Randomized trials, clinical significance & shared decision-making: Respiratory infection and beyond

Bruce Barrett MD PhD
Associate Professor
Department of Family Medicine
University of Wisconsin - Madison
Acknowledgements

- Roger Brown PhD, Marlon Mundt PhD, Zhengjun Zhang PhD
- Shari Barlow, Michelle Gassman, Tola Ewers, Gay Thomas MS, etc
- Mary Beth Plane PhD et al. Department of Family Medicine at the University of Wisconsin School of Medicine and Public Health
- National Center for Complementary and Alternative Medicine at the National Institutes of Health (NIH NCCAM)
  - K-23 Patient-Oriented Career Grant
  - R01 Physician Echinacea Placebo RCT
  - R01 Meditation & Exercise for Prevention of Respiratory Infection
- Robert Wood Johnson Foundation Generalist Physician Faculty Scholars Program
- Gordon Guyatt MD – EBM guru, MID originator & RWJF mentor
Where it all began (for me)
THE HERBAL MEDICINE BOOM
It's great business, but is it good for what ails us?
Medical profession applauds safe herb miracle!

ECHINACEA CURES Colds & FLU

'Try it as a front-line treatment for colds, flu & sore throats'
- TOP NATURAL HEALTH AUTHOR DR. ANDREW WEIL

'Echinacea helps you get over colds & flu much faster than any other treatment'
- PURDUE UNIVERSITY RESEARCH GROUP

'Protects against colds, flu & boosts immune system'
- PRESIDENT CLINTON'S COMMISSION ON DIETARY SUPPLEMENT LABELS

'Supercharge Echinacea’s amazing healing power with Goldenseal'
- DR. MICHAEL MURRAY, BASTYR COLLEGE, SEATTLE
The Common Cold
Patient-oriented outcomes

- Health-related quality-of-life (HRQoL) measures are increasingly used as primary outcomes in RCTs
- Items assessing symptomatic and functional domains are self-reported using general or illness-specific HRQoL instruments
- Patient-Reported Outcome (PRO) questionnaires

General Health Measures

- EuroQual (EQ-5D)
- Feeling thermometer (100mm VAS)
- Sickness Impact Profile
- Medical Outcomes Survey
  - SF-36 (4 week recall)
  - SF-12
  - SF-8 (24 hour recall)
- Subscales:
  - Physical health
  - Mental health
Illness specific patient-reported outcomes (PROs)

- Hamilton, Beck (depression, anxiety)
- PHQ-9 (depression)
- Connor’s (ADHD)
- Vanderbilt (ADHD)
- Mini-mental status exam (dementia; cognitive function)
- Glasgow coma scale
- WOMAC - Western Ontario McMaster Universities Osteoarthritis Index
- AIMS - Arthritis Impact Measurement Scale
- RSOM-31 - Rhinosinusitis outcome measure
- SNOT - SinoNasal Outcomes Test
Wisconsin Upper Respiratory Symptom Survey (WURSS)

- In 1998 there was no HRQoL instrument for common cold
  - Jackson’s 8-item scale assessed symptoms as mild/mod/severe
- Using local expert opinion and some field testing, we created a questionnaire instrument for use in a
  1999 RCT testing echinacea as a cold treatment

Wisconsin Upper Respiratory Symptom Survey (WURSS)

- After the initial echinacea trial, we combined qualitative and quantitative methods to create/develop/validate WURSS.
- WURSS is an illness-specific quality-of-life outcomes instrument designed to evaluate cold symptoms, functional impact, and quality of life during cold/flu illness.
- Open-ended questions with cold-sufferers gave us >150 terms.
- Some 44 were mentioned by multiple participants (WURSS-44).

Validity, reliability & responsiveness

- Initial validity/reliability testing
- 151 participants, 1,681 days of cold symptoms
- Frequency, severity, importance-to-patients, responsiveness, and dimensional cohesion
- External (convergent) validity was assessed with comparisons to the Jackson scale and the SF-8 (24 hour recall version)

Using importance-to-patients and responsiveness* as guides, we selected items for a short version, the WURSS-21

*responsiveness = sensitivity to change over time, important property for clinical trial outcome measurements

