Running title: A validation study of the SCREENIVF

Title: A validation study of the SCREENIVF: an instrument to screen women or men on risk for emotional maladjustment before the start of a fertility treatment

Authors

Henrietta D.L. Ockhuijsen, Ph.D[a], Maarten van Smeden, Ph.D[b], Agnes van den Hoogen, Ph.D[c], Jacky Boivin, Ph.D[d]

[a]Department of Reproductive Medicine and Gynaecology, University Medical Centre Utrecht, the Netherlands

[b]Julius Centre for Health Science and primary care, University Medical Centre Utrecht, the Netherlands

[c]Department of Neonatology, Wilhelmina Children's Hospital and University Medical Centre Utrecht, the Netherlands

[d]School of Psychology, Cardiff University, United Kingdom

[a]Corresponding author: e-mail: h.d.l.ockhuysen@umcutrecht.nl, Heidelberglaan 100, 3508 GA UTRECHT, The Netherlands, telephone: +31 88 75 536 28

Capsule

The SCREENIVF is an objective instrument to talk about emotional maladjustment but should not be used as a predictive tool until more research is done to improve the predictive value.

Abstract

Objective: To examine construct and criterion validity of the Dutch SCREENIVF among women and men undergoing a fertility treatment.

Design: A prospective longitudinal study nested in a Randomized Controlled Trial.

Setting: University hospital in the Netherlands.

Patients: Couples, 468 women and 383 men, undergoing an IVF/ICSI treatment in a fertility clinic completed the SCREENIVF.

Main Outcome Measure(s): Construct and criteria validity of the SCREENIVF.

Results: The comparative fit index and root mean square error of approximation for women and men show a good fit of the factor model. Across time, the sensitivity for Hospital Anxiety and Depression Scale subscale in women ranged from 61%-98%, specificity 53%-65%, positive predictive value 13%-56%, negative predictive value 70%-99%. The sensitivity scores for men ranged from 38%-100%, specificity 71%-75%, positive predictive value 9%-27%, negative predictive value 92%-100%. A prediction model revealed that for women 68.7 % of the variance in the Hospital Anxiety and Depression Scale on time 1 and 42.5% at time 2 and 38.9% was explained by the predictors, the sum score scales of the SCREENIVF. For men, 58.1% of the variance in the Hospital Anxiety and Depression Scale on time 1 and 46.5% at time 2 and 37.3% at time 3 was explained by the predictors, sum score scales of the SCREENIVF.

Conclusion: The SCREENIVF has good construct validity but the concurrent validity is better than the predictive validity. SCREENIVF will be most effectively used in fertility clinics at the start of treatment and should not be used as a predictive tool.

Key words: screening, validation studies, infertility, psychology, assisted reproductive techniques

Introduction

Women and men experience distress and lower quality of life during fertility treatments (1-3). There is premature discontinuation in fertility treatments like an In Vitro Fertilization (IVF) and IntraCytoplasmic Sperm Injection (ICSI) (4-9). Reasons for discontinuation are multifactorial but can be distinguished in being financial, hospital/clinic related, medical, physical and/or psychological reasons (4, 8, 10, 11).

Psychological burden for discontinuation represent 14%, physical burden 6% and relational and personal problems almost 17% of cases of treatment discontinuation (7). Early detection of psychological vulnerability can be achieved by screening (12). A screening test is not intended to be diagnostic but to sort patients who possibly have psychological vulnerabilities during a fertility treatment from those who possibly do not. Early detection (case-finding) of psychological vulnerability aims at discovering conditions which already produced pathological change but which have not so far reached a stage at which psychological aid is sought spontaneously (12). The SCREENIVF was designed for use at the start of a fertility treatment to identify women or men at risk for emotional maladjustment after treatment. Additionally patients who are identified as at risk could be provided with additional psychosocial support, to prevent them from discontinuation of treatment. The advantages of the SCREENIVF, in relation to other already existing screening instruments, is that it identifies five risk areas in the field of emotional maladjustment (13, 14).

The development of the SCREENIVF is based on two studies which revealed the following risk factors for increased emotional problems: pre-treatment anxiety, depression, helplessness, less acceptance regarding fertility problems, and lack of social support. Based on these studies, the

SCREENIVF was developed in a 34 item questionnaire with five risk factors (13, 14). Two additional studies investigated the methodological quality of SCREENIVF (15, 16) and one study focused on the feasibility of the instrument (17).

Several checklists exist to investigate the methodological quality of instruments for example the COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) (18-21). The COSMIN is a checklist to assess the methodological quality of studies evaluating health measurement tools (18). Although COSMIN was developed for health-related patient reported outcomes it can also be used to evaluate studies about the reliability and validity of the SCREENIVF. According to COSMIN checklist an instrument can be based on a reflective or formative model of health measurement. The models lead to different specifications for model testing (22). In a reflective model all items are a manifestation of the same underlying construct and the items are expected to be highly correlated and interchangeable. In a formative model the items together form the construct and these items do not need to be correlated. The subscales of the SCREENIVF are based on reflective models (i.e., anxiety subscale has 5 anxiety items) but each of the SCREENIVF subscales (i.e., anxiety, depression, helplessness, acceptance, social support) contributes to risk of psychological vulnerability, and this aspect of the tool makes it a formative model. As stated by the COSMIN checklist, studies about the methodological quality of an instrument should investigate the following: measurement properties (internal consistency, reliability, measurement error), content validity (including face validity), construct validity (including structural validity, hypotheses testing, and cross-cultural validity), criterion validity, responsiveness, and interpretability (18).

