Exploring Physical and Mental Health Component Summaries of the SF-12 Quality of Life Questionnaire

Greg Schaefer
Plan B Project
University of Minnesota – Twin Cities
2 September 2013

Abstract

The SF-12 is a twelve question questionnaire used in clinical trials to assess certain patient quality of life outcomes. Among other reasons, health-related quality of life outcomes (QoL) are important because they can be the tie-breaker in a clinical trial when choosing between two equally effective treatments. The SF-12 is an abbreviated version of the SF-36, a thirty-six question form. Both instruments are summarized by a physical health component score (PCS) and a mental health component score (MCS). The first aims to be a summary of patients' perceptions of aspects of their physical health; the second a summary of patients' perceptions of aspects of their mental health, including mood and energy levels. The SF-36 PCS and MCS summaries are derived from principal components analysis on data collected from many different clinical trials. By contrast, the SF-12 summaries are weighted linear combinations of the twelve questions, with weights designed to estimate the PCS and MCS that would have been obtained if the patient had filled out the SF-36.

The weights on the SF-12 are not easy to interpret. For example, higher PCS and MCS summary scores aim to reflect more positive perceptions of health, but the weight assignment for the SF-12 is such that the more pain one has, the higher one's MCS score, and the more "downhearted and blue" one is, the higher one's PCS score. This leads one to question whether the SF-12 PCS and MCS are accurately tracking better and worse perceptions of health. A related concern is whether the SF-12 PCS is measuring a distinctly physical dimension of patients' perceptions of QoL, and whether the SF-12 MCS is measuring a distinctly mental dimension.

This Plan B project explores these questions by analyzing SF-12 data from 1,225 HIV-positive patients who resided in the United States and were enrolled in the Strategies for Management of Antiretroviral Therapy (SMART) trial between January 2002 and January 2006. The present analyses show that, despite the counter-intuitive weights, the SF-12 PCS and MCS do a very good job tracking changes in patients' perceptions of their health and do in fact measure, respectively, perceptions that are distinctly related to physical health and perceptions that are distinctly related to mental health.

List of Summary Scores Discussed in this Paper

allqol : the arithmetic mean of the *twelve* numeric scores associated with a patient’s responses to the questions on the SF-12. (See Figure 1 for the numeric scores used to weight each question on the SF-12).

mypcs : the arithmetic mean of the *five* numeric scores associated with a patient’s responses to 5 questions on the SF-12 that are clearly related to physical health (questions 2, 3, 4, 5, and 8).

mymcs : the arithmetic mean of the *four* numeric scores associated with a patient’s responses to 4 questions on the SF-12 that are clearly related to mental health (questions 6, 7, 9, and 11).

QMpcs : Quality Metric’s physical component summary score for the SF-12. This is calculated by applying the PCS weights in Figure 1 to a patient’s responses on the SF-12.

QMmcs : Quality Metric’s mental component summary score for the SF-12. This is calculated by applying the MCS weights in Figure 1 to a patient’s responses on the SF-12.

pc1 through pc4 : the first four principal components (specific to a measurement period), arrived at by applying PCA to the values of the 8 domain variables in Figure 3.

curhlth : A question about overall health added to the SF-12 forms used for the patients in the SMART trial. Because this summary measure is not involved in the computation of any of the above summary scores, it provides a benchmark against which they can be compared.

Introduction

Health-related quality of life (QoL) is an important metric in assessing effects of medical treatments. If two treatments are similarly effective in treating a particular disease, the treatment with a more positive or less negative impact on QoL would often be preferred. The patient with cancer, for example, will prefer the least toxic course of treatment if it is as effective as the other available options. One instrument that has been used for the past two decades to measure QoL outcomes in clinical trials is the SF-36, a 36-question self-administered survey that is licensed by the Medical Outcomes Trust (MOT), Health Assessment Lab (HAL), and QualityMetric, Inc. It asks about patients' physical and mental health, general health and level of energy, the level of pain, and whether the patient is limited in their social and physical functioning due to their health condition. In large clinical trials, however, this 36-question survey is difficult to administer. It might take ten minutes or more of a patient's time, and the patient may be asked to complete it at each assessment point in the study. This leads to fewer patients actually completing the entire survey. To address this problem, QualityMetric designed the SF-12, a shortened version of the SF-36 containing only 12 questions (see Figure 1). In large clinical trials the response rate greatly improves with the shorter survey, and the cost of administering the QoL instrument decreases.

Questions on the SF-36 address various domains of QoL. The most well-known summary of the 36 questions combines questions pertaining to physical health into a "physical health component score" (PCS), and those pertaining to mental health, mood and energy into a "mental health component score" (MCS). The two scores provide us with a way to quickly assess two basic dimensions of a patient's quality of life. By shrinking the number of questions from 36 to 12, the SF-12 provides less information. More importantly, the reduction means that a simple average of "physical health questions" and "mental health questions" on the SF-12 will not yield PCS and MCS summary scores that are comparable to those of the SF-36. Therefore, the official PCS and MCS scores licensed by QualityMetric, Inc. for the SF-12 are weighted linear combinations of the 12 questions, with weights designed to estimate the PCS and MCS that would have been obtained if the patient had filled out the SF-36. While the SF-12 doesn't capture quite as much information about QoL as the SF-36 does, it is reported that it accounts for more than 90% of the variance of the PCS and MCS scores that QualityMetric uses for the SF-36 **[1,2]**.

One reason for relying on the SF-36 as the "gold-standard" rather than deriving new PCS and MCS scores for the SF-12 is the desire to use extensive normative data collected from the general population using the SF-36. A second reason is to be able to compare SF-12 results to studies that used the SF-36. Given how hard it is to properly design and execute clinical trials, large or small, such

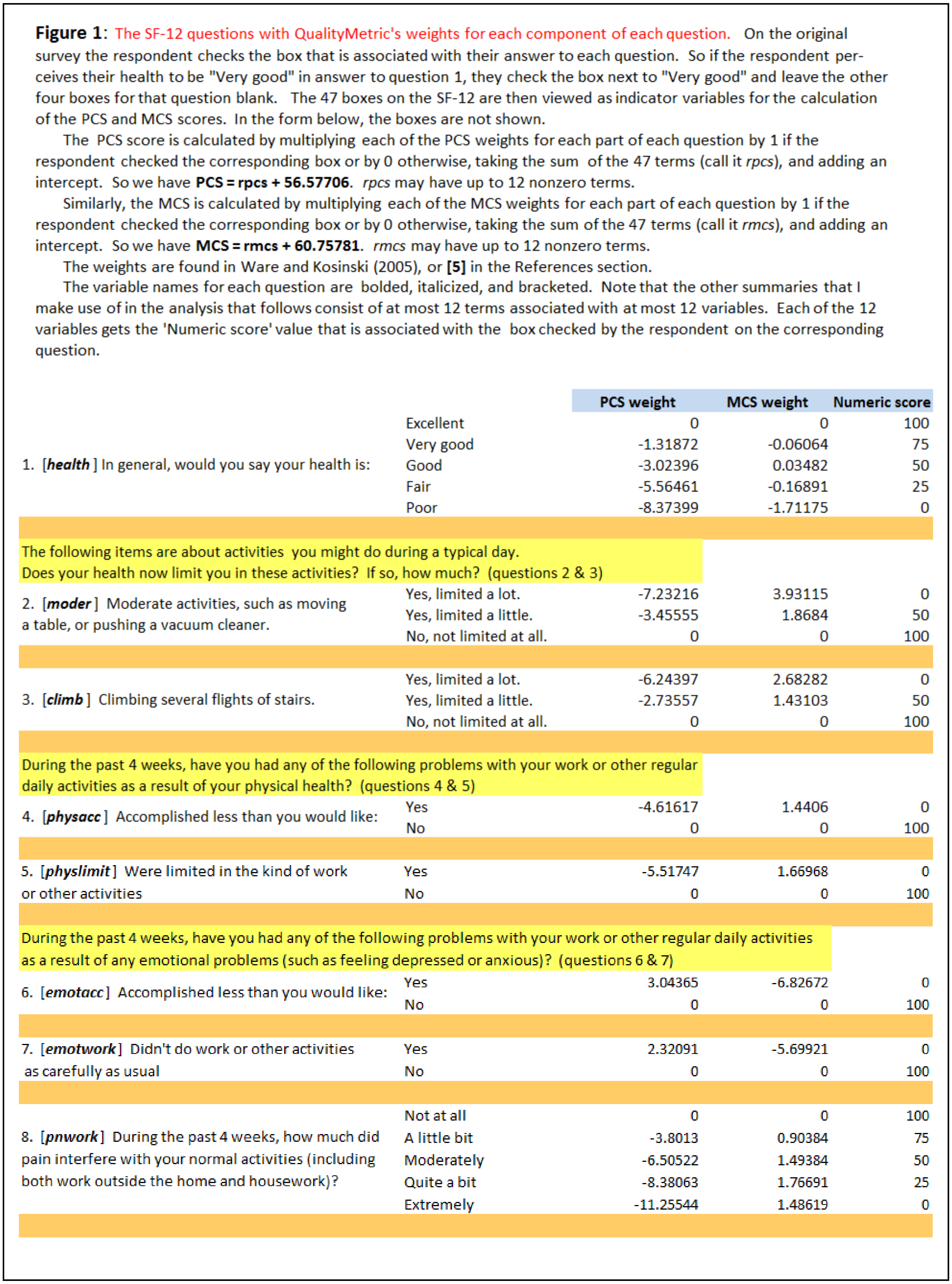


Figure 1: The SF-12 questions with QualityMetric's weights for each component of each question. On the original survey the respondent checks the box that is associated with their answer to each question. So if the respondent perceives their health to be "Very good" in answer to question 1, they check the box next to "Very good" and leave the other four boxes for that question blank. The 47 boxes on the SF-12 are then viewed as indicator variables for the calculation of the PCS and MCS scores. In the form below, the boxes are not shown.