<table>
<thead>
<tr>
<th>Item</th>
<th>Domain</th>
<th>Frequency</th>
<th>Severity*</th>
<th>Importance*</th>
<th>MID</th>
<th>S.E. stable</th>
<th>Responsiveness</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Global</td>
<td>100</td>
<td>3.92 ± 1.17</td>
<td>3.79 ± 0.90</td>
<td>0.721</td>
<td>0.516</td>
<td>0.709</td>
</tr>
<tr>
<td>2</td>
<td>Cough</td>
<td>94</td>
<td>3.04 ± 1.33</td>
<td>3.21 ± 1.10</td>
<td>0.389</td>
<td>0.839</td>
<td>0.300</td>
</tr>
<tr>
<td>3</td>
<td>Cough</td>
<td>80</td>
<td>2.77 ± 1.22</td>
<td>3.13 ± 1.06</td>
<td>0.308</td>
<td>1.283</td>
<td>0.193</td>
</tr>
<tr>
<td>5</td>
<td>Throat</td>
<td>90</td>
<td>3.19 ± 1.46</td>
<td>3.39 ± 0.95</td>
<td>0.407</td>
<td>1.072</td>
<td>0.278</td>
</tr>
<tr>
<td>6</td>
<td>Throat</td>
<td>91</td>
<td>3.10 ± 1.37</td>
<td>3.16 ± 0.95</td>
<td>0.404</td>
<td>0.863</td>
<td>0.307</td>
</tr>
<tr>
<td>9</td>
<td>Nasal</td>
<td>96</td>
<td>3.55 ± 1.47</td>
<td>3.60 ± 0.94</td>
<td>0.509</td>
<td>1.140</td>
<td>0.337</td>
</tr>
<tr>
<td>10</td>
<td>Nasal</td>
<td>95</td>
<td>2.78 ± 1.27</td>
<td>2.72 ± 1.13</td>
<td>0.423</td>
<td>1.103</td>
<td>0.285</td>
</tr>
<tr>
<td>18</td>
<td>Tired</td>
<td>99</td>
<td>3.98 ± 1.50</td>
<td>3.90 ± 0.91</td>
<td>0.627</td>
<td>1.276</td>
<td>0.392</td>
</tr>
<tr>
<td>26</td>
<td>None</td>
<td>72</td>
<td>2.38 ± 1.23</td>
<td>2.69 ± 1.10</td>
<td>0.216</td>
<td>0.641</td>
<td>0.191</td>
</tr>
<tr>
<td>29</td>
<td>Chest</td>
<td>73</td>
<td>2.52 ± 1.26</td>
<td>3.12 ± 1.00</td>
<td>0.279</td>
<td>0.767</td>
<td>0.226</td>
</tr>
<tr>
<td>32</td>
<td>Tired</td>
<td>97</td>
<td>3.94 ± 1.51</td>
<td>3.98 ± 0.87</td>
<td>0.614</td>
<td>1.808</td>
<td>0.323</td>
</tr>
<tr>
<td>34</td>
<td>Activity</td>
<td>85</td>
<td>3.09 ± 1.41</td>
<td>4.04 ± 0.97</td>
<td>0.431</td>
<td>1.327</td>
<td>0.265</td>
</tr>
<tr>
<td>39</td>
<td>Activity</td>
<td>85</td>
<td>3.02 ± 1.42</td>
<td>3.86 ± 0.99</td>
<td>0.506</td>
<td>0.813</td>
<td>0.397</td>
</tr>
<tr>
<td>40</td>
<td>Activity</td>
<td>76</td>
<td>3.04 ± 1.46</td>
<td>3.37 ± 0.98</td>
<td>0.398</td>
<td>1.196</td>
<td>0.257</td>
</tr>
<tr>
<td>41</td>
<td>Activity</td>
<td>76</td>
<td>3.03 ± 1.38</td>
<td>3.56 ± 1.00</td>
<td>0.437</td>
<td>0.724</td>
<td>0.363</td>
</tr>
<tr>
<td>42</td>
<td>Activity</td>
<td>87</td>
<td>2.78 ± 1.34</td>
<td>3.50 ± 1.07</td>
<td>0.485</td>
<td>1.015</td>
<td>0.340</td>
</tr>
<tr>
<td>43</td>
<td>Activity</td>
<td>86</td>
<td>3.09 ± 1.42</td>
<td>3.91 ± 0.99</td>
<td>0.509</td>
<td>1.004</td>
<td>0.359</td>
</tr>
<tr>
<td>Symptoms</td>
<td>Symptoms</td>
<td>Symptoms</td>
<td>Functional impairments</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>----------</td>
<td>----------</td>
<td>----------</td>
<td>-----------------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10. Sneezing</td>
<td>21. Sinus pressure</td>
<td>32. Lack of energy</td>
<td>43. Live your personal life</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11. Headache</td>
<td>22. Sinus drainage</td>
<td>33. Loss of appetite</td>
<td>44. Compared to yesterday, I feel...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The WURSS-21 has now been independently validated in a study of N=230 participants monitored for 2,457 person-days.

