described is highly recommended (Dirksen & Evers, 2016).

Secondly, non-randomized data need sophisticated modelling or statistical techniques to control for limitations like uncertainty in the parameter estimates and differences in observed baseline variables. Although our statistical methods were carefully chosen, they are not the only methods available to adjust for data limitations. For example, using information about the parameters available only in the dataset when modelling the cost-effectiveness and estimating the value of conducting further research using a Markov model is one way to fill in the model. An alternative would have been to first systematically search the literature and fill in the model parameters based on this information, then modelling the uncertainty, and then updating the model with the information from the dataset available. An advantage would have been that all available evidence would have been submitted in the model. A disadvantage would have been that the evidence used came from different datasets with different contexts. Then, the question would arise how the conclusions would relate to the specific situation in which the data were gathered. In addition, there are available alternatives to the PS method to adjust for observed baseline differences between intervention groups. This method only controls for measured baseline differences and it is not the only method that can control for such differences. Alternatives such as instrumental variables and multivariate matching methods are also plausible options (Borah et al, 2014; Kreif, Grieve, Radice, Sadique, Ramsahai, & Sekhon, 2012). In addition, among other methods to test the robustness of the findings, like using sensitivity analyses or repeating the analyses in different subsets, one could think of using two or three different statistical techniques to find out whether the results found were robust (Borah et al., 2014; Duncan, Engel, Claessens, & Dowsett, 2014). It could overcome misinterpretations and false conclusions from analyses, especially when only the p-value is used to draw inferences from the analyses (Greenland, Senn, Rothman, Carlin, Poole, Goodman et al., 2016). Results could be presented as a probability that the findings are likely given the selected dataset, and the robustness of the findings can be represented in the method, in the interpretations, and in the selected datasets (Nuzzo, 2014).

Conclusion

Although the methods used were not new in all aspects (e.g., in applying a cost-effectiveness framework or using statistical methods to control for allocation bias), the studies in this thesis showed the practical applicability of these methods and of the use of clinical practice data to answer relevant research questions. More importantly, this thesis showed that different research fields can learn from each other. Health economic evaluations are not yet widely applied in youth care. Therefore, youth care can learn from and adopt these modelling techniques. The other way around, health economics can learn from youth care practice and adopt strategies from that field when deciding on guidelines. Issues that are initially thought to be specific for a certain research area are probably not that specific and can be used in related fields as well.

Thus, this thesis showed that cost-effectiveness analyses provide valuable information that can be used to allocate public budgets on available youth care interventions wisely. Furthermore, using clinical practice data in youth care that are routinely gathered is needed to evaluate interventions on their effectiveness, and is needed to ultimately evaluate these intervention on their costs and effectiveness in every day practice setting. When this information is available, money saved by not reimbursing a costly and ineffective treatment can be spend wisely by helping adolescents and their families with an intervention that is cost-effective.

References

Aos, S., Lieb, R., Mayfield, J., Miller, M., & Pennucci, A. (2004). *Benefits and costs of prevention and early intervention programs for youth*. Olympia: Washington State Institute for Public Policy.

APA Presidential Task Force on Evidence-based practice (2006). Evidence-based practice in psychology. *American Psychologist, 61*, 271-285.

Baglivio, M. T., Jackowski, K., Greenwald, M. A., & Wolff, K. T. (2014). Comparison of Multisystemic Therapy and Functional Family Therapy effectiveness: A multiyear statewide propensity score matching analysis of juvenile offenders. *Criminal Justice and Behavior, 41*, 1033-1056.

Baltussen, R., & Niessen, L. (2006). Priority setting of health interventions: The need for multi-criteria decision analysis. *Cost Effectiveness and Resource Allocation, 4*, 14.

Berger, M. L., Dreyer, N., Anderson, F., Towse, A., Sedrakyan, A., & Normand, S. L. (2012). Prospective observational studies to assess comparative effectiveness: The ISPOR good research practices task force report. *Value in Health, 15*, 217-230.

Borah, B. J., Moriarty, J. P., Crown, W. H., & Doshi, J. A. (2014). Applications of propensity score methods in observational comparative effectiveness and safety research: Where have we come and were should we go? *Journal of Comparative Effectiveness Research, 3*, 63-78.

Busschbach, J. J. V., & Delwel, G. O. (2010). *Het pakketprincipe kosteneffectiviteit: Achtergrondstudie ten behoeve van de 'appraisal' fase in pakketbeheer (publicatienummer 291) [A background study on the 'costeffectiveness' package principle for the benefit of the appraisal phase in package management]*. Diemen: College voor zorgverzekeringen.

Caspi, A., Houts, R. M., Belsky, D. W., Harrington, H., Hogan, S., Ramrakha, S., Poulton, R., & Moffitt, T. E. (2016). Childhood forecasting of a small segment of the population with large economic burden. *Nature Human Behaviour, 1*, 0005.

Cohen, M. A., Rust, R. T., Steen, S., & Tidd, S. T. (2004). Willingness-to-pay for crime control programs. *Criminology, 42*, 89-109.

Cohen, M. A., & Piquero, A. R. (2009). New evidence on the monetary value of saving a high risk youth. *Journal of Quantitative Criminology, 25*, 25-49.

Dirksen, C. D., & Evers, S. M. A. A. (2016). *Broad consultation as part of the standardization of economic evaluation research in the youth sector*. Maastricht: Maastricht University Medical Centre & Maastricht University.

Duncan, G. J., Engel, M., Claessens, A., & Dowsett, C. J. (2014). Replication and robustness in developmental research. *Developmental Psychology, 50*, 2417-2425.

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, P values, confidence intervals, and power: A guide to misinterpretations. *European Journal of Epidemiology, 31*, 337-350.

Hakkaart – van Roijen, L., van der Linden, N., Bouwmans, C., Kanters, T., & Tan, S. S. (2015). *Kostenhandleiding: Methodologie van kostenonderzoek en referentieprijzen voor economische evaluatie in de gezondheidzorg [Dutch costing manual for health care]*. Diemen: Zorginstituut Nederland.

