Chronic fatigue syndrome and myalgic encephalomyelitis: towards an empirical case definition

Leonard A. Jasona*, Bobby Kot,a Madison Sunnquista, Abigail Brown,a, Meredyth Evansa, Rachel Jantkea, Yolonda Williamsa, Jacob Furst,a and Suzanne D. Vernona

aDePaul University, Chicago, IL, USA; bSolve ME/CFS Initiative, Los Angeles, CA, USA

(Received 19 November 2014; accepted 19 January 2015)

Current case definitions of myalgic encephalomyelitis and chronic fatigue syndrome (CFS) have been based on consensus methods, but empirical methods could be used to identify core symptoms and thereby improve the reliability. In the present study, several methods (i.e. continuous scores of symptoms, theoretically and empirically derived cut off scores of symptoms) were used to identify core symptoms best differentiating patients from controls. In addition, data mining with decision trees was conducted. Our study found a small number of core symptoms that have good sensitivity and specificity, and these included fatigue, post-exertional malaise, a neurocognitive symptom, and unrefreshing sleep. Outcomes from these analyses suggest that using empirically selected symptoms can help guide the creation of a more reliable case definition.

Keywords: myalgic encephalomyelitis; chronic fatigue syndrome; biomarkers; case definitions

Chronic fatigue syndrome and myalgic encephalomyelitis: towards An empirical case definition

Considerable controversy surrounds the illnesses known as chronic fatigue syndrome (CFS), myalgic encephalomyelitis (ME) and myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS). The terms CFS and ME were introduced to describe outbreaks of illness based upon their symptoms. The degree to which they overlap or are the same is currently under debate, and considerable controversy surrounds the term CFS. Patients experience debilitating fatigue in addition to other physical and cognitive symptoms, and substantial recovery from having CFS or ME occurs in less than 10% of cases (Cairns & Hotopf, 2005). The estimated annual direct and indirect costs of this illness to society are over 18 billion dollars (Jason, Benton, Johnson, & Valentine, 2008). Community-based CFS prevalence estimates range from 0.42% (Jason et al., 1999) to 2.54% (Reeves et al., 2007), and these discrepancies might be due to criteria variance.

The most widely used consensus-based CFS case definition is based on the Fukuda et al. (1994) criteria, which require four of the eight core symptoms. Because of a polythetic method of selecting four out of the eight symptoms, it is possible that some individuals who meet these criteria do not have core symptoms of the illness, such as post-exertional malaise, memory and concentration problems, and unrefreshing sleep (Jason, Brown, Evans, Sunnquist,
In contrast, Carruthers et al. (2003) developed what are known as the Canadian ME/CFS consensus-based clinical criteria, which require seven core symptoms. Carruthers et al. (2011) have recently developed another consensus-based case definition called the Myalgic Encephalomyelitis International Consensus Criteria (ME-ICC), which further increased the number of required symptoms to eight. We use the terms ME, CFS, and ME/CFS and suggest that these syndromes may be different from each other. Each has a case definition and a different set of criteria. These terms have been used to describe multi-symptom outbreaks of syndromes in disparate geographic areas. Whether they are same or different is an argument that continues to this day. As for the term ME/CFS, it was first proposed by patient advocate groups who wished to have ME precede the name of their illness (CFS) to counter the stigma that has become associated with CFS and/or a fatiguing illness which might be more in one’s mind than in one’s abnormal physiology. The term ME/CFS was granted legitimacy when it was recommended by the Chronic Fatigue Syndrome Advisory Committee, and later when Dennis Mangan, the National Institutes of Health (NIH) Moderator of the NIH Chronic Fatigue Syndrome State of Knowledge Workshop event stated that out of deference to the patients, the NIH would now and henceforth refer to the illness as ME/CFS.

The Canadian ME/CFS consensus criteria (Carruthers et al., 2003) and the ME-ICC criteria (Carruthers et al., 2011) do identify a smaller subset of patients with more severe symptoms and physical functioning impairment (Jason, Brown, et al., 2013), but both are consensus based rather than empirical. In addition, case definitions that have higher numbers of core symptoms can contribute to higher rates of psychiatric comorbidity (Brown, Jason, Evans, & Flores, 2013; Katon & Russo, 1992). In a recent systematic review, Brurberg, Fønhus, Larun, Flottorp, and Malterud (2014) identified 20 case definitions. While the Fukuda et al. (1994) criteria were the most frequently used, validation of any of the case definitions is inconsistent, and no studies rigorously assessed these case definitions’ reliability to accurately capture people with the illness.

Sources of diagnostic unreliability include subject, occasion, and information variance account, but criteria variance, differences in the formal inclusion, and exclusion criteria used by clinicians to classify patients’ data into diagnostic categories, accounts for the largest source of diagnostic unreliability (Jason & Choi, 2008). Criteria variance is most likely to occur when there are varying criteria for contrasting case definitions. When diagnostic categories lack reliability and accuracy, the validity of a diagnostic category is inherently limited by its reliability. Problems of criteria variance have plagued case definitions involving CFS, ME/CFS, and ME.