“Validation of a Short Form Wisconsin Upper Respiratory Symptom Survey (WURSS-21)”

Bruce Barrett MD PhD, Roger Brown PhD, Marlon Mundt PhD, Gay Thomas MS, Shari Barlow, Alex Highstrom, Mozhdeh Bahrainian MD

*Health and Quality of Life Outcomes* 7 (76), 2009
The top and bottom of the notched boxes indicate the 25% and 75% percentiles, respectively.

The center of the notched boxes is the median summed score for that day.

The notches can be compared to assess difference at the P = 0.05 level of significance.

The ends of the vertical lines indicate the last actual data point within 1.5 (IQR) from the 25%ile and 75%ile.

The symbols above and below these lines are actual outlying data points.
Frequency, severity, MID, MSE and responsiveness of WURSS-21 Items

Frequency = present at least once during first seven days

Severity = Mean on 7-point scale averaged over first three days n ± SD

MID = minimal important difference

Responsiveness = MID / (2*MSE)\(^{1/2}\)

MSE = mean square error

<table>
<thead>
<tr>
<th>Symptom</th>
<th>Frequency</th>
<th>Severity</th>
<th>MID</th>
<th>MSE</th>
<th>Responsiveness</th>
</tr>
</thead>
<tbody>
<tr>
<td>How sick</td>
<td>100.0</td>
<td>4.13 ± 1.46</td>
<td>0.77</td>
<td>0.78</td>
<td>0.62</td>
</tr>
<tr>
<td>Runny nose</td>
<td>98.3</td>
<td>3.70 ± 1.77</td>
<td>0.56</td>
<td>1.48</td>
<td>0.33</td>
</tr>
<tr>
<td>Plugged nose</td>
<td>96.5</td>
<td>4.00 ± 1.79</td>
<td>0.57</td>
<td>1.54</td>
<td>0.32</td>
</tr>
<tr>
<td>Sneezing</td>
<td>95.7</td>
<td>3.34 ± 1.76</td>
<td>0.50</td>
<td>1.20</td>
<td>0.32</td>
</tr>
<tr>
<td>Sore throat</td>
<td>92.6</td>
<td>3.76 ± 1.85</td>
<td>0.49</td>
<td>1.00</td>
<td>0.35</td>
</tr>
<tr>
<td>Scratchy throat</td>
<td>96.1</td>
<td>3.82 ± 1.81</td>
<td>0.50</td>
<td>1.25</td>
<td>0.32</td>
</tr>
<tr>
<td>Cough</td>
<td>92.2</td>
<td>3.80 ± 1.84</td>
<td>0.46</td>
<td>1.76</td>
<td>0.25</td>
</tr>
<tr>
<td>Hoarseness</td>
<td>86.1</td>
<td>3.38 ± 2.01</td>
<td>0.41</td>
<td>1.29</td>
<td>0.26</td>
</tr>
<tr>
<td>Head congestion</td>
<td>93.0</td>
<td>4.03 ± 1.70</td>
<td>0.64</td>
<td>1.54</td>
<td>0.37</td>
</tr>
<tr>
<td>Chest congestion</td>
<td>75.7</td>
<td>3.76 ± 1.88</td>
<td>0.38</td>
<td>0.97</td>
<td>0.27</td>
</tr>
<tr>
<td>Feeling tired</td>
<td>99.6</td>
<td>4.33 ± 1.80</td>
<td>0.82</td>
<td>1.41</td>
<td>0.49</td>
</tr>
<tr>
<td>Think clearly</td>
<td>91.3</td>
<td>3.53 ± 1.68</td>
<td>0.54</td>
<td>1.02</td>
<td>0.38</td>
</tr>
<tr>
<td>Sleep well</td>
<td>93.9</td>
<td>4.17 ± 1.82</td>
<td>0.66</td>
<td>1.69</td>
<td>0.36</td>
</tr>
<tr>
<td>Breathe easily</td>
<td>96.5</td>
<td>3.84 ± 1.86</td>
<td>0.60</td>
<td>1.08</td>
<td>0.41</td>
</tr>
<tr>
<td>Walk/Climb stairs</td>
<td>89.6</td>
<td>3.75 ± 1.81</td>
<td>0.50</td>
<td>0.88</td>
<td>0.38</td>
</tr>
<tr>
<td>Accomplish daily activities</td>
<td>90.4</td>
<td>3.57 ± 1.74</td>
<td>0.57</td>
<td>1.08</td>
<td>0.39</td>
</tr>
<tr>
<td>Work outside the home</td>
<td>82.2</td>
<td>3.80 ± 1.84</td>
<td>0.48</td>
<td>1.16</td>
<td>0.32</td>
</tr>
<tr>
<td>Work inside the home</td>
<td>87.0</td>
<td>3.52 ± 1.81</td>
<td>0.51</td>
<td>0.80</td>
<td>0.40</td>
</tr>
<tr>
<td>Interact with others</td>
<td>86.5</td>
<td>3.50 ± 1.73</td>
<td>0.53</td>
<td>0.93</td>
<td>0.39</td>
</tr>
<tr>
<td>Live your personal life</td>
<td>88.3</td>
<td>3.58 ± 1.74</td>
<td>0.58</td>
<td>0.92</td>
<td>0.43</td>
</tr>
</tbody>
</table>
Exploratory factor analysis for WURSS-21 using 2 to 7 dimensions

