



ELSEVIER

Journal of Clinical Epidemiology 58 (2005) 609–617

**Journal of
Clinical
Epidemiology**

The Wisconsin Upper Respiratory Symptom Survey is responsive, reliable, and valid

Bruce Barrett*, Roger Brown, Marlon Mundt, Nasia Safdar, Leota Dye,
Rob Maberry, Jennifer Alt

Department of Family Medicine, University of Wisconsin Medical School, 777 South Mills Street, Madison, WI 53715, USA

Accepted 25 November 2004

Abstract

Objective: To assess reliability, responsiveness, importance to patients, and convergent validity for the Wisconsin Upper Respiratory Symptom Survey (WURSS-44) and to develop a short-form WURSS.

Study Design and Setting: Community-based recruitment of participants with colds. Prospective monitoring from within 48 hours of first symptom until 2 days after end of cold. The WURSS-44 includes 1 global illness severity item, 32 symptom-based items, 10 functional quality-of-life items, and 1 item assessing global change. The SF-36, SF-8, and the Jackson cold scale were used as external comparators.

Results: Participants included 104 women and 45 men, aged 18 to 80 years, self-reporting on 1,681 person-days of illness. Factor analysis suggested 10 dimensions, with reliability coefficients from 0.62 to 0.93. Comparing daily WURSS-44 to Jackson and SF-8 yielded Pearson correlation coefficients from 0.73 to 0.93, and from -0.60 to -0.84 , respectively. Importance to patients and responsiveness assessment yielded a short version, the WURSS-21. Guyatt's responsiveness index was 0.54 for the SF-8, 0.61 for the Jackson, 0.71 for the WURSS-44, and 0.80 for the WURSS-21, suggesting that a two-armed trial would require 74 participants for the WURSS-21, 92 for the WURSS-44, 124 for the Jackson scale, and 156 for the SF-8.

Conclusions: The construct validity of WURSS-44 is supported by measures of reliability, responsiveness, importance to patients, and convergence. A shorter version, the WURSS-21, may be even more responsive. © 2005 Elsevier Inc. All rights reserved.

Keywords: Clinical significance; Common cold; Controlled trials; Evidence-based medicine; Minimal important difference; Psychometrics; Quality of life; Questionnaires; Symptom measurement; Upper respiratory infection; Validation

1. Introduction

Viral infection of the upper respiratory tract leads to a syndrome known as the common cold. More than half of acute illness episodes may fit within this nosologic category [1]. Adults average two to three colds per year and children average four to six, with 70% of the population reporting cold symptoms [2–4]. The annual economic impact of noninfluenza viral respiratory infection is estimated at \$40 billion, with 40 million days of work and school lost [5].

A variety of laboratory and questionnaire-based outcomes have been used in trials testing cold remedies [6,7]. None have been systematically tested for validity, perhaps due in part to the fact that there is no gold standard for assessing the common cold. Various laboratory measures used in research (e.g., mucus weight, viral titer, nasal neutrophils, and cytokine assays including IL-1, IL-6, and IL-8) have major limitations, and none are part of clinical practice. The most

commonly used symptom severity measures are the Jackson criteria, developed in the late 1950s [8,9]. A Jackson score is based on a simple sum of severity points (0 = absent, 1 = mild, 2 = moderate, 3 = severe) for eight cold symptoms: sneezing, nasal obstruction, nasal discharge, sore throat, cough, headache, chilliness, and malaise. Discriminative (diagnostic) validity was shown by Jackson et al. [8,9], using community-acquired and coxsackievirus-induced colds, and later by Gwaltney [10], using rhinovirus. The Jackson scale has several limitations, including lack of systematic assessment of reliability, responsiveness, factorial structure and external validity. Perhaps more important, the Jackson scale does not assess the “functional” and “quality of life” domains that are important to most cold sufferers: impairment of activities of daily living, breathing, sleeping, working, and interpersonal relationships.

2. Materials and methods

The Wisconsin Upper Respiratory Symptom Survey (WURSS) is an empirically derived patient-oriented illness-specific quality-of-life evaluative outcomes instrument.

* Corresponding author. Tel.: 608-263-2220; fax: 608-263-5813.

E-mail address: bruce.barrett@fammed.wisc.edu (B. Barrett).

Development methods are reported elsewhere [11]. First, WURSS was designed to reflect outcomes important to people with colds. Second, WURSS was developed as an evaluative instrument, designed to measure items and domains that change over time. Recognizing the diversity of approaches to instrument development and validation, we tend to follow the conceptual framework described by McDowell and Newell [12] and to use methods advanced by Guyatt and colleagues at McMaster [13–15]. As there is no gold standard for measuring the common cold, we chose to assess construct validity, in which evaluation of reliability, responsiveness, and importance to patients is complemented by assessment of associations with external measures, all guided by appropriate conceptual frameworks [12].

The WURSS-44 was designed and developed to comprehensively measure all significant health-related dimensions that are negatively affected by the common cold [11]. The WURSS-44 includes 1 global severity item (“How sick do you feel today?”), 32 symptom-based items, 10 functional quality-of-life items, and 1 global change item (“Compared to yesterday, I feel that my cold is”). See Table 1. All items are based on seven-point Likert-type severity scales, with “very mild,” “mild,” “moderate,” and “severe” aligned with the numbers 1, 3, 5, and 7. A cumulative score is calculated by simply summing the severity scores of the first 43 items. (The last item, assessing global change, is conceptually distinct and is formatted differently, and so is not included in the simple sum cumulative score.)

The present study was designed as a prospective observational validation research project. People with new-onset common cold were invited to meet for informed consent and enrollment. Participants responded to community advertising by calling a listed phone number. Answering research personnel screened for inclusion and exclusion criteria. Exclusion criteria were (1) any nasal or throat symptom present

for more than 48 hours, (2) a history of allergy along with current eye or nose itching or sneezing, (3) if either the participant or the enroller thought that any symptoms might be due to allergy, or (4) age below 18. Participants had to have a Jackson score of 3 or higher to be enrolled. Time of onset of first symptom was assessed first during phone screening, then again in person at the consent and enrollment intake interview. Consent procedures were approved by the University of Wisconsin–Madison institutional review board’s human subjects committee. Following the intake interview, participants were asked to fill out questionnaires every day until they answered “Not sick” two days in a row to the question, “How sick do you feel today?” We attempted daily telephone contact in order to enhance adherence to protocol. Participants were met for an exit interview within a few days of the end of their cold.