The PCS score is calculated by multiplying each of the PCS weights for each part of each question by 1 if the respondent checked the corresponding box or by 0 otherwise, taking the sum of the 47 terms (call it *rpcs*), and adding an intercept. So we have **PCS = rpcs + 56.57706**. *rpcs* may have up to 12 nonzero terms.

Similarly, the MCS is calculated by multiplying each of the MCS weights for each part of each question by 1 if the respondent checked the corresponding box or by 0 otherwise, taking the sum of the 47 terms (call it *rmcs*), and adding an intercept. So we have **MCS = rmcs + 60.75781**. *rmcs* may have up to 12 nonzero terms.

The weights are found in Ware and Kosinski (2005), or **[5]** in the References section.

The variable names for each question are bolded, italicized, and bracketed. Note that the other summaries that I make use of in the analysis that follows consist of at most 12 terms associated with at most 12 variables. Each of the 12 variables gets the 'Numeric score' value that is associated with the box checked by the respondent on the corresponding question.

		PCS weight	MCS weight	Numeric score
	Excellent	0	0	100
	Very good	-1.31872	-0.06064	75
1. [***health***] In general, would you say your health is:	Good	-3.02396	0.03482	50
	Fair	-5.56461	-0.16891	25
	Poor	-8.37399	-1.71175	0
The following items are about activities you might do during a typical day. Does your health now limit you in these activities? If so, how much? (questions 2 & 3)				
2. [***moder***] Moderate activities, such as moving a table, or pushing a vacuum cleaner.	Yes, limited a lot.	-7.23216	3.93115	0
	Yes, limited a little.	-3.45555	1.8684	50
	No, not limited at all.	0	0	100
	Yes, limited a lot.	-6.24397	2.68282	0
3. [***climb***] Climbing several flights of stairs.	Yes, limited a little.	-2.73557	1.43103	50
	No, not limited at all.	0	0	100
During the past 4 weeks, have you had any of the following problems with your work or other regular daily activities as a result of your physical health? (questions 4 & 5)				
4. [***physacc***] Accomplished less than you would like:	Yes	-4.61617	1.4406	0
	No	0	0	100
5. [***physlimit***] Were limited in the kind of work or other activities	Yes	-5.51747	1.66968	0
	No	0	0	100
During the past 4 weeks, have you had any of the following problems with your work or other regular daily activities as a result of any emotional problems (such as feeling depressed or anxious)? (questions 6 & 7)				
6. [***emotacc***] Accomplished less than you would like:	Yes	3.04365	-6.82672	0
	No	0	0	100
7. [***emotwork***] Didn't do work or other activities as carefully as usual	Yes	2.32091	-5.69921	0
	No	0	0	100
	Not at all	0	0	100
8. [***pnwork***] During the past 4 weeks, how much did pain interfere with your normal activities (including both work outside the home and housework)?	A little bit	-3.8013	0.90384	75
	Moderately	-6.50522	1.49384	50
	Quite a bit	-8.38063	1.76691	25
	Extremely	-11.25544	1.48619	0

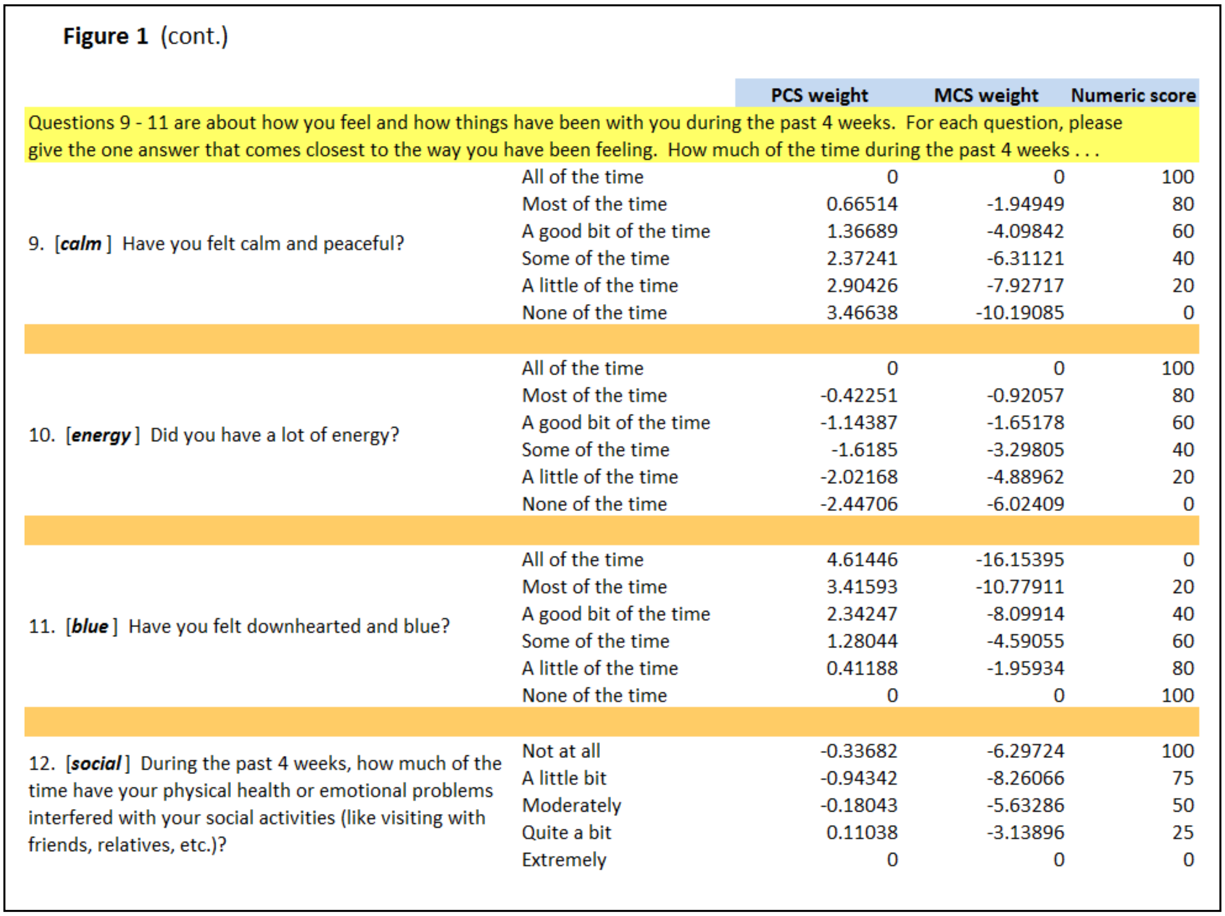


Figure 1 (cont.)

		PCS weight	MCS weight	Numeric score
Questions 9 - 11 are about how you feel and how things have been with you during the past 4 weeks. For each question, please give the one answer that comes closest to the way you have been feeling. How much of the time during the past 4 weeks . . .				
9. [*calm*] Have you felt calm and peaceful?	All of the time	0	0	100
	Most of the time	0.66514	-1.94949	80
	A good bit of the time	1.36689	-4.09842	60
	Some of the time	2.37241	-6.31121	40
	A little of the time	2.90426	-7.92717	20
	None of the time	3.46638	-10.19085	0
10. [*energy*] Did you have a lot of energy?	All of the time	0	0	100
	Most of the time	-0.42251	-0.92057	80
	A good bit of the time	-1.14387	-1.65178	60
	Some of the time	-1.6185	-3.29805	40
	A little of the time	-2.02168	-4.88962	20
	None of the time	-2.44706	-6.02409	0
11. [*blue*] Have you felt downhearted and blue?	All of the time	4.61446	-16.15395	0
	Most of the time	3.41593	-10.77911	20
	A good bit of the time	2.34247	-8.09914	40
	Some of the time	1.28044	-4.59055	60
	A little of the time	0.41188	-1.95934	80
	None of the time	0	0	100
12. [*social*] During the past 4 weeks, how much of the time have your physical health or emotional problems interfered with your social activities (like visiting with friends, relatives, etc.)?	Not at all	-0.33682	-6.29724	100
	A little bit	-0.94342	-8.26066	75
	Moderately	-0.18043	-5.63286	50
	Quite a bit	0.11038	-3.13896	25
	Extremely	0	0	0

comparisons are obviously extremely important for making good assessments of whether a particular treatment does better or worse in terms of QoL than other treatment options. But the downside of relying on the SF-36 for generating SF-12 PCS and MCS summaries is that it makes these summary scores hard to interpret. Because the PCS and MCS for an SF-12 are computed using weights for each part of each question in the survey instrument (see Figure 1), they are each really a summary of *all* questions on the SF-12, whether or not they are physically-related (in the case of the PCS) or mentally-related (in the case of the MCS). This leads one to wonder then whether the SF-12 PCS actually measures what we might reasonably understand as a distinctly physical dimension of a patient's QoL, and whether the SF-12 MCS is a good measure of what we might consider to be a distinctly mental dimension of a patient's QoL. Further, when we look at the weights themselves (there are 47 altogether for the 12 questions), it is hard to see how they are capturing better and worse QoL perceptions. For example, higher PCS and MCS summary scores aim to reflect more positive perceptions of health, but the weight assignment for the SF-12 is such that the more pain one has, the higher one's MCS score (indicating better mental "quality of life"), and the more "downhearted and blue" one is, the higher one's PCS score.