A Dutch study investigated sensitivity, specificity and negative and positive predictive value of the SCREENIVF among 279 women in their first IVF treatment cycle (16). Sensitivity was 69%

showing that 46 out of 67 women who were classified as being at risk by the SCREENIVF at T1 had emotionally maladjusted at time 2. Specificity was 77% showing that 163 out of 212 that were correctly classified as not being at risk at T1 had no emotional maladjustment at time 2. The negative predictive value was 89% and the positive predictive value was 48% meaning that SCREENIVF is better in identifying patients without emotional maladjustment than with emotional maladjustment. The Portuguese study focused on reliability and construct validity among 291 women and 92 men undergoing an Intra-Uterine Insemination (IUI) or IVF/ICSI treatment (15). The set of analyses provided reasonable evidence for the reliability and construct validity of SCREENIVF in a different cultural setting.

Both studies concluded that the SCREENIVF was a reliable and valid instrument to identify people at risk for emotional maladjustment. However, reliability and validity analyses to date do not fully conform to COSMIN requirements for making this conclusion. Not all areas of validity and reliability of the SCREENIVF have been investigated and there is a need especially for further testing of construct and criterion validity (18). Construct validity refers to the scores of an instrument being consistent with the hypothesised structure of the instrument (23). Although shown in a Portuguese population of women and men, this has not yet been evaluated in a population of Dutch women and men. Criterion validity can be subdivided in concurrent and predictive validity (23). Concurrent validity refers to the degree to which the scores of an instrument are an adequate reflection of the scores on a criterion measure assessed at the same time. Predictive validity refers to the extent to which scores on one instrument can be used to predict future scores on a criterion measure. A criterion measure is an instrument that is reliable and valid that can be compared with the new instrument (23).

The aim of the present prospective study was to test the construct and criterion validity of the Dutch SCREENIVF among women and men about to undergo a fertility treatment. The criterion measure of emotional maladjustment was the Hospital Anxiety and Depression Scale (HADS) (24). Participants (men and women) completed SCREENIVF and the HADS at the start of an IVF/ICSI treatment (to assess concurrent validity), and again during the waiting period after embryo transfer and six weeks after embryo transfer (to assess predictive validity) because the latter two time points are risk periods for emotional maladjustment (25, 26)

METHOD

Procedure

The present prospective longitudinal study was nested in a three-armed Randomized Controlled Trial (RCT) and an additional fourth group added after the RCT (Clinical Tials.gov: NCT01701011). The ethical committee of the University of Utrecht provided ethical review and approval for this study (protocol number 10-174/K). The RCT investigated the effects of the Positive Reappraisal Coping Intervention (PRCI) on psychological well-being of women waiting for the outcome of their fertility treatment cycle compared to daily emotional monitoring or routine care (27-29). The present study used the baseline data from all the participating women but used only the data from subsequent measurements (during and after treatment) of women in the control group who did not receive the intervention or did not daily monitor emotions. The PRCI intervention was aimed at women and not at their partners. Partners of the women participating in the RCT were invited to take part in the study of the SCREENIVF. All data of the men could be used because they did not receive the intervention.

Recruitment for the overall study took place by sending an information letter regarding this study to all couples on the waiting list to start an IVF treatment. All couples undergoing a stimulated or cryopreserved IVF/ICSI treatment cycle were included. Excluded were women not speaking the Dutch language. Couples had to send a reply form or email if they were interested in the study. A researcher contacted the interested couples to inform them about the study and to answer any questions. Those who decided to participate were sent a written information sheet and an informed consent to return in a pre-addressed stamped envelope.

Couples completed questionnaires at three time points. The first time point was at the start of an IVF/ICSI treatment before the first ultrasound, the second time point was on day ten after the embryo transfer during the waiting period for the pregnancy test and the last time point was 6 weeks after the embryo transfer.

Materials

The Background Information Form (BIF) was only completed by women and contains 16 items about demographic (e.g. age, educational status), medical (e.g. previous illness) and gynaecological (e.g. infertility diagnosis, previous infertility treatment) characteristics.

The SCREENIVF contains 34 items on a 4-point Likert scale divided into five domains: state anxiety and trait anxiety derived from a short version of Spielberger State and Trait Anxiety Inventory (5 items each) (30), depression derived from the short Beck depression Inventory (7 items) (31), social support derived from the Inventory of Social Involvement (5 items) (32) and helplessness (6 items) and acceptance (6 items) derived from the Illness Cognition Questionnaire for IVF patients (13, 33). Use of SCREENIVF results in an at risk score when the score on one

of the five risk factors shows clinically relevant problems. In the Dutch population the cut-off scores for risk on anxiety are 24 and above, for depression four or higher, social support 15 and less, helplessness 14 and above and acceptance 11 and less. The cut-off scores were one standard deviation above or below the mean scores of IVF-patients. The SCREENIVF thresholds can differ according to population (15).