Husereau, D., Drummond, M., Petrou, S., Carswell, C., Moher, D., Greenberg, D., … , Loder, E. (2013). Consolidated Health Economic Evaluation Reporting Standards (CHEERS) Statement. *Value in Health, 16*, 231-50.

Kreif, N., Grieve, R., Radice, R., Sadique, Z., Ramsahai, R., & Sekhon, J. S. (2012). Methods for estimating subgroup effects in cost-effectiveness analyses that use observational data. *Medical Decision Making, 32*, 750-763.

Kremer, I. E. H., Kann, D., van den Berg, G., Dirksen, C. D., Hiligsmann, M., & Evers, S. M. A. A. (2016). *Welke jeugdinterventies in Nederland zijn kosteneffectief? Systematische review naar de huidige stand van zaken [in Dutch]*. Utrecht: Nederlands Jeugdinstituut.

Netherlands Youth Institute (2016). *Databank effectieve jeugdinterventies: Over de databank [Databank Effective youth interventions: About the databank]*. Retrieved from http://www.nji.nl/nl/Databank/Databank-Effectieve-Jeugdinterventies-Over-de-databank.html.

Nuzzo, R. (2014). Statistical errors: P values, the 'gold standard' of statistical validity, are not as reliable as many scientists assume. *Nature, 506*, 150-152.

Pomp, M., Schoemaker, C. G., & Polder, J. J. (2014). *Themarapport Volksgezondheid Toekomst Verkenning. Op weg naar maatschappelijke kosten-baten analyses voor preventie en zorg [Societal cost-benefit analyses within prevention and health care]*. Bilthoven: Rijksinstituut voor Volksgezondheid en Milieu (RIVM).

Roscam Abbin, H. D. C. (1991). Kiezen en delen; rapport van de commissie Keuzen in de zorg (Commissie-Dunning) [Choosing or sharing: A report of the committee on choices in health care]. *Nederlands Tijdschrift voor Geneeskunde, 135*, 2239-2241.

Thokala, P., & Duenas, A. (2012). Multiple criteria decision analysis for health technology assessment. *Value in Health, 15*, 1172-1181.

Veerman, J. W., van Yperen, T., Bijl, B., Ooms, H., & Roosma, D. (2008). Praktijkgestuurd effectonderzoek maakt hulpverlening beter [Practice-driven effectiveness research improves youth care]. *Jeugd en Co Kennis, 2*, 8-18.

Veerman, J. W., & van Yperen, T. A. (2008). Wat is praktijkgestuurd effectonderzoek? In: T. A. van Yperen, & J. W. Veerman (Eds.), *Zicht op effectiviteit. Handboek voor praktijkgestuurd effectonderzoek in de jeugdzorg [Effectiveness in practice: Handbook to practice-driven effectiveness research in youth care]* (pp. 17-34). Delft: Eburon.

Velthoven, B. C. J. (2008). Kosten-batenanalyse van criminaliteitsbeleid [Cost-benefit analyses in crime policy]. *Tijdschrift voor strafrechtspleging, 87*, 108-120.

Williams, R. J., Tse, T., Harlan, W. R., & Zarin, D. A. (2010). Registration of observational studies: Is it time?. *Canadian Medical Association Journal, 182*, 1638-1642.

ZonMw (2016). *Programma 'Effectief werken in de jeugdsector' [Program 'Working effectively within youth care']*. Retrieved from http://www.zonmw.nl/nl/onderzoek-resultaten/jeugd/programmas/programma-detail/effectief-werken-in-de-jeugdsector/.

Zorginstituut Nederland (2015). *Richtlijn voor het uitvoeren van economische evaluaties in de gezondheidszorg [Guideline for economic evaluations in health care]*. Diemen: Zorginstituut Nederland.

Zwaap, J., Knies, S., van der Meijden, C., Staal, P., & van der Heiden, L. (2015). *Kosteneffectiviteit in de praktijk (volgnummer 2015076142) [Costeffectiveness in practice]*. Diemen: Zorginstituut Nederland.

Addendum

Summary

In health care, economic evaluation studies using state of the art decision analytic methods are common practice. The data used in such studies are often gathered within randomized clinical trials, but observational data are considered a valid alternative. In contrast, in Dutch youth care   cost-effectiveness research and the use of available, non-randomized data are not common practice. Therefore, this thesis addressed the use and feasibility of state of the art decision analytical methods in youth care and showed how available, non-randomized data can be used when evaluating youth care.

In **Chapter 2**, we constructed a probabilistic Markov model to assess the cost-effectiveness of systemic interventions in youth care. To illustrate model functioning and the interpretation of the results, Functional Family Therapy (FFT) was compared to Treatment as Usual (TAU). The assumptions and parameters normally used to evaluate health care interventions were adjusted to better fit the characteristics of the systemic interventions. The treatment outcome, for example, was defined as Criminal Activity Free Years (CAFY) to address the clinical and the societal effect of the interventions. In addition, costs of resource use of the adolescents and of one of their parents were taken into account in the model, because systemic interventions are not only aimed at the referred client (i.e., the adolescent), but also at the system surrounding him or her. Resource use was defined broader than health care costs alone; The assignment to a foster home, a residential institution, or several contacts within the criminal justice system, for example, were also measured and expressed in costs per adolescent. Because the interventions were provided to youth aged 12 to 18 years and treatment effects can be expected to last into adulthood, long-term cost-effectiveness was estimated. The results of this model showed that (slightly adjusted) common economic evaluation methods are applicable in evaluating youth care. Moreover, the results can be interpreted in the same way as other economic evaluation results. These findings are an important first step towards a more systematic application of this method in youth care. The findings also led to important recommendations, such as defining an outcome that allows comparisons with other youth care interventions and taking into account costs outside the health care system (e.g., victim costs and reduced costs of avoided crimes).

