

# Comparison of anchor-based and distributional approaches in estimating important difference in common cold

Bruce Barrett · Roger Brown · Marlon Mundt

Accepted: 29 October 2007 / Published online: 20 November 2007  
© Springer Science+Business Media B.V. 2007

## Abstract

**Context** Evaluative health-related quality-of-life instruments used in clinical trials should be able to detect small but important changes in health status. Several approaches to minimal important difference (MID) and responsiveness have been developed.

**Objectives** To compare anchor-based and distributional approaches to important difference and responsiveness for the Wisconsin Upper Respiratory Symptom Survey (WURSS), an illness-specific quality of life outcomes instrument.

**Design** Participants with community-acquired colds self-reported daily using the WURSS-44. Distribution-based methods calculated standardized effect size (ES) and standard error of measurement (SEM). Anchor-based methods compared daily interval changes to global ratings of change, using: (1) standard MID methods based on correspondence to ratings of “a little better” or “somewhat better,” and (2) two-level multivariate regression models.

**Participants** About 150 adults were monitored throughout their colds (1,681 sick days.): 88% were white, 69% were women, and 50% had completed college. The mean age was 35.5 years ( $SD = 14.7$ ).

**Results** WURSS scores increased 2.2 points from the first to second day, and then dropped by an average of 8.2 points per day from days 2 to 7. The SEM averaged 9.1 during these 7 days. Standard methods yielded a between day MID of 22 points. Regression models of MID projected 11.3-point daily changes. Dividing these estimates

of small-but-important-difference by pooled SDs yielded coefficients of .425 for standard MID, .218 for regression model, .177 for SEM, and .157 for ES. These imply per-group sample sizes of 870 using ES, 616 for SEM, 302 for regression model, and 89 for standard MID, assuming  $\alpha = .05$ ,  $\beta = .20$  (80% power), and two-tailed testing.

**Conclusions** Distribution and anchor-based approaches provide somewhat different estimates of small but important difference, which in turn can have substantial impact on trial design.

**Keywords** Clinical significance · Common cold · Evidence-based medicine · Health status · Minimal important difference · Psychometrics · Quality of life · Questionnaires · Respiratory tract infections · Severity of illness index · Symptom measurement · Treatment outcome · Upper respiratory infection

## Introduction

Randomized controlled trials (RCTs) inform and influence the practice of medicine. RCTs and systematic reviews are increasing in number, quality, and impact. Nevertheless, applicability to decision-making at the individual patient level remains limited. One of the greatest challenges is the assessment of clinical significance. While some outcomes are intuitively meaningful, such as death or hospitalization, most benefits come in a range of magnitudes, from trivial to truly important. For instance, reduction in pain, increased exercise tolerance, or improved mood or outlook on life can be so small as to be barely detectable, or so large as to constitute a major change in health status. Even significant events (e.g., heart attack or stroke) come in a range of magnitudes of severity and importance.

B. Barrett (✉) · R. Brown · M. Mundt  
Department of Family Medicine, University of Wisconsin  
Medical School, 777 South Mills, Madison, WI 53715, USA  
e-mail: bruce.barrett@fammed.wisc.edu

Health-related quality-of-life questionnaire instruments (HRQoL) are designed to measure symptomatic and functional outcomes for both acute and chronic illness [1]. HRQoLs can be classified as predictive, discriminative, or evaluative, according to their purpose and use [2]. Predictive instruments are designed to predict subsequent events. Discriminative instruments help distinguish, classify, or diagnose. Evaluative instruments are “used to measure the magnitude of longitudinal change in an individual or group on the dimension of interest” [2]. Representation of domains important to patients and clinicians (content validity), reliability (proportion of measurement not due to chance or error), and responsiveness (sensitivity to change over time) remain the key parameters used to assess HRQoL instruments.

Recognizing that very small, imperceptible, and/or clinically irrelevant changes can be demonstrated by sensitive instruments in large trials, and that small but clinically important changes can be missed by insensitive instruments and/or small studies, Jaeschke, Singer, and Guyatt introduced the concept of “minimal clinically important difference” in 1989, defining it as “the smallest difference in score in the domain of interest which 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” [3]. Since then, dozens of studies have sought to assess “minimal important difference” (MID) for a variety of conditions, using a variety of methods [4–9]. The MID concept has special appeal, as it corresponds to the magnitude of benefit for which a RCT should be powered in order to minimize risks of false positive and false negative trials. Additionally, MID may serve as a benchmark of clinical significance when interpreting trial results, creating health policy, or making treatment decisions.

Two general approaches are used to estimate small-but-important-difference: anchor-based and distribution-based [10–17], also described as external and internal [18]. The most widely used anchor-based approach was developed by Guyatt et al. [19], Jaeschke et al. [3], and Juniper et al. [20, 21] and compares interval changes in HRQoL scores to global rating of change (GRC) scores. The magnitude of interval change corresponding to a self-assessed GRC of “a little better” or “somewhat better” is interpreted as the standard MID. Another patient-centered method pioneered by Wells et al. [22] and Redelmeier et al. [5] allows patients diagnosed with the same condition to meet and discuss symptoms and functional impairments, then compare their health status with their conversational partner(s). Alternatively, physician assessments [9, 23] or other external measures [24–26] can be used as anchors. In order to control for possible confounders, and to take account of within-person-over-time

dependencies, formal regression-based statistical methods have been proposed [11], but are not widely implemented.

Distribution-based assessments take a variety of formats, but are all based on either within-group change-over-time or between-group comparison [27–29]. Comparisons may include: (1) before versus after treatment, (2) before vs. after natural change, or (3) between treatment and control groups at one or more points in time. Observed absolute differences are often standardized by dividing by standard deviation, yielding the standardized effect size (ES). Various ES ranges have been attributed varying levels of clinical significance. In an influential 1969 text, Cohen designated an ES of up to 0.2 as “small,” 0.5 as “medium,” and 0.8 or more as “large” [30]. This is problematic in that population variability, conceptually independent from clinical significance, strongly influences ES estimation. Nevertheless, the ES method has appeal, as it is time-tested, widely understood and central to many psychometric indices [31].