The HADS measures general anxiety and depression and contains 14 items (7 items for each subscale) that are rated on a 4-point Likert scale (24). Scores on each subscale can be interpreted in ranges: normal (0–7), mild (8–10), moderate (11–14) and severe (15–21) anxiety and depression (24). Psychometric properties in the Dutch population show a reliable and valid instrument (34).

DATA ANALYSES

IBM SPSS Statistics 20 and R with the lavaan package were used to perform the statistical analysis (35). Descriptive statistics were used to describe baseline variables of the BIF. The difference between women and men in age at baseline was compared using a paired t-test. McNemar's Chi-square test for categorical, nominal outcomes in related samples (i.e., two spouses of same couple) was used for all other baseline variables.

To investigate the construct validity of the SCREENIVF a confirmatory factor analysis (CFA) was conducted. CFA is special case of a structural equation model that estimates the relationship between the observed (manifest) variables or indicators and the (unobserved) latent variables or factors (36). The CFA was used to test the hypothesised reflective measurement model underlying the SCREENIVF against the data obtained to evaluate the hypothesised structural

relations. In accordance with the theoretical model of the SCREENIVF, a CFA was conducted with 5 factors representing: anxiety, depression, social support, helplessness and acceptance. The CFA was estimated by full-information maximum likelihood (37). The root mean square error of approximation (RMSEA) and the comparative fit index (CFI) were used, to test whether the model fits the manifest data. When RMSEA < 0.05 and CFI > 0.95 the model is said to have a good fit to the data (38). One of the assumptions in CFA is (multivariate) normally distributed item distributions, which can be difficult to achieve using single items as indicators (39). In such cases the same latent construct can be estimated from multi-item parcels that are the mean of several items. The selection of items can be done at random, splitting all odd and even items into two parcels or theory based (39-41). The most effective way to create parcels is theory based, other methods like at random parcelling should be based on the assumption that all items are interchangeable, large sample size, many items, high item communality, low item diversity (40). The parcels developed for the CFA of the SCREENIVF were theory based, taking into account the conceptual relatedness (that the parcel reflected the different components of the construct e.g., depression there was vegetative symptoms, cognitive symptoms) of the items grouped within a parcel. See Figure 1.

Measurement invariance (MI) was tested to investigate whether the construct of the SCREENIVF has the same structure across gender. With MI it can be demonstrated that men and women understand the questions of the SCREENIVF and the fundamental latent factor in the same way (42), and that the pattern of interrelationships among the latent factors are similar. The threshold values to assess MI are: CFI>0.95 and RMSEA <0.05 (42). If the model with constrained parameters (loading, intercept, mean) differs significantly from the unconstrained model this is indicative of variance according to gender.

To test the criterion validity (concurrent and predictive) of the SCREENIVF the sensitivity, specificity, positive predictive value and negative predictive value were measured and a prediction model was developed. According to COSMIN the criterion validity can be assessed by the use of a gold standard (43) but this assumes measurement without error, which is unrealistic for measures of anxiety and depression. A good alternative is to use instead a 'reference standard' and well validated criterion measure of the diagnosis or condition in question (19). In this study the HADS was used as a criterion measure for emotional maladjustment. The cut off score of the HADS for the anxiety and depression scales were ≥ 8. Sensitivity refers to the ability of the SCREENIVF to correctly identify those patients with emotional maladjustment while specificity refers to the ability of the test to correctly identify those patients without emotional maladjustment (44). Positive predictive value (PPV) refers to the likelihood that a patient has emotional maladjustment given that the test result is positive. In this study, this means that patients identified as 'at risk' at the start of treatment (T1), will have high scores for anxiety and depression during the waiting period (time 2) or six weeks after treatment failure (T3). Negative predictive value (NPV) refers to the likelihood that a patient does not have emotional maladjustment given that the test result is negative (44). In this study, this means patients identified as 'not at risk' at the start of treatment (time 1), will not have high scores for anxiety and depression during the waiting period (time 2) or six weeks after treatment failure (time 3).

To further explore the concurrent and predictive validity of the SCREENIVF, and assist its use in clinics, prediction models were developed to evaluate its efficacy to predict patients' levels of emotional maladjustment during or after the course of IVF treatment. Concurrent validity is the degree to which the scores of the SCREENIVF are an adequate reflection of the scores on the HADS measured at the same time (T1). Predictive value is the degree to which the scores of the

SCREENIVF predict the scores on the HADS in the future (time 2, time 3). The prediction models were developed using linear (ordinary least square) regression for men and women separately. All variables were entered at the same time. The predictor variables were the sum scores on each of the 5 subscales of the SCREENIVF according to the development of the questionnaire (16). The outcome used for the prediction model was the HADS score at baseline, day 10 of the waiting period and 6 weeks after embryo transfer. To account for missing data for the prediction model, multiple imputation was performed using R with the multivariate imputation by chained equations (MICE) package using default settings; 20 imputation data sets were generated (45). Multiple imputation is preferable over other analysis in several circumstances (46). In this study data was missing at random, and variables chosen for imputation have a correlation with the incomplete variables, furthermore the percentage missing values varied between 3-41%. The variables used for imputation were: having previous children (yes/no), age (years), knowing why one was infertile (yes/no), any previous miscarriage (yes/no) and previous use of counselling for infertility (yes/no). For the imputation, only the data were used for males and females where data was available on at least one subscale of SCREENIVF. Imputation for males and females was done separately. When one subscale was not completed the case was deleted. See for an overview of the analysis supplemental table 1.

