Methods Articles

Triangulating Patient and Clinician Perspectives on Clinically Important Differences in Health-Related Quality of Life among Patients with Heart Disease

Kathleen W. Wyrwich, Stacie M. Metz, Kurt Kroenke, William M. Tierney, Ajit N. Babu, and Fredric D. Wolinsky

Objective. To use triangulation methodology to better understand clinically important differences (CIDs) in the health-related quality of life (HRQoL) of patients with heart disease.

Data Sources/Study Setting. We used three information sources: a nine-member expert panel, 656 primary care outpatients with coronary artery disease (CAD) and/or congestive heart failure (CHF), and the 46 primary care physicians (PCPs) treating these outpatients. From them, we derived CIDs for the Modified Chronic Heart Failure Questionnaire (CHQ) and the Medical Outcomes Study Short Form 36-Item Health Status Survey, Version 2 (SF-36).

Study Design. The expert physician panel employed Delphi and consensus methods to obtain CIDs. The outpatients received bimonthly HRQoL interviews for 1 year that included the CHQ and SF-36, as well as retrospective assessments of HRQoL changes. Their PCPs assessed changes in the patient's condition at follow-up clinic visits that were linked to HRQoL assessments to determine change over time.

Data Collection/Extraction Methods. Patient- and PCP-assessed changes were categorized as trivial (no change), small, moderate, or large improvements or declines. Moderate or large changes in HRQoL reflect the added risk or investment associated with some treatment modifications. Estimates for each categorization were calculated by finding the mean change scores within anchored change classifications.

Principal Findings. The small CID for the CHQ domains was consistently one to two points using the patient-assessed change categorizations, but small CIDs varied greatly for the SF-36. PCP-assessed changes differed substantially from patient estimates for both the CHQ and SF-36, while the panel-derived estimates were generally larger than those derived from patients.

Conclusions. Triangulation methodology provides a framework for securing a deeper understanding of each informant group's perspective on CIDs for these patient-reported outcome measures. These results demonstrate little consensus and suggest that the derived estimates depend on the rater and assessment methodology.

The reliance on multiple measurement strategies in health services research is motivated by the widely held belief that "the most persuasive evidence comes through a triangulation of measurement processes, as well as through minimizing the error contained in each instrument" (Bowling 2002, p. 202). This process is analogous to the precision that a surveyor or global positioning system uses to accurately pinpoint a specific location based on three information sources. Denzin (1970, p. 300) argues that triangulation can raise researchers "above the personalistic biases that stem from single methodologies. By combining methods and investigators in the same study, observers can partially overcome the deficiencies that flow from one investigator and/or one method."

The interpretation of health-related quality of life (HRQoL) measures, and an individual patient's changes over time on these measures, has received much attention as researchers seek to validate the effectiveness of treatments or interventions on HRQoL (Guyatt et al. 2002). Knowing the amount of change on HRQoL measures that an individual should change so that the observed shift can be considered clinically important, not merely statistically significant, rests at the heart of the HRQoL interpretation issue. Yet, each of the methods used to better understand and interpret clinically important change thresholds for HRQoL measures demonstrates some form of methodological deficiency (Wyrwich and Wolinsky 2000). Therefore, a triangulation of informants and methods could "secure an in-depth understanding of the phenomenon in question" (Denzin and Lincoln 1994, p. 2).

The first informant group for our triangulation study was an expert panel of physicians familiar with the use of a disease-specific and a generic HRQoL

Address correspondence to Kathleen W. Wyrwich, Ph.D., Department of Research Methodology, Saint Louis University, St. Louis, MO. Stacie M. Metz, M.P.H., Ph.D., is with the Department of Health, West Chester University, West Chester, PA. Kurt Kroenke, M.D., and William M. Tierney, M.D., are with the Regenstrief Institute for Health Care, Indianapolis, IN, and the Indiana University School of Medicine, Indianapolis, IN. Ajit N. Babu, M.B.B.S., M.P.H., is with the Veterans Affairs Medical Center, St. Louis, MO, and the Amrita Institute of Medical Sciences, Kochi, India. Fredric D. Wolinsky, Ph.D., is with the Department of Health Management and Policy, University of Iowa, Iowa City, IA; and the Center for Research in the Implementation of Innovative Strategies in Practice, Iowa City VA Health Care System, Iowa City, IA; the Saint Louis University School of Public Health, St. Louis, MO.

instrument for patients with heart disease. Using consensus methods, the expert panelists recommended threshold levels for small, moderate, and large clinically important differences (CIDs) in each scale or domain within the HRQoL instrument (Wyrwich et al. 2004, 2005). Our second informant group was 656 patients with heart disease, either coronary artery disease (CAD) or congestive heart failure (CHF), who participated in bimonthly HRQoL interviews for 1 year. The third informant group was the 46 primary care physicians (PCPs) of these patients with heart disease who assessed their patients' health at baseline, as well as changes in their condition when these patients returned for office visits during their participation in the 1-year study. These informant groups considered not only small-but-important differences in HRQoL over time, but also moderate and large HRQoL changes. These larger improvement and decline thresholds go beyond the small or minimally important differences (MID) or minimal CIDs (MCID) and reflect the need for standard measures of expected change magnitudes for which treatment modifications are warranted. The resulting small, moderate, and large CIDs for improvements and declines determined from these different informant groups provide not only a better understanding to interpret individual HRQoL changes in a clinical decision-making context, but also greater opportunities to use these findings toward improved patient and clinician communication.