The standard error of measurement (SEM) has been proposed as a measure for powering and interpreting RCTs. The SEM is a theoretically fixed psychometric property that expresses magnitude in the same units as the original measure. SEM is defined as  $\sigma_x(1-r_{xx})^{1/2}$ , and thus incorporates both estimates of both reliability ( $r_{xx}$ ) and variability ( $\sigma_x$ ). Wyrwich and coauthors investigated the relationship between SEM and MID, reporting substantial consistencies between these theoretically distinct entities [29, 32, 33]. Across a reasonably wide range of chronic conditions, magnitudes of SEM and MID appeared similar, hence the recommendation that one SEM could be used as a benchmark when investigators want a distributional approach consistent with patient-centered anchor-based methods. In a similar vein, Norman, Sloan and Wyrwich reviewed 38 studies suitable for comparing MID to standardized ES, found “remarkable uniformity,” and concluded that “the threshold of discrimination for changes in health-related quality of life for chronic diseases appears to be approximately half a SD.” They found only two suitable studies of acute conditions, where “meaningful change” corresponded to .80 and 1.38 SD, a “somewhat larger magnitude of change than those in the other [chronic] studies considered.” This finding agrees with our impression that very little data is available regarding assessment of important difference and responsiveness for acute conditions.

The purpose of the current paper is to compare distributional and anchor-based approaches to assessment of important difference and responsiveness in acute upper respiratory infection (common cold). We do this using a data set generated by people with colds who self-reported symptoms on a validated questionnaire instrument once daily throughout the length of their illness.

## Methods

### Instrument

The Wisconsin Upper Respiratory Symptom Survey (WURSS) was developed as an evaluative illness-specific quality-of-life instrument [34]. A predecessor questionnaire (15 items, 9-point response range) was used in a randomized trial testing echinacea as treatment for common cold [35]. After that trial, using mixed qualitative and quantitative methods, the questionnaire was developed into the 44-item WURSS-44 instrument used in the present study. Development methods included open-ended elicitation of symptoms and dysfunctions from people with colds, followed by iterative assessment of content validity and ease-of-use assessed with in-person interview and focus group methods [34].

The WURSS-44 includes 1 global health item, 32 items rating specific symptoms, 10 functional quality-of-life items, and 1 GRC question (Table 1). Using responsiveness and importance-to-patient as guides, we selected best items for a shorter version, the WURSS-21, which is now undergoing prospective validation. Items of the WURSS-21 are a subset of WURSS-44 items, and include 1 item rating global health, 10 items rating symptoms, 9 functional quality-of-life items, and 1 item rating global change. All use similar 7-point Likert-type scales. Both instruments are available free-of-charge for educational, public health, and nonprofit purposes: <http://www.fammed.wisc.edu/wurss/>. The construct validity of the WURSS-44 is supported by assessments of face validity, importance-to-patients, responsiveness, and dimensional cohesion [36]. External (convergent; concurrent) validity was assessed through comparisons to the Jackson scale [37], the SF-36 [38], and the SF-8 [39] (24 h recall version), and laboratory-assessed biomarkers including viral titer, nasal neutrophils, mucus weight, and interleukin-8 [40]. Development [34], and validation [36, 40] details are more fully described elsewhere.

### Participants: enrollment and monitoring

To be eligible for this study, adults had to answer “Yes” to either “Do you think that you have a cold?” or “Do you think that you are coming down with a cold?” Participants were required to have a Jackson score [37] of three or more, calculated by summing severity scores 1 = mild, 2 = moderate, 3 = severe for eight symptoms: sneezing, runny nose, nasal congestion, sore throat, cough, headache, malaise, and chilliness. At least one of the first 4 (cold specific) symptoms had to be present. None could be present for more than 48 h. Allergy was excluded by

asking about itchy eyes, sneezing, and previous allergy diagnosis. Callers were screened by phone, then met in person for informed consent and enrollment, following a protocol approved by the University of Wisconsin Institutional Review Board. Participants filled out the WURSS-44 at enrollment, then once each day until they indicated that they were “Not sick” for 2 days in a row. Telephone contact was attempted each day in order to enhance adherence to protocol. Participants were met for an exit interview within a few days after their colds had ended.

### Anchor-based methods

The anchor-based methods used here compare retrospective self-assessment of improvement or worsening to prospectively assessed WURSS interval changes. All comparisons are on consecutive days, approximately 24 h apart. The GRC item starts with the question, “Compared to yesterday, I feel that my cold is...,” with “Better,” “The Same,” and “Worse” as initial response options. Those who feel that their colds have improved (“better”) or deteriorated (“worse”) are then asked to rate the degree of change using the following format:

If Better	If Worse
1. Almost the same, hardly any better at all	1. Almost the same, hardly any worse at all
2. A little better	2. A little worse
3. Somewhat better	3. Somewhat worse
4. Moderately better	4. Moderately worse
5. A good deal better	5. A good deal worse
6. A great deal better	6. A great deal worse
7. A very great deal better	7. A very great deal worse

The originators of this standard MID method suggest that response options starting with “Almost the same...” be excluded from analysis, and that options “a little better” and “somewhat better” be lumped together and interpreted to correspond to “MID” [3, 19–21]. Options starting with “moderately” and “a good deal” are interpreted as “moderate important difference,” and options starting with “a great deal” and “a very great deal” are interpreted as “large important difference.”

In order to assess the relationship of GRC to WURSS scores across time, across individuals, and across the entire spectrum of the GRC scale, we selected a multi-level mixed effect multivariate regression approach. Drawing on works of Brant et al. [11] and Yang and Goldstein [41] we developed multi-level multivariate linear regression models, providing variance-covariance matrices at the

**Table 1** Content of the Wisconsin Upper Respiratory Symptom Survey

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

The WURSS-44 includes all items. Items selected for WURSS-21 are displayed in italics

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

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

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

fixed-occasion level (daily increments), and at patient-levels, with initial intercepts modeled as random variables. Daily assessments are treated as repetitions at level 1 (indicated by  $t$ ) nested under patients (indicated by  $i$ ). Let  $Z_t$  be a vector of indicator variables for  $t = 1, 2, 3, 4, 5, 6,$  and  $7$  for daily change from day 2–3, 3–4, 4–5, 5–6, 6–7, 7–8, and 8–9, respectively, and  $y$  = the change in the WURSS score for each patient at each time period. The general model for our data may be written as:

$$y_{ii} = \sum_{t=1}^7 \beta_{0,t} Z_{it} + \sum_{t=1}^7 \sum_{h=1}^H \beta_{h,t} Z_{it} x_{h,ii} + \sum_{t=1}^7 \mu_t Z_{it} + e_{ii} \\ \times y_{ii} \approx N(\mathbf{XB}, \Omega),$$

where  $x_{h,ii}$  are covariates, with  $\mu_t \sim N(0, \Omega)$ , and  $e_{ii} \sim N(0, \Omega)$ , where  $\Omega$  are the variance-covariance matrices for the various levels. The model projects the amount of day-to-day WURSS interval change that corresponds to a single point difference on the GRC scale. We chose a single point GRC difference as reference point for three reasons: (1) there is a long line of research suggesting that people can reliably discriminate at this level [1, 42–44], (2) several MID studies suggest that a difference of .5–1.0 points on a 7-point scale is significant [3, 5, 21], and (3) simplicity/interpretability.

