**Appendix 1**
**Analysis of trials that were still in progress**
According to the reviewers, there were six trials still in progress (Bleijenberg, 2008; Carney, 2008; Gibson-Saxty, 2002; Vissers, 2008; Wearden, 2006; White, 2005), which will help to strengthen the evidence base on CBT interventions for CFS. They wrote that in 2008. These trials should have been published by now and we will analyse them here.

*Bleijenberg (2008) (study protocol), published as Wiborg et al. (2015)*
This trial was not properly controlled: 14 sessions (two hours of CBT over a six-month period; split into groups of four and eight patients) and 0 (waiting-list control group). The authors acknowledged that this could artificially inflate the treatment effect. 37.3% (124/328) of patients who were asked to engage in group therapy refused that so that only those who were willing and motivated to have group therapy were enrolled into the trial. 19% (32/168; CBT) and 11.8% (8/68; waiting-list) of participants dropped out. Missing data were imputed using mean proportions of improvement based on the outcome scores of similar patients with a second assessment. This might have artificially inflated the results, because participants who do not respond to treatment or are negatively affected by it, are more likely to drop out or be lost to follow-up (Lilienfeld et al., 2014).

Psychological distress improved by 16.7% (27/162, CBT) and 10.5% (18/171, waiting-list); suggesting that the subjective improvement in the treatment group might be a reflection of improvement in psychological well-being.

The study used a post hoc definition of clinically significant improvement and recovery using "rigorous criteria for normal functioning". They concluded that more people had recovered after CBT (15.4%) then in the waiting-list group (1.5%). However, in a well-designed trial, such a definition should be defined before it starts, to avoid that the results can influence it. Also they did not use rigorous criteria. One was deemed to have recovered with the following scores: 80 or more (physical functioning), less than 27 (CIS-fatigue) and less than 203 (SIP overall impairment), even though healthy people, with a similar mean age, have scores of 93.1 (PF), 17.3 (CIS-fatigue) and 65.5 (SIP) according to Knoop et al. (2007a), which was used by the authors.

In view of the above-mentioned problems, and not using objective outcomes, one cannot safely conclude that group CBT is effective and leads to recovery in one in eight patients.

*Carney & Jones-Alexander (2008)*
This project received two grants, totalling $418,335 (NIH Project Reporter, 2008), but no results have been published yet.

*Vissers (2008)*
This study has not been published yet.

*Wearden et al. (2006) (study protocol), published as Wearden et al. (2010b)*
Wearden et al. (2010b) is often referred to as the FINE trial **(F**atigue **I**ntervention by **N**urses **E**valuation). This trial was not properly controlled. The average number of sessions in the trial was as follows: 9.63 (pragmatic rehabilitation group (PR), 9.5 (supportive listening group (SL) and 0 (no treatment group (treatment as usual by their own GP when needed). Number of GP consultations: 2 (PR), 3 (SL) and 3 (no treatment).

The entry criteria were changed from the Fukuda to the even wider Oxford criteria, eight months into a four-year lasting trial. No reason was given (Wearden, 2001). According to the FINE trial protocol (Wearden et al., 2006), primary outcomes were to be self-reported physical functioning and fatigue at one year. Yet in the 2010 paper, this was changed to 20 weeks (end of treatment) and 70 weeks from recruitment. 20.3% (60/296) of the participants suffered from anxiety and 17.9% (53/296) from depression.

The only objective secondary outcome measure (step test) was omitted from the 2010 paper even though not publishing results jeopardizes the validity of a study (Heneghan et al., 2017). When the results were published three years later (Wearden et al., 2013), there were no differences between pragmatic rehabilitation and no treatment. The fatigue scores were changed from bimodal (0-11) to Likert (0-33) in a Rapid Response in the BMJ (Wearden et al., 2010a) and in Wearden et al., 2013. This change was made despite the fact that two of the authors (including the Principal Investigator) concluded in a paper, devoted to analysing the use of the Chalder Fatigue scale in CFS (Morriss et al., 1998), that near-maximal scoring on six physical fatigue scale items from the total of 14 items (five if the 11 item scale is used), supports using the two-point bimodal, rather than the four-point Likert scoring. Re-scored there was now a clinically modest, but statistically significant effect of PR compared with no treatment at both outcome points. However, altering measures in this way after the trial to find a small effect suggests a form of p-hacking.

The entry criteria, outcome switching and null objective improvement in this trial mean that it is unsafe to claim any effect for the interventions.

**The PACE trial**

*White (2005), the corresponding reference is White et al. (2007) (study protocol), published as White et al. (2011)*

This trial is the largest CBT and GET study conducted so far. It included 640 patients, compared to 1008 in the Cochrane review. This study used the Oxford criteria, 47% had a comorbid depression or anxiety disorder and 80% of the screened patients were not selected. The control group did not have the same number of contact hours: 16 sessions (CBT), 17 sessions GET (both including 3 sessions of SMC), yet the SMC group only had 5. This imbalance creates serious biases toward finding a positive effect for the intervention, regardless of whether it's effective or not (Lilienfeld et al., 2014).