The SF-12's summary scores are less intuitive than the SF-36's PCS and MCS. In order to calculate the latter, the questions on the SF-36 were first summarized into eight "scales", or "domains", where each question appeared in exactly one of the domains.[1] The eight domains were then summarized into two principal components using principal components analysis (PCA). The first principal component had high loadings on domains relating to physical health; the second had high loadings on domains relating to mental health. From these components are derived the PCS and MCS, respectively. Figure 2 shows a schematic of SF-36 question items, their summary to scales/domains, and which of the domains are relevant for physical and mental QoL measures.[2]

The PCA loadings for each of these summaries were derived from data from SF-36 surveys taken by people in the United States. The scores associated with the first principal component were then centered on 50 and scaled to a standard deviation of 10. After the centering and scaling we have the PCS score. Similarly, the scores associated with the second principal component were centered on 50 and scaled to a standard deviation of 10, yielding the MCS scores. Thus, for the general population of the United States the average PCS score should be 50 and the average MCS score should be 50. Again, higher scores reflect more positive perceptions of health.

The questions on the SF-12 (specifically, version 1 of the SF-12) come directly from the SF-36, verbatim, except for questions 2 and 12.[3] QualityMetric does not describe in detail how they arrived at the specific weights given in Figure 1. We obtained the weights from manuals sold by QualityMetric in 2005, manuals which have since been discontinued [**4**]. The PCS weights for each question of the SF-12 were found, it seems, by making the individual components of the SF-12 questions predictors of the PCS scores computed from the respective SF-36 surveys. In this regression there would have been 47 predictors rather than 12 because there are 47 individual boxes which might have been checked on the SF-12 (although for each patient only 12 of the 47 boxes should have been checked). The same approach may have been used to get the weights for the SF-12 MCS scores. There may have been some adjustment of the weights based on the variance of the variable to which the weight was attached; for example,

[1] Presumably, each question associated with a domain contributes equally to that domain.

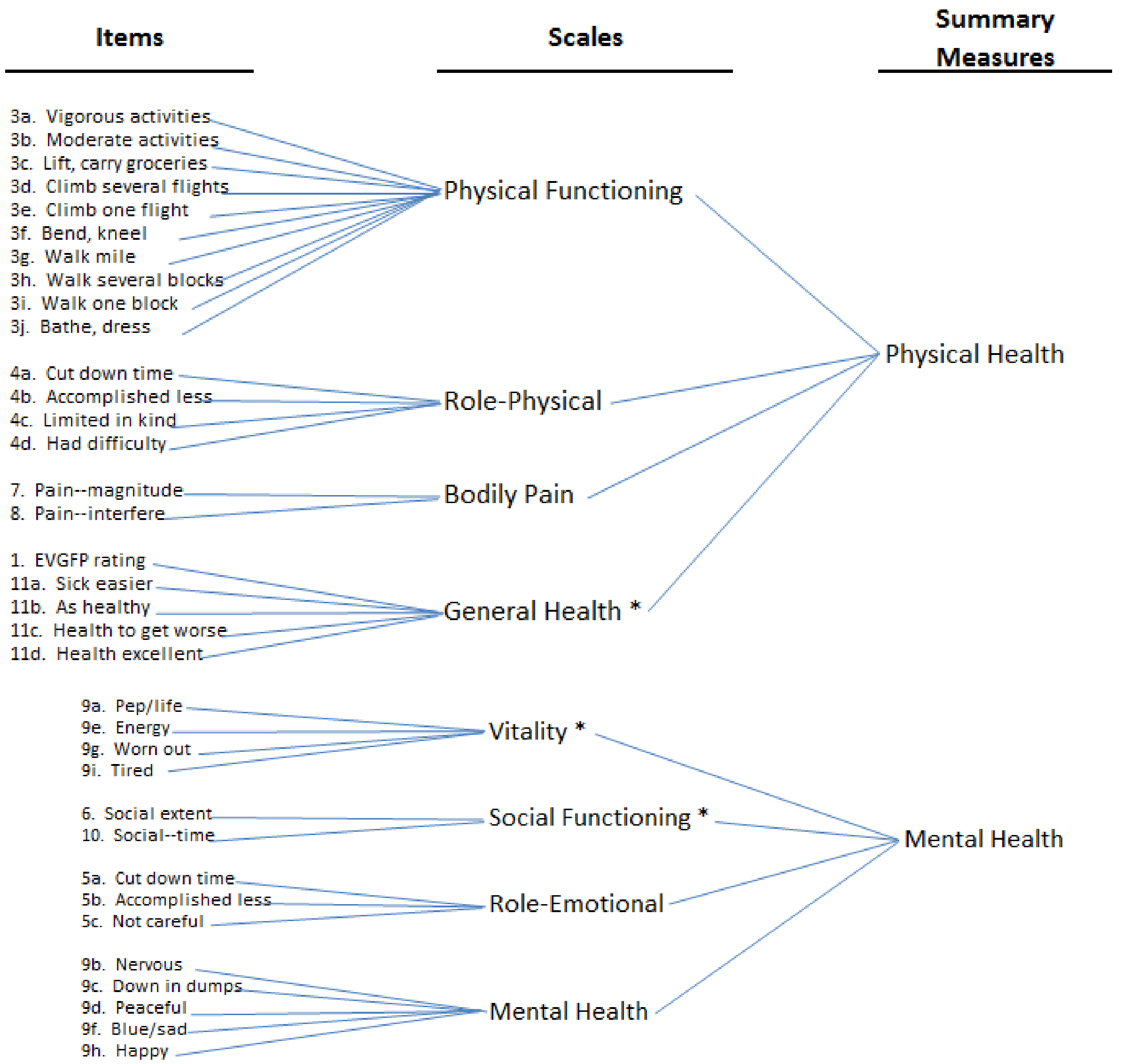
[2] Figure 2 is a bit deceiving because it leaves one with the impression that the Physical Health summary is derived solely from the first four domains shown and that the Mental Health summary is derived solely from the remaining four domains. Each summary is actually a linear combination of all eight domains, the Physical Health summary being the first principal component and the Mental Health summary being the second principal component. Thus, Figure 2 is telling us which of the domains had the highest loadings for the first and second principal components.

[3] Question 2 is a slightly truncated version of question 3(b) on the SF-36. Question 12 joins both question 6 and question 10 from the SF-36.

Figure 2: The SF-36 Measurement Model. (Illustration is from p. 6 of Ware & Kosinski, 2005: "SF-36 Physical and Mental Health Summary Scales: A Manual for Users of Version 1, Second Edition".)

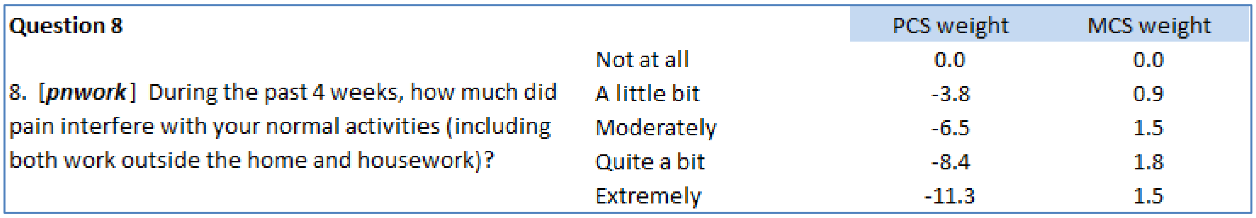
This diagram shows the connection between the SF-36 PCS and MCS summary scores and the individual questions on the SF-36. The eight "scales", or "domains", are summarized into the two summary scores using principal components analysis. Ware and Kosinski write:

> Three scales (Physical Functioning, Role-Physical, and Bodily Pain) correlate most highly with the physical component and contribute most to scoring of the PCS measure of that component. The mental component correlates most highly with the Mental Health, Role-Emotional, and Social Functioning scales, which contribute most to the scoring of the MCS measure of that component. Three of the scales have noteworthy correlations with both components: the Vitality scale correlates substantially with both; General Health correlates with both but higher with the physical component; and Social Functioning correlates much higher with the mental component.



perhaps the higher the variance, the lower the weight. [4]

The difficulty of interpreting the SF-12's summary scores is made concrete when we look at the weights assigned to the five components of the question on pain, Question 8. The variable associated with Question 8 is *pnwork*. Given that higher PCS and MCS scores aim to reflect more positive perceptions of health, one would expect that higher scores for *pnwork* mean that the patient has *less* pain in their daily activities. The Question 8 weights for computing the PCS and MCS (rounded to one decimal place) are as follows:



Question 8		PCS weight	MCS weight
	Not at all	0.0	0.0
8. [***pnwork***] During the past 4 weeks, how much did pain interfere with your normal activities (including both work outside the home and housework)?	A little bit	-3.8	0.9
	Moderately	-6.5	1.5
	Quite a bit	-8.4	1.8
	Extremely	-11.3	1.5

While the PCS weights for Question 8 are consistent with the requirement that lower scores reflect a lower quality of life, the MCS weights—at least when they are considered in isolation from the 42 other MCS weights for the SF-12—violate this simple interpretive rule. The first four weights are increasing, meaning that the more pain one has, the higher the MCS score. Yet it is counter-intuitive that the more pain one has, the higher one's mental quality of life.