METHODS

Study Goals and Instruments

Our objective for this study was to develop a methodology that incorporated information from three groups—expert physicians, patients, and the clinicians treating these patients—on how much change in a HRQoL measure is needed for that change to be considered a trivial, small, moderate, or large clinically important improvement or decline. The HRQoL instruments were the Modified Chronic Heart Failure Questionnaire (CHQ) (Guyatt et al. 1989; Wolinsky et al. 1998) and the Medical Outcomes Study Short Form 36-Item Health Status Survey, Version 2 (SF-36, Version 2.0) (Ware, Kosinski, and Dewey 2000). The CHQ contains three domains—activities, emotional function, and fatigue—with five, seven, and four items, respectively. All item response categories have a seven-point scale where 1 indicates the worst and 7 the best HRQoL response. The original CHQ was modified to include shortness of breath *or chest pain* symptoms during important activities so that this measure captured the HRQoL of patients with CHF *and/or CAD*. The SF-36 contains

eight scales that measure physical functioning (10 items), role physical (four items), general health (five items), vitality (four items), bodily pain (two items), social functioning (two items), role emotional (two items), and mental health (five items). To adjust for the differing number of items and response categories in each scale, all scale scores are reported on a 0 (worst) to 100 (best) metric. Version 2.0 was used because of its improved item wordings, response categories, and scoring increments (Ware, Kosinski, and Dewey 2000).

Expert Panel

We convened a nine-member expert physician consensus panel using the following process. First, we performed a Medline database search from 1995 to 1999 to find relevant literature on HRQoL changes in patients with heart disease measured using the CHQ and/or the SF-36. We reviewed the articles that our search yielded and prepared a list of physician authors from these ($n = 39$). Subsequently, we sent invitations to the listed physicians seeking their interest and availability to serve on a consensus panel to engage in determining the thresholds for CIDs that identify small, moderate, and large improvements and declines in the three domains of the CHQ and the eight scales of the SF-36. Based on geographic and specialty representation, a diverse nine-member panel was selected.

The panel pursued their task using two modified Delphi rounds to coalesce their reasoning for specific point estimates for the CIDs. Additionally, the panelists were asked to seek consensus on the criterion for selecting potential CAD and CHF patients using the available electronic database systems at our primary care clinic sites (the Indiana University School of Medicine and the St. Louis Veteran's Affairs Medical Center). After the results of the second Delphi round were distributed to panelists, they attended a half-day face-to-face meeting to discuss and debate their individual conclusions and seek consensus.

During the consensus process, our expert physician panel adopted a method for determining CIDs that centered on the use of *state changes*, defined to be the amount of change in a domain or scale score that results for one shift up or down on the response category for only one item (Wyrwich et al. 2004). For the domains of the CHQ, all *state change values* are one point; the SF-36 yields *state change values* of different magnitudes across the eight scales due to the varied number of items and response categories in each scale (Table 1). In addition, it is also important to note that for the general health and bodily pain scales of the SF-36, the reported state change represents the magnitude of one

Table 1: Expert Panel Results for Clinically Important Differences for the Modified Chronic Heart Failure Questionnaire (CHQ) and SF-36, Version 2.0

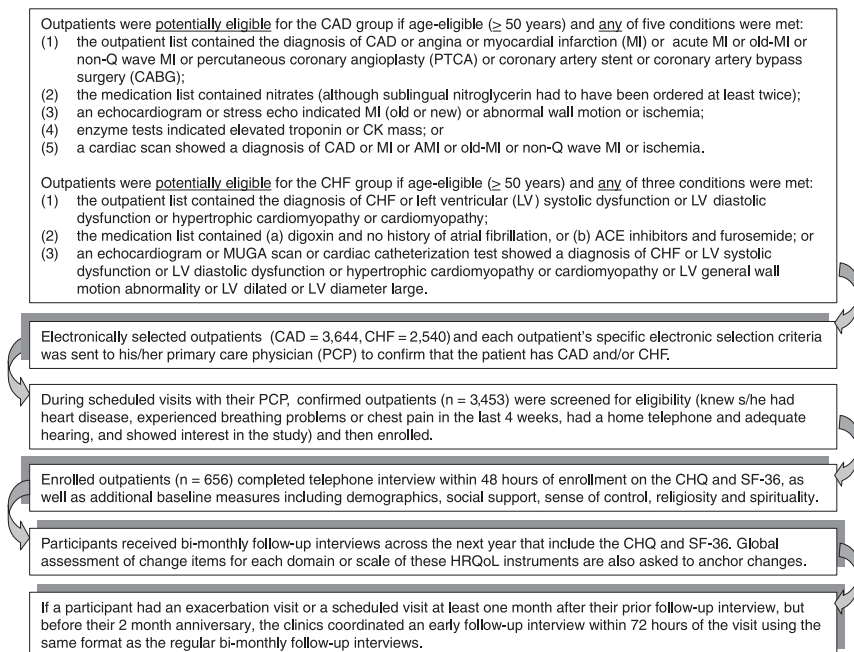|			Thresholds for Change in Scores		
HQRoL Measure	Number of Items	Points per One Category Change	Minimal Change	Moderate Change	Large Change
CHQ domains					
Activities	5	1	4	8	12
Fatigue	4	1	3	6	10
Emotional function	7	1	5	10	15
SF-36 scales					
Physical function	10	5	15	25	35
Role physical	4	6.25	18.75	31.25	50
Bodily pain	2	10	20	40	60
General health	5	5	15	30	45
Vitality	4	6.25	18.75	37.5	56.25
Social function	2	12.5	25	50	75
Role emotional	3	8.33	16.7	33.3	50
Mental health	5	5	15	30	45

HQRoL, health-related quality of life.

response category shift for *most* items on the scale. Differing item weights within these two scales, however, can yield individual change scores on these scales as low as one point on the 0–100 range. We defined consensus a priori as agreement among at least seven members, and this was reached for both the CID thresholds (Table 1) and the electronic selection criterion (Figure 1) (Wyrwich et al. 2004).