Advanced statistical methods could be used to evaluate these consensus criteria as well as suggest a more empirically based case definition, which could deal with the problem of criteria variance. For example, factor analytic studies have explored latent factors (Arroll & Senior, 2009; Brown & Jason, 2014; Friedberg, Dechene, McKenzie, & Fontanetta, 2000; Hickie et al., 2009; Jason, Corradi, & Torres-Harding, 2007), and the domains of neurocognitive impairments and post-exertional malaise are common, whereas fewer studies identify pain, autonomic, immune, and neuroendocrine factors.

Other statistical selection techniques can also help reveal which symptoms are the most useful in distinguishing between patients and healthy individuals and, hence, which symptoms are most characteristic of the illness. For example, using data mining, Jason, Skendrovic, et al. (2012) found that the core features of the illness, including the inability to concentrate, post-exertional malaise, and unrefreshing sleep, best discriminated patients from non-patients. However, this data set was limited in size, and efforts were not directed towards developing an empirical case definition. In another data set, Jason, Sunnquist, Brown, Evans, Vernon, et al. (2014) also found items involving fatigue, post-exertional malaise and neurocognitive problems as differentiating patients from controls. However, there were a number of limitations in that study. The way
these authors determined whether the frequency and severity of symptoms were severe enough to meet criteria was not derived empirically, so it is unclear if similar findings would occur when using more rigorous methods. Another limitation of this study was that only one data mining analysis was conducted and, to avoid any one sample from affecting the results of the analysis, it is important to create multiple sets. In addition, the sample sizes for patients and controls were not equivalent, and this poses additional problems for data mining. Finally, individuals with the identified core symptoms versus those without the core symptoms had not been compared on any symptom or disability measures, so it was unclear whether these core symptoms identified a more impaired group of patients.

The present study attempted to overcome these limitations by empirically developing symptom and frequency cutoff points, creating multiple data sets, using equivalent samples of patients and controls, and examining whether patients identified with core symptoms would evidence more impairment than those without those core symptoms. However, the intent of this study was to gain clarity of case definition for research purposes as opposed to clinical purposes. We hypothesized that symptoms assessing post-exertional malaise, neurocognitive problems, and unrefreshing sleep would best differentiate the patients from controls. Theoretical support for these symptoms includes the fact that each of the various case definitions does list these foundational symptoms, but some employ choice through a polythetic method (Fukuda et al., 1994) whereas others require them (Carruthers et al., 2003, 2011). Empirical support for these symptoms derives from factor analytic studies (e.g. Brown & Jason, 2014) as well as predictors differentiating CFS from major depressive disorder (Hawk, Jason, & Torres-Harding, 2006). We also hypothesized that empirical methods could identify case definition criteria involving fewer core symptoms than more consensus-based approaches.

Method

Participants

SolveCFS BioBank sample

Data from the SolveCFS BioBank were de-identified and shared with the DePaul research team by the Solve ME/CFS Initiative (SMCI). The SolveCFS BioBank has clinical information and blood samples on individuals who were diagnosed by a licensed physician using either the Fukuda et al. (1994) CFS criteria or the Carruthers et al. (2003) Canadian ME/CFS criteria. Individuals with medical or psychiatric reasons for their fatigue were excluded, as this is a requirement for current case definitions. Some patients who had been cared for and whose treatment may have had a beneficial effect on these patients were included, however, all patients still met case definition criteria for either CFS or ME/CFS. All individuals included in the present study were over 18 years of age. Participants were recruited by the SMCI through physician clinics. All participants who met eligibility criteria completed a written informed consent process. Control participants were recruited who were in generally good physical and mental health and did not have a substance use disorder or any disorder that could cause immunosuppression. Controls could not have any medical condition or mental health disorder that caused fatigue. Participants completed the study measures electronically or by hard copy.

Measures

The DePaul Symptom Questionnaire

All participants completed the DePaul Symptom Questionnaire (DSQ) (Jason et al., 2010), a self-report measure of symptomatology, demographics, and medical, occupational, and social
Participants were asked to rate the frequency and severity of 54 symptoms on a 5-point Likert scale. Symptom frequency was rated: 0 = none of the time, 1 = a little of the time, 2 = about half the time, 3 = most of the time, and 4 = all of the time. Likewise, severity was rated: 0 = symptom not present, 1 = mild, 2 = moderate, 3 = severe, and 4 = very severe. The DSQ has evidenced good test–retest reliability among both patient and control groups (Jason et al., in press). A factor analysis by Brown and Jason (2014) found a three-factor solution, with factors evidencing good internal consistency. The DSQ is available at REDCap’s shared library: https://redcap.is.depaul.edu/surveys/?s=tRxytSPVVw

RAND 36-Item Health Survey (Version 1.0). RAND-36 is a self-report questionnaire that measures the impact of physical and mental health on functioning (Ware & Sherbourne, 1992). Low scores indicate that an individual’s health is affecting his or her functioning; higher scores indicate less of an impact. Test construction studies have shown adequate internal consistency, significant discriminant validity among subscales, and substantial differences between patient and non-patient populations (McHorney, Ware, Lu, & Sherbourne, 1994).