#### Distribution-based methods

Calculation of means divided by pooled standard deviations is a straightforward, time-honored, and easily

understandable process for most researchers and many clinicians. For this analysis, we base our ES (standardized ES) calculations on interval changes corresponding to the passage of approximately 24 h of time. We do this for three reasons: (1) there are no proven effective treatments for common cold, hence no intervention-based standards are available for comparison; (2) one day (about 24 h) is the interval duration between WURSS assessments used for the GRC anchor-based methods described above; and (3) qualitative evidence from our experience with hundreds of cold-sufferers suggests that the natural average daily improvement in a cold is a small but significant change.

Although definition of the SEM as  $\sigma_x (1 - r_{xx})^{1/2}$  is relatively well-accepted, several distinct methods of estimating reliability ( $r_{xx}$ ) and variability ( $\sigma_x$ ) are available. For reasons of ease-of-interpretation and consistency with published literature, we selected Cronbach’s  $\alpha$  as the measure of reliability and standard deviation as the measure of variability for use when calculating SEM. Following Wyrwich et al. [29, 32, 33], we used a single unit change in the SEM for comparative estimation of clinically important difference. For all methods, calculation of sample size assumed two-tailed comparison and tolerance for Type I error rate of  $\alpha = .05$ , with tolerance for Type II error rate set at  $\beta = .20$  (80% power).

#### Results

Recruitment of study participants ran from March 2002 to August 2003, during which time 737 callers responded to

community advertising in Madison, Wisconsin. Of these, 167 made it through enough telephone screening to be deemed eligible for enrollment, and 157 of these were met in-person for informed consent and enrollment. Participants then filled out questionnaires once each day of illness, to a maximum of 14 days, then met study personnel for an exit interview. About 150 participants were followed throughout their colds, documenting 1,681 days of symptoms. Excluding days when they said they were no longer sick, and including the estimated length of time from first symptom until enrollment, the mean duration of illness was 9.1 days ( $SD = 3.7$ ). Of the 7 participants lost to follow-up, six could not be contacted after the initial interview despite multiple attempts. The 7th was unable to return from travel to another state, but did send back paperwork documenting the first three days of her illness. Characteristics of the 150 participants completing protocol are shown in Table 2.

The mean total WURSS-44 score was 91.3 at enrollment, with a  $SD$  of 48.5, and an inter-quartile range of 54–125. The mean total score increased slightly to 93.5 on day 2, then dropped by an average of 8.2 points per day over the next 6 days (range = 6.5–10.0 point decrease per day). The downward trend continued, albeit more slowly, after day 7, with scores stabilizing in the 40 s for days 10–14 (for participants whose colds lasted that long). Dividing interval changes from days 2 to 7 by two-day pooled  $SD$ s yielded  $ES$ s of .13–.19 (mean = .16), small by Cohen's arbitrary standard [45], but perhaps important if one accepts the argument that one day's average natural improvement in common cold illness is significant. See Table 3.

We analyzed a total of 704 consecutive between-day comparisons in which participants rated themselves as "better." On 335 between-day comparisons when participants said they were "a little" or "somewhat" better, WURSS-44 scores decreased by an average of 16.7 points (95%  $CI = 14.5, 18.9$ ). This represents the standard  $MID$ , using the  $GRC$ -based method described above. For the 208 between-day comparisons in which participants rated themselves as "moderately" or "a good deal" better ("moderate important difference"), WURSS-44 scores improved by an average of 23.1 points (95%  $CI = 20.1, 26.2$ ). For the 161 between-day comparisons rated as "a great deal" or "a very great deal" better ("large important difference"), WURSS-44 scores improved by an average of only 15.6 points. Possible reasons for this discrepant finding are provided in the Discussion section.

As colds tend to improve over time, there were fewer between-day comparisons in which participants rated themselves as "worse" ( $N = 246$ ). On 139 between-day comparisons when participants said they were "a little" or "somewhat" worse, WURSS-44 scores increased by an average of 11.1 points (95%  $CI = 7.5, 14.6$ ). For the 83

**Table 2** Participant characteristics

Response rate	Enrolled		Monitored to end of cold		Income Bracket (\$)		No./Total (%)	Age (Years)	Gender		No./Total (%)	
	Number of calls:	737	167	Monitored >3 days	157	151			150	Range		Mean
No. eligible	167	157	151	150	157	151	150	18–80 years	35.50 ( $SD$ 14.74)	Women	Men	104/149 (69.7)
Ethnicity	No./Total (%)	No./Total (%)	No./Total (%)	No./Total (%)	No./Total (%)	No./Total (%)	No./Total (%)	No./Total (%)	No./Total (%)	Tobacco Use		No./Total (%)
Am. Indian	5/151 (3.3)	42/151 (27.8)	28/151 (18.5)	24/151 (15.8)	42/151 (27.8)	28/151 (18.5)	21/151 (13.9)	1/151 (0.7)	18/151 (11.9)	Current	Past	22/151 (14.7)
Black	4/151 (2.6)	15 to <25 K/year	25 to <50 K/year	50 to <75 K/year	75 to <100 K/year	>100 K/year	No response	18/151 (11.9)	36/151 (23.8)	37/151 (26.2)	Non-smoker	37/151 (26.2)
Hispanic	6/151 (3.9)	15 to <25 K/year	25 to <50 K/year	50 to <75 K/year	75 to <100 K/year	>100 K/year	No response	36/151 (23.8)	4/151 (2.6)	88/151 (59.0)	No response	88/151 (59.0)
White	133/149 (88.0)	50 to <75 K/year	75 to <100 K/year	>100 K/year	No response			4/151 (2.6)	45/151 (29.8)	4/151 (1.3)		4/151 (1.3)
Asian	4/151 (2.6)	No response						45/151 (29.8)	24/151 (15.8)			
								24/151 (15.8)	7/151 (4.6)			
								7/151 (4.6)	16/151 (10.5)			