Patients

Patient data for determining CIDs for the domains of the CHQ and the SF-36 scales were solicited through telephone interviews with 656 clinic outpatients having heart disease who enrolled in this study. A trace of the selection, enrollment, and participation process is depicted in Figure 1. Using the electronic selection criterion established by our expert physician consensus panel, we created a list for each potential enrollee of the indicator(s) that triggered selection. The outpatient's PCP reviewed this list to confirm that the outpatient did indeed have CAD and/or CHF. Each confirmed outpatient was then approached at her/his next scheduled clinic appointment for a brief screening, and if she/he passed the screen and was interested in the study, the patient was enrolled. This in-clinic enrollment process also included the selection of five

Figure 1:   Patient Eligibility, Selection, Enrollment, and Follow-up
Participation Process

Outpatients were <u>potentially eligible</u> for the CAD group if age-eligible (≥ 50 years) and <u>any</u> of five conditions were met:
(1)   the outpatient list contained the diagnosis of CAD or angina or myocardial infarction (MI) or acute MI or old-MI or non-Q wave MI or percutaneous coronary angioplasty (PTCA) or coronary artery stent or coronary artery bypass surgery (CABG);
(2)   the medication list contained nitrates (although sublingual nitroglycerin had to have been ordered at least twice);
(3)   an echocardiogram or stress echo indicated MI (old or new) or abnormal wall motion or ischemia;
(4)   enzyme tests indicated elevated troponin or CK mass; or
(5)   a cardiac scan showed a diagnosis of CAD or MI or AMI or old-MI or non-Q wave MI or ischemia.

Outpatients were <u>potentially eligible</u> for the CHF group if age-eligible (≥ 50 years) and <u>any</u> of three conditions were met:
(1)   the outpatient list contained the diagnosis of CHF or left ventricular (LV) systolic dysfunction or LV diastolic dysfunction or hypertrophic cardiomyopathy or cardiomyopathy;
(2)   the medication list contained (a) digoxin and no history of atrial fibrillation, or (b) ACE inhibitors and furosemide; or
(3)   an echocardiogram or MUGA scan or cardiac catheterization test showed a diagnosis of CHF or LV systolic dysfunction or LV diastolic dysfunction or hypertrophic cardiomyopathy or cardiomyopathy or LV general wall motion abnormality or LV dilated or LV diameter large.

Electronically selected outpatients (CAD = 3,644, CHF = 2,540) and each outpatient's specific electronic selection criteria was sent to his/her primary care physician (PCP) to confirm that the patient has CAD and/or CHF.

During scheduled visits with their PCP, confirmed outpatients (n = 3,453) were screened for eligibility (knew s/he had heart disease, experienced breathing problems or chest pain in the last 4 weeks, had a home telephone and adequate hearing, and showed interest in the study) and then enrolled.

Enrolled outpatients (n = 656) completed telephone interview within 48 hours of enrollment on the CHQ and SF-36, as well as additional baseline measures including demographics, social support, sense of control, religiosity and spirituality.

Participants received bi-monthly follow-up interviews across the next year that include the CHQ and SF-36. Global assessment of change items for each domain or scale of these HRQoL instruments are also asked to anchor changes.

If a participant had an exacerbation visit or a scheduled visit at least one month after their prior follow-up interview, but before their 2 month anniversary, the clinics coordinated an early follow-up interview within 72 hours of the visit using the same format as the regular bi-monthly follow-up interviews.

activities that were important to the outpatient but limited due to shortness of breath or chest pain, which are primary HRQoL symptoms in persons with CHF or CAD.

Within 72 hours of the clinic enrollment, participating patients received a baseline telephone interview that solicited responses to demographics and psychosocial measures, as well as the CHQ and the SF-36. A tentative bi-monthly follow-up interview date was also set at the end of each interview, and participants continued to receive HRQoL interviews every other month over 1 year. However, if an outpatient returned to the clinic for either a scheduled or exacerbation visit with their PCP on a date that was at least 1 month after their latest interview but before their bimonthly interview anniversary date, this triggered an early follow-up interview within 48 hours of their office visit. These early follow-up HRQoL interviews provided an opportunity to compare the patient's perceived changes in HRQoL with the perceptions of change given at the same time by their PCP.

Transition Rating Index

Each follow-up interview included the CHQ and SF-36 measures, as well as retrospective items measuring patient-perceived changes for each HRQoL domain and scale. For example, for the fatigue domain, patients were asked, "Since your last interview on ⟨prior interview date⟩ has there been a change in your fatigue? Is it better, worse or about the same?" Those patients who chose "about the same" were given a transition rating index equal to 0, but those who replied "better" or "worse" were subsequently asked, "How much better (or worse)?" The transition rating index for this item ranged from ± 1 (hardly any better/worse) to ± 7 (a very great deal better/worse). These indices were later grouped in the following manner: -1, 0, and 1 represented "no change"; ± 2 and ± 3 corresponded to small improvements $(+)$ or declines $(-)$; ± 4 and ± 5 denoted moderate improvements or declines; and ± 6 and ± 7 signified large improvements or declines.