Statistics

Methods for replacing missing values

In examining the frequency and severity ratings of the 54 DSQ symptoms, participants missing responses to 10% or more items were removed. Of the participants who remained (233 individuals with CFS and 80 controls), there were 137 instances of missing values (about 0.4% of the total data). If there is a high rate of missing data, this could inflate Type I error. However, as Fidell and Tabachnick (2003) have argued, if less than 5% is missing, then there is not a problem with imputing data. In our study, the percentage of values that were missing was very low. The approach we used in this paper was used by Watson et al. (2014). These missing values were replaced using the following method: for the cases that had a score of 0 for either frequency or severity of a symptom and were missing the other field, the missing value was set to 0; the rationale was that a symptom should occur “none of the time” (frequency = 0) if and only if the symptom is “not present” (severity = 0). Otherwise, if a subject was missing data in only one of the two fields (frequency or severity) for a symptom, then the missing value was replaced with the mode value from the cases that had the same score for the non-missing field. When both fields were missing for a symptom, the values were replaced with the overall medians in those fields for that symptom.

Receiver operating characteristic curve analysis

For this analysis, a composite score for each symptom was created by averaging its frequency and severity scores and multiplying the result by 25; thus, possible scores ranged from 0 to 100.

Classification accuracy of individual symptoms

The symptoms listed in the DSQ were converted into binary variables for use in the next analysis, as we wanted to develop a method for determining when a symptom met a threshold that indicated it was a significant problem for the patient. In other words, binary variables for each symptom indicated whether or not the participant reported frequency and severity levels that met a minimum threshold. Initially, a threshold was applied that was defined in a prior study (Jason, Sunnquist, Brown, Evans, Vernon, et al., 2014): a symptom’s frequency and severity scores needed to be greater than or equal to 2 (symptoms of at least moderate severity that occur at
least half of the time). We assumed that symptoms that occurred at least half the time and were of moderate severity would be a reasonable threshold for discriminating somewhat serious symptoms from those that were relatively mild and not impairing (and below we provide a more empiric way of determining this threshold). The resulting binary symptoms derived from this 2,2 threshold were used to test the predictive accuracy of each of the symptoms in discriminating between patients and healthy controls. A benefit of this threshold is that it has some face validity and is theoretically appealing; furthermore, it is easier to interpret than a continuous score.

Next, as an alternative to applying a static, 2,2 threshold for all symptoms, empirical methods were used to determine the frequency and severity scores that best discriminate patients and controls for each individual symptom. The threshold was dynamically adjusted for each symptom based on observed frequency and severity scores, similar to Watson et al.’s (2014) use of unsupervised learning. Supervised machine learning techniques, as used in Hanson, Gause, and Natelson (2001), are only valid insofar as they reflect the initial diagnosis criteria. In order to develop an empirical definition, it is imperative to minimize any reliance on pre-existing case definitions so that results do not simply mirror selection biases of the prior definitions. In the current study, a k-means clustering approach was used. Generally speaking, the k-means algorithm iteratively divides coordinate points into a predetermined number of clusters based on which cluster centre the point lies closest to. In this case, the k-means clustering algorithm was set to find two clusters, based on the underlying assumption that the data consisted of patient and control groups. Frequency and severity scores for each symptom were treated as coordinate pairs for the purpose of cluster assignment, and the Euclidean distance was used to measure closeness to the cluster centres. After equilibrium was reached, the perpendicular bisector of the line between cluster centres was found. This bisecting line was used as the threshold; frequency-severity pairs above the threshold line were considered “symptom present”, whereas scores below the line were considered “symptom not present”.

Data mining: All symptoms were placed into the analyses, rather than one symptom at a time. In the current study, decision trees were used to determine which symptoms were the most effective at accurately classifying participants as either a patient or control. Decision trees consist of a series of successive binary choices (branch points) that result in an accurate classification of participants.

SPSS Statistics software was used to build our decision tree models. To build the models, a Classification and Regression Tree algorithm was applied to a training set consisting of 66% of the cases, stratified to reflect the distribution of patient and control groups. The value of the model was measured by evaluating its classification performance when applied to cases reserved for testing (34% of the data), allowing this technique the ability to be generalized to new data. Data mining in general, and decision trees specifically, is biased when label sets are not of equal size. We took a random subsample of 80 patients along with all the 80 controls, omitting the other 153 patients. To avoid any one training or testing subsample from affecting the result of the analysis, we created 100 such sets (random subsample of 80 of patients and all 80 controls) for analysis. For most analyses, only three to five variables were needed to classify participants.