Further, there is a connection between the mental and physical dimensions of quality of life that these weight assignments fail to capture. Although there is reason to say that pain should be more strongly associated with the physical rather than the mental dimension, we still expect that pain, especially chronic pain, subtracts from the mental dimension of quality of life. In fact, the connection is even broader: a noticeable decline in one's physical health is likely to negatively impact one's mental health, or one's perceptions of that health. Lower PCS scores, in other words, should correspond to lower MCS scores.[5] Yet we don't see this relationship represented in the PCS and MCS weight assignments for Question 8.

[4] The 47 weights, as opposed to just 12, better take into account the correlation between the individual components of the SF-12, which in turn must lead to better predictions of the SF-36 PCS and MCS scores.

Obviously, using 47 variables rather than just 12 lends itself to much greater refinement in the assignment of weights. Notice, for instance, the PCS weights QualityMetric assigns to question 12, a question asking about both the physical and mental dimensions of one's quality of life. The weights are not monotonically increasing or decreasing. But we do have a monotone direction for the PCS weights assigned to each of the components of the other 11 questions. In calculating the MCS there is a non-monotonic progression of weights in questions 1, 8, and 12, while the progression of weights is monotone in the other nine questions.

[5] I don't think the converse happens as quickly. That is, one's perceptions of mental health can decline without this leading, in the short term, to a decline in one's physical health.

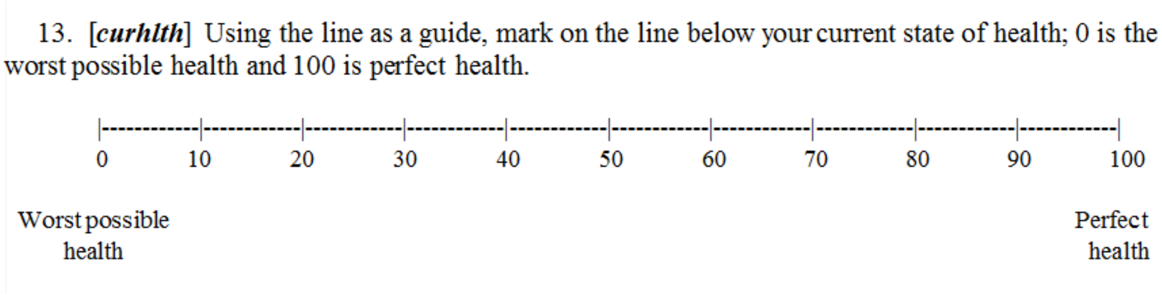
The weights on Question 8 are not uniquely problematic. We have the same problem for interpretation with the Question 11 weights. The more “downhearted and blue” one is, the higher one’s PCS score, albeit the lower one’s MCS score. One would expect both lower PCS and MCS scores, or the PCS score to at least not be higher. We have similar interpretive problems with questions 2 through 5, 9, and 12. Also, there are non-monotonicity issues with questions 1, 8, and 12. In all of these instances the weight assignments are conflicting with the understanding that higher PCS and MCS scores reflect more positive perceptions of health.

One might question, then, whether QualityMetric’s summary scores provide us with a good assessment of patients’ perceptions of their physical and mental health. What do these summary scores measure? What are their limitations? Is there a better way to summarize the data and track changes in patients’ perceptions of QoL?

Methodology

Of specific interest is whether the SF-12 PCS is measuring a distinctly physical dimension of QoL, and whether the SF-12 MCS is measuring a distinctly mental dimension of QoL. I investigate these questions by looking at SF-12 data from 1,225 HIV-positive patients who resided in the United States and were enrolled in the Strategies for Management of Antiretroviral Therapy (SMART) trial between January 2002 and January 2006. The SMART study was a large international randomized clinical trial that compared CD4 count-guided episodic antiretroviral therapy with continuous antiretroviral therapy **[3]**. As the data below show, the health of this set of patients, on average, did not change much during the period of time that they were enrolled in the study. This was true regardless of which treatment regimen the patients were on.

The 1225 patients were asked to fill out an SF-12 at each of their visits for the study, plus one additional question:



13. [***curhlth***] Using the line as a guide, mark on the line below your current state of health; 0 is the worst possible health and 100 is perfect health.

|------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|
0 10 20 30 40 50 60 70 80 90 100

Worst possible
health

Perfect
health

Because *curhlth* is not used in the computation of either PCS or MCS, we can use it as a benchmark for assessing the utility and interpretability of these summary scores.[6] The SF-12 PCS and MCS are doing some of what we want them to do if they are good predictors of *curhlth* and positively correlated with it. The two summaries become even more valuable if together they can predict the current health status of patients better than other candidate summary scores.

Two other such scores that I work with are *mypcs* and *mymcs*. They are simple averages of the scores for specific questions on the SF-12: in the case of *mypcs*, questions clearly related to physical health (questions 2, 3, 4, 5, and 8); in the case of *mymcs*, questions clearly related to mental health (questions 6, 7, 9, and 11).

mypcs = 0.20 * (moder + climb + physacc + physlimit + pnwork)

mymcs = 0.25 * (emotacc + emotwork + calm + blue)

The point of constructing *mypcs* is to see how it compares to QualityMetric's SF-12 PCS score. How well does the latter (what I will refer to as '*QMpcs*') track *mypcs*? If it does a great job tracking *mypcs* but does not do nearly as well tracking *mymcs*, then we have reason to say that *QMpcs* reflects perceptions of a distinctly physical dimension of health. Similarly, I want to know how well what I will call '*QMmcs*' (i.e., QualityMetric's SF-12 MCS score) tracks *mymcs* and what the degree of correlation is between *QMmcs* and *mypcs* to see if the former reflects a distinctly mental dimension of health.

I am also interested in other comparisons, such as whether *mypcs* and *mymcs* do a better job predicting *curhlth* than the combination of *QMpcs* and *QMmcs*. If comparisons and tests like this show that *mypcs* and *mymcs* perform better on average than QualityMetric's ('QM's') scores, there is reason to question the utility of the latter. It won't do us much good to know how the SF-12 PCS and MCS compare to national norms or to compare trials in which the SF-12 is used with trials in which the SF-36 is used if the SF-12 scores are not giving us what we might reasonably consider to be accurate assessments of patients' perceptions of their physical and mental health.

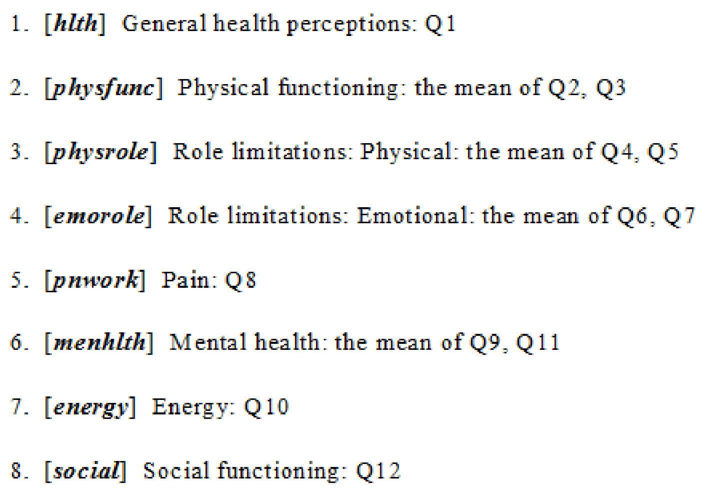
Another summary that I look at is *allqol*. This variable is the arithmetic mean of the scores from each of the 12 questions on the SF-12. It is a linear combination of the same 12 variables used in the computation of *QMpcs* and *QMmcs*, but with none of the complexity in the weights. One would hope

[6] However, *curhlth* provides only a very rough benchmark. It would be far better if we had independent measures of patients' perceptions of their physical health and patients' perceptions of their mental health for each of the completed SF-12 surveys in our dataset.

that QM's two summaries, in combination, can do more work for us than *allqol* alone, work such as predicting and tracking *curhlth*. Otherwise the additional complexity in QM's summaries has no payoff.