A null effect at long-term follow-up was spun as positive. Outcomes with SMC alone or adaptive pacing therapy (APT) improved from the 1 year outcome and were similar to CBT and GET at long-term follow-up, but it was claimed the data should be interpreted in the context of additional therapies having been given after the 1 year trial final assessment (Sharpe et al., 2015). However, the Supplementary appendix long-term follow-up shows the majority of participants did not have any additional CBT (76%) or GET (83%) after the trial. It also shows that patients in all four groups, who did not receive additional treatment subsequent to trial completion, exhibited lower fatigue and higher physical functioning scores relative to those of patients who received additional treatment (Vink, 2016).

Baseline figures were used for one objective test, an actometer, a reliable measure of activity to assess improvement objectively (Scheeres et al., 2009), but were not recorded at the end of the trial. The reason given was that it would be too great a burden for patients (Vink, 2016), even though they had consented to use it, they had completed moderately effective

treatment (White (2011) and 22% of those in the CBT and GET groups had recovered according to the investigators (White et al., 2013).

An extensive number of endpoint changes were made (Sharpe et al., 2015; Vink, 2016; Wilshire et al., 2018b; White, 2011). The timing of the changes to the primary outcomes - several months after trial completion - was highly problematic (Wilshire et al., 2018b). As a result there was suddenly an overlap in entry and recovery criteria: 13.3% of participants were already recovered according to one (12.8%) or two (0.5%) of the recovery criteria at trial entry (Vink, 2017a). That is before receiving any treatment and without a change in their medical situation.

These changes affected both the physical function scores (PF) and the fatigue scores. The minimum PF required to qualify as recovered was reduced from 85 to 60 (White et al., 2011). The maximum score for trial entry was increased from 60 to 65 (0-100; higher scores indicating better functioning). Even though according to the PACE trial's recovery article, a score of 65 or less represents "abnormal levels of physical function" (White et al., 2013) and severe disability according to the literature (Stulemeijer et al., 2005). Participants with a score of 60 to 65 (inclusive) where thus considered ill enough to participate and to have an abnormal level of physical functioning, yet were also recovered and severely disabled. Three participants (0.45%) saw their physical functioning score go down from 65 to 60, reflecting deterioration, and three others (0.45%) had unchanged physical functioning scores, but all (0.9%) were still classed as recovered, according to the physical functioning recovery criterion (Vink, 2017b).

Something similar happened to the fatigue scores. When PACE was registered with the ISRCTN on 22 May 2003, participants needed a Chalder Fatigue Questionnaire (CFQ) score of four or more to be classed as ill enough to take part (White, 2003). The CFQ entry criterion was changed to six or more before the trial started and then during an nonblinded trial switched from bimodal to Likert, 18 or more to qualify. To be classed as recovered, a bimodal score of ≤3 out of 11, which represented a screening threshold for abnormal fatigue, was changed to a Likert score of 18 or less (0-33) (White et al., 2013). Consequently, with a Likert score of 18, one was simultaneously classed as disabled and recovered. These endpoint changes increased recovery rates of CBT and GET 4-fold. Had the PACE trial stuck to the protocol defined endpoints than there would have been no statistically significant difference in recovery rates between the four treatment groups (Wilshire et al., 2018b).

The net improvement of the quality of life scores (EQ-5D) after CBT at 52 weeks over APT was 1.8% (0.09/0.63 - 0.06/0.48) (McCrone et al., 2012). A study by Olesen et al. (2016) of 20,220 adult patients, found a mean quality of life score of 0.84 for the total population and 0.93 for people without a chronic condition. Yet the quality of life at 52 weeks in the CBT group (0.63) (McCrone et al., 2012) was similar to the score for cerebral thrombosis (0.62) and still worse than in rheumatoid arthritis and angina (0.65), AMI (acute myocardial infarction) (0.66) (Olesen et al., 2016), MS (0.67), lung cancer and people with 4 or more chronic health conditions (0.69), stroke (0.71) or ischemic heart disease (0.72) (higher scores indicating a better quality of life) (Falk Hvidberg et al., 2015). Also, there was no statistically significant difference in the improvement in CFS symptom count between CBT and APT ($p=0.0986$) at 52 weeks.

These flaws in the trial render unsafe any conclusion that CBT is effective.

**Review of the objective outcomes**

Two of the trials that were still in progress when the Cochrane review was published (Wearden et al., 2010b; White et al., 2011), used objective outcomes too. They form an important part of the evidence base as combined they included 936 (640 + 296) participants. The scores for the only

objective outcome used in Wearden et al. (2010b), the step test, showed no differences between the pragmatic rehabilitation and no treatment (GP treatment as usual) groups on any of the step test measures at 20 or 70 weeks (Wearden et al., 2013).