Results

Sample characteristics

The baseline characteristics are described in Table 1. More women (n=487) than men (n=426) agreed to participate in the study, and all participants were heterosexual couples. Mean age was

higher for men (t=10.67, df=415, p=0.000). More women (52.1%) than men (31.9%) were at risk for emotional maladjustment before the start of an IVF/ICSI treatment on one or more of the subscales of the SCREENIVF (x^2=57.5, df=1, p=0.009)

The number of women who returned the SCREENIVF at Time 1 was n=468 (96.1%) and men n=383 (90%). The number of women who returned the HADS at time 1 was n=471 (97%) and men n=387 (91%), at time 2 women n=398 (82%) and men n=326 (77%), at time 3 women n=354 (73%) and men n=285 (67%). The number of women in the control group who returned the HADS at time 2 was n=102 (82%) and at time 3 was n=90 (73%).

Construct validity

The confirmatory factor analysis was done separately for women and men (see Table 2 and Figure 1). The model fit statistics showed good fit based on CFI and RMSEA for men and women (for women: CFI=0.992; RMSEA=0.038 [90% CI: 0.022-0.054]; for men: CFI=0.994; RMSEA =0.026 [90% CI: 0.000-0.046]). As shown in Figure 1, all standardized factor loadings of the parcels were high between 0.700-0.953 and significant (p=0.000) and all five latent factors of the SCREENIVF were significantly interrelated (p=0.000). The correlations between the factors anxiety and depression were the highest (women=0.842, men=0.726) and between social support and helplessness the lowest (women=0.215, men=0.150). Consistent with this analysis, the fit indices for measurement invariance across gender showed a RMSEA > 0.05 for intercepts and means which indicated that measurement is not invariant (see supplemental table 2).

Criterion validity

The criterion validity was analysed separately for women and men based on the differential results of the CFA. Two methods are used to test the criterion validity of the SCREENIVF, the sensitivity, specificity, positive predictive, negative predictive value were measured and a prediction model was developed.

Sensitivity, specificity, positive and negative predictive value

On T1 the results for concurrent validity are displayed and the predictive validity at time 2 and time 3. The results for sensitivity, specificity, positive and negative predictive value are presented in Table 3. Across time, sensitivity scores for HADS subscale in women ranged from 61%-98%, specificity 53%-65%, PPV 13%-56%, NPV 70%-99%. The sensitivity scores for men ranged from 38%-100%, specificity 71%-75%, PPV 9%-27%, NPV 92%-100%. Overall, the scores for sensitivity and NPV are the highest at time 1. This means that the SCREENIVF completed by women and men is better in identifying patients at risk for emotional maladjustment (sensitivity) and better in correctly screening patients who have not emotional maladjustment (NPV) at the start of the treatment (time 1) than during (time 2) or after treatment (time 3).

The NPV is better at time 1, time 2, and time 3 than de PPV for both women and men. This means the SCREENIVF is better in correctly screening patients who do have not emotional maladjustment than screening patients who have emotional maladjustment. The use of the SCREENIVF leads to more false positive then false negative results. (See supplemental Figures 1 and 2).

Prediction model

The results of the prediction model are presented in Supplemental table 3. At time 1 the results for concurrent validity are displayed, at time 2 and time 3 of predictive validity. A linear regression model investigated whether the subscales of the SCREENIVF at time 1 predicted anxiety and depression levels as measured with the HADS at time 1, time 2 and time 3. The results for women show that sum score subscales of the SCREENIVF explained 68.7 % (R^2=0.687) of the variance in the HADS on time 1, 42.5% (R^2=0.425) at time 2 and 38.9% (R^2=0.389) at time 3. As shown in Table 5, the predictors anxiety and depression subscales have a significant contribution to the model at time 1 ($p<0.05$), the predictor anxiety and helplessness subscales have a significant contribution to the model at time 2 ($p<0.05$) and the predictor anxiety has a significant contribution to the model at time 3 ($p<0.05$). The results for men show that 58.1% ($R^{2=}$0.581) of the variance in the HADS on time 1, 46.5% (R^2=0.465) at time 2 and 37.3% (R^2 =0.373) at time 3 is explained by the predictors, sum score subscales of the SCREENIVF. The predictors anxiety and depression subscales have a significant contribution to the model at time 1, 2 and 3 ($p<0.05$). The predictor social support subscale has a significant contribution to the model at time 1 and helplessness at time 2 ($p<0.05$).