<table>
<thead>
<tr>
<th>Dimension(s)</th>
<th>Chi-square</th>
<th>df</th>
<th>( \chi^2/df )</th>
<th>CFI</th>
<th>TLI</th>
<th>RMSEA</th>
<th>SRMR</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1547.9</td>
<td>134</td>
<td>11.5</td>
<td>0.986</td>
<td>0.982</td>
<td>0.215</td>
<td>0.064</td>
</tr>
<tr>
<td>3</td>
<td>866.5</td>
<td>117</td>
<td>7.4</td>
<td>0.993</td>
<td>0.989</td>
<td>0.167</td>
<td>0.043</td>
</tr>
<tr>
<td>4</td>
<td>580.3</td>
<td>101</td>
<td>5.7</td>
<td>0.995</td>
<td>0.992</td>
<td>0.144</td>
<td>0.032</td>
</tr>
<tr>
<td>5</td>
<td>381.8</td>
<td>86</td>
<td>4.4</td>
<td>0.997</td>
<td>0.994</td>
<td>0.123</td>
<td>0.023</td>
</tr>
<tr>
<td>6</td>
<td>254.8</td>
<td>72</td>
<td>3.5</td>
<td>0.998</td>
<td>0.996</td>
<td>0.105</td>
<td>0.017</td>
</tr>
<tr>
<td>7</td>
<td>136.1</td>
<td>59</td>
<td>2.3</td>
<td>0.999</td>
<td>0.998</td>
<td>0.076</td>
<td>0.012</td>
</tr>
</tbody>
</table>

CFI = Comparative Fit Index

TLI = Tucker-Lewis Index

RMSEA = Root Mean Square Error of Approximation

SRMR = Standardized Root Mean Square Residual

Hu and Bentler (1999)\[58\] suggest the following cut off values for good fit, CFI > .95, TLI > .95, RMSEA < .06, and SRMR < .08
Best fit factorial model for WURSS-21

| CFA Final model structure of the WURSS-21 at day 3 |
|---|---|---|---|---|
| Dimensions 3 restricted | Chi-square | df | $\chi^2$/df | CFI |
| | 245.7 | 37 | 6.6 | 0.949 |
| Number of items used = 20 | TLI | RMSEA | WRMR |
| | 0.990 | 0.157 | 1.074 |

Time invariance (configural invariance) Days 2 to 7

<table>
<thead>
<tr>
<th>Day</th>
<th>CFI</th>
<th>TLI</th>
<th>RMSEA</th>
<th>WRMR</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.903</td>
<td>0.978</td>
<td>0.170</td>
<td>1.234</td>
</tr>
<tr>
<td>3</td>
<td>0.949</td>
<td>0.990</td>
<td>0.157</td>
<td>1.074</td>
</tr>
<tr>
<td>4</td>
<td>0.962</td>
<td>0.993</td>
<td>0.157</td>
<td>1.047</td>
</tr>
<tr>
<td>5</td>
<td>0.970</td>
<td>0.995</td>
<td>0.145</td>
<td>0.973</td>
</tr>
<tr>
<td>6</td>
<td>0.983</td>
<td>0.995</td>
<td>0.147</td>
<td>1.030</td>
</tr>
<tr>
<td>7</td>
<td>0.980</td>
<td>0.995</td>
<td>0.132</td>
<td>0.909</td>
</tr>
</tbody>
</table>

CFI = Comparative Fit Index
TLI = Tucker-Lewis Index
RMSEA = Root Mean Square Error of Approximation
WRMR = Weighted Root Mean Square Residual

Hu and Bentler suggest the following cut off values for good fit, CFI > .95, TLI > .95, RMSEA < .06, and WRMR < .90

In tribute to Roger Brown PhD, statistician/psychometrician/methodologist extraordinaire