PCPs

The anchors for the third leg of our triangulated CID estimates were provided by each enrollee's PCP (31 PCPs in Indianapolis and 15 in St. Louis). As mentioned, these PCPs had confirmed the existence of either CHF and/or CAD in their respective patients after the initial electronic selection process. In addition, each PCP completed a baseline assessment of her/his patients at enrollment that identified: the likelihood of hospitalization or death in the next year due to heart disease; how the patient compared with others that the PCP treats for heart disease; and whether any tests, medications, or referrals to specialists had been ordered in the past year due to the enrollee's heart disease.

If during their year of follow-up an enrolled outpatient returned to the clinic for either a scheduled or exacerbation visit with their PCP on a date that was at least 1 month after their latest interview but before their bimonthly interview anniversary date, the PCP also completed a follow-up assessment of change (none, small, moderate, or large improvement or decline) in the patient's heart disease since enrollment or the latest early follow-up visit. It is important to note that this clinically important change assessment consistent of only one item where the PCP evaluated change (no change, improvement or decline in heart disease and whether the improvement or decline was small, moderate or large) to accommodate real-time assessment during the PCPs' busy clinic schedules. These PCP change assessments were used to classify corresponding patient change scores in each HRQoL measure. In addition, if the PCP indicated change, she/he was asked to list any orders for laboratory

tests, additional medications, or referrals to specialists due to heart disease ordered during the linked clinical encounter.

Statistical Procedures

We began by stratifying our patient sample into those with CHF, CAD, or both diagnoses to see if CIDs would differ based on the type of heart disease. Within these three strata, the CID estimates for patient and PCP assessments were calculated separately using mean change scores within change classifications. That is, we classified each patient's change intervals (e.g., from fourth to fifth follow-up, or baseline to first follow-up) using the transition rating index (patients) or the PCP change assessment (if an early follow-up visit occurred). The change score (Time 2 to Time 1) for the corresponding interval on each HRQoL measure was also calculated. We then computed the average change scores within each classification for our CID estimates (Jaeschke, Singer, and Guyatt 1989).

The CIDs for each of the three heart disease patient strata (CAD, CHF, or both CAD and CHF) were compared. Because the CIDs were similar, we further investigated the appropriateness of pooling patients from all three strata by conducting confirmatory factor analyses of each CHQ domain and SF-36 scale within each stratum to examine factorial invariance. Because there was little invariance, and given the observed similarities in strata-specific CIDs, pooling the data was appropriate and created a simpler approach for evaluating patient-perceived and PCP-rated change classifications.

Finally, to better understand the level of agreement between the patients' and PCPs' classification of change for the linked office visits, we conducted a weighted κ analysis. Across the seven levels of change (large decline to large improvement) patient and PCP ratings were cross-tabulated and quadratic weights were applied to empirically evaluate agreement between these two rater groups.

RESULTS

Expert Panel

Applying the state change concept, the expert panel considered how many shifts they considered it took to reflect trivial, small, moderate, and large changes, and then computed the resulting change score for each HRQoL measure. For the sake of simplicity, CIDs for corresponding improvements and declines were judged by the panels to be at the same absolute magnitude.

Moreover, the panel applied a simple doubling and the tripling of the small CIDs for estimating moderate and large CIDs, respectively, for most CHQ domains and SF-36 scales (Table 1).

These panelists' consensus results for the SF-36 were generally larger than the CIDs agreed on by very similar asthma and chronic obstructive pulmonary disease (COPD) expert panels convened separately but within 1 month of the heart disease panel meeting (Wyrwich et al. 2005). Likewise, this panel's CID estimates for the activities and fatigue domains on the CHQ were larger than those reported by the COPD panel for these same domains, which are identically measured in the Chronic Respiratory Disease Questionnaire (CRQ) (Guyatt et al. 1987). In contrast, the CAD/CHF and COPD panels had equivalent results for the emotional functioning domain that is present in both disease-specific instruments (CHQ and CRQ).

Patients

Table 2 presents the demographic and psychosocial characteristics, as well as baseline HRQoL scores, of our sample of outpatients with heart disease. As shown, participants in the CAD, CHF, and both CAD and CHF strata had very similar demographic distributions. That is, over half of each outpatient class was older than 65 years, and nearly half did not have a high school education or an annual household income of $20,000 or more. Few (<18 percent) worked for wages, and most (>65 percent) had fair or poor self-reported health, and had smoked (over 77 percent). We observed a notable difference, however, in the gender and race percentages of the CHF-only class—this group included more women (52.5 percent) and more black (43.8 percent) outpatients.