Finally, I apply principal components analysis to the 8 domain variables for all observations at each measurement period and explore the first four principal components—what I refer to as *pc1* through *pc4*. Each question on the SF-12 belongs to exactly one of the eight domains or scales shown in Figure 2 above. The mapping of SF-12 questions to domains is given in Figure 3. Running a principal components analysis on the SF-12 values of these variables at each measurement period mimics the procedure outlined in Figure 2, the procedure QualityMetric used to derive the PCS and MCS for the SF-36.[7]

Figure 3: Domains for the SF-12, the questions they are comprised of, and their variable names. Obtained from Reference **[6]**. ("Mean" refers to the numeric score for each question. The numeric scores are given in Figure 1.)



1. [***hlth***] General health perceptions: Q1
2. [***physfunc***] Physical functioning: the mean of Q2, Q3
3. [***physrole***] Role limitations: Physical: the mean of Q4, Q5
4. [***emorole***] Role limitations: Emotional: the mean of Q6, Q7
5. [***pnwork***] Pain: Q8
6. [***menhlth***] Mental health: the mean of Q9, Q11
7. [***energy***] Energy: Q10
8. [***social***] Social functioning: Q12

I explore whether *QMpcs* and *QMmcs* have any resemblance to *pc1* and *pc2*. (See Tables 1 and 2 below for the *pc1* and *pc2* weights.) Does *QMpcs* do much the same work as *pc1* (or one of the other principal components)? For instance, how do they compare in terms of tracking *mypcs*? Similarly, does *QMmcs* resemble *pc2* in terms of tracking *mymcs*? Do the QM summaries look like *pc1* and *pc2* in terms of correlations with the 8 domain variables? Such comparisons help to increase our understanding of the nature of *QMpcs* and *QMmcs*.

[7] Specifically, *pc1* through *pc4* for the baseline data are created by applying principal components analysis solely to the values of the 8 domain variables for each of the patients at baseline; *pc1* through *pc4* for the month 4 data are created solely from the values of the 8 domain variables for each of the patients at month 4; and so forth.

Further, we gain valuable information about QM's two summaries if they beat *pc1* and *pc2* in predicting and tracking *curhlth*. For, by one measure at least, the first two principal components are optimal at summarizing the information contained in the 8 domain variables. No other pair of standardized linear combinations of the 8 domain variables can capture, or explain, a greater proportion of the variance in the data.[8] (The last row in Table 1 shows the proportion of variance in the data explained by *pc1*; the last row in Table 2 shows the proportion of variance explained by *pc1* and *pc2* combined.) So it would be quite meaningful if *QMpcs* and *QMmcs* do a better job than *pc1* and *pc2* in terms of tracking the general health of patients.

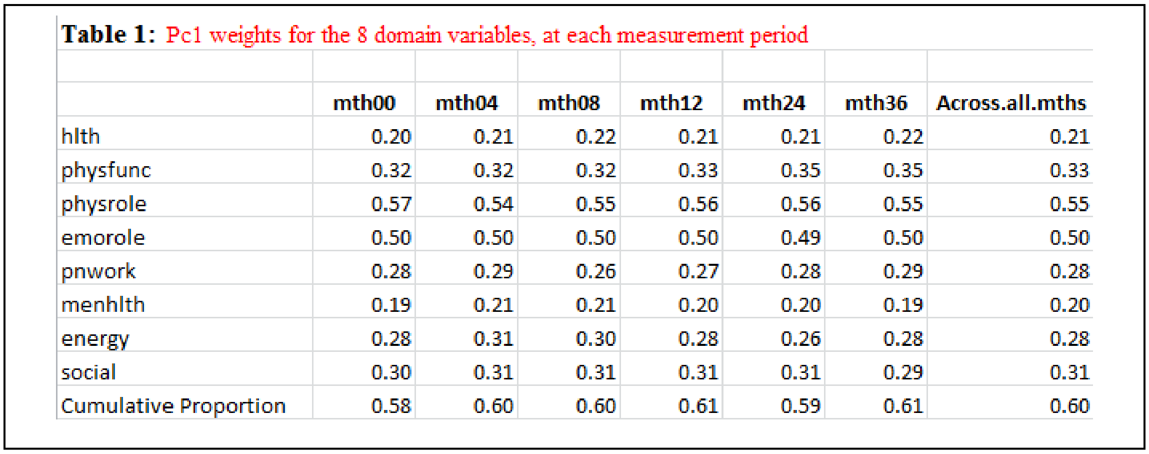


Table 1: Pc1 weights for the 8 domain variables, at each measurement period

	mth00	mth04	mth08	mth12	mth24	mth36	Across.all.mths
hlth	0.20	0.21	0.22	0.21	0.21	0.22	0.21
physfunc	0.32	0.32	0.32	0.33	0.35	0.35	0.33
physrole	0.57	0.54	0.55	0.56	0.56	0.55	0.55
emorole	0.50	0.50	0.50	0.50	0.49	0.50	0.50
pnwork	0.28	0.29	0.26	0.27	0.28	0.29	0.28
menhlth	0.19	0.21	0.21	0.20	0.20	0.19	0.20
energy	0.28	0.31	0.30	0.28	0.26	0.28	0.28
social	0.30	0.31	0.31	0.31	0.31	0.29	0.31
Cumulative Proportion	0.58	0.60	0.60	0.61	0.59	0.61	0.60

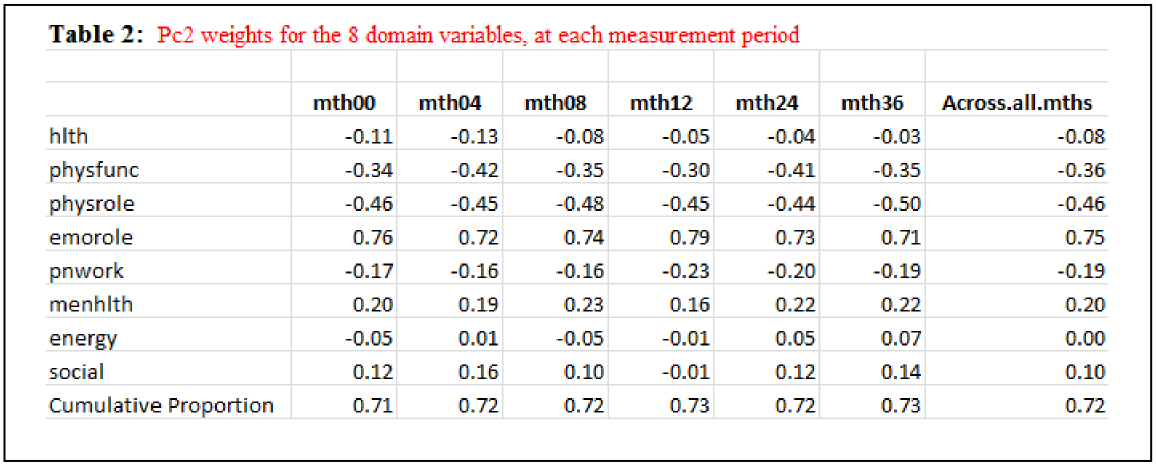


Table 2: Pc2 weights for the 8 domain variables, at each measurement period

	mth00	mth04	mth08	mth12	mth24	mth36	Across.all.mths
hlth	-0.11	-0.13	-0.08	-0.05	-0.04	-0.03	-0.08
physfunc	-0.34	-0.42	-0.35	-0.30	-0.41	-0.35	-0.36
physrole	-0.46	-0.45	-0.48	-0.45	-0.44	-0.50	-0.46
emorole	0.76	0.72	0.74	0.79	0.73	0.71	0.75
pnwork	-0.17	-0.16	-0.16	-0.23	-0.20	-0.19	-0.19
menhlth	0.20	0.19	0.23	0.16	0.22	0.22	0.20
energy	-0.05	0.01	-0.05	-0.01	0.05	0.07	0.00
social	0.12	0.16	0.10	-0.01	0.12	0.14	0.10
Cumulative Proportion	0.71	0.72	0.72	0.73	0.72	0.73	0.72

[8] By "standardized linear combination" of the 8 domain variables I mean one which can be written as $\sum_{j=1}^{p} \delta_j X_j$ where $\sum_{j=1}^{p} \delta_j^2 = 1$. Here X_1, X_2, . . . , X_8 represent the domain variables; the δ_j are the weights that *pc1* and *pc2* assign, respectively, to each of the domain variables. Of course, *QMpcs* and *QMmcs* are not standardized linear combinations of the 8 domain variables or of any other set of variables. But they do have a strong connection to summaries which are themselves derived through PCA.