Because the input in a cost-effectiveness analysis can be uncertain due to imperfect or incomplete estimates of costs and effects, the decision whether or not to reimburse an intervention, based on this cost-effectiveness analysis, is marked by uncertainty. Further research may reduce this uncertainty, but is probably not without costs. The added value of future cost-effectiveness research is estimated in a value of information analysis. The aim of **Chapter 3** was to investigate whether a value of information analysis, which is commonly applied in health care evaluation studies, is feasible and meaningful in youth care as well, and especially applicable to systemic interventions aimed at reducing criminal activity of adolescents. The cost-effectiveness model of Chapter 2 was used to illustrate such a value of information analysis, in which FFT was compared to the Course House (i.e., a comparative intervention on which we had literature data). Model parameters were grouped to identify those parameters that contributed most to the uncertainty and, therefore, would be most worthwhile for future research. Thereby, it was needed to assume a ‘willingness-to-pay (WTP) value’ for one CAFY (i.e., what is society willing to pay for one year without criminal activity of one adolescent?), because the expected value of further information depends on this WTP value. The illustrative WTP value was based on and averaged over different values society wants to pay to prevent crimes like robbery and vandalism. The illustrative value of information analysis revealed that at a societal willingness-to-pay of €71,700 per criminal activity free year, further research to eliminate parameter uncertainty was valued at €176 million. This means that society should be willing to spend a maximum of €176 million in reducing decision uncertainty in the cost-effectiveness of the two interventions. In particular, most of the uncertainty was found in the effects of the interventions, which was translated to transition probabilities of adolescents moving from one state to another in the Markov model, and, to a lesser extent, in the intervention costs of the Course House and the direct non health-care costs in both model states. This illustrative analysis showed that the results were meaningful and can be interpreted according to health care evaluation studies. Moreover, it showed that this analysis can be helpful in justifying additional research funds to further inform the reimbursement decision with regard to youth care interventions. Finally, this study made important recommendations for applying this method more systematically in youth care, like defining a WTP value to the defined outcome and, as in the cost-effectiveness analysis itself, determining the range of effects and costs that should be taken into account.

In **Chapters 4** and **5**, the aim was to investigate and illustrate the use of available, non-randomized data when evaluating treatments in clinical practice. The propensity score (PS) method was used to control for initial differences due to the non-random assignment of adolescents to the treatments evaluated. In **Chapter 4**, the feasibility of the univariate and multivariate PS method was demonstrated in subgroup analyses of outcomes research. The performance of using the univariate PS was tested using Monte Carlo simulations with additional adjustment on the subgroups. . The multivariate PS was estimated by combining the treatment groups and subgroup categories. The treatment effect and subgroup effects were estimated in a linear regression model adjusting for either of the two PS estimations. The bias and mean squared error showed minor differences between both PS methods, with marginally lower values of the bias and mean squared error when using the multivariate PS. Clinical practice data from a large effectiveness study on psychotherapy in personality disorders were used to compare the two methods. Using these data, the differences between short-term and long-term treatment were compared using the severity of patients' problems as the subgroup of interest. Both the univariate and multivariate PS estimations yielded similar results. The results of this study support the use of the multivariate PS with slightly less biased estimated treatment effects. The choice of the subgroup of interest, however, should be clinically relevant and influences the choice for analyses and interpretations as well.

In **Chapter 5**, two youth care interventions, FFT and Multisystemic Therapy (MST) were compared on their effectiveness using non-randomized, clinical practice data from adolescents assigned to either one of these interventions (422 MST; 275 FFT) at the Viersprong, institute for personality disorders and behavioral problems in the Netherlands. Data were gathered within Routine Outcome Monitoring and the effectiveness of the two interventions was estimated using the PS method to control for initial measured differences between the treatment groups. The primary outcome was externalizing problem behavior. The secondary outcomes were the proportion of adolescents who were living at home, who were engaged in school or work after treatment, and who lacked police contact during treatment. No difference was found between MST and FFT regarding externalizing problem behavior, but the adolescents who received MST were more likely to be engaged in school or work after treatment compared with FFT. Because the risk-need-responsivity (RNR) model guided treatment assignment, effectiveness was also estimated in youth with and without a court order, as an indicator of their risk level. For adolescents without a court order, MST yielded a larger effect on externalizing problems. The propensity score could not balance the treatment groups in the subsample of adolescents with a court order, and, therefore, MST and FFT could not be compared on their effectiveness in this subsample. Though treatment assignment was based on the RNR model, results in the group without a court order were not in accordance with this model, while higher-risk adolescents with a court order were indeed more often assigned to the more intensive treatment, namely MST. Although MST is expected to be the more expensive treatment because it is more intensive, estimating the cost-effectiveness of these interventions seems only relevant in the subgroup of adolescents without a court order.

In the general discussion in **Chapter 6** we summarized the findings of this thesis and we conclude that cost-effectiveness analyses provide valuable information to allocate public budgets to available youth care interventions wisely. In addition, we conclude that clinical practice data in youth care, that is routinely gathered, is needed to evaluate interventions on their effectiveness, and ultimately on their cost-effectiveness. Furthermore, the implications for youth care and for future research and policy are discussed. From a youth care perspective, this thesis implicates that research findings from evaluation studies in every day practice should be brought to clinical practice by effectively communicating these findings to clinicians, patients, and policymakers to enable them to assign youth to evidence-based interventions. Data on health care costs besides the intervention itself should be routinely gathered to evaluate interventions on their cost-effectiveness. Cost-effectiveness research is especially important since a large number of municipalities have to decide on reimbursing interventions and current standards may differ between municipalities. From a research and policy perspective, it is recommended, whenever possible, to follow guidelines in reporting findings from studies using clinical practice data, since in such studies the context of the interventions is even more important than in randomized controlled trials. Furthermore, the criteria to decide upon the level of effectiveness and cost-effectiveness evidence should be made more explicit and transparent, especially if this evidence is used to decide upon the reimbursement of youth care interventions. To improve the implementation of health economic evaluation studies in youth care, guidelines on cost-effectiveness research in youth care should be developed. When developing these guidelines, one should be prone to learning from other related research fields and combine relevant perspectives and methods.

