Response to the Referees' report on the paper "Study protocol for the evaluation of a vocational rehabilitation"

Philip Fowler, Xavier de Luna, Per Johansson,
Petra Ornstein, Sofia Bill and Peje Bengtsson.

We thank the reviewers for their constructive feedback and suggestions for revision as well as for the speedy review process. We have revised the paper as described below. Reviewers' comments are reproduced in italics, followed by our answers.

As part of the revision, some errors were found in our R code. We have corrected these and updated all the corresponding tables and figures. The results are very similar to the previous ones. We apologise for this inconvenience.

Reviewer 1

The stratification strategy for waiting time seems arbitrary. Were the cutoffs based on quantiles? (Doesn't appear to be based upon sample sizes in table 4) If not, how were they chosen?

Table 4 shows both the treated and the controls in each of the strata. The stratification strategy was based upon the deciles of waiting times for the treated (that fulfilled our inclusion criteria). Using those quantiles would give the following cutpoints:
$[31, 70)$, $[70, 90)$, $[90, 109)$, $[109, 126)$, $[126, 146)$, $[146, 166)$, $[166, 187)$, $[187, 214)$, $[214, 249)$ and $[249, 361)$.
However, some of these strata do not respect the fact that the regulations for sick leave change by Swedish law once an individual exceeds a waiting time of 90, 180 or 365 days (Socialförsäkringsbalk, 2010, 27 kap. 47-49 §). Since the regulations might change the probability of being treated as well as affect the outcome, the decile-based cutpoints were corrected to take this into account.

We now clarify this in the paper on page 4, lines 39-42.

Are subjects aware of their prognosis? If so, assumption 3i seems very strong to me. A longer prognosis (whether or not the prognosis is appropriate) might be abused by a subject who could really return to work earlier (e.g. had a speedier then average recovery).

This is an important question. The subjects were not aware of their prognosis. Furthermore, the caseworkers were informed that the prognosis variable would not be used to evaluate their performance as caseworkers. We clarify this in the paper on page 5, lines 22-25.

Algorithm steps 2 and 3 - The discarding strategy in these two steps are at odds. Step 3 proceeds to restrict the number of discarded treatments, but step 2 discards at the arbitrary cutoff of needing at least 5 controls. Why not be more flexible with that threshold?

We agree that it would be possible to have a variable number of matches (5 or less) within a given caliper instead of discarding treated not having five matches. While perhaps more appealing

than simply discarding the observations for which five controls could not be found, it has not been implemented in the `Matching` package used to perform our analysis. We do not expect that a change in implementation would make much of a practical difference however. In Step 3, we need to put a restriction on how many treated observations are allowed to be deleted to make sure that best balance is not obtained by simply deleting a large number of treated in Step 2.

Reviewer 2

Assignment to the treatment group. There is limited information provided concerning how individuals are assigned to the JAM tx group, and no information for the post-match activities of the control group. A richer description of both would allow for a better assessment of the appropriateness of the proposed research design.

Caseworkers are trained to make decisions regarding eligibility to get access to sickness benefits. This training is a priority at the SIA and the process is monitored on a yearly basis, ending in a report of how to make improvements, see, e.g., Swedish Social Insurance Agency (2016). Thus while the decision of JAM assessment is at each caseworker's discretion, and thus might vary between caseworkers, it is expected to be based on similar principles. This is now clarified in the paper in the first paragraph of Section 2.3.1. Also, see our response to the next comment for information about the post-match activities of the control group.

Rationale for conceptualizing JAM alone as the treatment. The authors describe two types of interventions - work preparatory and work oriented - that are delivered following a JAM if the individual is determined to be ready for services from PES. Assuming that not all individuals who attend a JAM are either (a) determined ready for services or (b) receive services even when deemed eligible, wouldn't receipt of these services constitute a distinct form of treatment? Are these services accessible to individuals from the control group? The answer to the latter question is unclear in the current manuscript.

The referee is correct that receipt of the services from PES do constitute a distinct form of treatment, and this for different reasons, e.g.: those called to a JAM will not necessarily receive services; those called to a JAM are only sick benefit recipients while PES services are given to unemployed irrespective of them being sick benefit recipients or not; and the services are accessible to all individuals even those from the control group. This was not explicit enough in the manuscript which has been revised (page 4, lines 24-30).