Roger

Bruce

I can prove it or disprove it, what do you want me to do?
University of California San Francisco (2002-2008); University of Queensland, Australia (2002); University of Wolverhampton, England (2002-04); Case Western University (2003); Harvard University (2003); University of Southern Indiana (2003); Novartis Consumer Health, Hungary (2003); Baylor University in Texas (2003-04); University of Brisbane, Australia (2004-05); Bayer-sponsored URI trial in Italy (2004-06); University of Iowa (2004); Altamira Family Medicine, San Juan, Puerto Rico (2005); UCLA - Harbor UCLA (2005-06); Laboratoires Goemar, Saint Malo, France (2005-06); University of Leuven, Belgium, with field work in Peru (2005-2007); Northwestern University (2005-06); Unilever, the Netherlands (2006); Hindustan Lever Research Centre, Bangalore, India (2006); Baskent University, Ankara, Turkey (2006); California School of Professional Psychology, San Diego (2006); McMaster, Hamilton, Ontario, Canada (2006); University of North Carolina, Chapel Hill (2006); Navy Medical Center Portsmouth (2006); Appalachian State University (2006); Bastyr University, Seattle (2006); Philadelphia Center for Health Equity Research & Promotion (2006); University of Rochester (2007); Duke University Clinical Research Institute (2007); California School of Professional Psychology, Sacramento (2007); Washington University School of Medicine (2007); Georgetown University (2007); East Carolina University (2007); Research Institute for Sport and Exercise Sciences, Liverpool, U.K. (2007); Universiti Teknologi MARA Shal Alam, Malaysia (2007); Daejeon University, South Korea (2007); University of Wales at Bangor, Wales, U.K. (2007); Washington University School of Medicine, St. Louis MO (2007); Georgetown University, Washington DC (2007); SouthHamptom University Hospital, England (2007); Fundacion Santafe, Bogota, Columbia (2007); University of Western Australia (2007); McNeil Consumer HealthCare, Fort Washington, Pennsylvania (2007); University of Iowa, Coralville, Iowa (2007); Research Institute, Lancaster General Hospital, PA (2007); Davao Medical School, Davao City, Philippines (2007); Northwestern University, Chicago (2007); Georgetown University, Washington DC (2007); Pharma Projects, Czech Republic (2007); Castle Hill, Cottingham, U.K. (2007); Northcentral University, Pembroke Pines, FL (2007); Northcentral University, Pembroke Pines, Florida (2007); Queen Mary University, London, England (2007); University Hospital Heidelberg, General Practice and Health Services Research, Heidelberg, Germany (2007); University of California San Francisco (2007); Department of Physiology, Medical University of Vienna, Austria (2007); Université de Montreal (2008); St John Fisher College, Rochester, N.Y. (2008); Massey University / Zespri, Mt Maunganui, New Zealand (2008); Burgmann College, Canberra, Australia (2008); Coventry University, Coventry, U.K. (2008); University of South Carolina (2008); University of Texas, San Antonio (2008); University of Georgia (2008); University of Western Australia (2008); Washington University, St. Louis (2008); James Cook University, Australia (2008). Research Institute, Lancaster General Hospital (2008); University of Jyväskylä, Finland (2008); University of Exeter, U.K. (2008); Griffith University, Gold Coast Campus, Australia (2008); University of Massachusetts, Amherst (2008); Shinobu University, Shinobu, Japan (2008); Georgetown University (2009); Murrayville Family Practice, Langley, Canada (2009); University of Montana, Missoula (2009); Georgetown University, Smithtown, New York (2009); University of Torino, Italy (2009); Sports Medicine, Technical University of Munich, Germany (2009); Faculdade de Educacao Fisica de Limeira, Sao Paulo, Brasil (2009); University of KwaZulu-Natal, Durban, South Africa (2009); Allergy & Immunology, Vzsoké Tatry, Slovakia (2009); University College, London, England (2009); Texas Women’s University, Houston, Texas (2009); Clinical Consulting, Quebec, Canada (2009); University of Sumatera Utara, Medan, Indonesia (2009); Family Practice, Trinidad and Tobago (2009); Georgetown University (2009); Biopharmacopae Design International, Quebec, Canada (2009); Seattle Children's Hospital, Seattle, Washington (2009); University of Alberta, Alberta, Canada (2009); Reckitt Benckiser Inc., Long Valley, New Jersey (2009); University of Torino, Italy (2009), Pediatric allergy and immunology; University of Wisconsin Hospitals (2009), Asthma Bhawan, Rajasthan, India (2009), Chinese Clinical Trials Coordination Center, Beijing, China (2009), University Technology MARA, Selangor, Malaysia (2009), Preventive and Rehab. SportMedicine, Technical University of Munich, Germany (2009), Medicus Research, LLC, Midlothian, VA (2009), Northcentral University, Mesa, AZ (2009), GAMD Inc, Los Angeles, CA (2009)
Evidence Based Medicine

- Clinical medicine is increasingly guided by evidence, mostly from randomized controlled trials (RCTs).
- RCTs provide the least biased estimates of the effects of medical interventions (drugs, surgery, counseling, lab tests, screenings, etc).