Samenvatting

Economische evaluaties op basis van besliskundige modellen worden binnen de gezondheidszorg veelvuldig toegepast om interventies te onderzoeken op kosteneffectiviteit. De gegevens die gebruikt worden in deze evaluaties, worden veelal verzameld binnen gerandomiseerd onderzoek. Observationeel of quasi experimenteel onderzoek, waarin deelnemers niet op basis van toeval zijn toegewezen aan interventies, kunnen als een valide alternatief worden beschouwd, onder de voorwaarden dat met statistische technieken op een valide manier gecorrigeerd wordt voor verschillen in baseline kenmerken tussen de groepen. In tegenstelling tot de gezondheidszorg, staat binnen de jeugdzorg het onderzoek naar kosteneffectiviteit van interventies en het gebruik maken van observationele data nog in de kinderschoenen. In dit proefschrift staan deze onderwerpen dan ook centraal: 1) het gebruik en de interpreteerbaarheid van economische evaluatiemethoden om de kosteneffectiviteit van interventies binnen de jeugdzorg te kunnen bepalen en 2) het gebruik van observationele en reeds beschikbare, niet gerandomiseerde gegevens in onderzoek naar de effectiviteit van jeugdinterventies.

Om de toepassing van een kosteneffectiviteitsanalyse in de jeugdzorg te illustreren, wordt in **Hoofdstuk 2** een Markovmodel gepresenteerd. Als voorbeeld wordt Functionele Gezinstherapie (Functional Family Therapy, FFT) vergeleken met andere beschikbare behandelingen. De parameters en aannames in het model werden specifiek toegespitst op de vergelijking van deze jeugdinterventies. Onder andere werd de uitkomstmaat gedefinieerd als crimineel-vrije-jaren (CAFY) om daarmee het klinische en maatschappelijke effect van de interventies weer te kunnen geven. Daarnaast werd niet alleen het zorggebruik van de jongere, maar ook dat van een van de ouders meegenomen, omdat systemische interventies zich ook richten op de verbetering van het gezinsfunctioneren en de relaties in het systeem rond een jongere. Omdat deze interventies zich richten op meer dan alleen gezondheidswinst, zijn ook kosten buiten de gezondheidszorg nadrukkelijk betrokken in het model, zoals de kosten van een uithuisplaatsing of tijdelijke plaatsing in een pleeggezin en de kosten binnen het justitieel kader. Vanwege het verwachte lange termijn effect van de interventies werden in het Markovmodel effecten tot 30 jaar later gemodelleerd. De resultaten tonen aan dat kosteneffectiviteitsanalyses op basis van besliskundige modellen bruikbaar en toepasbaar zijn binnen de jeugdzorg en dat de resultaten op eenzelfde manier kunnen worden geïnterpreteerd als binnen de gezondheidszorg. De resultaten vormen een belangrijke stap in het toepassen van economische evaluaties, en meer specifiek kosteneffectiviteitsanalyses, binnen de jeugdzorg. Het verdient aanbeveling in toekomstige economische evaluaties in de jeugdzorg rekening te houden met bijvoorbeeld het definiëren van een meer generieke uitkomstmaat die vergelijkingen met andere jeugdinterventies mogelijk maakt. Ook kosten van zorggebruik buiten de gezondheidszorg die in ons onderzoek niet gemeten zijn, zoals mogelijke kosten van slachtoffers van criminaliteit gepleegd door de jongere en kosten van criminaliteit die voorkomen zijn door de interventies, zouden in toekomstige analyses meegenomen kunnen worden.

De parameters in een Markovmodel zijn schattingen op basis van data uit de klinische praktijk of op basis van literatuur. Deze schattingen hebben een mate van onzekerheid die is weergegeven in termen van de verdelingen van de parameters. Als op basis van die onzekerheid een beslissing rond het vergoeden van jeugdinterventies wordt genomen, wordt die beslissing eveneens gekenmerkt door onzekerheid. Verder onderzoek kan de onzekerheid rond de schatter van de parameter reduceren, en daarmee de onzekerheid in het model. Met dergelijk aanvullend onderzoek zijn echter ook kosten gemoeid. Om een goede afweging te kunnen maken met betrekking tot investeringen in verder onderzoek, kan een zogenaamde 'value-of-information' analyse uitgevoerd worden. In deze analyse worden de kosten berekend van het nemen van een verkeerde beslissing als gevolg van onzekerheid in de gebruikte parameters. Op die manier kan de waarde van het verzamelen van aanvullende informatie worden bepaald, de zogenaamde 'verwachte waarde van perfecte informatie'. Als deze perfecte informatie voorhanden zou zijn, dan zou dit de onzekerheid in een model kunnen beperken. Het doel van **Hoofdstuk 3** was te bepalen of een value-of-information analyse zinvol en interpretabel is in aanvulling op de kosteneffectiviteitsanalyse uit Hoofdstuk 2. Dergelijke analyses worden in de gezondheidszorg namelijk al gebruikt, maar in de jeugdzorg nog niet. In de value-of-information analyse werd FFT vergeleken met het Kursushuis (een vergelijkbare interventie). In kosteneffectiviteitsanalyses in het algemeen, en in een value-of-information analyse in het bijzonder, is het van belang om een zogenaamde kosteneffectiviteitsgrenswaarde te bepalen. Dat is de waarde die de budgethouder of de maatschappij bereid is te betalen voor in dit geval één crimineel-vrij jaar. De waarde van het nog uit et voeren onderzoek is afhankelijk van deze grenswaarde. Ter illustratie is deze kosteneffectiviteitsgrenswaarde berekend als gemiddelde van verschillende grenswaarden ter voorkoming van verschillende soorten criminaliteit, zoals vandalisme en diefstal. De value-of-information analyse liet zien dat bij een kosteneffectiviteitsgrenswaarde van €71,700 per crimineel-vrij-jaar, verder onderzoek om onzekerheid rond de modelparameters te reduceren gewaardeerd werd op €176 miljoen. Dit betekent dat de maatschappij maximaal €176 miljoen zou moeten willen uitgeven om de onzekerheid rond de vergoedingsbeslissing van de jeugdinterventies FFT en Kursushuis te willen reduceren en perfecte informatie te hebben. De onzekerheid in het model zat met name in de geschatte effectiviteit van de interventies. In mindere mate zat de onzekerheid in de geschatte kosten van het Kursushuis en de directe kosten buiten de gezondheidszorg. Concluderend werd in Hoofdstuk 3 aangetoond dat de resultaten van een value-of-information analyse zinvol en bruikbaar zijn in de jeugdzorg. Het verdient aanbeveling een kosteneffectiviteitsgrenswaarde te definiëren voor een uitkomstmaat die breed toepasbaar is in het jeugdveld en om de relevante kosten en effecten van jeugdinterventies breed te definiëren en mee te wegen in een kosteneffectiviteitsanalyse en in een value-of-information analyse.