As an evaluative instrument, WURSS was designed to measure patient-valued domains that change over time, and whose course might be modifiable by medical interventions. The term “minimal important difference” is used differently by different authors [16–18], but in general refers to the minimal amount of positive change that “patients perceive as beneficial, and which would mandate, in the absence of troubling side effects and excessive cost, a change in the patient’s management” [19]. Using methods developed by Guyatt, Jasechke, Juniper, and Redelmeier and colleagues [13–15,20], we asked participants to rate their perceived global change (improvement or deterioration). First, participants were asked to select “Better,” “The Same,” or “Worse” in response to “Compared to yesterday, I feel that my cold is....” Next, they were asked to rate the degree of change using the following scale: 1, almost the same, hardly any better [or worse] at all; 2, a little better [or worse]; 3, somewhat better [or worse]; 4, moderately better [or worse]; 5,

Table 1
Content of the Wisconsin Upper Respiratory Symptom Survey (WURSS-44)

Symptoms ^a	Symptoms	Symptoms	Functional impairments ^b
1. How sick do you feel today?	12. Body aches	23. Swollen glands	34. Think clearly
2. Cough	13. Feeling “run down”	24. Plugged ears	35. Speak clearly
3. Coughing stuff up	14. Sweats	25. Ear discomfort	36. Sleep well
4. Cough interfering with sleep	15. Chills	26. Watery eyes	37. Breathe easily
5. Sore throat	16. Feeling feverish	27. Eye discomfort	38. Walk, climb stairs, exercise
6. Scratchy throat	17. Feeling dizzy	28. Head congestion	39. Accomplish daily activities
7. Hoarseness	18. Feeling tired	29. Chest congestion	40. Work outside the home
8. Runny nose	19. Irritability	30. Chest tightness	41. Work inside the home
9. Plugged nose	20. Sinus pain	31. Heaviness in chest	42. Interact with others
10. Sneezing	21. Sinus pressure	32. Lack of energy	43. Live your personal life
11. Headache	22. Sinus drainage	33. Loss of appetite	44. Compared to yesterday, I feel...

The WURSS-44 and WURSS-21 are available for viewing and PDF download at <http://www.fammed.wisc.edu/wurss/>. Educational and nonprofit users may use WURSS without charge, but should notify us of any use. Pharmaceutical companies and other for-profit entities must obtain permission and negotiate a user fee through the Wisconsin Alumni Research Foundation.

Items selected for the WURSS-21 are highlighted in **bold italics**.

^a Directions for symptom-based items (2–33) ask respondents to: “Please rate the average severity of your cold symptoms over the last 24 hours by marking the appropriate circle for each of the following symptoms.”

^b Directions for functional impairment items (34–43) ask: “Over the last 24 hours, how much has your cold interfered with your ability to....”

a good deal better [or worse]; a great deal better [or worse]; and 7, a very great deal better [or worse].

Prospectively monitored day-to-day changes on the WURSS-44 corresponding to assessments of global change of either “a little better” or “somewhat better” were used to calculate the minimal important difference (MID). Dividing MID by the square root of twice the mean squared error of stable participants (those responding “the same” to the global change question) yields Guyatt’s index of responsiveness:

$$\text{Responsiveness Index} = \text{MID}/(2 \text{ MSE})^{1/2}$$

This in turn can be used directly as a hypothesized clinically significant effect size when doing sample size and power calculations [15,21].

For this study, we added a five-point importance scale to WURSS-44 items: “How important is this to you?,” with “Not,” “Somewhat,” and “Very” aligned with the numbers 1, 3, and 5. We asked participants to fill out the importance scale every day for every item. We told them, “Some people may rate one symptom as fairly severe, but not think it is very important, while other, milder symptoms may really bother them. When answering the question, “How important is this to you?” please think about how bothersome a symptom is, or how much you dislike having it.”

In addition to the WURSS-44, participants filled out the general health-related quality-of-life instrument known as the SF-36 (version 2; 4-week recall) at enrollment and exit. A 24-hour recall version of the SF-8 became available after our study had begun. We started using it in October 2002, less than a quarter of the way into our study. The SF-8 was filled out beginning the second day and continuing until the participant indicated “not sick” for two days in a row. The primary comparison instrument we used was the Jackson scale, described above, which was filled out at intake, and during every day of symptom monitoring, but not at the exit interview.

Our analytic strategy emphasized importance to patients, MID, and responsiveness [15], as described above. Reliability and dimensional cohesion were prospectively chosen as measures of internal validity. Comparisons of WURSS to the Jackson scale, the SF-36, and the SF-8 were used to assess convergent validity, using both standard measures of association and comparison of the instruments’ ability to detect change over time. In general, analysis was planned to begin with tabular and graphic portrayal of descriptive data, then progress to more complex methods such as factor analysis, regression-based modeling, and partial correlation analysis.

3. Results

The first participant for this study was enrolled on March 22, 2002; the last was exited on August 12, 2003. Of 167 people whose eligibility was documented, 157 were enrolled, and 150 were monitored through the duration of their colds,

for a total of 1,681 person-days. Of 570 documented callers, at least 160 couldn’t be screened, usually due to inability to achieve telephone contact within the enrollment time limit. Reasons for exclusion during screening included symptom duration greater than 48 hours (97), allergic symptoms (43), and insufficient Jackson score (29). Eleven prospective participants said it would take too much time, and two said we weren’t paying enough. For the seven lost to follow-up, six could not be found despite multiple attempts, and one was unable to return from another state, but sent her data booklet documenting 2 days of symptoms past enrollment. For the 151 participants monitored past enrollment, sociodemographic data are portrayed in Table 2. We obtained a fairly diverse sample in terms of age, gender, income, education, and smoking status; representation of minorities, however, was limited by the homogeneity of the study area.