Comparison of groups: To further explore the results of the decision tree analyses, three groups of participants were compared: healthy controls, participants diagnosed with CFS or ME/CFS who met the 2,2 frequency and severity criteria for the symptoms identified in the decision tree analyses, and participants diagnosed with CFS or ME/CFS who did not meet these 2,2 criteria. As we had unequal sample sizes and unequal variances, we selected statistical tests which would accommodate these data problems. Welch’s F tests and Games-Howell post hoc tests were conducted to compare the RAND-36 subscale scores and 100-point symptom scores of these groups. Additionally, a total symptom score was computed by summing each participant’s frequency and severity scores for the 54 DSQ symptoms; a Welch’s F test and Games-Howell post hoc test were conducted to compare the total symptom scores of the three groups.
Results

Demographics

Jason, Sunnquist, Brown, Evans, Vernon, et al. (2014) report demographic characteristics of this sample. About three quarters of the sample were female and 98–99% were white, and these demographics are comparable to much of the published literature. Significant differences existed in work status between the control group and those that met the Fukuda et al. (1994) CFS criteria \( p < 0.000 \), two-tailed Fisher’s exact test, the Carruthers et al. (2003) Canadian ME/CFS criteria \( p < 0.000 \), two-tailed Fisher’s exact test, and the Carruthers et al. (2011) ME-ICC criteria \( p < 0.000 \), two-tailed Fisher’s exact test. Most of the individuals in the control group were working, while about 70% of the patients were on disability. Additionally, a significant difference was found when comparing the marital status of those meeting the Fukuda et al. CFS criteria and control groups \( p = 0.03 \), two-tailed Fisher’s exact test, as a larger proportion of the Fukuda et al. CFS group were single.

Receiver operating characteristic curve analysis

Table 1 shows the AUCs for the 10 most accurate symptoms using continuous scores in the DSQ. Accuracy was determined based on the individual symptom’s ability to correctly predict CFS or healthy control status, and it was used to generate the receiver operating characteristic (ROC) curves (and area under the curve (AUC) of .90 or better is considered as very good). It is apparent that fatigue, post-exertional malaise, and neurocognitive symptoms are among the most accurate items. Unrefreshing sleep was also among the top 10 items. These findings provided evidence that our hypothesized symptoms would be among the most accurate predictors.

Classification accuracy of individual symptoms

Table 2 provides the most accurate symptoms when this 2,2 threshold was applied, and among the most accurate were fatigue, post-exertional malaise, neurocognitive, and sleep symptoms. Using the unsupervised learning system, where the threshold was dynamically adjusted for each symptom based on observed frequency and severity scores, as evident in Table 2, we found comparable results to the 2,2 criteria analysis, thus confirming the usefulness of the simpler-to-use 2,2 criteria.

Table 1. AUC for top 10 symptoms.

| Individual symptoms                                      | AUC  |
|----------------------------------------------------------|------|
| Fatigue/extreme tiredness                               | 0.978|
| Physically drained/sick after mild activity             | 0.971|
| Minimum exercise makes you physically tired             | 0.968|
| Next-day soreness after non-strenuous activities        | 0.952|
| Dead, heavy feeling after starting to exercise          | 0.951|
| Unrefreshing sleep                                      | 0.946|
| Mental tiredness after slightest effort                 | 0.945|
| Slowness of thought                                     | 0.943|
| Muscle weakness                                         | 0.940|
| Difficulty finding the right word to say or expressing thoughts | 0.931|
Data mining

In Table 3, the data mining analyses suggested the selection of four symptoms (using the 2,2 criteria): fatigue or extreme tiredness, difficulty finding the right word to say or expressing thoughts, physically drained/sick after mild activity, and unrefreshing sleep. In particular, these were all symptoms that appeared in a majority of the 100 classification trees. Figure 1 shows that 62% of patients referred by medical specialists had these four symptoms, and these criteria are referred to as the four-symptom criteria.

| 2.2 Threshold individual symptoms | Accuracy (%) |
|-----------------------------------|--------------|
| Fatigue/extreme tiredness         | 94.9         |
| Minimum exercise makes you physically tired | 89.8 |
| Unrefreshing sleep                | 89.5         |
| Physically drained/sick after mild activity | 87.5 |
| Next-day soreness after non-strenuous activities | 86.6 |
| Dead, heavy feeling after starting to exercise | 85.9 |
| Problems remembering things       | 83.4         |
| Difficulty finding the right word to say or expressing thoughts | 79.2 |
| Muscle weakness                   | 78.9         |
| Can only focus on one thing at a time | 78.0 |

| Dynamic threshold individual symptoms | Accuracy (%) |
|--------------------------------------|--------------|
| Fatigue/extreme tiredness            | 95.5         |
| Dead, heavy feeling after starting to exercise | 88.2 |
| Minimum exercise makes you physically tired | 88.2 |
| Physically drained/sick after mild activity | 87.9 |
| Next-day soreness after non-strenuous activities | 87.5 |
| Unrefreshing sleep                   | 87.2         |
| Problems remembering things          | 83.7         |
| Muscle weakness                      | 81.8         |
| Difficulty finding the right word to say or expressing thoughts | 80.5 |
| Can only focus on one thing at a time | 78.9 |