Discussion

In this study the construct and criterion validity (concurrent and predictive) of the SCREENIVF were investigated among women and men undergoing fertility treatment. Results show that the concurrent validity of the SCREENIVF is better than the predictive validity. Analyses revealed that the circumstances in which SCREENIVF would be most effectively used in fertility clinics

are at the start of treatment, in women and men to rule out emotional maladjustment or to predict people NOT at risk for depression during the waiting period or after treatment. The SCREENIVF is not good in predicting emotional maladjustment during or after treatment. The advantage of using the SCREENIVF, with respect to other instruments like the HADS, is that it covers cognitions, emotions and support using items worded in a way that will resonate with the lived experience of people undergoing treatment. The future evaluation studies proposed comprise prospective longitudinal qualitative and quantitative studies to better understand why men and women respond differently to items, why negative predictive values are better than positive predictive values.

The construct validity of the SCREENIVF is good. The CFA shows that the hypothesised five factor structure of the SCREENIVF with subscales of depression, anxiety, helplessness, acceptance and social support has a good fit to the observed data.

Men and women do not interpret or respond to the SCREENIVF items in the same way. Our study confirms that the measurement model for women and men is different and that criterion and predictive validity are different for women and men. The Portuguese validation study of the SCREENIVF revealed a significant structural and measurement invariance for gender (15). In a cross sectional study where 946 women and 670 partners completed the SCREENIVF women were significantly more at risk than their partner on all subscale of the SCREENIVF except on social support (2). SCREENIVF was originally developed based on research among women in an IVF treatment (13). Our results suggest that more in-depth qualitative research is necessary to identify why gender affects responses to items, and whether the differences suggest a need for scoring and cut off values of the SCREENIVF to be gender based.

The concurrent validity of the SCREENIVF is better than the predictive validity. Sensitivity, NPV and the results of the prediction model show that results of time 1 are better than at time 2 and time 3. Results are somewhat different from another Dutch validation study of the SCREENIVF with a sensitivity of 69%, specificity of 77%, PPV 48% and NPV 89% (16). In both studies the PPV was very low. Differences could be due to the fact that in our study the HADS was used as a criterion measure and separate analyses were done for anxiety and depression while in the other Dutch validation study the subscales anxiety and depression of the SCREENIVF were used with a dichotomous variable of yes or no high scores on anxiety and/or depression (16). We would argue that using subscales of the SCREENIVF at both times confounds reliability and prediction, and makes prediction seem better than it actually is. The results show that fertility clinic staff can generally feel confident that on average when men and women do not score as at risk on SCREENIVF they will not have depression now or later in the treatment cycle (i.e., during waiting or after results) because all negative predictive values were more than 96% for depression. At a practical level predicting absence of maladjustment is useful because it means that these patients will not require support for depression at a clinic, conserving resources for others.

Whilst one can feel confident about the absence of disorder (on average), the same cannot be said about the presence of disorder. Positive predictive values of the SCREENIVF were poor. In the present study people judged to be at risk at T1 were unlikely to have emotional maladjustment at time 2 or time 3. Several reasons could exist for this outcome. Positive and negative predictive values are influenced by the prevalence of the disease in the population tested (44, 47). Several studies suggest a low prevalence of anxiety and depression among women and men attending an IVF clinic (48, 49). Positive and negative predictive values are also influenced

by the sensitivity and the specificity and these are influenced by the cut off scores of the SCREENIVF. In our study the cut off scores were based on the first Dutch validation study of the SCREENIVF, one standard deviation above or below the mean scores (16). It could be that cut-off scores were not correctly set. In the Portuguese validation study more patients scored above the cut-off scores (15). Perhaps there are transcultural differences for cut off scores. More research has to be conducted about the correct cut-off scores for instance with Area Under the Curve and Receiver Operating Characteristic. It should also be investigated, whether adding prognostic variables available at the start of IVF (e.g., age, previous pregnancy, with/without IVF) to the SCREENIVF, improves positive prediction rates.

The prediction model showed that not all subscales of the SCREENIVF had a significant contribution to the model. The question arises whether all five risk factors are necessary to detect emotional maladjustment. A reason why not all subscales contribute to the model could be due to the fact that prediction models are difficult to develop because of the reactive and dynamic nature of emotional maladjustment. Distress levels fluctuate according to events during treatments such as oocyte retrieval and embryo transfer (50). External factors like stressful life events can also affect distress during treatment (51). These influences make it difficult to predict with high accuracy emotional maladjustment.

In the analysis the SCREENIVF was investigated as a stand-alone tool without considering other variables that could influence emotions during and after treatment. A stand-alone tool means it does not take into account any other psychosocial or biomedical variables that could influence (post) treatment emotions. For example, events that occur during the cycle (e.g., number of oocytes) or after it (e.g., pregnancy outcome) because these data are not available at the start of treatment when SCREENIVF is administered. Analytically this was necessary to investigate the

claim that the SCREENIVF could predict future emotional maladjustment from the start of treatment. To include these future events at the start of treatment, future research should consider including prognostic information, for example patient's predicted chance of success at baseline using predictive models (52). The inclusion of such variables may improve the predictive ability of SCREENIVF. Alternatively, it could be argued that the SCREENIVF should (and did not) predict emotional maladjustment regardless of outcome, for example predict the level of anxiety in response to pregnancy (e.g., fear about miscarriage). A recent review confirmed the importance of the five risk factors of the SCREENIVF associated with emotional adjustment (53).