We defined the end of the cold as the last time the participant scored their global illness severity above zero, followed

Table 2
Participant characteristics

Variable	Value
Response rate, no.	
Calls made	737
Eligible	167
Enrolled	157
Monitored >3 days	151
Monitored to end of cold	150
Age, years	
Mean (SD)	35.50 (14.74)
Range	18–80
Gender, no./total (%)	
Female	104/149 (69.7)
Male	45/149 (30.2)
Ethnicity, no./total (%)	
American Indian	5/151 (3.3)
Black	4/151 (2.6)
Hispanic	6/151 (3.9)
White	133/149 (88.0)
Asian	4/151 (2.6)
Income bracket, \$1000, no./total (%)	
<15/yr	42/151 (27.8)
15 to <25/yr	28/151 (18.5)
25 to <50/yr	25/151 (16.5)
50 to <75/yr	21/151 (13.9)
75 to <100/yr	24/151 (15.8)
>100/yr	9/151 (5.9)
No response	2/151 (1.3)
Education, highest, no./total (%)	
Some HS	1/151 (0.7)
HS diploma or GED	18/151 (11.9)
Some college	36/151 (23.8)
Associate degree	4/151 (2.6)
BA or BS	45/151 (29.8)
MA or BS	24/151 (15.8)
PhD or professional degree	7/151 (4.6)
Other or no response	16/151 (10.5)
Tobacco use, no./total (%)	
Current	22/151 (14.7)
Past	37/151 (26.2)
Nonsmoker	88/151 (59.0)
No response	4/151 (1.3)

by 2 days of global severity marked zero. If global severity was marked above zero on the day after a single day of zero, the cold was defined to be continuing; however, if the day marked above zero was the 13th or 14th day, this was defined as the end of the cold. Symptom duration prior to enrollment varied from 0 to 49 hours (mean = 30.9 hours). (One person was inadvertently enrolled 49 hours after first symptom began, and was monitored to end of cold. We chose to include this data in analysis.) Adding these hours to the total time from enrollment until last above-zero global severity, mean duration of colds was calculated to be 9.3 days (223 hours), with a standard deviation of 3.7 days (89 hours). A total of 24 (16%) of 151 participants continued to report being sick at the end of their 14-day monitoring period.

Results confirmed the diversity of common cold symptoms, and supported the breadth of the WURSS-44 instrument. In this sample, at least once during the first 7 days of monitoring, 99% of participants reported nasal symptoms, 91% reported sore or scratchy throat, and 94% reported cough. Sinus pain or pressure (80%), chest congestion (73%), headache (88%), body ache (84%), feverishness (69%), sweats (55%), and chills (57%) were reported less frequently. Generalized discomfort was measured by the terms “feeling run down,” “feeling tired,” “lack of energy,” and “loss of appetite,” which were scored above zero at least once by 97%, 99%, 97%, and 81% of participants, respectively. Interference with thinking (86%), speaking (80%), sleeping (94%), and exercise (84%) was common, as were difficulties with accomplishing daily activities (86%), work outside the home (76%), work inside the home (76%), interactions with other people (87%), and living one’s personal life (86%). These data support our contention that colds frequently include functional, quality-of-life domains that the Jackson scale does not address.

Using retrospectively assessed Jackson criteria at enrollment, 59 (39%) of our sample had a sore or scratchy throat as their first symptom, 61 (40%) had colds starting with a nasal symptom, and only 8 (5%) started with cough. Simply summing severity points from the 43 items scored on a seven-point scale (excluding global change) the mean total WURSS score was 91.3 was at enrollment, with a SD of 48.5, and an interquartile range of 54 to 125. For those with continuing colds, this mean total score increased slightly to 93.5 on day 2, then dropped to 87.5 on day 3, to 79.2 on day 4, and to 74.1 on day 5, to 63.4 on day 6, to 57.5 on day 7, and to 53.8 on day 8. This downward trend continued, with scores stabilizing in the 40s for days 10 through 14 of illness (see Fig. 1). Responses to the global severity item followed this trend, with initial mean scores rising from 3.84 on day 1 to 4.03 on day 2, then falling gradually to just under 2 points by day 9. Individual items had mean severities ranging from 2.17 (feeling dizzy) to 3.98 (feeling tired), averaging over the first 3 days for those with these symptoms present.

Although importance ratings were significantly associated with symptom severity and functional impairment, only a

small proportion of the variance was explained, with Pearson correlation coefficients ranging from 0.25 to 0.48. (Because participants reported some difficulty rating the importance of symptoms that they were not experiencing, importance ratings reported here only include data when item severities were scored above zero.) The mean importance of items ranged from 2.54 to 4.22, with generalized symptoms and functional impairments generally reported as more important than specific symptoms. Table 3 displays mean importance by item. To arrive at these values, we first averaged within-person over time, then averaged among participants with valid data, in order to equally represent all participants.

Starting without any a priori grouping of items, and seeking to discover and empirically verify illness dimensions, we performed a series of factor analyses, which led to the 10-dimensional structure displayed in Table 4. Analysis followed procedures suggested by Kroonenberg and Lewis [22], utilizing both exploratory and confirmatory common factor analysis using maximum likelihood estimation as a guide. Composite construct reliability was estimated using procedures originally suggested by Joreskog [23], with details provided by Bollen [24]. Of the 42 items used in the models, a total of 36 fit neatly within the 10 dimensions, with 2 to 10 items per dimension. Reliability coefficients ranged from 0.62 to 0.93 (mean reliability was 0.83), and were all significant at $P < 0.01$ using Wald testing [25,26]. Assessment of structural stability over time was evaluated by conducting two analyses. First, the 10-dimensional model pattern was constrained to be equal in a confirmatory factor analysis and assessed at four different time periods, two days from start of the study, three days, four days, and five days, the time frame when overall severity change was greatest. The 10-dimensional structure was supported by the lack of significant change in fit indices in this temporal analysis. Finally, the data from all five days were aggregated and a confirmatory factor analysis based on the 10-dimensional structure was derived.