**Table 3** Between-day responsiveness using distributional and anchor-based methods

Days	2–3	3–4	4–5	5–6	6–7	7–8	8–9	Days 2–9	Days 2–7
<b>WURSS-44</b>									
Raw change	<b>6.46</b>	<b>9.98</b>	<b>6.91</b>	<b>9.92</b>	<b>7.51</b>	<b>5.13</b>	<b>2.89</b>	<b>6.97</b>	<b>8.16</b>
SD <sub>pool</sub>	51.1	51.3	53.6	52.9	49.6	48.6	46.0	50.4	51.7
ES	0.13	0.19	0.13	0.19	0.15	0.11	0.062	0.136	0.157
Standard MID	<b>25.3</b>	<b>25.5</b>	<b>20.7</b>	<b>22.8</b>	<b>16.8</b>	<b>16.5</b>	<b>8.5</b>	<b>18.9</b>	<b>22.0</b>
MID coefficient	0.49	0.50	0.39	0.43	0.34	0.34	0.18	0.374	0.425
GRC estimate	<b>12.44</b>	<b>13.48</b>	<b>9.99</b>	<b>12.19</b>	<b>8.41</b>	<b>6.53</b>	<b>5.44</b>	<b>9.78</b>	<b>11.30</b>
GRC coefficient	0.24	0.26	0.19	0.23	0.17	0.13	0.12	0.194	0.218
Intercept	6.68	8.12	5.07	8.93	4.75	3.74	2.04		
Slope	5.76	5.36	4.92	3.27	3.66	2.79	3.41		
Variance	440.5	489.6	513.1	524.1	278.7	464.1	359.5		
Sample size	147	146	146	138	131	113	101		
MID Sample size	24	31	34	39	31	30	40		
<b>WURSS-21</b>									
Raw change	<b>3.40</b>	<b>5.35</b>	<b>4.48</b>	<b>5.57</b>	<b>3.83</b>	<b>2.21</b>	<b>1.95</b>	<b>3.93</b>	<b>4.53</b>
SD <sub>pool</sub>	27.2	27.7	28.3	28.0	27.2	27.2	26.4	27.4	27.7
ES	0.13	0.19	0.16	0.20	0.140	0.081	0.073	0.138	0.162
Standard MID	<b>16.5</b>	<b>13.7</b>	<b>11.4</b>	<b>13.3</b>	<b>9.5</b>	<b>8.3</b>	<b>5.1</b>	<b>10.8</b>	<b>12.7</b>
MID coefficient	0.61	0.49	0.40	0.48	0.35	0.31	0.19	0.394	0.458
GRC estimate	<b>6.47</b>	<b>7.27</b>	<b>6.36</b>	<b>5.83</b>	<b>3.44</b>	<b>3.49</b>	<b>3.14</b>	<b>5.14</b>	<b>5.87</b>
GRC coefficient	0.24	0.26	0.23	0.21	0.13	0.13	0.12	0.187	0.212
Intercept	3.00	4.61	3.99	3.76	0.97	1.55	1.07		
Slope	3.48	2.67	2.37	2.07	2.47	1.94	2.08		
Variance	145.0	162.7	137.9	111.5	94.3	126.4	106.9		

Raw change is the mean difference in simply summed total WURSS scores

Effect sizes are standardized by dividing raw change by the pooled SD

Standard MID corresponds to the mean interval change for all participants with GRC ratings of 2 or 3 (i.e., reporting that they are “a little better” or “somewhat better”)

MID coefficient divides standard MID by pooled SD

GRC estimates derive from hierarchical multivariate models relating WURSS score changes to single point differences in the 15-point GRC scale

GRC coefficient divides GRC estimate by pooled SD

Sample size refers to the number of participants used for ES and GRC model calculations (this is the largest number for any of the calculations; differential missing data means that many of the parameter estimates are based on smaller numbers)

MID sample size is smaller because it only includes those with GRC ratings of 2 or 3

Bolded values are main comparators

between-day comparisons in which participants rated themselves as “moderately” or “a good deal” worse, WURSS-44 scores deteriorated by an average of 20.5 points (0.48 points per item; 95% CI = 14.5, 26.6) For the 24 between-day comparisons rated as “a great deal” or “a very great deal” worse, WURSS-44 scores increased by an average of 33.6 points (0.78 points per item; 95% CI = 16.9, 50.2).

Standard MIDs were calculated for all between day comparisons from day 2 to day 9 (Table 3). The interval from day 1 to day 2 was excluded because participants’ symptoms tended to worsen slightly during the first 24 h after enrollment. We excluded data from after day 9

because of small sample size, reduced rate of symptom severity decline, and concern that recovery might be introducing recall bias into GRC ratings. Day-to-day standard MIDs from day 2 to 7 ranged from 16.5 to 25.5 (mean 22.0). The standard MID for the day 8 to day 9 interval was 8.5, somewhat lower than previous intervals. Given these factors, our judgment is that the time range from day 2 to day 7 was best for estimating MID and responsiveness for this data set. Dividing the mean standard MID of 22.0 for days 2 to 7 by the pooled SD of 51.7 for this period provides a standardized coefficient of .425.

Two-level regression models described above were used to project the magnitude of interval change on the WURSS

instrument corresponding to single point differences on the GRC scale. A series of models sequentially adding and removing variables demonstrated that the covariates age, gender, education, ethnicity, smoking status and Jackson score did not significantly effect the projections. Hence, these variables were not included in the final models yielding the indicators shown in Table 3. These models showed that single point differences on the GRC scale corresponded to interval changes of 8.4–12.4 points (mean = 11.3) on the WURSS-44 over days 2–7.

The model-based estimate of 11.3 points yields a coefficient = .218, implying that some 302 participants would be needed in each arm of a trial (assuming two-tailed testing,  $\alpha = .05$  and  $\beta = .20$ .) For the standard MID estimate of 22.0 points (coefficient = .426), only 89 people would be needed for each group, using the same assumptions.

The GRC item is worth focused attention, as it is the anchor for both standard and regression-based MID methods. We present its over-time distribution in Fig. 1. It is readily apparent that at enrollment on day 1, the majority of people felt that they had worsened, with a substantial number describing themselves as “the same” (Retrospective assessment at enrollment indicated that the mean time from first symptom to enrollment was 30.9 h.). On day 2, slightly fewer people, but still a majority, indicated worsening, with a

few more in “the same” and “better” groups. By day 3 and subsequently, a gradually increasing number of participants rated themselves as improved, with the sample size gradually declining as declarations of “not sick” exited people from the study. These results are consistent with previous research and theory [46], and considered along with expected associations of GRC data with WURSS scores, serve to support validity of the GRC item.

Distribution-based ES methods provide the smallest estimates of small-but-significant change. Effect sizes for between day comparisons for days 2–9 are shown in Table 3, and range from 0.06 to 0.19 for the WURSS-44, and from 0.07 to 0.20 for the WURSS-21. Averaging ES over days 2–7 provides coefficients of 0.157 for the WURSS-44, implying that a two-armed trial would require 870 participants per group, again assuming two-tailed testing, with  $\alpha = .05$  and  $\beta = .20$ . For the WURSS-21, the average ES coefficient was .162, which implies a needed sample size of 858.