Although the predictive validity is not as good as the concurrent validity there may still be value in using SCREENIVF. First, SCREENIVF has face validity in that constructs and items were selected and tested with patients undergoing fertility treatment, which probably gives it more face validity than solely generic versions but also items can resonate with patients validating their own experiences of, for example, helplessness. Second, using all the SCREENIVF subscales at the start of a fertility treatment gives healthcare professionals a good profile of areas of vulnerability (cognition, emotion, support). This is a good starting point to talk about the specific support women and men need from healthcare professionals working in a fertility clinic, helping individualized care to be delivered. Practical aspects such as the easy availability, clear scoring guidelines and possibility to rule out people needing help also makes it worthwhile.

This study had some limitations that should be mentioned. A limitation was the HADS, with two subscales, that was used as a criterion measure for the SCREENIVF with five subscales. The HADS was used because it is a well validated questionnaire and no comparable instrument like the SCREENIVF exists. The present study was nested in a RCT and only women of the control

group who were not assigned to the intervention could be included for the second and third time measurement while all the data of the men could be used because they did not participate in the RCT. This could represent some bias in that women who received the intervention could have had an influence on their respective male partner's scores across time. Although the intervention was given to women and not men, it could be that women shared the PRCI with their partner. Analyses showed there were no significant differences between the anxiety and depression scores of men according to their partner's group assignment but such bias cannot be ruled out. Another weakness could be that only patients willing to participate in a RCT about a psychological intervention were sampled, and this could also have introduced bias that minimized the SCREENIVF to predict maladjustment. Another limitation was the missing data. In this study the method of multiple imputation was used, a method that is likely to be superior over other methods to deal with missing data (46).

The strength of this study is the use of COSMIN. The COSMIN checklist provides a systematic way of not only assessing the quality of other studies on measurement properties but it can also be helpful in designing or reporting an own study on measurement properties (18). Another strength is the prospective and longitudinal design with men and women focusing on points in the treatment trajectory where patients with psychological vulnerability could experience emotional maladjustment, namely, at the start of treatment, during the waiting period and after treatment results are known.

The conclusion of this study is that the SCREENIVF can be used to identify women and men with emotional maladjustment before the start of a fertility treatment. The results of the SCREENIVF give healthcare workers an objective instrument to talk about emotional maladjustment in different areas with their patients. Based on these results SCREENIVF should

not be used as a predictive tool until more research is done to improve the predictive value of SCREENIVF.

Author's roles

H.O. designed the study, monitored data collection for the whole study, wrote the statistical analysis plan, cleaned and analysed the data, and drafted and revised the paper; M.S. contributed to the statistical analytic plan, cleaned and analysed the data and edited the paper; A.H. monitored and edited the paper, J.B. initiated the project, designed the study, contributed to the statistical analytic plan, and interpretation of the data, drafted and revised the paper.

Acknowledgements

We thank the participating practices, staff, and patients for their contribution. We thank Sofia Gameiro and Chris Verhaak for providing feedback on the drafts of this article.

Funding

The study was funded by the division Women and Baby of the University Medical Centre Utrecht.

Conflicts of interest

None declared

Table 1 Baseline demographic, medical and psychological characteristics women and men

Table 2 Summary statistics for confirmatory factor analysis of SCREENIVF in female (n=487) and male patients (n=426)

Table 3 Criterion validity testing the concurrent validity (time 1: at the start of treatment) and predictive validity (time 2: during the waiting period and time 3: six weeks after embryo transfer) of the SCREENIVF by using the scores of the Hospital Anxiety and Depression Scales as a reference test

Figure 1 Confirmatory factor analysis model SCREENIVF

Supplemental Table 1 Overview analysis SCREENIVF

Supplemental table 2: Testing measurement invariance by gender

Supplemental Table 3 Overview analysis SCTEENIVF Summary statistics for the prediction model with the sum score subscales of the SCREENIVF as predictors and the Hospital Anxiety and Depression Scales as the outcome variable at time 1, 2 and 3

Supplemental Figure 1 Percentage of women with false negative, correct en false positive scores of the SCREENIVF

Supplemental Figure 2 Percentage of men with false negative, correct en false positive scores of the SCREENIVF

References

1. Wichman CL, Ehlers SL, Wichman SE, Weaver AL, Coddington C. Comparison of multiple psychological distress measures between men and women preparing for in vitro fertilization. Fertil Steril 2011;95:717-21.

2. Huppelschoten AG, van Dongen AJ, Verhaak CM, Smeenk JM, Kremer JA, Nelen WL. Differences in quality of life and emotional status between infertile women and their partners. Hum Reprod 2013;28:2168-76.

3. Volgsten H, Skoog Svanberg A, Ekselius L, Lundkvist O, Sundstrom Poromaa I. Risk factors for psychiatric disorders in infertile women and men undergoing in vitro fertilization treatment. Fertil Steril 2010;93:1088-96.