External validity was assessed by comparison to the SF-36 (4-week recall, reported at intake and exit), the SF-8 (24-hour recall), and the Jackson scale, both reported daily. (The 24-hour recall version of the SF-8 was added after the study began, with 118 of 151 participants contributing SF-8 data.) Unadjusted pair-wise Pearson correlation coefficients of WURSS-44 with SF-36 were -0.42 at intake and -0.35 at exit ($P < 0.01$). (For the SF instruments, 100 corresponds to perfect health, with lower scores indicating worse health. For WURSS and Jackson, 0 would correspond to perfect health, with higher scores indicating worse health. Therefore, some coefficients are negative.) Comparing WURSS-44 with Jackson led to correlation coefficients ranging from 0.73 to 0.93 ($P < 0.001$). Comparing WURSS with SF-8 yielded correlation coefficients from -0.60 to -0.84 ($P < 0.001$). These associations were stronger than those between Jackson and SF-8, for which coefficients ranged from -0.55 to -0.78 ($P < 0.001$) (see Fig. 2). A partial correlation analysis [27] adjusting for participant age, gender, education, income,

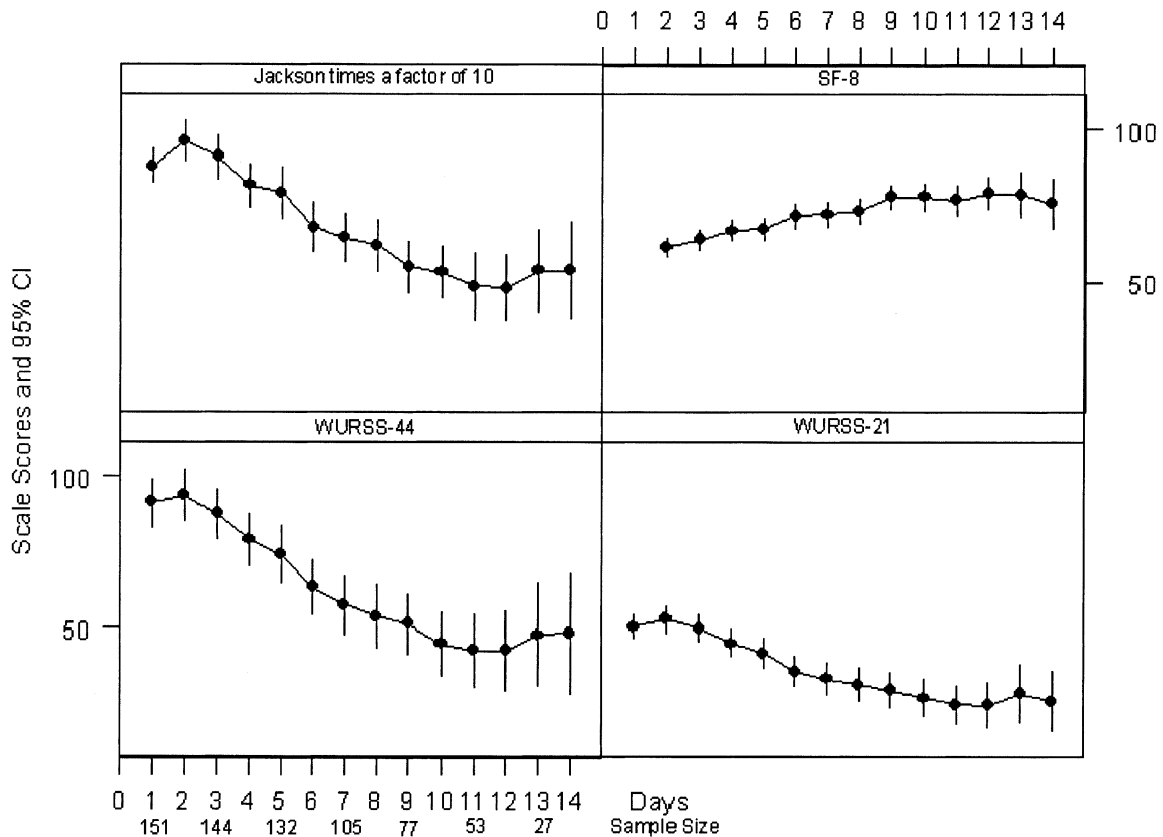


Fig. 1. Mean severity scores over time for the WURSS-44, WURSS-21, Jackson, and SF-8 instruments. Error bars indicate 95% confidence intervals. The WURSS-44, WURSS-21 and Jackson scores are simple sums of severity points (0 = perfect health). The SF-8 scores are calculated according to published methods [35] (100 = perfect health).

smoking status, and illness severity as covariates yielded very similar results, overall, and over time, without indication of need for adjustment. Thus, the data supported our expectations that the WURSS-44 would display stronger associations with the illness-specific Jackson criteria and with general health quality of life instruments, than either of these would with each other. Furthermore, these associations appeared to be linear, stable over time, and not significantly affected by possible confounders.

The range of the simply summed WURSS-44 (43 items summed) severity scores varies from 0 to 301 points theoretically (43 items times maximum of 7 points per item), and from 0 to 259 in our sample. Means of total daily scores were in the 90s in the first days of colds, when symptoms were worst, dropping rapidly after the second day of monitoring. As noted above, WURSS scores varied proportionately with Jackson, and inversely with the SF-8 and the SF-36. For example, a 10-point change on the WURSS-44 corresponded to changes of 3.3 points on the SF-8, and 0.89 points on the Jackson. Conversely, a 10-point change on the SF-8 corresponded to about 29 points on the WURSS, and a 1-point change on the Jackson scale corresponded to about 11 points on the WURSS-44.