Accepting one SEM as an indicator of small-but-important change provides slightly larger ES coefficients than ES, and thus slightly smaller required sample sizes to detect these changes. Cronbach’s  $\alpha$ , within-day standard deviation, and resulting SEMs for days 1–7 are portrayed in Table 4, along with standardized coefficients and sample

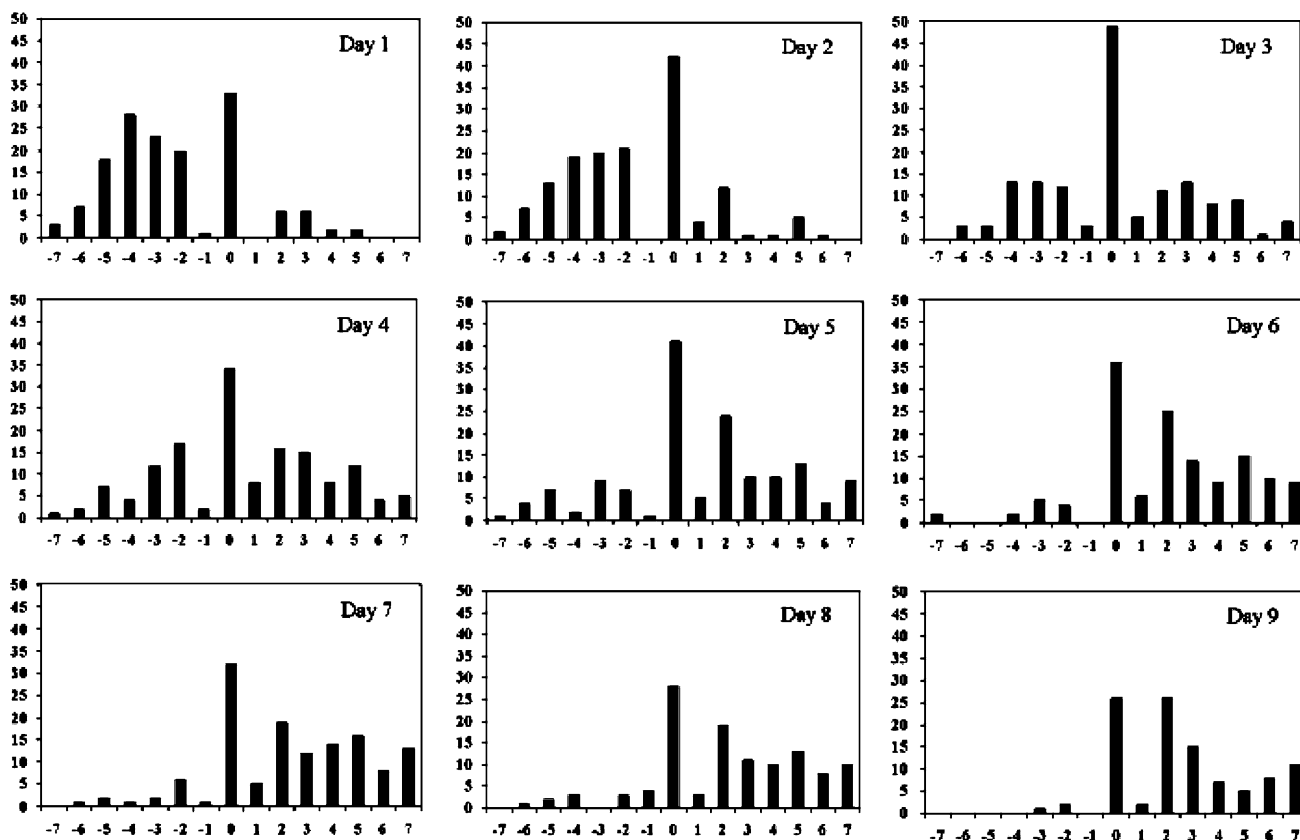


Fig. 1 Distribution of GRC scores by day of illness

**Table 4** Reliability, variability, and standard error of measurement

Day	Cronbach's $\alpha$	SD	SEM	Item SEM	Effect size for 1 SEM	Projected sample size
WURSS-44						
1	.950	48.5	10.74	0.249	0.222	518
2	.963	51.5	9.85	0.228	0.191	585
3	.968	51.6	9.13	0.212	0.177	616
4	.972	52.6	8.70	0.202	0.166	639
5	.976	55.9	8.59	0.199	0.154	666
6	.972	51.1	8.50	0.197	0.166	639
7	.971	48.8	8.27	0.192	0.169	633
7 day mean	.967	51.4	9.10	0.211	0.177	616
WURSS-21						
1	.916	24.8	7.17	0.358	0.289	372
2	.943	26.8	6.39	0.319	0.239	481
3	.951	26.2	5.79	0.289	0.221	520
4	.953	28.3	6.10	0.305	0.215	533
5	.961	29.5	5.77	0.288	0.195	576
6	.962	27.7	5.38	0.269	0.194	579
7	.962	27.4	5.35	0.267	0.195	576
7 day mean	.949	27.2	5.99	0.299	0.221	520

Table 4 portrays within-day and averaged-over-7-day SEM-related parameter estimates

Cronbach's  $\alpha$  and standard deviation (SD) are used to compute standard error of the mean (SEM) using the following formula, where  $\sigma_x$  = standard deviation and  $r_{tt}$  = Cronbach's alpha for the scale:

$SEM = \sigma_x \sqrt{1 - r_{tt}}$ . Both full instrument SEM and per item SEM are portrayed, along with the standardized ES coefficient for 1 SEM calculated by dividing the full instrument SEM by the SD for that day

The projected sample size is the per group size needed to detect one SEM, assuming two-tailed testing and  $\alpha$  (tolerance for Type I error) of .05, and  $\beta$  (tolerance for Type II error) of .20

sizes. A seven day average standardized SEM is .177 for the WURSS-44 and .221 for the WURSS-21. This suggests that a trial would need group sizes of 616 and 520, respectively, for power-to-detect-difference.

It may be worth pointing out that the shorter instrument, the WURSS-21, appeared to be more responsive than the WURSS-44 in the original validation study [36]. Using the estimates of small but important change of 12.7, 5.87, 5.99, and 4.53 points for standard MID, regression-based MID, SEM, and ES, respectively, for the WURSS-21, treatment group sample sizes of 80, 306, 520, and 858 would be needed. These corresponding projected sample sizes for the WURSS-44 are 89, 302, 616, and 870. We consider these to be conservative estimates, as a well-designed RCT would look at outcomes over the full duration of illness rather than a single between-day comparison, which would presumably increase power to detect between-group difference.