Initial Look at the Data

Aim of this Initial Look

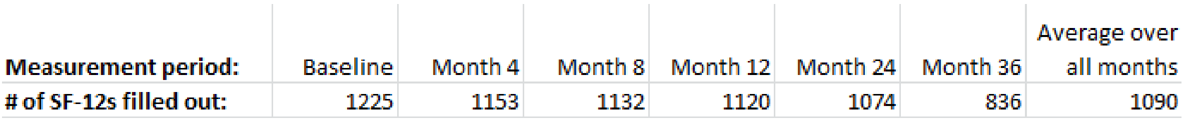
We want to know how well QualityMetric's SF-12 PCS and MCS summary scores do in capturing, respectively, individual's perceptions of their physical and mental health. The problem is that we lack external, or independent, benchmarks for the "mental QoL" and "physical QoL" of the 1,225 SMART trial participants—we have only the SF-12 data itself. Nonetheless, we can begin to get an idea of whether *QMpcs* and *QMmcs* are performing as advertised by looking at the data directly and seeing whether what the data tell us about patients' perceptions of their health is consistent with what QM's summaries seem to be telling us. Further, if variables like age and gender are predictors for perceptions of mental QoL and/or for perceptions of physical QoL, do they similarly predict *QMmcs* and/or *QMpcs*?

A second aim of this initial look at the data is to note some of the similarities and differences that exist between QM's summary scores and the other summary scores (i.e., *allqol, mypcs, mymcs, pc1,* and *pc2*). Doing so will give us a better sense of the nature of *QMpcs* and *QMmcs*.

Details

The data on the 1,225 HIV-positive patients is over six measurement periods: at baseline and at 4, 8, 12, 24, and 36 months out from baseline. There is a fair amount of missing data. The main reason is due to staggered enrollment and the fact that the data were right-censored, the SMART study having been stopped by the study's data and safety monitoring board due to the inferiority of the intermittent anti-retroviral therapy group with respect to clinical outcomes. With complete data, every variable would have 7,350 values (1225 patients times 6 repeated measures) but, as the first column of Table 5 shows, many of the variables are missing around 820 values. *QMpcs* and *QMmcs* have 1001 missing values since neither summary score can be calculated for a patient at a measurement period if any of the 12 questions on that patient's survey is unanswered. The average number of visits per patient was 5.34; the average number of patients per visit was 1090; 15 patients had only one visit; at least 720 patients made

Table 3: Showing the number of SF-12's filled out at each measurement period.

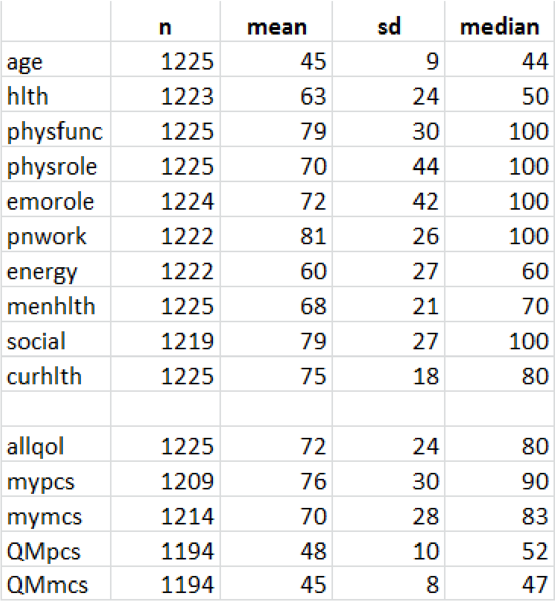


Measurement period:	Baseline	Month 4	Month 8	Month 12	Month 24	Month 36	Average over all months
# of SF-12s filled out:	1225	1153	1132	1120	1074	836	1090

all six visits.

Table 4 shows summary statistics at baseline for the variables associated with the eight domains represented in the SF-12, as well as for the age of the patients, their *curhlth* scores, and for five summary scores. As we saw in Figure 1, the numeric score range for each of the eight domain variables is 0 to 100. Note that the median for five of the variables is 100, whether we look at just the baseline data or across all measurement periods (Table 5). These five domain variables represent 8 of the 12 questions on the SF-12. Thus, at least one-half of the patients are evaluating their QoL using the best possible score for two-thirds of the SF-12.

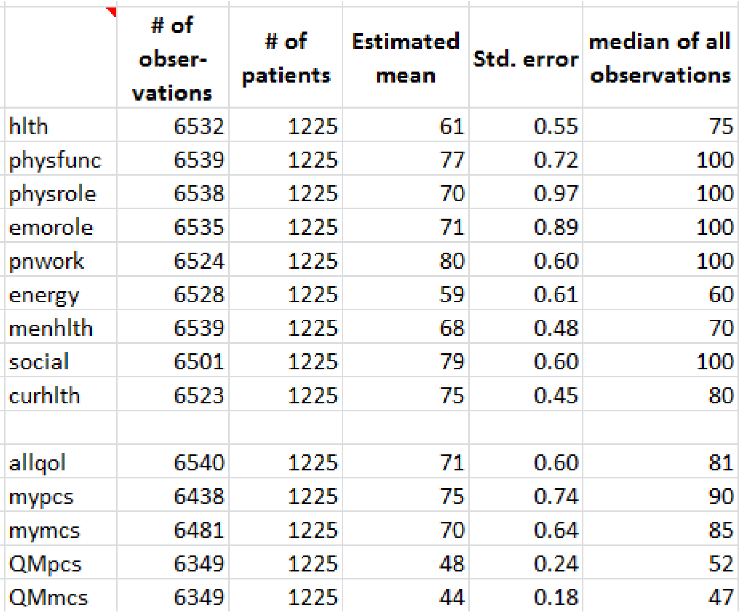
Table 4: Summary statistics for the 8 domain variables at baseline, as well as statistics on *age*, *curhlth*, and five derived summary scores. Each domain variable can have a numeric score from 0 to 100. The numeric score assignment for each question on the SF-12 is shown in Figure 1, and how the scores for the 8 domain variables are derived from the 12 numeric scores is given in Figure 3.



	n	mean	sd	median
age	1225	45	9	44
hlth	1223	63	24	50
physfunc	1225	79	30	100
physrole	1225	70	44	100
emorole	1224	72	42	100
pnwork	1222	81	26	100
energy	1222	60	27	60
menhlth	1225	68	21	70
social	1219	79	27	100
curhlth	1225	75	18	80
allqol	1225	72	24	80
mypcs	1209	76	30	90
mymcs	1214	70	28	83
QMpcs	1194	48	10	52
QMmcs	1194	45	8	47

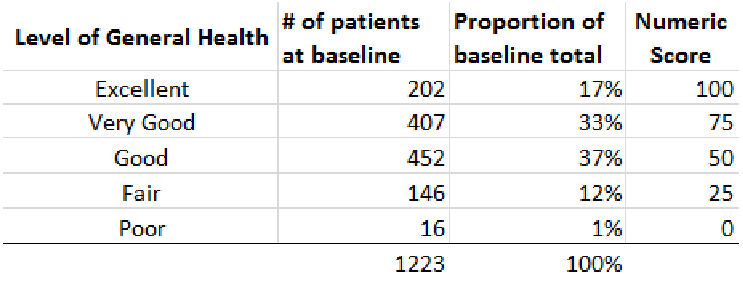
Table 5: Summary statistics on the 8 domain variables, plus *curhlth* and five summary scores, across all measurement periods. The estimated mean for each variable is a point estimate at month 18 given by the linear trajectory that is the average of all individual patients' linear trajectories for the given variable. For example, the R code used to obtain the mean for *allqol* was:

```
lme(allqol~ I(month - 18), data=dat, random= ~month | factor(pid), method="REML",
correlation = corCAR1(form= ~I(month - 18) |factor(pid)), na.action=na.omit, control= iter).
```



	# of obser-vations	# of patients	Estimated mean	Std. error	median of all observations
hlth	6532	1225	61	0.55	75
physfunc	6539	1225	77	0.72	100
physrole	6538	1225	70	0.97	100
emorole	6535	1225	71	0.89	100
pnwork	6524	1225	80	0.60	100
energy	6528	1225	59	0.61	60
menhlth	6539	1225	68	0.48	70
social	6501	1225	79	0.60	100
curhlth	6523	1225	75	0.45	80
allqol	6540	1225	71	0.60	81
mypcs	6438	1225	75	0.74	90
mymcs	6481	1225	70	0.64	85
QMpcs	6349	1225	48	0.24	52
QMmcs	6349	1225	44	0.18	47

Table 6: Frequency of patient responses to Question 1 (*hlth*) at baseline.



Level of General Health	# of patients at baseline	Proportion of baseline total	Numeric Score
Excellent	202	17%	100
Very Good	407	33%	75
Good	452	37%	50
Fair	146	12%	25
Poor	16	1%	0
	1223	100%	

Table 6 helps explain the low median score of 50 for variable *hlth* (Question 1 of the survey) at baseline when five of the other domain variables have median scores of 100 and when *curhlth*—a reformulation, so-to-speak, of Question 1—has a median score of 80. We see that almost exactly one-half of the patients at baseline are rating their general health status as either ‘Excellent’ or ‘Very Good’ while only 13% of the patients are rating their general health status as either ‘Fair’ or ‘Poor’. In general, the patients enrolled in the study perceive themselves to be in fairly good health and, on average, they continue to rate their health as very good throughout the study. Also, as seen in Table 12 below, which treatment regimen patients were on made no difference to patients’ perceptions of their health.