Comparing retrospective assessments of global change to prospective changes reported on WURSS, a minimal important difference (MID) of 16.7 points per day was calculated for the sum total score. Dividing this estimate of small but important difference by the square root of twice the mean square error of stable patients (280.0) yields a Guyatt responsiveness index of 0.71, suggesting that a randomized trial with two arms would require 92 participants to have ~90% chance of detecting this level of daily change (assuming $\alpha = 0.05$ and $\beta = 0.10$; two-tailed testing). Using identical methods, MID and responsiveness were calculated for individual items, which were then used as criteria for item reduction. (See Table 3, and next paragraph.) Using a simply summed score, a MID of 1.56 was calculated for the Jackson scale, yielding a responsiveness index of 0.61 for the Jackson scale. This indicates that a two-armed trial would need 124 participants to detect this level of minimally important daily change. Similarly, an MID of 4.52 and a responsiveness index of 0.54 was calculated for the SF-8, indicating that an SF-8 powered RCT would need 156 participants to detect this level of daily change.

To create a shorter instrument with equal or better psychometric properties, we used responsiveness and importance

Table 3
Validity markers of WURSS-44 items

Item ^a	Domain	Frequency ^b	Severity ^c	Importance ^c	MID ^d	SE stable	Responsiveness
1	Global	100	3.92 ± 1.17	3.79 ± 0.90	0.721	0.516	0.709
2	Cough	94	3.04 ± 1.33	3.21 ± 1.10	0.389	0.839	0.300
3	Cough	80	2.77 ± 1.22	3.13 ± 1.06	0.308	1.283	0.193
4	Cough	77	2.75 ± 1.48	3.90 ± 1.10	0.353	1.020	0.247
5	Throat	90	3.19 ± 1.46	3.39 ± 0.95	0.407	1.072	0.278
6	Throat	91	3.10 ± 1.37	3.16 ± 0.95	0.404	0.863	0.307
7	Throat	79	2.71 ± 1.35	2.82 ± 1.17	0.396	1.034	0.276
8	Nasal	99	3.70 ± 1.46	3.41 ± 0.95	0.527	1.014	0.370
9	Nasal	96	3.55 ± 1.47	3.60 ± 0.94	0.509	1.140	0.337
10	Nasal	95	2.78 ± 1.27	2.72 ± 1.13	0.423	1.103	0.285
11	Sinus	87	3.08 ± 1.44	3.56 ± 1.04	0.450	1.527	0.258
12	Aches	83	3.03 ± 1.42	3.57 ± 0.93	0.376	1.523	0.216
13	Tired	97	3.83 ± 1.47	3.99 ± 0.90	0.611	1.318	0.376
14	Sweats	55	2.50 ± 1.26	2.86 ± 0.95	0.183	0.723	0.152
15	Sweats	57	2.72 ± 1.37	3.08 ± 1.06	0.263	0.987	0.188
16	Sweats	69	2.42 ± 1.22	3.05 ± 1.05	0.264	0.859	0.202
17	None	62	2.17 ± 1.26	3.26 ± 1.13	0.168	0.977	0.120
18	Tired	99	3.98 ± 1.50	3.90 ± 0.91	0.627	1.276	0.392
19	None	87	2.93 ± 1.49	3.31 ± 1.08	0.405	1.425	0.240
20	Sinus	68	2.97 ± 1.43	3.28 ± 1.09	0.289	1.246	0.183
21	Sinus	80	2.91 ± 1.43	3.17 ± 1.14	0.317	1.254	0.200
22	Sinus	80	3.22 ± 1.47	3.18 ± 1.07	0.298	1.037	0.207
23	Aches	64	2.62 ± 1.37	2.54 ± 1.11	0.232	0.653	0.203
24	Ears	66	2.65 ± 1.43	2.94 ± 1.13	0.172	0.761	0.139
25	Ears	65	2.42 ± 1.31	3.02 ± 1.15	0.232	0.893	0.174
26	None	72	2.38 ± 1.23	2.69 ± 1.10	0.216	0.641	0.191
27	None	66	2.32 ± 1.24	2.96 ± 1.17	0.212	1.025	0.148
28	None	94	3.35 ± 1.42	3.50 ± 0.92	0.523	1.623	0.290
29	Chest	73	2.52 ± 1.26	3.12 ± 1.00	0.279	0.767	0.226
30	Chest	65	2.43 ± 1.37	3.08 ± 1.09	0.230	0.788	0.183
31	Chest	62	2.59 ± 1.33	2.98 ± 1.22	0.224	0.789	0.179
32	Tired	97	3.94 ± 1.51	3.98 ± 0.87	0.614	1.808	0.323
33	None	80	3.03 ± 1.44	2.59 ± 1.25	0.341	1.004	0.241
34	Activity	85	3.09 ± 1.41	4.04 ± 0.97	0.431	1.327	0.265
35	Activity	79	2.58 ± 1.31	3.43 ± 1.10	0.296	1.054	0.204
36	Activity	94	3.54 ± 1.56	4.22 ± 0.91	0.521	1.512	0.300
37	Activity	92	3.37 ± 1.51	3.92 ± 1.04	0.446	1.307	0.276
38	Activity	83	3.02 ± 1.48	3.55 ± 0.97	0.512	0.757	0.416
39	Activity	85	3.02 ± 1.42	3.86 ± 0.99	0.506	0.813	0.397
40	Activity	76	3.04 ± 1.46	3.37 ± 0.98	0.398	1.196	0.257
41	Activity	76	3.03 ± 1.38	3.56 ± 1.00	0.437	0.724	0.363
42	Activity	87	2.78 ± 1.34	3.50 ± 1.07	0.485	1.015	0.340
43	Activity	86	3.09 ± 1.42	3.91 ± 0.99	0.509	1.004	0.359

To weight each person's responses equally, data were first averaged within-person-over-time, then averaged among participants

Items selected for the WURSS-21 are displayed in **bold italics**.

^a The 44th item on the WURSS-44 assesses global change (change since yesterday), and hence cannot be assessed in the same way as the 43 items portrayed here.

^b Frequency = Scored above zero at least once in first 7 days of monitoring.