## Discussion

In 1957, Cronbach described "the two disciplines of scientific psychology" as "correlational" and "experimentalist"

[47]. Correlational psychologists used multivariate and factor analysis techniques to investigate relationships within cross-sectional and prospective cohorts. Experimentalists used RCTs to isolate the effects of interventions, to establish causality, and to estimate the magnitude of change attributable to interventions. For correlationists, diversity was an ally, as it increased power to detect between-person relationships using within-person variables. For experimentalists, diversity was an obstacle. The greater the diversity across individuals, the more difficult it was to detect changes over time attributable to interventions. In this article and others [48, 49], Cronbach correctly pointed out that experimentalists could say little or nothing about the effects of interventions on individuals. While the average score of the treated group in an RCT might change by a significant amount compared to the control group, this could be due either to large changes in a few individuals or smaller changes across a greater number. For the clinician and the patient, the likelihood and magnitude of change *for the individual* was the key issue. Because individual trajectories could not be known a priori, neither correlational nor experimental science could accurately predict individual response to treatment.



In 1969, Cohen published “Statistical Methods for the Behavioural Sciences” [45], which became a popular resource for investigators designing and interpreting RCTs. In that text proposed the standardized ES an appropriate coefficient to assess change-over-time (later termed responsiveness). Dividing the absolute difference (“raw gain” in the words of Cronbach) by the standard deviation yielded the ES, which, when complemented by tolerance for type I and type II error, defined power and sample size requirements. As mentioned earlier, standardized ESs up to 0.2 were described as “small,” 0.5 as “medium,” and 0.8 as “large.”

The 1970s, 1980s, and 1990s saw the widespread introduction of HRQoL instruments as RCT outcome measures. This, along with a rising interest in the assessment of clinical significance, led to a variety of methods for assessing important difference and responsiveness. Working within the ability-to-detect-change framework, Deyo and colleagues compared HRQoL responsiveness to diagnostic test performance [12, 50]. Kazis followed a similar track, concentrating on ES [27]. Guyatt et al. [19], Jaeschke et al. [3], Juniper et al. [20, 21], Redelmeier et al. [5], and Wells et al. [22] adopted an entirely different approach, using patient’s value judgments to rate degree-of-change. Global ratings of change (GRC) were introduced, refined, and used as anchors for assessing interval changes. Perhaps due to simplicity and face validity, these methods were picked up by several research groups, leading to a body of work that arguably defines the state-of-the-art for assessing important difference and responsiveness [10–17]. Wyrwich has compared resulting MIDs from these studies to corresponding SEM calculations, and has noted a consistent relationship [29, 32, 33].

As Norman, Stratford and Regehr [15] have noted, there are several potential problems with these approaches. The GRC item itself has not been extensively validated. As a single item, it cannot be assessed by reliability measures such as Cronbach’s  $\alpha$ . While it could be assessed by test-retest methods, interpretation would be problematic due to the time-specific nature of the rating. There are also doubts regarding people’s ability to retrospectively assess severity change over time. Effects of context, framing, concurrent severity, implicit theory, and duration of recall could bias these judgments [51, 52]. Brant, Sutherland, and Hilsden have argued that simply reporting means and standard errors of interval changes corresponding to specific GRC responses is “naive,” and have proposed a formal system of regression that would take into account potential confounders as well as within-person-over-time dependencies [11]. Finally, as Norman et al. argue using both real and simulated data [15], the GRC scale may be somewhat insensitive to prospectively assessed interval changes.

Our data suggest that the most commonly used distributional and GRC anchor-based approaches yield somewhat different results, with substantive implications for RCT design. The amount of daily change that occurs during the natural resolution of a cold is less than the amount of change assessed as minimal but important by standard MID methods, or projected by GRC-based regression models. Nevertheless, we are impressed as much by the consistencies as the discrepancies. We interpret these results to support the construct validity of the WURSS instrument and the utility of the GRC scale as a useful measure. Relationships between WURSS and GRC ratings were consistent across individuals and over time, and were in general unaffected by potential confounders. The sole anomalous finding was that the WURSS-44 interval changes corresponding to GRC ratings of “a great deal better” or “a very great deal better” were somewhat less than those corresponding to “moderately better” or “a good deal better.” This was perhaps due in part to the tendency of people at the end of their colds to rate daily GRC self-improvement highly, even when prospectively measured WURSS end-of-cold interval changes are small.

We are indebted to the editor and a primary reviewer of this article for asking us to include SEM as a comparator. We were surprised and pleased to see that SEM calculation and corresponding coefficients and implications for power and responsiveness portrayed were reasonably consistent with other methods, and in general supportive of Wyrwich and Norman’s findings [16, 29, 32, 33, 53, 54]. While the sample sizes needed to detect MID, SEM, and ES may appear discrepant, the absolute amount of between-assessment change these various methods project is surprisingly consistent, given the theoretical influences of reliability, variance, and ability of individuals to accurately assess HRQoL over time. We are pleased that the WURSS-44 and the embedded WURSS-21 items appear to function equally well in these analyses. We expect that most users will prefer the WURSS-21, as it is shorter and easier to use.

There are, of course, several limitations to this study: Participants were volunteers, representing a limited socioeconomic and cultural profile. Our sample size was modest, hence confidence intervals are wide. All data were self-reported, subject to recall and other biases. While the underlying constructs of illness severity and degree-of-change are continuous, the self-reported data came in the form of ordered categorical measures, raising questions of normality and interval equivalence. Analyses reported here are based on simple-summed WURSS scores, implying unity weighting, which may or may not be appropriate. Day-to-day changes in severity scores may not be ideal outcome measures. There are

reasonably strong theoretical arguments to instead use global area-under-the-curve measures, which would compound severity and duration into a single variable, and thus obviate the need for responsiveness assessment. Finally, and perhaps most importantly, there are neither known effective treatments nor independent gold standards for common cold severity assessment.