Figure 1. Individuals referred by medical specialists in CFS and ME/CFS.
Comparison of groups

Table 4 displays the mean RAND-36 subscale scores of healthy controls, individuals diagnosed with CFS or ME/CFS who did not meet the four-symptom criteria, and individuals diagnosed with CFS or ME/CFS who met the four-symptom criteria. Welch’s F-tests indicated that these groups were significantly different on all eight subscales: physical functioning \[ F(2, 179.3) = 535.54, p < .001 \], role physical \[ F(2, 148.5) = 639.57, p < .001 \], bodily pain \[ F(2, 182.7) = 172.53, p < .001 \], general health \[ F(2, 172.8) = 462.87, p < .001 \], social functioning \[ F(2, 179.9) = 452.26, p < .001 \], mental health \[ F(2, 192.1) = 18.67, p < .001 \], Role Emotional \[ F(2, 190.1) = 18.17, p < .001 \], and Vitality \[ F(2, 154.2) = 355.23, p < .001 \]. Games-Howell post hoc tests revealed significant differences between the control group and both patient groups on all eight subscales. The group that met the four-symptom criteria showed significantly worse physical functioning, bodily pain, general health, social functioning, and vitality scores than the patient group that did not meet criteria.

Table 4 also displays the three groups’ mean scores for the symptoms included in the four-symptom criteria as well as the total symptom score. As expected, Welch’s F-tests evidenced significant differences among groups for all four symptoms: fatigue / extreme tiredness \[ F(2, 154.4) = 381.37, p < .001 \], physically drained/sick after mild activity \[ F(2, 173.8) = 858.18, p < .001 \], difficulty finding the right word to say or expressing thoughts \[ F(2, 173.7) = 394.24, p < .001 \], and unrefreshing sleep \[ F(2, 151.5) = 225.23, p < .001 \], as well as the total symptom score \[ F(2, 192.3) = 443.97, p < .001 \]. In order to control for Type I error, Games-Howell post hoc tests were used to show whether each group was significantly different from all other groups. The patient group that met the four-symptom criteria had significantly worse scores than the patient group that did not meet criteria, and both groups had significantly worse scores than controls.
Conclusion

The findings of this study suggest that core symptoms of this illness are fatigue, post-exertional malaise, a neurocognitive symptom, and unrefreshing sleep. These findings were consistent when using continuous scores, theoretically and empirically derived cutoff scores, and data mining analyses. These results are theoretically compatible with other studies, such as Hawk, Jason, and Torres-Harding’s (2006) investigation which found that these domains were able to successfully differentiate patients with CFS from major depressive disorder. Factor analytic studies also suggest these are among the most common domains found for this illness (Brown & Jason, 2014). Other symptoms, such as pain, autonomic, immune, and neuroendocrine symptoms are less prevalent, but still important; and scores on these domains could be specified as secondary areas of assessment. The present study suggests that empirical methods can be used to help determine which symptoms to include in the case definition.

For the Canadian ME/CFS case definition (Carruthers et al., 2003), seven symptoms need to be present for a patient to meet criteria, whereas eight are required for the ME-ICC (Carruthers et al., 2011). However, by using data mining empirical methods, only four symptoms were required to differentiate patients from controls. These have the advantage of referring to specific core symptoms rather than using the polythetic method of four out of eight symptoms of the Fukuda et al. (1994) criteria. Using the same data set as the present study, Jason, Brown, Evans, Sunnquist, and Newton (2013) found that the Fukuda et al. (1994) criteria identified 93% of the referred sample, whereas the Canadian ME/CFS clinical criteria (Carruthers et al., 2003) identified 73% of the sample. In addition, our best estimate for the ME-ICC criteria (Carruthers et al., 2011) from two other patient data sets (Jason, Sunnquist, Brown, Evans, & Newton, 2014) indicated that approximately 58% of cases would be identified. Figure 1 graphically portrays how the use of the four-symptom

Table 4. Comparison of RAND-36 and symptom scores.