RCTs powered to detect specific effect sizes

- Small (underpowered) RCTs can fail to detect benefits large enough to be considered clinically significant
- Very large (overpowered) trials can detect benefits too small to be considered clinically significant

Correctly powered trials

- RCTs should be powered to detect the smallest effect size that is considered to be clinically significant
- This way, we minimize chances of both false-positive and false-negative trials

240 subjects per group provides more than 80% power ($1 - \beta = .80$) to detect a 20% difference in area-under-the-curve severity, with $\alpha = 0.05$ and assuming one-sided testing and proportionally stable standard deviation.
Clinical significance

“How much of an effect is important?”

Conventionally, any statistically significant effect was considered important

Hence, important was equated with detectable

Clinical & statistical significance were confused


Clinical significance

- Cohen’s $d$ standardized effect size (ES) = absolute effect size / standard deviation (SD)
- SD is a measure of background variability
- ES’s up to 0.2 were classified as “small,” 0.5 as “medium,” and 0.8 as “large”
- Unsatisfactory. Degree of benefit considered important (clinically significant) should not depend on background variability

Minimal important difference

MID originally defined 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”


Minimal important difference

- Can be measured by comparing self-reported global improvement to interval changes on HRQoL outcome instruments
- “Compared to yesterday, I feel that my [cold] is…”
  Better  The Same  Worse


Beaton DE, Boers M, Wells GA. Many faces of the minimal clinically important difference (MCID): a literature review and directions for future research. *Current Opinion in Rheumatology.* 2002;**14**:109-14
<table>
<thead>
<tr>
<th>If better</th>
<th>If worse</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Almost the same, hardly any better at all</td>
<td>1. Almost the same, hardly any worse at all</td>
</tr>
<tr>
<td>2. A little better</td>
<td>2. A little worse</td>
</tr>
<tr>
<td>3. Somewhat better</td>
<td>3. Somewhat worse</td>
</tr>
<tr>
<td>4. Moderately better</td>
<td>4. Moderately worse</td>
</tr>
<tr>
<td>5. A good deal better</td>
<td>5. A good deal worse</td>
</tr>
<tr>
<td>6. A great deal better</td>
<td>6. A great deal worse</td>
</tr>
<tr>
<td>7. A very great deal better</td>
<td>7. A very great deal worse</td>
</tr>
</tbody>
</table>
Minimal important difference

Minimal important difference (MID) is usually operationalized as the average amount of change between two consecutive HRQoL assessments for everyone who responds “a little better” or “somewhat better.”

Responsiveness

Also known as sensitivity to change, responsiveness is defined as the ability to
detect change over time

Guyatt’s responsiveness index =

\[
\frac{\text{Minimal Important Difference}}{(2\times\text{MSE})^{1/2}}
\]

G. H. Guyatt, S. Walter, and G. Norman. Measuring change over time: Assessing the

G. H. Guyatt, R. A. Deyo, M. Charlson, M. N. Levine, and A. Mitchell. Responsiveness and
validity in health status measurement: a clarification. *Journal of Clinical
## MID, Responsiveness, Power

<table>
<thead>
<tr>
<th></th>
<th>WURSS-21</th>
<th>WURSS-44</th>
<th>Jackson</th>
</tr>
</thead>
<tbody>
<tr>
<td>MID</td>
<td>9.48</td>
<td>16.7</td>
<td>1.56</td>
</tr>
<tr>
<td>Responsiveness</td>
<td>0.80</td>
<td>0.71</td>
<td>0.54</td>
</tr>
<tr>
<td>Trial size (# subjects)</td>
<td>74</td>
<td>92</td>
<td>124</td>
</tr>
</tbody>
</table>

(Assuming 2 arms, alpha = 0.05, beta 0.90, 2-tailed testing)
<table>
<thead>
<tr>
<th>β=</th>
<th>one-tailed α = 0.005 (two-tailed α = 0.01)</th>
<th>one-tailed α = 0.025 (2-tailed α = 0.05)</th>
<th>one-tailed α = 0.05 (two-tailed α = 0.10)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Power</td>
<td>95%</td>
<td>95%</td>
<td>95%</td>
</tr>
<tr>
<td></td>
<td>90%</td>
<td>90%</td>
<td>90%</td>
</tr>
<tr>
<td></td>
<td>80%</td>
<td>80%</td>
<td>80%</td>
</tr>
</tbody>
</table>

Sample size per group needed to detect day-to-day MID (using Guyatt’s responsiveness coefficient)

<table>
<thead>
<tr>
<th>WURSS-21</th>
<th>72</th>
<th>60</th>
<th>47</th>
<th>52</th>
<th>43</th>
<th>32</th>
<th>44</th>
<th>35</th>
<th>25</th>
</tr>
</thead>
<tbody>
<tr>
<td>WURSS-44</td>
<td>64</td>
<td>53</td>
<td>42</td>
<td>47</td>
<td>38</td>
<td>28</td>
<td>39</td>
<td>31</td>
<td>22</td>
</tr>
</tbody>
</table>