In **Hoofdstuk 4** en **5** stond het gebruiken van beschikbare, observationele en niet gerandomiseerde gegevens centraal bij de evaluatie van jeugdinterventies. De 'propensity score' werd gebruikt om te corrigeren voor verschillen tussen jongeren in de verschillende behandelingen. Zulke verschillen kunnen ontstaan door het onwillekeurig toewijzen van jongeren aan een behandeling, of anders gezegd, doordat alleen jongeren met specifieke kenmerken in aanmerking komen voor een bepaalde interventie. Met de propensity score wordt gecorrigeerd voor de kans op toewijzing aan een behandelgroep, gegeven een set van gemeten baseline kenmerken. In **Hoofdstuk 4** is de bruikbaarheid van de univariate propensity score (toepasbaar om twee behandelgroepen te vergelijken) en de multivariate propensity score (toepasbaar om meerdere behandelgroepen te vergelijken) aangetoond in subgroep analyses. Met Monte Carlo simulaties zijn fictieve datasets gegenereerd. Hierin is met de univariate propensity score gecorrigeerd voor baseline kenmerken terwijl daarnaast een subgroep effect berekend is. De multivariate propensity score corrigeerde voor de baseline verschillen tussen de behandelgroepen gecombineerd met de subgroepen om het behandeleffect in subgroepen te kunnen berekenen. Beide methoden lieten minimale verschillen zien in de berekende behandeluitkomst, met iets lagere berekende afwijkingen van de juist voorspelde waarde voor de multivariate propensity score. Wanneer beide propensity scores toegepast werden op data van een omvangrijke studie naar de effectiviteit van psychotherapie bij persoonlijkheidsstoornissen (SCEPTRE), lieten beide propensity scores vergelijkbare resultaten zien. Concluderend lijkt de behandeluitkomst iets zuiverder geschat te kunnen worden door te corrigeren met de multivariate propensity score. De keuze voor de analyse, en dus voor het toepassen van de univariate of multivariate propensity score, is echter ook afhankelijk van de het subgroep effect waarin men geïnteresseerd is.

In **Hoofdstuk 5** is de effectiviteit van twee jeugdinterventies, FFT en Multisysteem Therapie (MST) vergeleken. Hierbij is gebruik gemaakt van data uit de klinische praktijk van de Viersprong, een hoogspecialistische GGZ instelling. De jongeren werden volgens de reguliere klinische praktijk toegewezen aan een van beide interventies (422 jongeren aan MST; 275 aan FFT). De gegevens werden verzameld met 'Routine Outcome Monitoring'. De effectiviteit is onderzocht door de univariate propensity score toe te passen om te corrigeren voor gemeten baseline verschillen tussen de behandelgroepen. De primaire uitkomstmaat was externaliserend probleemgedrag, gemeten na afloop van de behandeling. De secundaire uitkomstmaten was het percentage jongeren dat thuis woonde na afloop van de behandeling, het percentage jongeren dat een zinvolle dagbesteding had na de behandeling, zoals werk of school, en het percentage jongeren dat geen politiecontact heeft gehad tijdens de behandeling. Kijkend naar de gehele behandelde groep verschilden MST en FFT niet wat betreft hun effect op externaliserend probleemgedrag. Wel had na afloop van MST een groter percentage van de jongeren een zinvolle dagbesteding dan na FFT. Omdat het 'risk-need-responsivity' (RNR) model de basis vormde om jongeren aan MST of FFT toe te wijzen, leek het zinvol ook naar subgroepen te kijken. Volgens het RNR-model zouden jongeren met meer risicofactoren namelijk meer moeten profiteren van een intensievere behandeling. Daarom werd de effectiviteit van MST en FFT ook onderzocht in subgroepen van jongeren met een hoog of laag risico. Als indicatie voor een hoog risico werd de aanwezigheid van een civielrechtelijke of strafrechtelijke maatregel gebruikt. Wanneer geen sprake was van een maatregel, werd het risico laag geacht. Uit de subgroep analyses bleek dat als jongeren voorafgaand aan de behandeling geen strafrechtelijke of civielrechtelijke maatregel hadden, MST een beter effect behaalde op externaliserend probleemgedrag dan FFT. Dit is niet in overeenstemming met het RNR-model waarin verwacht zou worden dat een minder intensieve behandeling, FFT, voldoende effectief zou zijn om deze jongeren met een laag risico te behandelen. In de groep jongeren die voorafgaand aan de behandeling wel een maatregel had, was het niet mogelijk om met de propensity score te corrigeren voor baseline verschillen tussen de behandelgroepen; MST en FFT konden niet worden vergeleken in deze subgroep. In lijn met het RNR-model bleek dat jongeren met een maatregel voorafgaand aan de behandeling inderdaad vaker toegewezen waren aan MST dan aan FFT. Omdat verwacht wordt dat MST een meer intensieve en duurdere behandeling is dan FFT, lijkt aanvullend onderzoek naar de kosteneffectiviteit van MST en FFT op basis van deze resultaten alleen relevant in de groep jongeren zonder maatregel voorafgaand aan de behandeling.