| RAND-36 subscale² | Control (n = 80) M (SD) | Did not meet criteria¹ (n = 88) M (SD) | Met criteria¹ (n = 145) M (SD) | Sig. |
|-------------------|------------------------|--------------------------------------|-------------------------------|-----|
| Physical functioning | 94.5 (9.3)ab | 46.6 (24.6)ac | 32.0 (20.9)bc | *** |
| Role physical | 93.4 (20.7)ab | 7.5 (18.3)a | 2.8 (12.5)b | *** |
| Bodily pain | 86.0 (15.2)ab | 55.4 (24.7)ac | 40.5 (21.5)bc | *** |
| General health | 82.2 (13.9)ab | 30.1 (19.5)ac | 23.0 (14.9)bc | *** |
| Social functioning | 93.2 (14.3)ab | 41.2 (28.5)ac | 22.4 (21.5)bc | *** |
| Mental health functioning | 80.1 (15.0)ab | 69.9 (14.5)a | 65.9 (20.3)b | *** |
| Role emotional | 92.0 (20.8)ab | 78.2 (37.3)a | 66.7 (43.3)b | *** |
| Vitality | 71.3 (17.7)ab | 20.2 (16.3)ac | 11.8 (12.1)bc | *** |
| Symptom scores³ | | | | |
| Fatigue/extreme tiredness | 17.6 (19.3)ab | 71.6 (19.7)ac | 84.6 (13.3)bc | *** |
| Physically drained/sick after mild activity | 2.8 (10.8)ab | 56.6 (31.7)ac | 80.0 (17.3)bc | *** |
| Difficulty finding the right word to say or expressing thoughts | 15.0 (15.1)ab | 39.8 (21.9)ac | 76.0 (17.0)bc | *** |
| Unrefreshing sleep | 23.9 (22.6)ab | 69.8 (25.8)ac | 84.6 (15.9)bc | *** |
| Total symptom score | 37.9 (37.7)ab | 151.1 (49.3)ac | 221.8 (55.4)bc | *** |

a,b,cHorizontal lines that have similar letters are significantly different.
¹“Criteria” indicate frequency and severity scores of 2 or greater for the four symptoms identified in the decision tree analyses.
²Higher scores indicate better functioning.
³Higher scores indicate worse symptoms.
***p < .001.
criteria identified in the present study classified 62% as meeting the new empirical criteria. In other words, using these four-symptom criteria, fewer patients were identified than by the Fukuda et al. (1994) CFS criteria or the ME/CFS Canadian criteria (Carruthers et al., 2003), and slightly more than by the ME-ICC criteria (Carruthers et al., 2011).

Table 4 indicates that participants who met the four-symptom criteria showed significantly more impairment than healthy controls and individuals with CFS or ME/CFS who did not meet these criteria. Of interest, those who met the four-symptom criteria did not show worse role emotional or mental health functioning than those who did not meet the four criteria. Furthermore, mean RAND-36 scores of individuals who met these criteria were similar to those who met the Canadian ME/CFS criteria (Jason, Brown, et al., 2013). The Canadian ME/CFS criteria require information on 54 symptoms in order to determine whether individuals have symptoms from the seven required domains. The results from the current study indicate that individuals identified using fewer, but empirically selected, symptoms can evidence comparable disability to those who meet other case definitions that require more symptoms.

The present study was methodologically stronger than the prior one by Jason, Sunnquist, Brown, Evans, Vernon, et al. (2014), as the current study first identified core symptoms by using continuous scores with ROC curve analyses, and then compared the unsupervised learning system to determine threshold with more theoretically driven 2,2 criteria. The current study found that the 2,2 threshold was comparable to the continuous method as well as the empirically defined threshold, thus providing support for the 2,2 criteria. In addition, the current study was different from the prior Jason, Sunnquist, Brown, Evans, Vernon, et al. (2014) study by conducting mining with equal sample sizes for patient and controls, reporting on 100 sets of data mining analyses as opposed to just one, and comparing those who met the four-symptom criteria to those that did not on disability measures. Findings identified four-symptom criteria that seem to differentiate patients with this illness.

Although the identified criteria in this paper resulted in a group of participants who were statistically significantly worse scores on impairment measures, some of the differences between patients who met the four-symptom criteria and those who did not were small. For example, those who met the four-symptom criteria had a 2.8 on the role physical subscale, those who did not have a 7.5, while the controls had a score of 93.4. It would seem that those who did not meet criteria were still experiencing a clinically significant impairment on that subscale that was fairly comparable to those who did meet criteria. However, in Table 4, there are only three areas where there is not a significant difference, and, in all cases, those who meet criteria have worse scores. It is possible that we have identified two groups of patients, and future work can focus on better understanding their differential characteristics.

It is possible that even what we have shown to be primary symptoms are not present at all stages of the illness. In addition, some patients were being cared for – therefore the management/treatment might have reduced the severity of symptoms as assessed in the current work. It is also possible that some of these patients did not have post-exertional malaise as defined by our questions, and then they would have been misclassified. In addition, it is important to note that whether one uses the four-item criteria or other criteria such as the Fukuda et al. (1994), those with other causes for CFS or ME (e.g. cancer, medications, etc.) need to be excluded, and those criteria also need to be developed using more empirical rather than just consensus methods.