WURSS-21 - Sample size per group needed to detect between group AUC differences of:

<table>
<thead>
<tr>
<th>Difference (%)</th>
<th>10%</th>
<th>20%</th>
<th>30%</th>
<th>40%</th>
<th>50%</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2348</td>
<td>1961</td>
<td>1540</td>
<td>1712</td>
<td>1385</td>
</tr>
<tr>
<td></td>
<td>1035</td>
<td>1426</td>
<td>1129</td>
<td>815</td>
<td>815</td>
</tr>
<tr>
<td></td>
<td>1426</td>
<td>1129</td>
<td>815</td>
<td>815</td>
<td>815</td>
</tr>
<tr>
<td></td>
<td>1129</td>
<td>815</td>
<td>815</td>
<td>815</td>
<td>815</td>
</tr>
<tr>
<td></td>
<td>815</td>
<td>815</td>
<td>815</td>
<td>815</td>
<td>815</td>
</tr>
</tbody>
</table>

WURSS-44 - Sample size per group needed to detect between group AUC differences of:

<table>
<thead>
<tr>
<th>Difference (%)</th>
<th>10%</th>
<th>20%</th>
<th>30%</th>
<th>40%</th>
<th>50%</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2787</td>
<td>2328</td>
<td>1828</td>
<td>2033</td>
<td>1644</td>
</tr>
<tr>
<td></td>
<td>1228</td>
<td>1693</td>
<td>1340</td>
<td>967</td>
<td>967</td>
</tr>
<tr>
<td></td>
<td>1340</td>
<td>967</td>
<td>967</td>
<td>967</td>
<td>967</td>
</tr>
<tr>
<td></td>
<td>967</td>
<td>967</td>
<td>967</td>
<td>967</td>
<td>967</td>
</tr>
<tr>
<td></td>
<td>967</td>
<td>967</td>
<td>967</td>
<td>967</td>
<td>967</td>
</tr>
</tbody>
</table>
Every medical intervention has:

- Potential *benefits* (improved symptoms and functions over time) AND
- Potential *harms* (side effects, adverse events, monetary costs, opportunity costs)

What about costs, side effects, and other potential harms?

- MID methods only assess benefit
- Shouldn’t RCTs be designed to detect, (and interpreted in the light of) net benefit, where likely positive effects are sufficient to outweigh potential negative effects?
Sufficiently Important Difference

SID = “the smallest amount of patient-valued benefit that an intervention would require in order to justify associated costs, risks, and other harms”


SID Methodology

- To estimate the magnitude of benefit that patients consider sufficient to justify costs and risks of treatments
- Common cold as model
- **Benefit harm trade-off** interviews were developed, tested & utilized
Benefit Harm Trade-Offs

- Interviews of people with illness
- Portray evidence-based treatment scenarios including benefits, costs, and risks
- Systematically vary one dimension until respondent changes orientation from “would take treatment” to “would not take treatment” (or vice-versa)
Sampling Framework

- People with self-identified colds in Madison, Wisconsin, responding to community advertisement
- At least 1 of 4 symptoms: sneezing, runny nose, nasal obstruction, or sore throat; Jackson score ≥ 3
- A) Enrolled within 48 hours of first symptom, monitored until cold is resolved. Interviewed at intake, when sick, and again at exit, after illness resolved  N=151
- B) Additionally, 162 people with self-identified colds were interviewed by telephone
Benefit Harm Trade Offs

- First, we describe costs, potential side effects, and possible symptomatic benefits. Then we ask:
  - “Would you take this treatment?”
  [Along with qualitative “Why or why not?”]
- Next, we say “On average, colds last about 6 days”
  Then ask, “If this treatment could reduce the length of your cold by 1 day, would you take it?”
Benefit Harm Trade Offs

- If the participant responds “Yes,” we decrease hypothetical duration benefit to “12 hours,” then if still “Yes” to “6 hours,” then if still “Yes” to “Any duration at all.”

- If the participant first responds “No,” we increase hypothetical duration benefit to “2 days,” then “3 days,” then “4 days,” then finally “Any duration at all.”
Tx scenario based on vitamin C evidence

“A 10-cent vitamin pill must be taken 3 times daily for the first 3 days of your cold. There are no significant risks or side effects to this treatment. Severity of symptoms might be reduced by as much as 20%.”

Douglas RM, Chalker EB, Treacy B. Vitamin C for preventing and treating the common cold. *Cochrane Database of Systematic Reviews* 2003.
Tx scenario based on echinacea evidence

“A 50-cent dropperful of an herbal extract must be taken by mouth 3 times each day for the first 3 days of your cold. Side effects are limited to bad taste. Severity of symptoms might be reduced by as much as 30%.”