In **Hoofdstuk 6** zijn de bevindingen uit de voorgaande hoofdstukken samengevat en als geheel nader beschouwd. Hieruit kan geconcludeerd worden dat kosteneffectiviteitsanalyses bruikbaar en waardevol zijn binnen het jeugdveld en dat de resultaten uit deze analyses gebruikt zouden kunnen worden om beschikbare budgetten in het jeugdveld te verdelen. Daarnaast hebben de onderzoeken naar de effectiviteit en kosteneffectiviteit van jeugdinterventies aangetoond dat het gebruik van data verzameld in de klinische praktijk, zoals via Routine Outcome Monitoring, valide en zinvol is. In Hoofdstuk 6 zijn ook de implicaties van dit proefschrift voor praktijk, onderzoek en beleid beschreven. De bevindingen uit wetenschappelijk onderzoek dienen actief bij hulpverleners, cliënten en beleidsmakers onder de aandacht gebracht te worden, zodat alle partijen over dezelfde informatie beschikken om te kiezen voor effectieve en kosteneffectieve interventies. Om de kosteneffectiviteit van interventies in kaart te brengen, zou het zorggebruik van jongeren routinematig gemeten moeten worden, bijvoorbeeld in Routine Outcome Monitoring. Onderzoek naar de kosteneffectiviteit van interventies lijkt immers een grotere rol te gaan spelen nu gemeenten verantwoordelijk zijn voor de jeugdzorg budgetten. Onderzoekers die gebruik maken van data uit de klinische praktijk, zouden richtlijnen moeten volgen die aangeven op welke manier transparant gerapporteerd kan worden over het gebruik van observationele data in evaluatiestudies. Ten slotte, als onderzoeksresultaten gebruikt worden om een jeugdinterventie al dan niet te vergoeden, zouden de vergoedingscriteria inzichtelijk en transparant moeten zijn. Er is grote behoefte aan richtlijnen voor kosteneffectiviteitsonderzoek in het jeugdveld om dergelijk onderzoek in het jeugdveld te bevorderen en te verbeteren. Bij het ontwikkelen van deze richtlijnen kan het noodzakelijk zijn over de grenzen van het eigen (jeugd)werkveld te kijken en gebruik te maken van andere relevante perspectieven en methoden.

PhD Portfolio

Name PhD student:	Hester van Eeren
Erasmus MC Department:	Psychiatry, section Medical Psychology and Psychotherapy
Research school:	Netherlands Institute for Health Sciences (NIHES)
PhD period:	2010-2016
Promotoren:	Prof.dr. J.J. van Busschbach
	Prof.dr. R.H.J. Scholte
Supervisor:	Dr. R.E.A. van der Rijken

1. PhD training

	Year	Workload (Hours/ ECTS)
General courses		
Quality of Life Measurement (NIHES, Erasmus MC, Rotterdam)	2010	0.9 ECTS
Interdisciplinary Postgraduate Training in Mental Health Policy and Economics Research (International Centre of Mental Health Policy and Economics, Milan, Italy)	2010	2 ECTS
Cost-effectiveness modelling methods (Maastricht University, Maastricht)	2010	6 ECTS
Basiscursus didactiek (teach-the-teacher) (Erasmus MC, Rotterdam)	2010	1 ECTS
Advanced Modelling Methods for Health Economic Evaluation (York University, York, UK)	2011	24 hours
Minicursus Methodologie van Patiëntgebonden Onderzoek en Voorbereiding van Subsidieaanvragen (Erasmus MC, Rotterdam)	2012	8 hours
Basiscursus Regelgeving en Organisatie van Klinische trials (BROK) (Erasmus MC, Rotterdam)	2012	8 hours
Academic writing in English (Erasmus MC, Rotterdam)	2013	4 ECTS
Psychiatric Epidemiology (NIHES, Erasmus MC, Rotterdam)	2013	1.1 ECTS
Cursus wetenschappelijke integriteit (Erasmus MC, Rotterdam)	2014	8 hours

Specific courses		
Master in Health Science, specialization 'Epidemiology' (NIHES, Erasmus MC, Rotterdam)	2011-2013	70 ECTS

Seminars and workshops		
Workshop Formuleren subsidieaanvragen (ZonMw, Den Haag)	2013	16 hours
Workshop Tentamen vragen maken (Erasmus MC, Rotterdam)	2014	4 hours
Workshop Hoorcollege geven (Erasmus MC, Rotterdam)	2014	4 hours
Workshop Individueel begeleiden (Erasmus MC, Rotterdam)	2014	4 hours
Workshop Feedback geven (Erasmus MC, Rotterdam)	2014	4 hours
Workshop Coachen van toekomstige Erasmusartsen (Erasmus MC, Rotterdam)	2015	4 hours

Presentations at national and international meetings

Finding treatmen teffects within subgroups when using the propensity score to control for selection bias: A monte carlo simulation study (oral), 14th International Society for Pharmacoeconomics and Outcomes Research (ISPOR), Madrid, Spain	2011	1 ECTS
The use of the propensity score (oral), Afdeling Ouderenzorg AMC, Amsterdam	2012	1 ECTS
Subgroup analysis when using the propensity score: A Monte Carlo simulation study (oral), Society for Medical Decision Making, Oslo, Norway	2012	1 ECTS
An illustrative Value of Information analysis (VoI) applied to systemic interventions aimed to reduce youth delinquency (oral), department of Psychiatry, Erasmus MC, Rotterdam	2013	1 ECTS
An illustrative value of information analysis applied to systemic interventions aimed to reduce youth delinquency (oral), ICMPE, Venice, Italy	2013	1 ECTS
How can we reduce uncertainty in modeling the cost-effectiveness of two systemic interventions aimed at reducing juvenile delinquency? (oral), MST Europe conference, London, UK	2014	1 ECTS
Comparing the effectiveness of Functional Family Therapy and Multisystemic Therapy using the propensity score method (oral), EUSARF, Oviedo, Spain	2015	1 ECTS
Various presentations at internal research meetings of de Viersprong and section Medical Psychology and Psychotherapy	2010-2015	2 ECTS