Future studies might also be directed towards determining what we have classified as secondary symptoms may contribute less to the illness burden. For example, pain is the major contributor to incapacity and self-reported symptoms; thus, some might challenge its relegation to a secondary role. Others might feel that autonomic dysfunction would occur in a larger proportion of patients using additional measures to those that examine orthostatic intolerance, as autonomic dysfunction can also affect many organs and may lead to a wide variety of symptoms including gastrointestinal, genitourinary, temperature dysregulation, ocular, etc.
In addition, there is a need to include biological indices rather than just self-report data to confirm differences in diagnostic classifications. For example, Brenu et al. (2013) found natural killer cell activity significantly decreased for both the Fukuda et al. (1994) and the ME-ICC (Carruthers et al., 2011) case definitions, but only those diagnosed with the ME-ICC had significant correlations between physical status and some immune parameters. Finally, the results of the current study need to be replicated.

This study has implications for assessment science and practice. Criteria variance is most likely to occur when operationally explicit criteria do not exist for diagnostic categories, or when there are varying criteria for contrasting case definitions. If the current CFS, ME/CFS, and ME diagnostic categories lack reliability and accuracy, their validity is inherently limited. There is considerable debate ongoing now within the scientific community regarding how to deal with this criteria variance problem (Jason, Najar, Porter, & Reh, 2009), and research presented in this study suggests that empirical strategies have many advantages over more consensus-based approaches in dealing with this issue. Dealing with improving the case definition is critical for enabling investigators to better understand aetiology, epidemiology, pathophysiology, and treatment approaches for those with ME, ME/CFS, and CFS.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

Funding was provided by the Eunice Kennedy Shriver National Institute of Child Health and Human Development [Grant Number R01HD072208] and the National Institute of Allergy and Infectious Diseases [Grant Number AI105781]. The authors appreciate the Solve ME/CFS Initiative (formerly the CFIDS Association of America), which approved the use of de-identified SolveCFS BioBank registry data in this analysis.

References

Arroll, M. A., & Senior, V. (2009). Symptom typology and sub-grouping in chronic fatigue syndrome. *Bulletin of the IACFS/ME*, 17(2), 39–52.
Brenu, E. W., Johnston, S., Hardcastle, S. L., Huth, T. K., Fuller, K., Ramos, S. B., …, Marshall-Gradisnik, S. M. (2013). Immune abnormalities in patients meeting new diagnostic criteria for chronic fatigue syndrome/myalgic encephalomyelitis. *Molecular Biomarkers & Diagnosis*, 152. doi: 10.4172/2155-9929.1000152
Brown, A., & Jason, L. A. (2014). Validating a measure of myalgic encephalomyelitis/chronic fatigue syndrome symptomatology. *Fatigue: Biomedicine, Health & Behavior*, 2, 132–152.
Brown, A. A., Jason, L. A., Evans, M. A., & Flores, S. (2013). Contrasting case definitions: The ME International Consensus Criteria vs. the Fukuda et al. CFS Criteria. *North American Journal of Psychology*, 15(1), 103–120.
Brurberg, K. G., Fønhus, M. S., Larun, L., Flottorp, S., & Malterud, K. (2014). Case definitions for chronic fatigue syndrome/myalgic encephalomyelitis (CFS/ME): A systematic review. *BMJ Open 2014*, 4, e003973
Cairns, R., & Hotopf, M. (2005). A systematic review describing the prognosis of chronic fatigue syndrome. *Occupational Medicine*, 55(1), 20–31.
Carruthers, B. M., Jain, A. K., De Meirleir, K. L., Peterson, D. L., Klimas, N. G., Lerner, A. M., …, van de Sande, M. I. (2003). Myalgic encephalomyelitis/chronic fatigue syndrome: Clinical working case definition, diagnostic and treatments protocols. *Journal of Chronic Fatigue Syndrome, 11*, 7–115.
Carruthers, B. M., van de Sande, M. I., De Meirleir, K. L., Klimas, N. G., Broderick, G., Mitchell, T., …, Stevens, S. (2011). Myalgic encephalomyelitis: International Consensus Criteria. *Journal of Internal Medicine*. doi: 10.1111/j.1365-2796.2011.02428.x
Fukuda, K., Straus, S. E., Hickie, I., Sharpe, M. C., Dobbins, J. G., & Komaroff, A. (1994). The chronic fatigue syndrome: A working case definition. *Annals of Internal Medicine, 121*, 953–959.

Hanson, S. J., Gause, W., & Natelson, B. (2001). Detection of immunologically significant factors for chronic fatigue syndrome using neural-network classifiers. *Clinical and Diagnostic Laboratory Immunology, 8*, 658–662.