4. Van den Broeck U, Holvoet L, Enzlin P, Bakelants E, Demyttenaere K, D'Hooghe T. Reasons for dropout in infertility treatment. Gynecol Obstet Invest 2009;68:58-64.

5. Brandes M, van der Steen JO, Bokdam SB, Hamilton CJ, de Bruin JP, Nelen WL, *et al*. When and why do subfertile couples discontinue their fertility care? A longitudinal cohort study in a secondary care subfertility population. Hum Reprod 2009;24:3127-35.

6. McDowell S, Murray A. Barriers to continuing in vitro fertilisation--why do patients exit fertility treatment? Aust N Z J Obstet Gynaecol 2011;51:84-90.

7. Gameiro S, Boivin J, Peronace L, Verhaak CM. Why do patients discontinue fertility treatment? A systematic review of reasons and predictors of discontinuation in fertility treatment. Hum Reprod Update 2012;18:652-69.

8. Lande Y, Seidman DS, Maman E, Baum M, Hourvitz A. Why do couples discontinue unlimited free IVF treatments? Gynecol Endocrinol 2014:1-4.

9. Gameiro S, Verhaak CM, Kremer JA, Boivin J. Why we should talk about compliance with assisted reproductive technologies (ART): a systematic review and meta-analysis of ART compliance rates. Hum Reprod Update 2013;19:124-35.

10. Rajkhowa M, McConnell A, Thomas GE. Reasons for discontinuation of IVF treatment: a questionnaire study. Hum Reprod 2006;21:358-63.

11. Verberg MF, Eijkemans MJ, Heijnen EM, Broekmans FJ, de Klerk C, Fauser BC, *et al*. Why do couples drop-out from IVF treatment? A prospective cohort study. Hum Reprod 2008;23:2050-5.

12. Wilson JM, Jungner YG. Principles and practice of mass screening for disease. Bol Oficina Sanit Panam 1968;65:281-393.

13. Verhaak CM, Smeenk JM, van Minnen A, Kremer JA, Kraaimaat FW. A longitudinal, prospective study on emotional adjustment before, during and after consecutive fertility treatment cycles. Hum Reprod 2005;20:2253-60.

14. Verhaak CM, Smeenk JM, Evers AW, van Minnen A, Kremer JA, Kraaimaat FW. Predicting emotional response to unsuccessful fertility treatment: a prospective study. J Behav Med 2005;28:181-90.

15. Lopes V, Canavarro MC, Verhaak CM, Boivin J, Gameiro S. Are patients at risk for psychological maladjustment during fertility treatment less willing to comply with treatment? Results from the Portuguese validation of the SCREENIVF. Hum Reprod 2014;29:293-302.

16. Verhaak CM, Lintsen AM, Evers AW, Braat DD. Who is at risk of emotional problems and how do you know? Screening of women going for IVF treatment. Hum Reprod 2010;25:1234-40.

17. Van Dongen AJ, Kremer JA, Van Sluisveld N, Verhaak CM, Nelen WL. Feasibility of screening patients for emotional risk factors before in vitro fertilization in daily clinical practice: a process evaluation. Hum Reprod 2012.

18. Mokkink LB, Terwee CB, Knol DL, Stratford PW, Alonso J, Patrick DL, *et al*. The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: a clarification of its content. BMC Med Res Methodol 2010;10:22,2288-10-22.

19. Aaronson N, Alonso J, Burnam A, Lohr KN, Patrick DL, Perrin E, *et al*. Assessing health status and quality-of-life instruments: attributes and review criteria. Qual Life Res 2002;11:193-205.

20. Oliveira MR, Gomes Ade C, Toscano CM. QUADAS and STARD: evaluating the quality of diagnostic accuracy studies. Rev Saude Publica 2011;45:416-22.

21. Zumbo BD, Chan EKH. Validity and Validation in Social, Behavioral, and Health Sciences. : Springer International Publishing, 2014.

22. Coltman T, Devinney TM, Midgley DF, Venaik S. Formative versus reflective measurement models: Two applications of formative measurement. Journal of Business Research 2008;61:1250-62.

23. Polit DF, Beck CT. Nursing research : generating and assessing evidence for nursing practice. c2008:xviii, 796 p. :.

24. Zigmond AS, Snaith RP. The hospital anxiety and depression scale. Acta Psychiatr Scand 1983;67:361-70.

25. Boivin J, Lancastle D. Medical waiting periods: imminence, emotions and coping. Womens Health (Lond Engl) 2010;6:59-69.

26. Verhaak CM, Smeenk JM, Nahuis MJ, Kremer JA, Braat DD. Long-term psychological adjustment to IVF/ICSI treatment in women. Hum Reprod 2007;22:305-8.

27. Ockhuijsen H, van den Hoogen A, Eijkemans M, Macklon N, Boivin J. Clarifying the benefits of the positive reappraisal coping intervention for women waiting for the outcome of IVF. Hum Reprod 2014;29:2712-8.

28. Ockhuijsen H, van den Hoogen A, Eijkemans M, Macklon N, Boivin J. The impact of a self-administered coping intervention on emotional well-being in women awaiting the outcome of IVF treatment: a randomized controlled trial. Hum Reprod 2014.