All approaches to important difference and responsiveness using HRQoL instruments are influenced by the population sampled, and by the conditions under which the subjects participate, as well as by the instrument itself. Responsiveness and important difference, like reliability, are properties of an instrument-in-use, not of an instrument as an isolated entity. Assessment of the amount of instrument-measured change that constitutes an important difference is value-based, and hence influenced by individual, cultural, and contextual factors. While distribution-based calculations are useful for the design of RCTs, they are insufficient for the interpretation of results. While clinician assessments and population norms can reasonably be used as reference standards, they should not be used to set benchmarks of clinical significance. Instead, it is the values of affected individuals that should be used to weigh the risks and benefits of treatments. For these reasons, we feel that GRC-based methods highlighted here are reasonably sound, if not completely satisfactory. For overall understanding of an instrument's strengths and weaknesses, we suggest multiple complementary approaches. In addition to the methods reported here, direct between-person severity comparison [5, 11, 22], clinician assessments [9, 23], and benefit-harm trade-off methods [55, 56] may be useful. Depending on the data available for a given HRQoL instrument, investigators, clinicians and patients will have to make judgments when predicting or interpreting the magnitude of treatment effects. We believe that the quality of these judgments will improve when data from several approaches are available, and when the strengths and weaknesses of specific strategies are more widely known.

**Acknowledgements** The authors would like to acknowledge the Department of Family Medicine and the School of Medicine and Public Health at the University of Wisconsin – Madison for providing startup funds, an institutional base, and collegial support. This work was also 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 Clinical Research Feasibility Funds (CRFF) from the NIH-funded University of Wisconsin-General Clinical Research Center (MO1 RR03186). The Robert Wood Johnson Foundation Generalist Physician Scholars Program supported the analysis phase of this project, and is allowing this work to go forward. Intellectually, we are indebted to Gordon Guyatt, who pioneered this area and has provided direct mentorship to Bruce Barrett, and to Geoffrey Norman and colleagues, whose 1997 [15] and 2001 [16] articles were particularly influential.

## References

1. McDowell, I., & Newell, C. (1996). *Measuring health: A guide to rating scales and questionnaires*. Oxford & New York: Oxford University Press.
2. Kirshner, B., & Guyatt, G. H. (1985). A methodological framework for assessing health indices. *Journal of Chronic Diseases*, *38*, 27–36.
3. Jaeschke, R., Singer, J., & Guyatt, G. H. (1989). Measurement of health status: Ascertaining the minimal clinically important difference. *Controlled Clinical Trials*, *10*, 407–415.
4. Powell, C. V., & Kelly, A.-M. (2001). Determining the minimum clinically significant difference in visual analog pain score for children. *Annals of Emergency Medicine*, *37*, 28–31.
5. Redelmeier, D. A., Guyatt, G. H., & Goldstein, R. S. (1996). Assessing the minimal important difference in symptoms: A comparison of two techniques. *Journal of Clinical Epidemiology*, *49*, 1215–1219.
6. Santanello, N. C., Zhang, J., Seidenberg, B., Reiss, T. F., & Barber, B. L. (1999). What are minimal important changes for asthma measures in a clinical trial? *European Respiratory Journal*, *14*, 23–27.
7. Schunemann, H. J., Griffith, L., Jaeschke, R., Goldstein, R., Stubbings, D., & Guyatt, G. H. (2003). Evaluation of the minimal important difference for the feeling thermometer and the St. George's Respiratory Questionnaire in patients with chronic airflow obstruction. *Journal of Clinical Epidemiology*, *56*, 1170–1176.
8. van Stel, H. F., Maille, A. R., Colland, V. T., & Everaerd, W. (2003). Interpretation of change and longitudinal validity of the quality of life for respiratory illness questionnaire (QoLRIQ) in inpatient pulmonary rehabilitation. *Quality of Life Research*, *12*, 133–145.
9. van Walraven, C., Mahon, J. L., Moher, D., Bohm, C., & Laupacis, A. (1999). Surveying physicians to determine the minimal important difference: Implications for sample-size calculation. *Journal of Clinical Epidemiology*, *52*, 717–723.
10. Beaton, D. E., Bombardier, C., Katz, J. N., & Wright, J. G. (2001). A taxonomy for responsiveness. *Journal of Clinical Epidemiology*, *54*, 1204–1207.
11. Brant, R., Sutherland, L., & Hilsden, R. (1999). Examining the minimum important difference. *Statistics in Medicine*, *18*, 2593–2603.
12. Deyo, R. A., & Centor, R. M. (1986). Assessing the responsiveness of functional scales to clinical change: An analogy to diagnostic test performance. *Journal of Chronic Diseases*, *39*, 897–906.
13. Frost, M. H., Bonomi, A. E., Ferrans, C. E., Wong, G. Y., & Hays, R. D. (2002). Clinical Significance Consensus Meeting Group. Patient, clinician, and population perspectives on determining the clinical significance of quality-of-life scores. *Mayo Clinic Proceedings*, *77*, 488–494.
14. Guyatt, G. H., Osoba, D., Wu, A. W., Wyrwich, K. W., & Norman, G. R. (2002). Clinical Significance Consensus Meeting Group. Methods to explain the clinical significance of health status measures. *Mayo Clinic Proceedings*, *77*, 371–383.
15. Norman, G. R., Stratford, P., & Regehr, G. (1997). Methodological problems in the retrospective computation of responsiveness to change: The lesson of Cronbach. *Journal of Clinical Epidemiology*, *50*, 869–879.
16. Norman, G. R., Sridhar, F. G., Guyatt, G. H., & Walter, S. D. (2001). Relation of distribution- and anchor-based approaches in interpretation of changes in health-related quality of life. *Medical Care*, *39*, 1039–1047.
17. Samsa, G. (2001). How should the minimum important difference for a health-related quality-of-life instrument be estimated? *Medical Care*, *39*, 1037–1038.