Patients and PCP Ratings of Change

Among the 656 outpatients completing baseline HRQoL interviews, 3,336 bimonthly follow-up interviews were completed over 1 year. The left side of Table 3 presents the number of patient-perceived changes and the magnitude (small, moderate, or large) and direction of these changes (improvement or decline). Of the 3,336 bimonthly follow-up interviews, 435 were linked to PCP visits that occurred within 48 hours of administration of the follow-up questionnaire. The right half of Table 3 shows the sample sizes of the PCP-categorized change from these linked interviews in the PCP columns. Most patient and PCP assessments indicate no change, and all PCP-reported improvement or decline classification cells had very few patients. This limits the stability and credibility of the CID estimates for those cells. Moreover, limited

Table 2:   Baseline Demographics and Health Characteristics among Patients with Heart Disease Attending Midwest Medical Centers (August 2000 to November 2001)

	CAD Only N (%) or Mean (SD)	CHF Only N (%) or Mean (SD)	Both CAD/CHF N (%) or Mean (SD)	All Heart Patients N (%) or Mean (SD)
Age*				
<55	44 (11.1)	11 (13.8)	14 (7.9)	69 (10.5)
55–64	127 (31.9)	25 (31.3)	44 (24.7)	196 (29.9)
65–74	135 (33.9)	33 (41.3)	62 (34.8)	230 (35.1)
≥ 75	92 (23.1)	11 (13.8)	58 (32.6)	161 (24.5)
Gender				
Female	81 (20.4)	42 (52.5)	40 (22.5)	163 (24.8)
Male	317 (79.6)	38 (47.5)	138 (77.5)	493 (75.2)
Race				
White	274 (68.8)	40 (50.0)	116 (65.2)	430 (65.5)
Black	109 (27.4)	35 (43.8)	54 (30.3)	198 (30.2)
Other	15 (3.8)	5 (6.3)	8 (4.5)	28 (4.1)
Education				
<High school	161 (40.5)	38 (47.5)	79 (44.4)	278 (42.4)
High school	116 (29.1)	21 (26.3)	47 (26.4)	184 (28.0)
>High school	121 (30.5)	21 (26.3)	52 (29.2)	194 (29.6)
Annual household income				
<$10,000	61 (15.3)	23 (28.8)	35 (19.7)	119 (18.1)
$10,000–14,999	72 (18.1)	16 (20.0)	32 (18.0)	120 (18.3)
$15,000–19,999	86 (21.6)	15 (18.8)	41 (23.0)	142 (21.6)
$20,000–24,999	67 (16.8)	11 (13.8)	28 (15.7)	106 (16.2)
≥ $25,000	96 (24.1)	12 (15.0)	37 (20.8)	145 (22.1)
Unknown	16 (4.0)	3 (3.8)	5 (2.8)	24 (3.7)
Employment status				
Employed for wages	71 (17.8)	12 (15.0)	15 (8.4)	98 (14.9)
Unable to work	120 (30.2)	29 (36.3)	59 (33.1)	208 (31.7)
Retired	192 (48.2)	38 (47.5)	92 (51.7)	322 (49.1)
Other	15 (3.8)	1 (1.3)	112 (6.8)	28 (4.3)
Patient-reported health				
Excellent	9 (2.3)	2 (2.5)	2 (1.1)	13 (2.0)
Very good	28 (7.0)	6 (7.5)	15 (8.4)	49 (7.5)
Good	109 (27.4)	19 (23.8)	40 (22.5)	168 (25.6)
Fair	171 (43.0)	30 (37.5)	71 (39.9)	272 (41.5)
Poor	81 (20.4)	23 (28.8)	50 (28.1)	154 (23.5)
Smoking status				
Never	67 (16.8)	18 (22.5)	25 (14.0)	110 (16.8)
<20 peak years	134 (33.7)	28 (35.0)	55 (30.9)	217 (33.1)
≥ 20 peak years	197 (49.5)	33 (41.3)	98 (55.1)	328 (50.0)
Unknown	–	1 (1.3)	–	1 (0.2)
CHQ domains				
Activities	22.0 (5.9)	20.7 (6.6)	21.8 (5.8)	21.8 (6.0)
Emotional function	31.8 (8.4)	30.7 (8.3)	31.8 (8.6)	31.7 (8.4)
Fatigue	14.4 (4.6)	13.1 (5.1)	14.2 (4.8)	14.2 (4.7)

Table 2. *Continued*

	CAD Only N (%) or Mean (SD)	CHF Only N (%) or Mean (SD)	Both CAD/CHF N (%) or Mean (SD)	All Heart Patients N (%) or Mean (SD)
SF-36 domains				
Physical functioning	40.6 (23.9)	35.2 (22.6)	36.2 (22.0)	38.8 (23.3)
Role physical	50.9 (26.8)	45.1 (25.1)	47.0 (26.2)	49.1 (26.5)
Bodily pain	48.8 (24.2)	47.6 (24.2)	50.0 (24.2)	48.9 (24.2)
Vitality	43.6 (20.3)	42.7 (19.4)	43.1 (21.4)	43.4 (20.5)
General health	44.2 (20.5)	42.0 (19.7)	40.7 (19.0)	43.0 (20.0)
Social functioning	67.7 (27.6)	60.3 (29.1)	62.1 (29.7)	65.3 (28.4)
Role emotional	71.7 (26.1)	70.5 (23.7)	69.0 (27.4)	70.8 (26.2)
Mental health	65.3 (21.1)	65.9 (20.5)	64.7 (22.1)	65.2 (21.3)

*Measured in years, mean age 67.6 (SD = 9.0), range 50–91 for all heart disease patients.

CAD, coronary artery disease; CHF, congestive heart failure.

missing data for the measurement of change in some patients' CHQ activities domain slightly reduces the number of patients with these PCP-rated changes compared with other domains and scales.