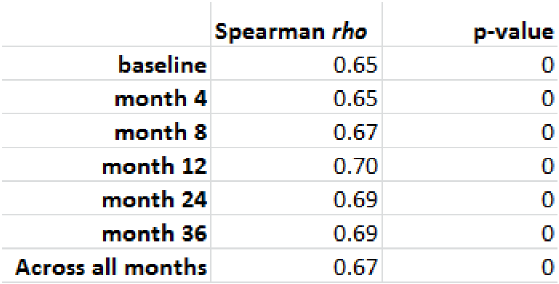
The average age of the 1225 patients at the beginning of the study is 45 years (Table 4). 75% of the patients are men.

As noted above, QualityMetric’s PCS and MCS scores are adjusted so that 50 is the national average for both, and they are scaled so that each has a standard deviation of 10. The patients in the present dataset are less than 2 points lower in average PCS score than the national average, but almost 5.5 points lower in their MCS average score. While the difference in perception of physical health is statistically significant (the 95% CI for the mean is (47.53, 48.47)), it may not be practically meaningful. A comparison with the general population is hard to make since average PCS scores depend on age, and any comparison to the general population should be adjusted for age. On the other hand, the 5.5 point difference in the MCS score is more likely to be practically meaningful; we will probably be able to notice that the HIV patients are less positive in this dimension of their lives.

A mean score of 75 for *curhlth* (Table 5) reflects that, while there are problems (or perceived problems) with the health of certain individuals, the population of patients is doing fairly well overall. The mean score for *curhlth* at baseline is the same as the mean score of *curhlth* across all measurement periods, indicating that—at least for those patients who continued to participate in the study—there is little change on average in their perceived health over the three-year period. Notice that the standard deviation for *curhlth* is about twice that of *QMpcs* and *QMmcs* (Table 4). The standard deviations for the other summary statistics (*allqol, mypcs,* and *mymcs*) are about 2.5 – 3.6 times as large as the standard deviations for *QMpcs* and *QMmcs*. In Table 5 note how close the average values of *mypcs* and *mymcs* are to the average value of *curhlth* and that, like the *QMpcs* and *QMmcs* scores, the mental dimension gets a lower score.

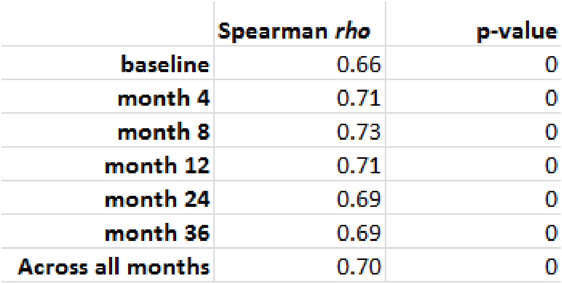
Questions 1 (*hlth*) and 13 (*curhlth*) are two different ways of asking about the state of one’s health in general. However, the correlations between these two variables at each measurement period are not as high as one might expect (Table 7). Perhaps the discrepancy is partly due to the fact that patients are answering Question 13 more thoughtfully than Question 1, for by the time they come to Question 13 they have had to answer eleven specific questions about their health, leading them to arrive at a more considered view of their health in general. Table 8 shows that *allqol* (the average of the numeric scores for the 12 questions) has about the same degree of correlation with *curhlth* as *hlth* does.

Table 7: Spearman correlations between *hlth* and *curhlth*. The p-value is on a test of no association between the two variables.



	Spearman *rho*	p-value
baseline	0.65	0
month 4	0.65	0
month 8	0.67	0
month 12	0.70	0
month 24	0.69	0
month 36	0.69	0
Across all months	0.67	0

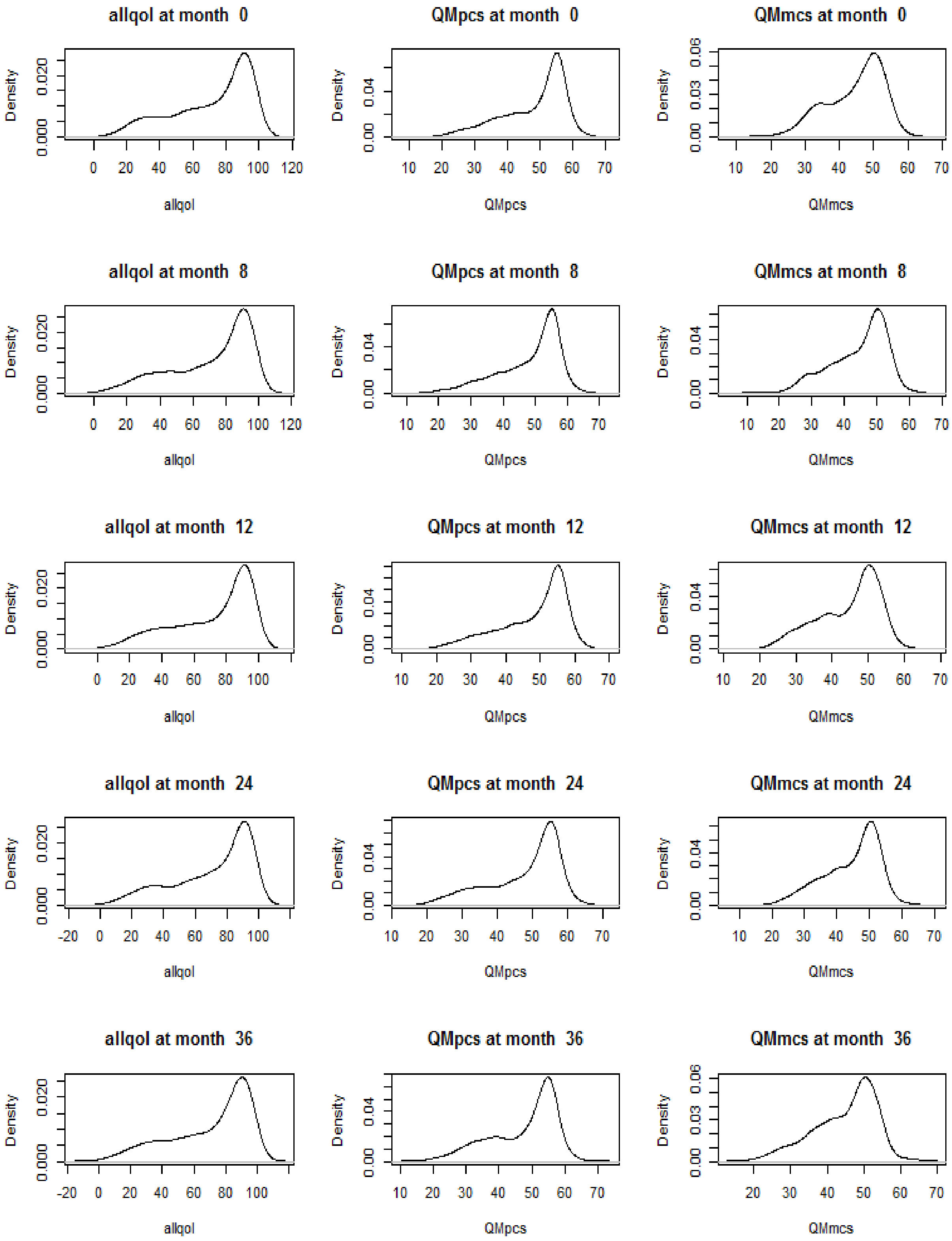
Table 8: Spearman correlations between *allqol* and *curhlth*.



	Spearman *rho*	p-value
baseline	0.66	0
month 4	0.71	0
month 8	0.73	0
month 12	0.71	0
month 24	0.69	0
month 36	0.69	0
Across all months	0.70	0

The histograms which follow give us a sense of the change in distributions across the six measurement periods for six SF-12 summary scores.