The aim of the evaluation study intended with this protocol is, however, to study the collaboration between SIA and PES (newly launched in 2012) which is targeted to sick benefit recipients. At the core of the collaboration is the call to JAM − at a JAM, individuals are assessed in terms of whether or not they are ready for services from the PES. Thus, while we agree that participation in the services themselves could be a treatment of interest per se, it has a different policy interest than the call to JAM. We have revised Section 2.3.1 in order to clarify the estimand of interest. In particular, we point out that the services (work preparatory and work oriented) are offered after an individual either has been called to a JAM or if he or she registers at the PES as in search of a job. The latter would then lead to him or her losing his/her sickness benefits.

Prognosis variable as proxy. The authors have elected to use caseworker prognosis as a proxy for unobserved confounders. I'm skeptical of the use of caseworker prognosis for this purpose given the possibility for variability across caseworkers in how this determination is made. What information does the caseworker collect that allows for an accurate prognosis that differs in some significant way from estimates based on the proportion of individuals who passed 30, 90 and 180

days of sick leave for a given ICD10 disease? Are you confident that caseworkers are influenced by the same factors and that these factors are accorded equal weight in determining the worker's prognosis? The authors rightly state in the discussion that the proxy property is not empirically testable but it should be theoretically grounded.

Individuals are eligible for sickness benefits if their health condition does not allow them to work. As health and work inability is difficult to observe this means that economic incentives and preferences for work also can be highly important for the sick leave (see e.g. Johansson and Palme, 1996; Hartman, Hesselius, and Johansson, 2013; Hesselius, Nilsson, and Johansson, 2009).

In the caseworker's telephone call with the individual, he/she can obtain otherwise unobserved information on the individual's motivation to work. As mentioned above, the caseworkers are trained to make decisions on the eligibility to sickness benefits. The training to guarantee consistent decisions across caseworkers is highly prioritised at the SIA and the process is monitored on a yearly basis ending with a report of how to make improvements (e.g. Swedish Social Insurance Agency, 2016). There is room for discretion as the referee points out though, and decisions processes may differ among caseworkers. It is likewise possible that caseworkers also differ in their predictions yielding the proxy variable. From a report (Swedish Social Insurance Agency, 2006) we know that an important predictor of the length of an individual's sick spell is his or her own assessment. In Swedish Social Insurance Agency (2014), the caseworker prediction (our proxy variable) was shown to be a significant predictor of the length of an individual sick spell. Based on these two reports we are inclined to believe that the caseworker assessment to a great deal stems from the individual's assessment rather than that the caseworker attitudes and/or preferences which potentially could bias the predictions across caseworkers. We are grateful to the referee for raising this issue, which we now discuss on page 5, lines 29-34. Even if we think we have ground to believe the prognosis variable is a good proxy, we agree that the proxy property must be questioned in the final analysis, since it is untestable, and we therefore plan to conduct a sensitivity analysis to this assumption as mentioned in Section 3.3.

Missing data. If the authors elect to continue using the prognosis variable then reporting out the chi-square value for Little's MCAR test related to prognosis would help to instill confidence that the data were in fact missing completely at random, a major issue given the volume of observations removed for this reason.

While we agree that a test of MCAR would be of use to the reader, Little's MCAR test requires multivariate normality and our data contains many categorical variables. However, Table 1 summarises the distribution of the various covariates (excluding Last County) for the treated individuals with (without) the prognosis registered. There were a total of 2024 treated with prognosis and 1266 without or where the prognosis was set after treatment assignment.

The distributions do not appear to be very dissimilar with regards to their means. The variances (of the numerical covariates) are not shown, but the largest variance ratio between the treated with prognosis and those without was 1.09. As for the distribution of Last County, the absolute difference in percentages between the treated with prognosis and those without, ranged between 0.03 and 5.48 percentage points, with a mean of 1.19. The MCAR assumption remains, however, a concern and we are grateful to the associate editor for suggesting to use those missing prognosis as a supplementary stratum, see below.