Of the 65 linked encounters where PCPs reported an improvement or decline, most (62 percent) resulted in at least one clinical action: a change in medication (45 percent), ordered laboratory tests or procedures (42 percent), or referral to a specialist (23 percent). Only 10 of the 42 patients with PCP-rated declines did not have any of these clinical actions. We understand that not all disease-related changes will necessarily result in a modification or reevaluation of treatment, yet we are encouraged by these clinical action-related reports for patients with PCP-assessed changes.

The CID estimates calculated using the mean of all heart disease patients' change scores in each respective categorization (Table 3) are given in Table 4. Patient-reported small improvements and declines on the domains of the CHQ, where the a state change is one point, yielded mean changes of one to two points in the expected direction and a general trend across magnitudes of change was observed across most CHQ domains. PCP estimates for small changes using mean change scores were less consistent and often moved in the wrong direction (i.e., mean increases in patient CHQ domains scores when the PCPs denoted a small decline). Mean PCP-rated estimates for moderate and large changes, both improvements and declines, displayed more volatile ranges likely due to the small number of change scores averaged within each cell.

Most SF-36 mean patient estimates for changes moved in the correct direction with positive mean change score estimates for improvements and

Table 3: Sample Size within Change Categorizations by Patient and Primary Care Physician Reports of Change on the CHQ Domains and SF-36 Scales

HRQoL Measure	Patient-Perceived Changes							PCP-Perceived Changes						
	Large Decline	Moderate Decline	Small Decline	No Change	Small Improve	Moderate Improve	Large Improve	Large Decline	Moderate Decline	Small Decline	No Change	Small Improve	Moderate Improve	Large Improve
CHQ														
Activities	67	154	251	2,053	244	272	132	2	18	17	322	11	6	3
Emotional	66	147	188	2,101	280	297	166	3	20	19	370	12	7	4
Fatigue	84	212	333	2,037	232	244	104	3	20	19	370	12	7	4
SF-36														
Physical functioning	110	198	240	2,230	205	196	70	3	20	19	370	12	7	4
Role physical	113	200	244	2,282	153	180	76	3	20	19	370	12	7	4
Bodily pain	163	302	364	1,943	168	176	124	3	20	19	370	12	7	4
General health	116	218	391	1,957	243	209	106	3	20	19	370	12	7	4
Vitality	111	221	371	1,993	248	218	84	3	20	19	370	12	7	4
Social functioning	100	173	184	2,399	150	152	84	3	20	19	370	12	7	4
Role emotional	78	137	185	2,417	170	164	90	3	20	19	370	12	7	4
Mental health	65	123	171	2,328	202	234	124	3	20	19	370	12	7	4

HQRoL, health-related quality of life; CHQ, Chronic Heart Failure Questionnaire.

negative estimates for declines. However, with the exception of declines in the role physical, social functioning, role emotional, and mental heath scales, the magnitude of nearly all of the small mean change estimates across all SF-36 scales did not reach the change associated with *state change values*, the amount of change in a scale score that occurs when shifting up or down one response category of only one item (see Table 1). PCP-perceived change estimates on the SF-36 scales for a small decline were larger than those of patients and many were negative in the small improvement classification. Other PCP cells for the SF-36 scale changes continued to demonstrate the instability of small sample sizes in the moderate and large change classifications.

Finally, weighted κ values empirically reflect the lack of agreement between the patient and PCP rating groups for the 435 linked change estimates. All of the weighted κ-values were poor, ranging from 0.23 for the CHQ activities domain down to 0.09 for the SF-36 bodily pain scale.

DISCUSSION

This rather complex triangulation study to derive CIDs estimates for heart disease patients in the domains of the CHQ and scales of the SF-36 followed Denzin's approach for combining informant groups and methods (Denzin 1970). We specifically targeted the three primary stakeholder groups: expert physicians, outpatients with heart disease, and the PCPs who care for those outpatients. Ultimately, however, the question comes down to this: do the resulting CID estimates in Tables 1 and 4 allow us to "overcome the deficiencies that flow from one investigator and/or one method," and if so, how? (Denzin 1970, p. 300). To answer that question, we must begin by reviewing the information obtained from each informant group.

Our expert physician panel used methods that may be subject to coercion and/or domination by one or more members (Stasser, Kerr, and Davis 1989), although we did not witness any evidence of this. Second, although prior patient-based measurement studies were referenced by panelists throughout the Delphi rounds and consensus meeting, the panelists made their final recommendation without reference to specific patient data, but through the use of the state change concept. Moreover, the simplicity of symmetry and even increments of change is attractive and easily communicated, yet "real" patient data may or may not behave in such a predictable pattern. In addition, we cannot know exactly what criteria were used by individual experts to assess their final and consensus values for the magnitude of each improvement or decline. It is important to note that our proposal for this study

Table 4:   Mean Change Scores Using Patient and Primary Care Physician Reports of Change on the CHQ Domains and SF-36 Scales (Scores Rounded to the Nearest Integer Values Due to Space Constraints)