^c Severity and Importance = Mean ± SD; averaged over first 3 days; only for those with symptom present

^d Minimal important difference MID: Mean day-to-day change for those indicating minimal improvement.

to patients as criteria for selecting the items highlighted in Tables 1 and 3. The WURSS-21 included 10 items assessing symptoms, 9 items assessing functional impairments, and 1 item each assessing global severity and global change. Assuming that the 21 items embedded within the WURSS-44 would have the same psychometric properties when arranged as an independent WURSS-21 instrument, we calculated the

MID to be 9.48, and the responsiveness index to be 0.80. This suggests that a two-armed trial of about 74 total participants would have ~90% power to detect this level of daily change. Because importance and responsiveness were used as selection criteria for the short-form items, it is possible that the WURSS-21 will not perform quite as well as a stand-alone instrument, due to selection bias, regression to

Table 4
Items, dimensions, and reliability of the WURSS-44

Item in dimension	Loading coefficient
Sore throat (reliability = 0.748) ^a	
Sore throat	0.927
Scratchy throat	0.704
Hoarseness	0.425
Nasal (reliability = 0.717)	
Runny nose	0.638
Plugged nose	0.738
Sneezing	0.645
Sinus (reliability = 0.872)	
Sinus pain	0.940
Sinus pressure	0.941
Sinus drainage	0.638
Headache	0.615
Ears (reliability = 0.916)	
Plugged ears	0.935
Ear discomfort	0.902
Sweats (Reliability = 0.799)	
Sweats	0.745
Chills	0.755
Feverish	0.767
Aches (reliability = 0.624)	
Body aches	0.706
Swollen glands	0.638
Cough (reliability = 0.828)	
Cough	0.836
Coughing stuff up	0.638
Cough interfering with sleep	0.866
Chest (reliability = 0.912)	
Chest congestion	0.812
Chest tightness	0.966
Heaviness	0.866
Tiredness (reliability = 0.937)	
Feeling “run down”	0.865
Feeling tired	0.919
Lack of energy	0.949
Activity and function (reliability = 0.934)	
Think clearly	0.735
Speak clearly	0.521
Sleep well	0.645
Breathe easily	0.699
Walk or climb stairs or exercise	0.803
Accomplish daily activities	0.919
Work outside the home	0.785
Work inside the home	0.834
Interact with others	0.822
Live your personal life	0.867

Factor analysis followed procedures outlined by Kroonenberg and Lewis [22]. Reliability and standardized loading coefficients estimated by methods of Joreskog [23] and Bollen [24]. Items measuring global severity and global change from the WURSS-44 were not included in the factor analyses. Of the 42 items included, the 36 shown here loaded into the 10 empirically assessed dimensions.

^a All reliability coefficients are significant at $P < 0.01$ using the Wald test [26].

the mean, or both. The WURSS-21 is now undergoing prospective validation as a stand-alone instrument. We advise potential users to power trials conservatively.

4. Discussion

Most studies evaluating remedies for the common cold have used variants of the Jackson scale for assessing outcome [8,9]. There has been no systematic attempt to assess reliability, responsiveness, or construct validity. We believe that the most glaring deficit in the Jackson method is the absence of functional or quality-of-life measures, which cold sufferers tend to value as important as or more important than specific symptoms. The development of the Wisconsin Upper Respiratory Symptom Survey [11], and its subsequent validation portrayed here, represents the first systematic attempt at creating an illness-specific quality-of-life instrument to measure the negative effects of the common cold.

The challenges in such an enterprise are many. The common cold is a syndrome characterized by abrupt onset, short but variable duration, and high diversity of symptomatic presentation. The rapid change over time precludes meaningful test–retest reliability assessments. The subjective nature precludes interpretable interobserver comparisons. The variability of symptomatic and functional presentation and course, as well as the lack of an adequate gold standard, further complicate validity testing. Currently available laboratory measures are inadequate. Upper respiratory infections may be caused by many different agents, including strains of adenovirus, coronavirus, enterovirus, influenza virus, parainfluenza virus, respiratory syncytial virus, and now metapneumovirus [28], as well as the prototype rhinovirus. Even the best laboratories still fail to identify etiological agents in anywhere from 25% to 75% of colds tested [29,30], however. Conversely, about 25% of those with documented infections fail to develop symptoms [31]. Although nasal and throat symptoms are present in most colds, their presence doesn't assure that it is a cold (it could be, for example, allergic rhinitis, or streptococcal pharyngitis), and their absence doesn't rule out the possibility (asymptomatic infection is possible). Other cold-related symptoms, such as cough, feverishness, chilliness, or general malaise are even less sensitive and specific. Severity, duration, and symptom presentation vary greatly among different populations, partially due to variability among the many pathogens, and partially due to differences in host response. The colds in our study, for example, are longer and more severe, with more cough, fever, and aches, than those of several previous studies [2–4,32,33]. Clearly, we are left with a syndrome that is truly an illness (a collection of experienced symptoms) rather than a disease (defined by verifiable biological criteria), using the model first put forth by Kleinman [34].

Given these realities, we chose to allow self-assessment to take front seat in both diagnosis (inclusion criteria) and assessment (severity ratings of items judged by cold-sufferers to be important). Nevertheless, we readily admit the limitations in this approach: Self-assessment is notoriously variable, and occasionally deceptive. Terms that cold-sufferers use may not correspond to scientific understanding (e.g., “sinus pain,” “chills”). The lack of a gold standard precludes formal