In White et al. (2011), the step test did not show any objective improvements, therefore fitness did not improve. This is matched by the net improvement of the quality of life scores after CBT over APT of only 1.8%. The number of patients who were unable to work and who were receiving benefits increased and the number of patients receiving income protection in the CBT group actually doubled (McCrone et al., 2012). In addition, there was no statistically significant difference in the improvement in CFS symptom count between CBT and APT ($p=0.0986$) at 52 weeks. There was no statistical significant difference in the 6-minute walking test outcome between CBT and SMC ($p=0.87$) and CBT and APT ($p=0.65$) after exercising for 24 weeks, at 52 weeks. According to these results - 354 m after CBT - patients would still be ill enough to be put on the waiting list for a lung transplant (≤400 m) (Vink, 2016). No one in the trial achieved actual recovery, where symptoms are eliminated and patients return to pre-morbid levels of functioning (Kennedy, 2002), which is the general public's understanding of the meaning of recovery (Vink, 2017a). The PACE trial protocol defined improvement as an increase of 50 per cent. According to the 6-minute walk test results, the only objective individual results that were released, this benchmark was matched by 3.7 per cent in the CBT group, but also by 5 per cent in the SMC group, implying a negative effect of CBT of 1.3 per cent (participants in all treatment groups also received SMC) (Vink, 2017a).

**Quality of life**

In White et al. (2011), participants improved 1.8% more after CBT then after APT at 52 weeks (McCrone et al. , 2012).

**Employment status**

In White et al. (2011), the number of patients who were unable to work and who were receiving benefits, due to illness or disability, increased and the number of patients receiving income protection in the CBT group doubled (McCrone et al., 2012).

**Appendix 2**

**Support for their recovery claims**

Two studies, Knoop et al. (2007a) and Flo & Chalder (2014), were used by White et al. (2011) as support for their recovery claims. We will therefore analyse these two studies too.

*Knoop et al. (2007a)*

This non-randomized cohort trial had no control group and it did not use objective outcomes. The authors concluded that 23% had "fully recovered" after CBT. They acknowledged that in the absence of a control group, it's difficult to attribute this to treatment with certainty. The recovery definition included having a maximum CIS-fatigue score of 27, a minimum physical function score of 80, a maximum SIP functional disability score of 203 and a minimum health perception score of 65, which is the score of people aged 65 to 74 (Twisk and Corsius, 2018). The mean age in this trial is 37.0. However, according to the authors, the scores for healthy people of a similar age, are the following: 17.3 (CIS-fatigue), 93.1 (PF), 80 (health perception, and 65.5 (SIP).

5.2% (5/96) had no PEM. According to the authors some suggest that this is the main characteristic feature of CFS. It's unclear why these people were not excluded from the study. Drop-out rate was 11%.

One cannot come to any meaningful conclusions about the efficacy of CBT in a nonblinded non-randomized trial without a control group, that uses a broad definition of recovery and does not use objective outcomes.

*Flo & Chalder (2014)*

This non-randomized cohort trial did not use a control group or objective outcomes; 28.9% already fulfilled the physical functioning recovery criterion at trial entry. It used the English NICE criteria (having fatigue for the last four months) - which are even wider then the Oxford criteria - 72.7% fulfilled the Oxford and 52.6% the Fukuda criteria.

The recovery criteria were very wide; for example, one was classified as recovered with a physical functioning score of 65 or more (0-100; higher score means better physical function). The problems with choosing a score that low, have been discussed earlier. 28.9% had already achieved this at trial entry. 12.3% already had a score of 83 or more at trial entry, just like 16.1% already fulfilled the Chalder fatigue questionnaire recovery score (less then 18; range 0-33) at trial entry. Drop-out rate was 27.8%. The authors used Knoop et al. (2007a; 23%) and White et al. (2011; 22%) as support for their claim that 18.3% had "fully recovered" after CBT. No objective outcomes or a control group were used in this nonblinded non-randomized cohort trial. One cannot sustain the recovery claim in view of all these flaws.

**Appendix 3**

*Núñez et al. (2011)*

This study was excluded from a recent Cochrane exercise review, as found by a reanalysis (Vink and Vink-Niese, 2018), because exercise therapy was a minor part of the intervention and it did not measure fatigue, viewed as primary outcome in the review.

The trial compared multidisciplinary treatment combining CBT, GET and pharmacological treatment with usual treatment, with one-year follow-up after the end of treatment. It concluded that at twelve months the interventions did not improve health-related quality of life scores, and led to worse physical function and bodily pain scores. Núñez et al. found that the combination of CBT and GET is ineffective and not evidence-based and may in fact be harmful.