HRQoL measure	Patient-Perceived Changes (n = 3,336)							PCP-Perceived Changes (n = 435)						
	Large Decline	Moderate Decline	Small Decline	No Change	Small Improve	Moderate Improve	Large Improve	Large Decline	Moderate Decline	Small Decline	No Change	Small Improve	Moderate Improve	Large Improve
CHQ														
Activities	−4	−3	−1	0	2	2	3	−9	3	1	2	−1	7	16
Emotional	−4	−2	−2	0	1	2	3	−13	3	−6	1	−2	−1	9
Fatigue	−2	−2	−1	0	2	2	2	−6	2	−4	1	0	3	12
SF-36														
Physical functioning	−6	−6	−2	0	2	6	7	−20	1	−4	1	5	16	40
Role physical	−4	−7	−10	1	4	11	7	−50	5	−12	3	−5	6	−63
Bodily pain	−11	−8	−4	2	6	7	5	−40	−2	−14	0	−3	8	16
General health	−3	−3	−2	−1	3	5	2	−20	5	−6	−2	−7	11	25
Vitality	−7	−7	−3	1	3	6	9	−13	8	−16	2	−3	5	25
Social functioning	−16	−9	−10	1	3	9	5	−50	0	−17	0	4	−22	0
Role emotional	−7	−8	−8	0	3	7	7	−42	5	−8	1	3	6	0
Mental health	−16	−10	−7	1	4	6	5	−20	4	−12	1	−6	−1	20

HQRoL, health-related quality of life; CHQ, Chronic Heart Failure Questionnaire.

stated a priori that expert panel-based CID estimates would be considered "of lowest evidence value because they do not reflect actual encounters between patients and their primary care physicians" (p. 102, Wolinsky et al., R01 HS10234).

The second group of informants—outpatients with heart disease—provided the most important source of data for our triangulation. However, their bimonthly cross-sectional and retrospective measurements are also subject to error. The retrospective change items for each CHQ domain and SF-36 scale ("Since your last interview on 'prior interview date,' has there been a change in your fatigue? Is it better, worse or about the same?") ask participants to revisit the day of their last interview, remember their HRQoL dimensional state (e.g., How much energy did I have at that time?) and then compare it with their current HRQoL state. These global assessments of change have long-been reported to be biased toward the patient's current health rating and may not reflect their health status before 2 months.

However, we had little (< 1 percent) incomplete data for these retrospective-anchoring items. Of course, as Table 3 demonstrates, most retrospective comparison responses were "about the same" and, indeed, this was the "easiest" manner to reply. If an interviewee chose "better" or "worse," she/he then had an extra follow-up item that solicited how much better or worse on a seven-point scale. However, nearly 33 percent of the retrospective change item responses did reflect patient-perceived changes. Our study design attempted to improve the accuracy of these responses through the use of a memory marker. Interviewers solicited an event or statement from respondents at the end of each interview that would help the respondent to remember that particular day at the next interview and noted it in the interview database. Examples of the memory markers include: "I went to church with my niece today" and "I dug up the garden for planting." These statements were then read back to the interviewees before they responded to the retrospective change items at their next interview. Although participants have expressed their delight in hearing their own descriptions of what was memorable on the prior day repeated back to them 2 months later, we have no empirical evidence that this practice decreased the well-known error associated with retrospective change assessments. Moreover, both our patients and PCPs provided repeated assessments, and therefore our ratings are not statistically independent. However, corrections for this statistical dependence, as well as the nesting of patients within PCPs, would have an effect on the standard deviation of the associated mean change score thresholds, but not on the value of the mean point estimates themselves. Thus, we have considerable confidence in the validity of the patient-based CIDs.

Measurements from the third-informant group reflected PCP assessments of change in the patients' heart disease, and these were then used to anchor change for all HRQoL measures. Unfortunately, this global item is not consistent with the specific domain change items used by patients. An alternative would have been to ask PCPs to separately assess changes in activities, fatigue, emotional functioning, physical functioning, role physical, bodily pain, etc., at each linked follow-up visit with each participant. This rather long list of relevant HRQoL dimensional changes would have burdened our PCPs and the busy practice settings where they treated our enrollees, and that would have made this study of clinically significant change unworkable. Nonetheless, it is clear that our PCP assessments reflect physician-perceived changes in the patients' heart disease, and that construct does not directly correspond to any specifically measured dimension of HRQoL in this study. Although this clinically informed evaluation is important in understanding CIDs, it is evident that PCPs are generally gauging a different theoretical construct than that which patients are evaluating (Detmar et al. 2001).

Those limitations notwithstanding our low-weighted κ results for seemingly related areas, like the CHQ activities domain ($\kappa = 0.23$) or the SF-36 physical functioning scale ($\kappa = 0.14$), demonstrate the great need for enhanced and improved dialogue between patients and their PCPs to improve clinical encounters and clinical decision making. It is possible that PCPs may have based their assessments on objective findings (e.g., evidence of CHF worsening on physical examination) or changes sufficient to trigger alterations in treatment, yet different changes are leading to patient-perceived improvements or declines, such as the inconvenience of moving to a new home with no stairs or selling one's farm (Velikova et al. 2004). Whether this is the explanation or others will surface in these necessary discussions, it is the primary result or insight gained from this triangulation process.

It is also important to note that our results for small patient-perceived changes reflect an average change score on each CHQ domain that meets or exceeds at least one state change. Results from the SF-36 did not perform as strongly. Indeed, they often yielded mean values that were smaller than a state change value.

Unlike the use of triangulation by a surveyor or navigator, the results from this study elucidate how different information sources and methods do not and should not necessarily point to a single best estimate. Instead, we seek to use these data to better understand the three stakeholders—experts, heart disease patients, and their PCPs—and their approach to the daunting challenge of estimating CID thresholds for patient-reported outcomes (PROs). We also seek to

provide a process that others can employ to determine CIDs for PROs. Toward that goal, we believe that our results demonstrate that it is very difficult in cohort studies of chronic disease to obtain sufficiently large enough samples of patient- and PCP-perceived "changers" at the moderate or large level for either improvements or declines. Instead, clinical trials involving effective interventions might be required, although this design would better assess moderate to large improvements rather than declines. Second, these results and others indicate that the SF-36 may not have the sensitivity needed to capture individual-level changes in a consistent manner as to yield a stable CID estimate. Therefore, we recommend the use of disease-specific instruments to demonstrate important HRQoL changes if the instrument also strives to capture those often overlooked mental or psycho-social domains of patient's health and well being.