National and international conferences

The Low Lands Health Economists' Study Group, Egmond aan Zee	2010	16 hours
13th Annual European Congress of the International Society for Pharmacoeconomics and Outcomes Research (ISPOR), Prague, Czech	2010	32 hours
The Low Lands Health Economists' Study Group, Soesterberg	2011	16 hours
Lustrum symposium Nederlandse Vereniging voor Technology Assessment in de Gezondheidszorg (NVTAG), Utrecht	2011	4 hours
14th Annual European Congress of the International Society for Pharmacoeconomics and Outcomes Research (ISPOR), Madrid, Spain	2011	32 hours
14th Biennial European meeting of the Society for Medical Decision Making, Oslo, Norway	2012	24 hours
Beyond the QALY (iMTA), Rotterdam	2012	4 hours
11th Workshop on costs and assessment in Psychiatry, ICMPE, Venice, Italy	2013	16 hours
MST Europe conference, London, UK	2014	16 hours
Symposium Maatschappelijke kosten-baten analyse en intersectorale kosten en baten: een stap voorwaarts!? (NVTAG), Den Haag	2014	4 hours
Zomersymposium (NVTAG), Utrecht	2015	4 hours

Other

Expertmeeting: Ophalen en verspreiden goede voorbeelden over kosteneffectieve zorg' (Nederlands Jeugd Instituut, Utrecht)	2014	16 hours
Deelname in werkgroep 'Onderzoek' (NVTAG)	2015-2016	10 hours

2. Teaching

Teaching activities

Teaching assistant SPSS practicals (Universiteit Leiden, Leiden)	2009	1 ECTS
Communicatie en Attitude onderwijs in het medische curriculum (Erasmus MC, Rotterdam)	2010-2015	0.2 fte
Vaardighedenonderwijs Onbegrepen pijn, Slecht nieuws, Counseling, Beslissingen rondom aangeboren afwijkingen (Erasmus MC, Rotterdam)	2010-2015	1 ECTS
Basiskwalificatie onderwijs (BKO) (Erasmus MC, Rotterdam)	2013-2014	5 ECTS
Supervising Master thesis Methodology & Statistics, Pyschology (Universiteit Leiden, Leiden)	2014-2015	2 ECTS
Supervising literature review medical students (Erasmus MC, Rotterdam)	2015	0.5 ECTS
Coaching medical students (Erasmus MC, Rotterdam)	2015-2016	1 ECTS

List of publications

This thesis

Eeren, H.V., Goossens, L.M.A., Scholte, R.H.J., Busschbach, J.J.V., & van der Rijken, R.E.A. (2016). Multisystemic Therapy and Functional Family Therapy compared on their effectiveness using the propensity score method. *Submitted.*

Eeren, H.V., Schawo, S.J., Scholte, R.H.J., Busschbach, J.J.V., & Hakkaart, L. (2015). Value of information analysis applied to the economic evaluation of interventions aimed at reducing juvenile delinquency: An illustration. *PLOS ONE, 10*, e0131255.

Eeren, H.V., Spreeuwenberg, M.D., Bartak, A., de Rooij, M., & Busschbach, J.J.V. (2015). Estimating subgroup effects using the propensity score method: A practical application in outcomes research. *Medical Care, 53*, 366-373.

Schawo, S.J., **van Eeren, H.**, Soeteman, D.I., van der Veldt, M.C., Noom, M.J., Brouwer, W., Busschbach, J.J.V., & Hakkaart, L. (2012). Framework for modelling the cost-effectiveness of systemic interventions aimed to reduce youth delinquency. *The Journal of Mental Health Policy and Economics, 15*, 187-196.

Other publications

Laurenssen, E.M.P., **Eeren, H.V.**, Kikkert, M.J., Peen, J., Westra, D., Dekker, J.J.M., & Busschbach, J.J.V. (2016). The burden of disease in patients eligible for Mentalization-Based Treatment (MBT): Quality of life and costs. *Health and Quality of Life Outcomes, 14*, 145.

Laurenssen, E.M.P., Smits, M.L., Bales, D.L., Feenstra, D.J., **Eeren, H.V.**, Noom, M.J., Köster, M.A., Lucas, Z., Timman, R., Dekker, J.J.M., Luyten, P., Busschbach, J.J.V., & Verheul, R. (2014). Day hospital mentalization-based treatment versus intensive outpatient mentalization-based treatment for patients with severe borderline personality disorder: protocol of a multicentre randomized clinical trial. *BMC Psychiatry, 14*, 301.

Laurenssen, E.M.P., Westra, D. Westra, D., Kikkert, M.J., Noom, M.J., **Eeren, H.V.**, Broekhuyzen van, A.J., Peen, J., Luyten, P., Busschbach, J.J.V., & Dekker, J.M. (2014). Day Hospital Mentalization-Based Treatment (MBT-DH) versus treatment as usual in the treatment of severe borderline personality disorder: protocol of a randomized controlled trial. *BMC Psychiatry, 14*, 149.

In preparation

Blankestein, A., van der Rijken, R.E.A., **Eeren, H.V.**, Lange, A., Scholte, R.H.J., Moonen, X, de Vuyst, K., Leunissen, J., & Didden, R. (2017). Evaluating the effectiveness of Multisystemic Therapy for adolescents with intellectual disabilities and their parents.

Goorden, M., Reckers, V., Dijkgraaf, M., **Eeren, H.V.**, McCollister, K., & Hakkaart – van Roijen, L. (2017). Unit costs of delinquent acts for use in economic evaluations.

Smits, M.L., Feenstra, D.J., **Eeren, H.V.**, Bales, D.L., Laurenssen, E.M.P., Blankers, M., Soons, M., Lucas, Z., Verheul, R., & Luyten, P. (2017). Results of a multicentre randomized clinical trial of Day Hospital versus Intensive Outpatient Mentalization-Based Treatment for borderline personality disorder.

van Geffen, M., **Eeren, H.V.**, Hutsebaut, J., & Brand, O. (2017). Denoting treatment outcome for personality disorders: Symptom severity or personality functioning?