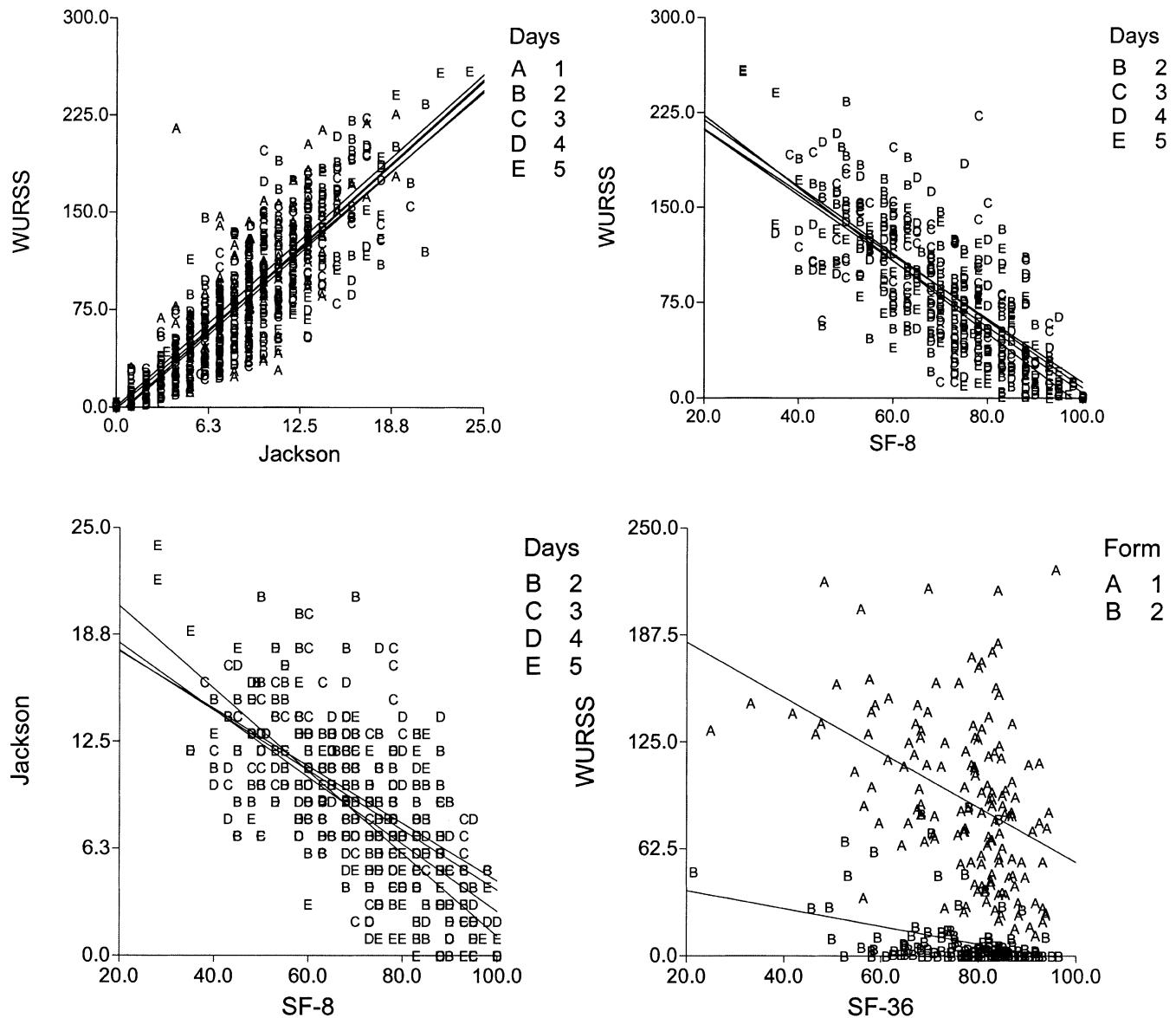


Fig. 2. Correlations among WURSS, Jackson, SF-8, and SF-36. The WURSS-44, Jackson, SF-8 and SF-36 scores are calculated as described in the text. WURSS-by-Jackson includes data from intake (day 1) through day 5. WURSS-by-SF8 and Jackson-by-SF8 include data from day 2 through day 5. WURSS-by-SF-36 includes data from intake (A, when sick) and exit (B, when recovered).

assessment of concurrent or predictive criterion validity, substituting instead the more problematic evaluations of convergent, and construct, validity [12]. Other limitations include the difficulty in recruiting a representative sample, variability in perception, understanding and response among different individuals, and the inherent difficulty in interpreting scales (indices) that are not solidly tied to universally understood reference standards. Additionally, there is considerable potential difficulty in the underlying assumption of equal distancing among the seven severity levels, in the assumptions of normal distributions, and in the assumption of equal item importance implied by our use of unity weighting. Finally, we should note that our sample size is only marginally adequate for many of the purposes at hand.

Despite these limitations, we feel that this attempt at creation and validation of WURSS has been fruitful. By face validity alone, WURSS is an important step forward. The items included use wordings supplied by cold-sufferers, and are reported to be present by most people at some time during their colds. Empirically conducted factor analysis yielded 10 dimensions that are internally coherent, and stable over time. The WURSS-44 correlates to Jackson (illness-specific) and to the SF-8 (general health) better than these two measures correlate to each other. These correlations appear to be stable over time, and are unaffected by covariates such as gender, age, and severity at presentation. Perhaps most important, WURSS-44 demonstrates greater responsiveness than Jackson, while at the same time expanding

both the measurement field and the response range. If the WURSS-21 performs as well by itself as the items embedded in the WURSS-44 appear to, it should be able to detect important change with even greater sensitivity, and with only minimal loss of content breadth. We expect WURSS to progress through further stages of development and are open to good ideas, conscientious criticism, or potentially, collaboration.

Acknowledgments

The authors would like to acknowledge the University of Wisconsin–Madison School of Medicine and Department of Family Medicine for providing startup funds, an institutional base, and collegial support. This work was partially supported by a Patient-Oriented Career Development Grant (K23 AT00051-01) from the National Center for Complementary and Alternative Medicine at the National Institutes of Health, and by the Clinical Research Feasibility Funds (CReFF) award from the NIH-funded University of Wisconsin–General Clinical Research Center (MO1 RR03186). A career development grant from the Robert Wood Johnson Foundation Generalist Physician Scholars Program supported the analysis phase of this project, and is allowing this work to go forward. Finally, we would like to thank Gordon Guyatt, MD, and Jack Gwaltney, MD, for inspiration, guidance, and constructive criticism.