18. Husted, J. A., Gladman, D. D., Cook, R. J., & Farewell, V. T. (1998). Responsiveness of health status instruments to changes in articular status and perceived health in patients with psoriatic arthritis. *Journal of Rheumatology*, *25*, 2146–2155.
19. Guyatt, G. H., Walter, S., & Norman, G. (1987). Measuring change over time: Assessing the usefulness of evaluative instruments. *Journal of Chronic Diseases*, *40*, 171–178.
20. Juniper, E. F., & Guyatt, G. H. (1991). Development and testing of a new measure of health status for clinical trials in rhinoconjunctivitis. *Clinical & Experimental Allergy*, *21*, 77–83.
21. Juniper, E. F., Guyatt, G. H., Willan, A., & Griffith, L. E. (1994). Determining a minimal important change in a disease-specific Quality of Life Questionnaire. *Journal of Clinical Epidemiology*, *47*, 81–87.
22. Wells, G. A., Tugwell, P., Kraag, G. R., Baker, P. R., Groh, J., & Redelmeier, D. A. (1993). Minimum important difference between patients with rheumatoid arthritis: the patient's perspective. *Journal of Rheumatology*, *20*, 557–560.
23. Todd, K. H., & Funk, J. P. (1996). The minimum clinically important difference in physician-assigned visual analog pain scores. *Academic Emergency Medicine*, *3*, 142–146.
24. Bruynesteyn, K., van der Heijde, H. D., Boers, M., Lassere, M., Boonen, A., Edmonds, J, et al. (2001). Minimal clinically important difference in radiological progression of joint damage over 1 year in rheumatoid arthritis: Preliminary results of a validation study with clinical experts. *Journal of Rheumatology*, *28*, 904–910.
25. Bombardier, C., Hayden, J., & Beaton, D. E. (2001). Minimal clinically important difference, low back pain: Outcomes measures. *Journal of Rheumatology*, *28*, 431–438.
26. Farrar, J. T., Portenoy, R. K., Berlin, J. A., Kinman, J. L., & Strom, B. L. (2000). Defining the clinically important difference in pain outcome measures. *Pain*, *88*, 287–294.
27. Kazis, L. E., Anderson, J. L., Meenan, R. F. (1989). Effect sizes for interpreting changes in health status. *Medical Care*, *27*(Suppl), S178–S189.
28. Ward, M. M., Marx, A. S., & Barry, N. N. (2000). Identification of clinically important changes in health status using receiver operating characteristic curves. *Journal of Clinical Epidemiology*, *53*, 279–284.
29. Wyrwich, K. W., Tierney, W. M., & Wolinsky, F. D. (2002). Using the standard error of measurement to identify important changes on the Asthma Quality of Life Questionnaire. *Quality of Life Research*, *11*, 1–7.
30. Cohen, J. (1988) *Statistical power analysis for the behavioral sciences*. Hillsdale, N.J.: Lawrence Erlbaum Associates.
31. Norman, G. R., Wyrwich, K. W., & Patrick, D. L. (2007). The mathematical relationship among different forms of responsiveness coefficients. *Quality of Life Research*, *16*(5), 815–822.
32. Wyrwich, K. W., Nienaber, N. A., Tierney, W. M., & Wolinsky, F. D. (1999). Linking clinical relevance and statistical significance in evaluating intra-individual changes in health-related quality of life. *Medical Care*, *37*, 469–478.
33. Wyrwich, K. W. (2004). Minimal important difference thresholds and the standard error of measurement: Is there a connection? *Journal of Biopharmaceutical Statistics*, *14*, 97–110.
34. Barrett, B., Locken, K., Maberry, R., Schwamman, J., Bobula, J., Brown, R., et al. (2002). The Wisconsin Upper Respiratory Symptom Survey: Development of an instrument to measure the common cold. *Journal of Family Practice*, *51*, 265–273.
35. Barrett, B. P., Brown, R. L., Locken, K., Maberry, R., Bobula, J. A., & D'Alessio, D. (2002). Treatment of the common cold with unrefined echinacea: A randomized, double-blind, placebo-controlled trial. *Annals of Internal Medicine*, *137*, 939–946.
36. Barrett, B., Brown, R., Mundt, M., Safdar, N., Dye, L., Maberry, R., et al. (2005). The Wisconsin Upper Respiratory Symptom Survey is responsive, reliable, and valid. *Journal of Clinical Epidemiology*, *58*, 609–617.
37. Jackson, G. G., Dowling, H. F., & Muldoon, R. L. (1962). Present concepts of the common cold. *American Journal of Public Health*, *52*, 940–945.
38. McHorney, C. A., Ware, J. E., & Raczek, A. E. (1998). The MOS 36-item short-form health survey (SF-36): II. Psychometric and clinical tests of validity in measuring physical and mental health constructs. *Medical Care*, *31*, 247–263.
39. Ware, J. E., Kosinski, M., Dewey, J. E., & Gandek, B. (2001) *How to score and interpret single-item health status measures: A manual for users of the SF-8 health survey*. Lincoln RI: Quality Metric.
40. Barrett, B., Brown, R., Voland, R., Maberry, R., & Turner, R. (2006). Relations among questionnaire and laboratory measures of rhinovirus infection. *European Respiratory Journal*, *28*, 358–363.
41. Yang, M., & Goldstein, H. (1996). Multilevel models for longitudinal data. In U. Engel & J. Tanner (Eds.), *Analysis of change: Advanced techniques in panel data analysis* (pp. 191–220). Berlin: Walter de Gruyter.
42. Jaeschke, R., Singer, J., & Guyatt, G. H. (1990). A comparison of seven-point and visual analogue scales. Data from a randomized trial. *Controlled Clinical Trials*, *11*, 43–51.
43. Miller, G. A. (1956). The magical number seven plus or minus two: some limits on our capacity for processing information. *Psychological Review*, *63*, 81–97.
44. Froberg, D. G., & Kane, R. L. (1989). Methodology for measuring health-state preferences-II: Scaling methods. *Journal of Clinical Epidemiology*, *42*, 459–471.
45. Cohen, J. (1969). *Statistical power analysis for the behavioural sciences*. London: Academic Press.
46. Gwaltney, J. M Jr., Hendley, J. O., & Patrie, J. T. (2003). Symptom severity patterns in experimental common colds and their usefulness in timing onset of illness in natural colds. *Clinical Infectious Diseases*, *36*, 714–723.
47. Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American Psychologist*, *12*, 671–684.
48. Gleser, G. C., Cronbach, L. J., & Rajaratnam, N. (1965). Generalizability of scores influenced by multiple sources of variance. *Psychometrika*, *30*, 395–418.
49. Cronbach, L. J., & Furby, L. (1970). How should we measure "change" – Or should we? *Psychological Bulletin*, *74*, 68–80.
50. Deyo, R. A., & Inui, T. S. (1984). Toward clinical applications of health status measures: sensitivity of scales to clinically important changes. *Health Services Research*, *19*, 277–289.
51. Ross, M. (1989). Relation of implicit theories to the construction of personal histories. *Psychological Review*, *96*, 341–347.
52. Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, *211*, 453–458.
53. Norman, G. R., Sloan, J. A., & Wyrwich, K. W. (2003). Interpretation of changes in health-related quality of life: The remarkable universality of half a standard deviation. *Medical Care*, *41*, 582–592.
54. Norman, G. R. (2005). The relation between the minimally important difference and patient benefit. *COPD*, *2*, 69–73.
55. Llewellyn-Thomas, H. A., Williams, J. L., Levy, L., & Naylor, C. D. (1996). Using a trade-off technique to assess patients' treatment preferences for benign prostatic hyperplasia. *Medical Decision Making*, *16*, 262–282.
56. Naylor, C. D., & Llewellyn-Thomas, H. A. (1994). Can there be a more patient-centred approach to determining clinically important effect sizes for randomized treatment trials? *Journal of Clinical Epidemiology*, *47*, 787–795.