About the Author

Hester van Eeren was born in Numansdorp on the 13th of August, 1986. She attended secondary school at the Willem van Oranje (Atheneum) in Oud-Beijerland, where she graduated in 2004. In the same year, she started her study Psychology at Leiden University, and continued in her second year with Criminology. In 2007 she obtained her Bachelor's degree in Criminology. She continued with a Master of Science in Criminology at Erasmus University Rotterdam, followed by a Master of Science in Methodology and Statistics in Psychology at Leiden University. She obtained her Master's degrees in 2010. In October 2010 she started working on the research described in this thesis, first at the Department of Psychiatry, section Medical Psychology and Psychotherapy and from 2011 also at the Viersprong Institute for Studies on Personality Disorders (VISPD). During her research she started a Master of Science in Epidemiology at the Netherlands Institute of Health Sciences (NIHES), of which she obtained her Master's degree in 2013. She also collaborated closely in research of dr. Leona Hakkaart at the institute of Medical Technology Assessment (iMTA) in Rotterdam from 2011 till 2013. In addition to her research activities, she was involved in teaching medical psychology and communication skills in the medical curriculum at the Erasmus MC. In 2014 she received her Basic Qualification in Education. Since 2014, she also assisted clinicians with statistics and methodological questions in their research. Since January 2017 she is working as a data analyst at DSW Zorgverzekeraar.

Dankwoord

Het is zover: ik mag het laatste stuk van dit proefschrift schrijven, het woord van dank. Dit proefschrift zou niet tot stand zijn gekomen zonder de hulp van velen. Daarom wil ik graag iedereen bedanken die op enige wijze betrokken is geweest bij dit proefschrift. Tot een aantal mensen wil ik mijn woord van dank in het bijzonder richten.

Allereerst wil ik alle jongeren en hun gezinnen bedanken die deelgenomen hebben aan het onderzoek naar FFT of die als onderdeel van hun MST of FFT behandeling bij De Viersprong tijd hebben vrijgemaakt om vragenlijsten in te vullen. Hun gegevens hebben geleid tot de datasets gebruikt in dit proefschrift. Ook wil ik de behandelaren en de collega's betrokken bij de dataverzameling danken voor hun hulp.

Prof.dr. Van Busschbach, beste Jan, zonder je blijvende vertrouwen en optimisme, en zonder je adviezen die betrekking hadden op het onderzoek en op allerlei bijkomende overwegingen en beslissingen, zelfs inclusief voedingsadviezen over eierkoeken, zou dit proefschrift hier niet gelegen hebben. Dank voor je waardevolle en leerzame commentaren, het laten zien en benadrukken van andere perspectieven en dank voor de mogelijkheden die je me gegeven hebt. Prof.dr. Scholte, beste Ron, dank voor het vertrouwen in mij, in het onderzoek en specifiek het vertrouwen in het gebruik van economische evaluaties in het jeugdveld. Je commentaren scherpten het waardevolle 'jeugdperspectief' aan en hielpen de bruikbaarheid van de bevindingen voor de praktijk van de jeugdzorg te onderstrepen. Dr. Van der Rijken, beste Rachel, ik ben ontzettend blij dat jij als copromotor betrokken bent geweest bij dit proefschrift. Het kunnen sparren over de keuzes in het onderzoek, de implicaties van deze keuzes op allerlei vlakken, en de onvermoeibare steun en relativering hebben mijn visie op het onderzoek en dit proefschrift goed gedaan. En dit heeft mij goed gedaan.

Dr. Hakkaart - van Roijen, beste Leona, graag wil ik je bedanken voor de mogelijkheid mee te kunnen werken aan het onderzoek naar FFT en voor de fijne en leerzame samenwerking. Dank je ook voor het in mij gestelde vertrouwen. Lieve Saskia, met jou heb ik dit onderzoek uit mogen voeren. Ik ben dankbaar voor de intensieve samenwerking en de bijzondere band die hieruit voortgekomen is.

Ik wil alle medeauteurs die meegewerkt hebben aan de artikelen bedanken voor hun kritische en bemoedigende commentaren op de manuscripten. In het bijzonder bedank ik Marieke voor de wijze woorden en hulp bij het schrijven van hoofdstuk 4. Beste Lucas, met jouw kennis en enthousiasme is hoofdstuk 5 nog scherper neergezet, dank daarvoor.

Collega's van de MPP, dank voor jullie steun de afgelopen jaren. Met jullie kunnen sparren (en spuien) over onderzoek of onderwijs was waardevol en leerzaam. En natuurlijk dank voor de gezelligheid en persoonlijke interesse. Hetty, dankjewel voor je ondersteuning en het meedenken in praktische zaken. Martijn, fijn dat jij mijn kamergenoot was al die jaren. Collega's van De Viersprong, dank voor jullie passie voor onderzoek en het vermogen het perspectief van de cliënt in het onderzoek nooit uit het oog te verliezen. Juist deze koppeling gaf mijn werk bij de Viersprong meer diepgang. En natuurlijk dank voor de gezelligheid. Ook jullie, (tijdelijke) collega's van het iMTA, wil ik graag bedanken voor jullie interesse en het delen van jullie kennis en inzichten. Collega's van DSW Zorgverzekeraar, ik wil jullie bedanken voor jullie belangstelling en medeleven tijdens de afrondende fase van dit proefschrift.

Bijzonder en speciaal vind ik het dat mijn paranimfen al jaren naast me staan, en ook op deze bijzondere dag naast me willen staan. Lieve Femke en Klarieke, erg bedankt voor jullie nuchtere kijk, steun, begrip en humor in het aanhoren van mijn 'proefschrift-verhalen' en al het andere. Lieve Nicolette, ook voor jouw relativering, steun en humor ben ik je dankbaar. Ik vind het erg bijzonder dat je de omslag van mijn proefschrift hebt ontworpen. Dank dat jullie er zijn.

Lieve familie en vrienden, zonder jullie geduld, relativering, steun en gezelligheid had dit proefschrift er ook zeker niet gelegen. In het bijzonder ben ik dankbaar dat mijn ouders, schoonouders en broer en zus me altijd met beide benen op de grond houden. Deze basis is voor mij erg waardevol. Dank dat jullie er zijn.

Lieve Ronald, in het moeten aanhoren van alle 'verhalen' span jij natuurlijk de kroon. Zonder jou naast me ben ik niet mezelf en mét jou is alles gewoon leuker. Dank je wel dat je bij mij bent. Lieve Tieme, met jou erbij is het extra leuk.