29. Ockhuijsen HD, van den Hoogen A, Macklon NS, Boivin J. The PRCI study: design of a randomized clinical trial to evaluate a coping intervention for medical waiting periods used by women undergoing a fertility treatment. BMC Womens Health 2013;13:35.

30. Spielberger C. Manual for the State-Trait Anxiety Scale: Palo Alto: Consulting; Psychologists Press, 1983.

31. Beck AT, Guth D, Steer RA, Ball R. Screening for major depression disorders in medical inpatients with the Beck Depression Inventory for Primary Care. Behav Res Ther 1997;35:785-91.

32. 'van Dam-Baggen R, Kraaimaat F'.
De inventarisatielijst sociale betrokkenheid
(ISB): een zelfbeoordelingslijst om sociale steun te meten [The inventarisation inventory to measure sociale integration: a self report inventory to assess social support] Gedragstherapie 1992;25:27-45.

33. Evers AW, Kraaimaat FW, van Lankveld W, Jongen PJ, Jacobs JW, Bijlsma JW. Beyond unfavorable thinking: the illness cognition questionnaire for chronic diseases. J Consult Clin Psychol 2001;69:1026-36.

34. Spinhoven P, Ormel J, Sloekers PP, Kempen GI, Speckens AE, Van Hemert AM. A validation study of the Hospital Anxiety and Depression Scale (HADS) in different groups of Dutch subjects. Psychol Med 1997;27:363-70.

35. Rosseel Y. lavaan: An R Package for Structural Equation Modeling. Journal of Statistical Software 2012;48.

36. Brown T. Confirmatory Factor Analysis for Applied Research. New York: The Guilford, 2006.

37. Collins LM, Schafer JL, Kam CM. A comparison of inclusive and restrictive strategies in modern missing data procedures. Psychol Methods 2001;6:330-51.

38. Kline RB. Principles and Practice of Structural Equation Modeling, Third Edition. 3rd ed. New York: Guilford Publications, 2010.

39. Hau K, Marsh HW. The use of item parcels in structural equation modelling: Non-normal data and small sample sizes. Br J Math Stat Psychol 2004;57:327-51.

40. Little TD, Rhemtulla M, Gibson K, Schoemann AM. Why the items versus parcels controversy needn't be one. Psychol Methods 2013;18:285-300.

41. Yang C, Nay S, Hoyle RH. Three Approaches to Using Lengthy Ordinal Scales in Structural Equation Models: Parceling, Latent Scoring, and Shortening Scales. Applied Psychological Measurement 2009.

42. Van de Schoot R, Lugtig P, Hox J. A checklist for testing measurement invariance. European Journal of Developmental Psychology 2012;9:486-92.

43. Henrica C. W. de Vet, Caroline B. Terwee, Lidwine B. Mokkink,Dirk L.Knol. Measurement in Medicine, A Practical Guide. Cambridge: Cambridge University Press, 2011.

44. Lalkhen AG, McCluskey A. Clinical tests: sensitivity and specificity. Continuing Education in Anaesthesia, Critical Care & Pain 2008;8:221-3.

45. Buuren S, Groothuis-Oudshoorn K. mice: Multivariate imputation by chained equations in R. Journal of statistical software 2011;45.

46. Lee KJ, Roberts G, Doyle LW, Anderson PJ, Carlin JB. Multiple imputation for missing data in a longitudinal cohort study: a tutorial based on a detailed case study involving imputation of missing outcome data. International Journal of Social Research Methodology 2016;19:575-91.

47. Glaros AG, Kline RB. Understanding the accuracy of tests with cutting scores: The sensitivity, specificity, and predictive value model. J Clin Psychol 1988;44:1013-23.

48. Volgsten H, Skoog Svanberg A, Ekselius L, Lundkvist O, Sundstrom Poromaa I. Prevalence of psychiatric disorders in infertile women and men undergoing in vitro fertilization treatment. Hum Reprod 2008;23:2056-63.

49. Chiaffarino F, Baldini MP, Scarduelli C, Bommarito F, Ambrosio S, D'Orsi C, *et al*. Prevalence and incidence of depressive and anxious symptoms in couples undergoing assisted reproductive treatment in an Italian infertility department. Eur J Obstet Gynecol Reprod Biol 2011;158:235-41.

50. Boivin J, Takefman JE. Stress level across stages of in vitro fertilization in subsequently pregnant and nonpregnant women. Fertil Steril 1995;64:802-10.

51. Ebbesen SM, Zachariae R, Mehlsen MY, Thomsen D, Hojgaard A, Ottosen L, *et al*. Stressful life events are associated with a poor in-vitro fertilization (IVF) outcome: a prospective study. Hum Reprod 2009;24:2173-82.

52. Nelson SM, Lawlor DA. Predicting live birth, preterm delivery, and low birth weight in infants born from in vitro fertilisation: a prospective study of 144,018 treatment cycles. PLoS Med 2011;8:e1000386.

53. Rockliff HE, Lightman SL, Rhidian E, Buchanan H, Gordon U, Vedhara K. A systematic review of psychosocial factors associated with emotional adjustment in in vitro fertilization patients. Hum Reprod Update 2014;20:594-613.