Associate Editor

The authors propose to use listwise deletion to address missingness of key covariates, including the case worker prognosis variable. Another strategy commonly used in the matching literature

Table 1: Covariate distribution of the treated with (without) prognosis registered

Covariate	Mean		Min		Max	
Origin of Birth:						
Sweden	81.92%	(80.09%)				
EU/Nordic Country	5.83%	(5.85%)				
Other	12.25%	(14.06%)				
Year of Birth	1968.79	(1968.79)	1948	(1948)	1993	(1993)
Sex:						
Female	61.41%	(62.8%)				
Male	38.59%	(37.2%)				
Marital Status:						
Married	35.18%	(34.12%)				
Unmarried	44.17%	(44.39%)				
Divorced	19.07%	(19.75%)				
Widow/Widower	0.84%	(1.26%)				
Missing	0.74%	(0.47%)				
Children	0.82	(0.81)	0	(0)	6	(7)
SBQI	206 540	(203 157)	0	(0)	999 900	(918 000)
Employment						
Unemployed	18.63%	(19.27%)				
Employed	70.31%	(67.85%)				
Missing	11.07%	(12.88%)				
Education:						
Level 1	2.77%	(3.63%)				
Level 2	15.76%	(15.09%)				
Level 3	57.76%	(58.85%)				
Level 4	5.34%	(4.58%)				
Level 5	18.23%	(17.61%)				
Missing	0.15%	(0.24%)				
ICD10 30 Day Probability	0.73	(0.73)	0.07	(0.07)	0.97	(0.97)
ICD10 90 Day Probability	0.42	(0.42)	0.01	(0.02)	0.84	(0.83)
ICD10 180 Day Probability	0.28	(0.28)	0.01	(0.01)	0.7	(0.68)
ICD10 Chapter:						
M or F	79.99%	(81.04%)				
Other	20.01%	(18.96%)				
Last TESL	0.93	(0.93)	0	(0)	1	(1)
TESL History	5.33	(5.63)	0	(0)	12	(12)

would be to separate out subjects missing the caseworker prognosis variable as a stratum unto themselves. The primary outcome analysis might be restricted to subjects who were not missing this covariate, while study of the no-prognosis stratum could be considered as a supplementary analysis, one with a greater potential for unmeasured confounding.

This was not something we had considered, so we thank the associate editor for this helpful suggestion. We now mention in Section 3.3 that we will use the same matching design (excluding the prognosis variable) to repeat the analysis on the group for which the prognosis is not observed thereby considering the no-prognosis category as a stratum in itself.

References

Hartman, L., P. Hesselius, and P. Johansson (2013). Effects of eligibility screening in the sickness insurance: Evidence from a field experiment. *Labour Economics 20*, 48–56.

Hesselius, P., J. P. Nilsson, and P. Johansson (2009). Sick of your colleagues' absence? *Journal of the European Economic Association 7*(2-3), 583–594.

Johansson, P. and M. Palme (1996). Do economic incentives affect work absence? Empirical evidence using Swedish micro data. *Journal of Public Economics 59*(2), 195–218.

Socialförsäkringsbalk (2010). SFS (2010:110). (Swedish legislation, Social insurance).

Swedish Social Insurance Agency, T. (2006). Prognosverktyget - ett stöd i det första vägvalet vid handläggning av sjukfall. `https://www.forsakringskassan.se/wps/wcm/connect/df25fd48-3750-4b36-bff3-437ab465693d/analyserar_2006_04.pdf?MOD=AJPERES` [Accessed 22 Mar 2017] (In Swedish).

Swedish Social Insurance Agency, T. (2014). Det förstärkta rehabiliteringssamarbetets effekter - utvärdering av arbetsförmedlingens och försäkringskassans förstärkta rehabiliteringssamarbete. `https://www.forsakringskassan.se/wps/wcm/connect/6d6c56ae-f7e7-4182-9209-c7df5f4001de/Rapport%2BAf_Fk%2Beffektutvardering_20140425.pdf?MOD=AJPERES` [Accessed 22 Mar 2017] (In Swedish).

Swedish Social Insurance Agency, T. (2016). Rehabiliteringsersättning (rättslig uppföljning 2016:2). `https://www.forsakringskassan.se/wps/wcm/connect/2627493b-c4a7-455d-96bb-04c003b578af/rattslig_uppfoljning_2016_2.pdf?MOD=AJPERES` [Accessed 22 Mar 2017] (In Swedish).