Hawk, C., Jason, L. A., & Torres-Harding, S. (2006). Differential diagnosis of chronic fatigue syndrome and major depressive disorder. *International Journal of Behavioral Medicine, 13*, 244–251.

Hickie, I., Davenport, T., Vernon, S. D., Nisenbaum, R., Reeves, W. C., Hadzi-Pavlovic, D., …, Lloyd, A. (2009). Are chronic fatigue and chronic fatigue syndrome valid clinical entities across countries and health-care settings? *Australian and New Zealand Journal of Psychiatry, 43*, 25–35.

Jason, L. A., Benton, M., Johnson, A., & Torres-Harding, S. (2008). The economic impact of ME/CFS: Individual and societal level costs. *Dynamic Medicine, 7*(6). PMCID: PMC2324078

Jason, L. A., Brown, A., Evans, M., Sunnquist, M., & Newton, J. L. (2013). Contrasting chronic fatigue syndrome versus myalgic encephalomyelitis/chronic fatigue syndrome. *Fatigue: Biomedicine, Health & Behavior, 1*, 168–183.

Jason, L. A., & Choi, M. (2008). Dimensions and assessment of fatigue. In Y. Yatanabe, B. Evengard, B. H. Natelson, L. A. Jason, & H. Kuratsune. (2008). *Fatigue science for Human Health* (pp. 1–16). Tokyo: Springer.

Jason, L. A., Corradi, K., & Torres-Harding, S. (2007). Toward an empirical case definition of CFS. *Journal of Social Service Research, 34*, 43–54. doi: 10.1300/J079v34n02_04

Jason, L. A., Evans, M., Porter, N., Brown, M., Brown, A., Hunnell, J., Anderson, V., …, Friedberg, F. (2010). The development of a revised Canadian Myalgic Encephalomyelitis-Chronic Fatigue Syndrome case definition. *American Journal of Biochemistry and Biotechnology, 6*(2): 120–135. doi: 10.3844/ajbbsp.2010.120.135

Jason, L. A., Najjar, N., Porter, N., & Reh, C. (2009). Evaluating the centers for disease control’s empirical chronic fatigue syndrome case definition. *Journal of Disability Policy Studies, 20*, 93–100.

Jason, L. A., Richman, J. A., Rademaker, A. W., Jordan, K. M., Plioplys, A. V., Taylor, R. R., …, Plioplys, S. (1999). A community-based study of chronic fatigue syndrome. *Archives of Internal Medicine, 159*, 2129–2137.

Jason, L. A., Skendrovic, B., Furst, J., Brown, A., Weng, A., & Bronikowski, C. (2012). Data mining: Comparing the empiric CFS to the Canadian ME/CFS case definition. *Journal of Clinical Psychology, 68*, 41–49.

Jason, L. A., So, S., Brown, A., Sunnquist, M., & Evans, M. (in press). Test-retest reliability of the DePaul Symptom questionnaire. *Fatigue: Biomedicine, Health, and Behavior.*

Jason, L. A., Sunnquist, M., Brown, A., Evans, M., & Newton, J. L. (2014). Are myalgic encephalomyelitis and chronic fatigue syndrome different illnesses? *Journal of Health Psychology. Advance online publication. doi: 10.1177/1359105313520335*

Jason, L. A., Sunnquist, M., Brown, A., Evans, M., Vernon, S. D., Furst, J., & Simonis, V. (2014). Examining case definition criteria for chronic fatigue syndrome and Myalgic Encephalomyelitis. *Fatigue: Biomedicine, Health, and Behavior, 2*(1), 40–56.

Katon, W., & Russo, J. (1992). Chronic fatigue syndrome criteria: A critique of the requirement for multiple physical complaints. *Archives of Internal Medicine, 152*, 1604–1609. doi: 10.1300/J092v13n02_01

McHorney, C. A., Ware, J. E., Lu, R. L., & Sherbourne, D. (1994). The MOS 36-item Short-Form Health Survey (SF-36): III. Tests of data quality, scaling assumptions, and reliability across diverse patient groups. *Medical Care, 32*, 40–66.

Reeves, W. C., Jones, J. F., Maloney, E., Heim, C., Hoaglin, D. C., Boneva, R. S., … , & Devlin, R. (2007). Prevalence of chronic fatigue syndrome in metropolitan, urban, and rural Georgia. *Population Health Metrics, 5*(5), 1–10.

Ware, J. E., & Sherbourne, C. D. (1992). The MOS 36-item Short-Form health survey (SF-36): Conceptual framework and item selection. *Medical Care, 30*, 473–483. PMID: 1593914.

Watson, S., Ruskin, A., Simonis, V., Jason, L., Sunnquist, M., & Furst, J. (2014). Identifying defining aspects of chronic fatigue syndrome via unsupervised machine learning and feature selection, *International Journal of Machine Learning and Computing, 4*, 133–138.