Tx scenario based on zinc evidence

“A 20-cent lozenge must be dissolved in the mouth every 2 to 3 hours while awake for the first 3 days of your cold. Side effects are limited to bad taste, and occasionally nausea. Severity of symptoms might be reduced by as much as 20%.”

Tx scenario based on evidence regarding pleconaril, an antiviral

“A $2 prescription-only pill must be taken 3 times daily for the first 3 days of the cold. Side effects are unknown. Severity of symptoms might be reduced by as much as 40%.”

Phase 1 - March 2002 to August 2003

WURSS Enrollments by month (Mar'02-Aug'03)
Age, gender & ethnicity of participants

Age range: 18 to 80 years
Mean = 35.5, SD = 14.7
Educational level of participants

Education completed

- Some HS
- HS degree/GED
- Some college
- BA/S
- MA/S
- PhD/Prof. Degree
- Other / No response
- Other / No response

- Other / No response
Income, smoking status

**Household income**
- No response
- >100K/yr
- 75 to <100K/yr
- 50 to <75K/yr
- 25 to <50K/yr
- 15 to <25K/yr
- <15K/yr
- Past smoker
- Current
- Non-smoker

**Smoking status**
- Current 15%
- Past smoker 26%
- Non-smoker 59%
Overall distribution of responses
(460 interviews; 1,840 scenarios)

RESULTS BY INTERVIEW TYPE

- PHONE
- INTAKE
- EXIT

ANY 6 HOURS 12 HOURS 24 HOURS 48 HOURS 72 HOURS 96 HOURS NEVER
Responses to vitamin scenario
Responses to herbal scenario

[Bar chart showing responses to herbal scenario over time.]
Responses to lozenge scenario

![Bar chart showing responses to lozenges over time.](chart.png)
Responses to antiviral scenario

![Graph showing responses to antiviral scenario](image_url)
These results were published as:

Using benefit harm tradeoffs to estimate sufficiently important difference: the case of the common cold.

Next we repeat the model in another cohort, asking about severity reduction instead of duration reduction.
SID severity cohort

- 983 people screened; 253 participated (2004-05)
- N= 91 in person – at beginning & after cold ends (182 interviews)
- N= 162 telephone interviews; in first 5 days of cold
- N= 344 interviews; 1,376 scenarios & responses
- Mean age = 34.8 years (SD 13.3)
- 68% women, 68% white
- 46% reported household income < $25,000/year

Bidding starts at 30% average overall severity reduction

“Severity of symptoms might be reduced by *as much as 30%.*”

Participants answering “Yes” are asked: “Would you take this [treatment] if it were able to reduce severity by 20%?”

If still “Yes,” hypothetical severity reduction is lowered to “10%,” then if still “Yes” reduced to “5%,” and, finally, “any?”

If the first answer was “No,” then severity reduction benefit was increased to “40%,” then if still “No” to “50%,” then “75%,” then “any?”
SID Severity reduction by treatment scenario

Vitamin
Herbal
Lozenge
Antiviral

Any 10% 20% 30% 40% 50% 75% Never

% symptom severity reduction

% sample.
Conclusions

- The SID framework seems valid
- Benefit harm trade-off interviews seem to work
- Response distribution is wide, but characteristic and reproducible
- Severity of illness and socioeconomic status have little impact on response distribution, at least for these people with common cold
Implications

No currently available cold treatment provides the average benefit that most respondents say they would want in order to justify costs and risks.

If similar results are found for other interventions with small effect sizes (e.g. anticholinergics for dementia, antihyperglycemics for type 2 diabetes, antidepressants for depression, statins for preventing heart attack and stroke, cancer screening, etc etc etc... ) there may be profound policy implications.

In terms of benefit harm trade-offs for medical decision-making, **one size does not fit all!**
Questions raised

Should trials be powered to detect SID effect sizes thought to be worthwhile by 50% of population? 75%? 25%? 10%?

Should an intervention be approved by FDA or other regulators when 75%, 90% or 99% of potential users say the benefits don’t justify the harms?

Should practice guidelines (and clinician report cards, bonuses, salaries) be based on effect sizes so small that most people wouldn’t want them?

Should manufacturers be required to assess and report SID?

Should FDA/regulators or non-conflicted health researchers be empowered (funded) to assess and report SID?
Limitations

- Quality of evidence behind scenarios
- Our ability to summarize evidence without bias
- Participants’ ability to weigh several hypothetical benefit and harm domains
- Potential bias arising from self-report of hypothetical intentions, not actual behaviors
- Loads of inter-person variability
Other related papers....