References

- [1] Douglas RM. Respiratory tract infections as a public health challenge. *Clin Infect Dis* 1999;28:192–4.
- [2] Dingle JH, Badger GF, Jordan WS Jr. *Illness in the home: a study of 25,000 illnesses in a group of Cleveland families*. Cleveland, OH: The Press of Western Reserve University; 1964.
- [3] Gwaltney JM, Hendley JO, Simon G, Jordan WS Jr. Rhinovirus infections in an industrial population. I. The occurrence of illness. *N Engl J Med* 1966;275:1261–8.
- [4] Monto AS, Ullman BM. Acute respiratory illness in an American community: the Tecumseh study. *JAMA* 1974;227:164–9.
- [5] Fendrick AM, Monto AS, Nightengale B, Sarnes M. The economic burden of non-influenza-related viral respiratory tract infection in the United States. *Arch Intern Med* 2003;163:487–94.
- [6] Smith MBH, Feldman W. Over-the-counter cold medications: a critical review of clinical trials between 1950 and 1991. *JAMA* 1993;269:2258–63.
- [7] Turner RB. The treatment of rhinovirus infections: progress and potential [Review]. *Antiviral Res* 2001;49:1–14.
- [8] Jackson GG, Dowling HF, Anderson TO, Riff L, Saporta J, Turck M. Susceptibility and immunity to common upper respiratory viral infections: the common cold. *Ann Intern Med* 1960;55:719–38.
- [9] Jackson GG, Dowling HF, Muldoon RL. Acute respiratory diseases of viral etiology. VII. Present concepts of the common cold. *Am J Public Health* 1962;52:940–5.
- [10] Gwaltney JM Jr. Rhinovirus colds: epidemiology, clinical characteristics and transmission. *Eur J Respir Dis Suppl* 1983;128:336–9.
- [11] Barrett B, Locken K, Maberry R, Schwamman J, Brown R, Bobula J, Stauffacher EA. The Wisconsin Upper Respiratory Symptom Survey (WURSS): a new research instrument for assessing the common cold. *J Fam Pract* 2002;51:265. Available at: <http://www.fammed.wisc.edu/wurss/>.
- [12] McDowell I, Newell C. *Measuring health: a guide to rating scales and questionnaires*. New York: Oxford University Press; 1996.
- [13] Juniper EF, Guyatt GH, Willan A, Griffith LE. Determining a minimal important change in a disease-specific Quality of Life Questionnaire. *J Clin Epidemiol* 1994;47:81–7.
- [14] Guyatt GH, Bombardier C, Tugwell PX. Measuring disease-specific quality of life in clinical trials. *CMAJ* 1986;134:889–94.
- [15] Guyatt G, Walter S, Norman G. Measuring change over time: Assessing the usefulness of evaluative instruments. *J Chron Dis* 1987;40:171–8.
- [16] Samsa G. How should the minimum important difference for a health-related quality-of-life instrument be estimated? *Med Care* 2001;39:1037–8.
- [17] Wells G, Anderson J, Beaton D, Bellamy N, Boers M, Bombardier C, Breedveld F, Carr A, Cranney A, Dougados M, Felson D, Kirwan J, Schiff M, Shea B, Simon L, Smolen J, Strand V, Tugwell P, van Riel P, Welch VA. Minimal clinically important difference module: summary, recommendations, and research agenda. *J Rheumatol* 2001;28:452–4.
- [18] Wright JG. The minimal important difference: who's to say what is important? *J Clin Epidemiol* 1996;49:1221–2.
- [19] Jaeschke R, Singer J, Guyatt GH. Measurement of health status: ascertaining the minimal clinically important difference. *Control Clin Trials* 1989;10:407–15.
- [20] Redelmeier DA, Guyatt GH, Goldstein RS. Assessing the minimal important difference in symptoms: a comparison of two techniques. *J Clin Epidemiol* 1996;49:1215–9.
- [21] Guyatt GH, Osoba D, Wu AW, Wyrwich KW, Norman GR; Clinical Significance Consensus Meeting Group. Methods to explain the clinical significance of health status measures. *Mayo Clin Proc* 2002;77:371–83.
- [22] Kroonenberg PM, Lewis C. Methodological issues in the search for a factor model: exploration through confirmation. *J Educ Stat* 1982;7:69–89.
- [23] Joreskog KA. Statistical analysis of sets of congeneric tests. *Psychometrika* 1971;36:109–33.
- [24] Bollen KA. *Structural equations with latent variables*. New York: Wiley, 1989.
- [25] Agresti A. *Categorical data analysis*. 1st ed. New York: Wiley, 1990.
- [26] Altman DG. *Practical statistics for medical research*. London: Chapman & Hall; 1991.
- [27] Thorndike RM. *Correlational procedures for research*. New York: Gardner Press, 1978.
- [28] Williams JV, Harris PA, Tollefson SJ, Halburnt-Rush LL, Pingsterhaus JM, Edwards KM, Wright PF, Crowe JE Jr. Human metapneumovirus and lower respiratory tract disease in otherwise healthy infants and children. *N Engl J Med* 2004;350:443–50.
- [29] Arruda E, Pitkäranta A, Witek TJ, Doyle CA, Hayden FG. Frequency and history of rhinovirus infections in adults during autumn. *J Clin Microbiol* 1997;35:2864–8.
- [30] Gwaltney JM Jr. Virology and immunology of the common cold. *Rhinology* 1985;23:265–71.
- [31] Gwaltney JM Jr. Rhinoviruses. In: Evans AS, Kaslow RA, editors. *Viral infections of humans: epidemiology and control*. 4th ed. New York: Plenum Medical Book Company, 1997;815–38.
- [32] Turner RB. Epidemiology, pathogenesis, and treatment of the common cold. *Ann Allergy Asthma Immunol* 1997;78:531–9.
- [33] Gwaltney JM Jr, Buier RM, Rogers JL. The influence of signal variation, bias, noise and effect size on statistical significance in treatment studies of the common cold. *Antiviral Res* 1996;29:287–95.
- [34] Kleinman A. Medicine's symbolic reality: on a central problem in the philosophy of medicine. *Inquiry* 1973;16:206–13.
- [35] Ware JE, Kosinski M, Dewey JE, Gandek B. How to score and interpret single-item health status measures: a manual for users of the SF-8 health survey. Lincoln, RI: QualityMetric, 2001.