A recent Food and Drug Administration report, *Innovation or Stagnation: Challenges and Opportunities on the Critical Path to New Medical Products*, focused on stemming the tide of products that are delayed from reaching those consumers who can benefit (Food and Drug Administration 2004). This document speaks directly to the need for community consensus between health professionals and patients "on appropriate outcome measures and therapeutic claims" (p. 24).

Beyond this FDA recommendation, the HRQoL of patients with heart disease should be a primary concern for their health care providers. Hence, awareness and appreciation of the multiple perspectives and the uniqueness of each perspective is necessary for assessing the impact of potential therapeutic interventions on maintaining and/or improving health. Like blind men and the elephant, all the points of view are essential. Dialogue, measurement, and respect for each of these stakeholder's perspectives are necessary to secure an in-depth understanding and eventually achieve an informed community consensus on the magnitude of an important change over time in PRO measures.

ACKNOWLEDGMENT

REFERENCES

Bowling, A. 2002. *Research Methods in Health: Investigating Health and Health Services Research*. Buckingham, UK: Open University Press.

Denzin, N. 1970. "Strategies of Multiple Triangulation." In *The Research Act in Sociology: A Theoretical Introduction to Sociological Method*, edited by N. Denzin, pp. 297–313. New York: McGraw-Hill.

Denzin, N. and Y. Lincoln. 1994. *Handbook of Qualitative Research.* Thousand Oaks, CA: Sage.

Detmar, S., M. Muller, L. Wever, J. Schornagel, and N. Aaronson. 2001. "Patient–Physician Communication during Outpatient Palliative Treatment Visits: An Observational Study." *Journal of the American Medical Association* 285: 1351–7.

Food and Drug Administration. 2004. *Innovation or Stagnation: Challenges and Opportunities on the Critical Path to New Medical Products.* Washington, DC: U.S. Department of Health and Human Services.

Guyatt, G., L. Berman, M. Townsend, S. Pugsley, and L. Chambers. 1987. "A Measure of Quality of Life for Clinical Trials in Chronic Lung Disease." *Thorax* 42: 773–8.

Guyatt, G., S. Nogradi, S. Halcrow, J. Singer, M. Sullivan, and L. Fallen. 1989. "Development and Testing of a New Measure of Health Status for Clinical Trials in Heart Failure." *Journal of General Internal Medicine* 4: 101–7.

Guyatt, G., D. Osoba, A. Wu, K. Wyrwich, G. Norman, and T. C. S. C. M. Group. 2002. "Methods to Explain the Clinical Significance of Health Status Measures." *Mayo Clinic Proceedings* 77: 371–83.

Jaeschke, R., J. Singer, and G. Guyatt. 1989. "Measurement of Health Status: Ascertaining the Minimal Clinically Important Difference." *Controlled Clinical Trials* 10: 407–15.

Stasser, G., N. Kerr, and J. H. Davis. 1989. "Influence Processes and Consensus Models in Decision-Making Groups." In *Psychology of Group Influence*, 2d edition, edited by P. Paulus. Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Velikova, G., L. Booth, A. B. Smith, P. M. Brown, P. Lynch, J. M. Brown, and P. J. Selby. 2004. "Measuring Quality of Life in Routine Oncology Practice Improves Communication and Patient Well-Being: A Randomized Controlled Trial." *Journal of Clinical Oncology* 22 (4): 714–2.

Ware, J., M. Kosinski, and J. Dewey. 2000. *How to Score Version Two of the SF-36 Health Survey.* Lincoln, RI: QualityMetric Inc.

Wolinsky, F. D., K. W. Wyrwich, N. Nienaber, and W. Tierney. 1998. "Generic vs. Disease-Specific Health Status Measures: An Example Using Coronary Artery Disease and/or Congestive Heart Failure Patients." *Evaluation and the Health Professions* 21: 216–43.

Wolinsky, F. D., K. W. Wyrwich, W. M. Tierney, A. N. Babu, and K. Kroenke. "Identification of clinically relevant changes in HRQoL" (Grant by Agency for Healthcare Research and Quality, Rockville, MD, 1999–2004, R01-HS10234).

Wyrwich, K. W., J. Spertus, K. Kroenke, W. Tierney, A. Babu, and F. D. Wolinsky. 2004. "Clinically Important Differences in Health Status of Life for Patients with Heart Disease: An Expert Panel Report." *American Heart Journal* 147: 615–22.

Wyrwich, K. W., W. Tierney, A. Babu, K. Kroenke, and F. D. Wolinsky. 2005. "A Comparison of Clinically Important Differences in Health-Related Quality of Life for Patients with Chronic Lung Disease, Asthma, or Heart Disease." *Health Services Research* 40: 577–92.

Wyrwich, K. W., and F. D. Wolinsky. 2000. "Identifying Meaningful Intra-Individual Change Standards for Health-Related Quality of Life Measures." *Journal of Evaluation in Clinical Practice* 6: 39–49.