Figure 4: Densities of *allqol*, *QMpcs*, and *QMmcs* over five of the six measurement periods.



allqol at month 0

QMpcs at month 0

QMmcs at month 0

allqol at month 8

QMpcs at month 8

QMmcs at month 8

allqol at month 12

QMpcs at month 12

QMmcs at month 12

allqol at month 24

QMpcs at month 24

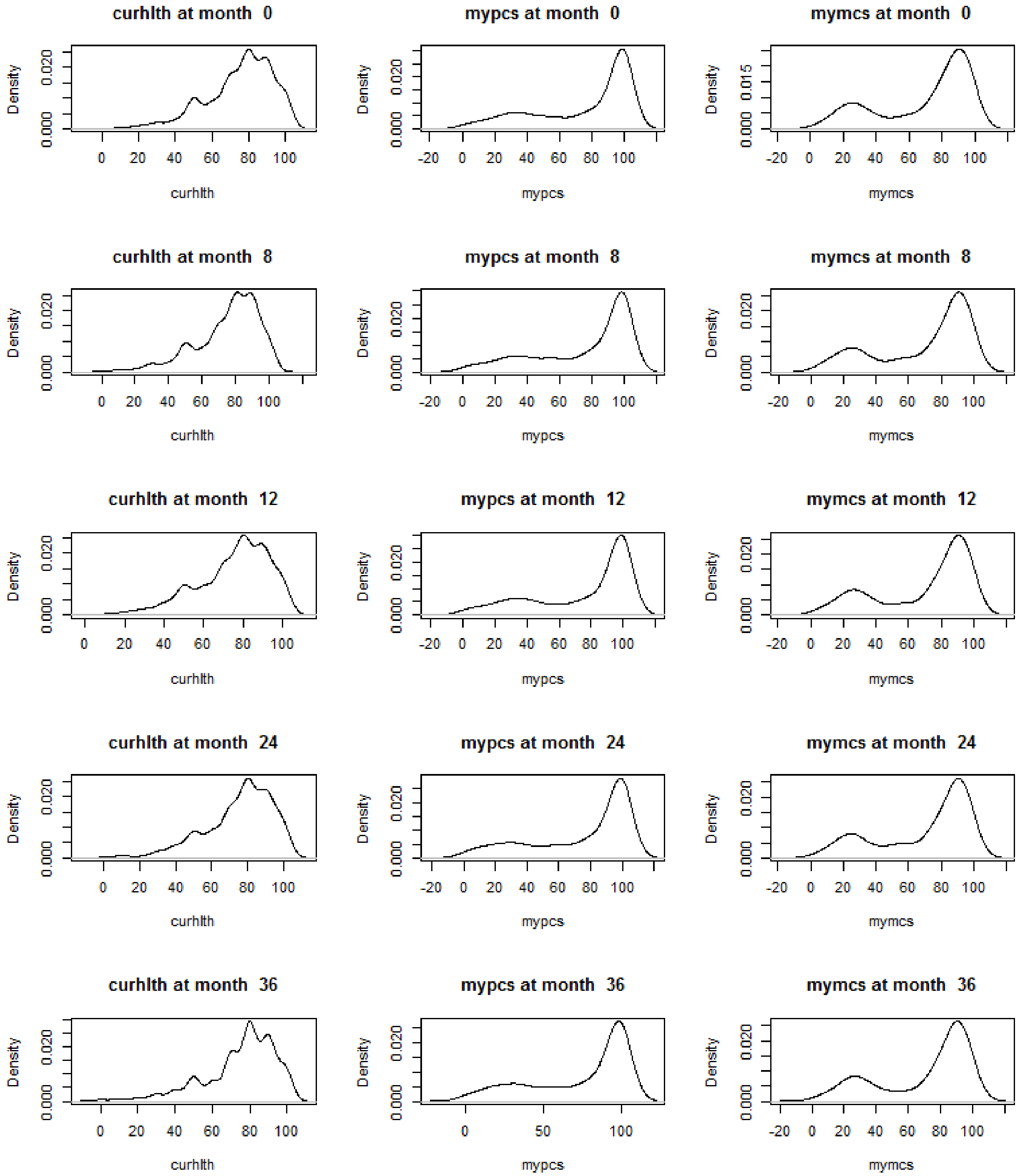
QMmcs at month 24

allqol at month 36

QMpcs at month 36

QMmcs at month 36

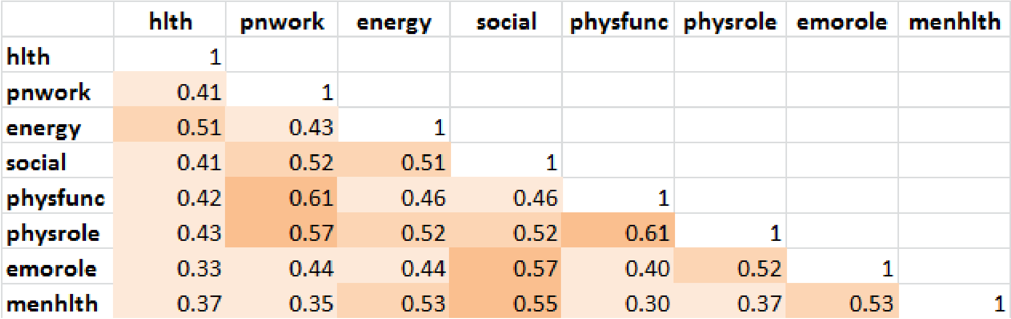
Figure 5: Densities of *curhlth*, *mypcs*, and *mymcs* over five of the six measurement periods. (The bimodal feature that we see in the densities of *mymcs* is not due to gender.)



I see no substantial changes in the distributions of any of the six variables across the visits. The six summary scores are in agreement in this respect.

The correlation tables which follow provide one measure of how the variables in our dataset are related to each other. Note in Table 9 that all of the domain variables are positively correlated, the weakest linear association being between *physfunc* and *menhlth* at 0.30. The strongest correlations among the eight domain variables are between *physfunc* and *physrole* (at 0.61) and between *physfunc* and *pnwork* (at 0.61). That all correlations are positive is unsurprising, since for every variable higher numeric scores are supposed to reflect perceptions of better health. It is interesting to see that *hlth* is most strongly correlated with *energy* and that *energy* has about the same degree of correlation with each of *social*, *menhlth*, and *physrole*. Further, while *social* has the strongest correlation with *emorole* and *menhlth*, it is almost as strongly correlated with *physfunc* and *physrole*—a result unlike what Ware and Kosinski observed in their data. They found that the social component they worked with had a much higher correlation with the mental component than the physical one (refer to the quoted paragraph in Figure 2). Like their results, however, we see that *hlth* is more strongly correlated with the physical component elements (*physrole* and *physfunc*) than with the mental ones (*emorole* and *menhlth*).

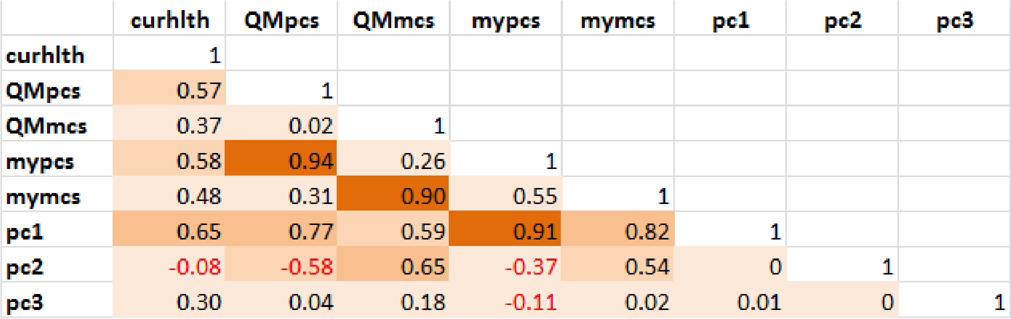
Table 9: Correlations between the 8 domain variables, at baseline. The darker the shading, the stronger the correlation. Note that all of the variables are positively correlated with one another.



	hlth	pnwork	energy	social	physfunc	physrole	emorole	menhlth
hlth	1							
pnwork	0.41	1						
energy	0.51	0.43	1					
social	0.41	0.52	0.51	1				
physfunc	0.42	0.61	0.46	0.46	1			
physrole	0.43	0.57	0.52	0.52	0.61	1		
emorole	0.33	0.44	0.44	0.57	0.40	0.52	1	
menhlth	0.37	0.35	0.53	0.55	0.30	0.37	0.53	1

Table 10 shows the degree of linear association between the summary variables. The correlation between *QMpcs* and *QMmcs* is near 0; in this respect, the two variables look like principal components *pc1* and *pc2*. We see too that, while *pc1* is more highly correlated with *QMpcs* than with *QMmcs*, it is highly correlated with both *mypcs* and *mymcs*. Hence, we should interpret it neither as especially associated with perceptions of physical health nor as especially associated with perceptions of mental health. We see only a few negative correlations; *pc2* is negatively correlated with *mypcs* and with *QMpcs* but positively correlated with *mymcs* and *QMmcs*. Finally, *mypcs* is much more strongly correlated with *QMpcs* than with *QMmcs*, and *mymcs* is much more strongly correlated with *QMmcs* than with *QMpcs*.

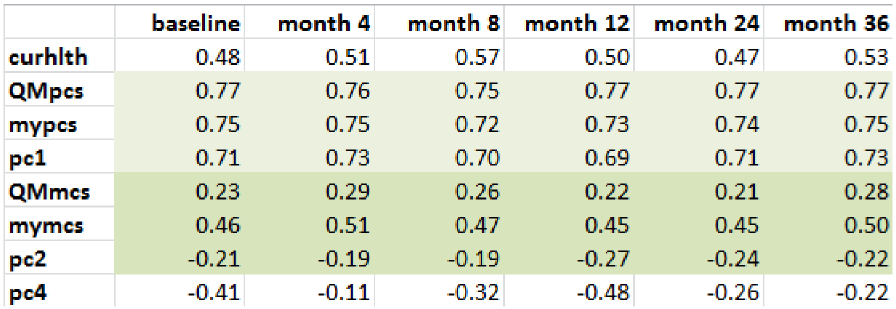
Table 10: Correlations between the summary variables, at baseline. The darker the shading, the stronger the correlation.



	curhlth	QMpcs	QMmcs	mypcs	mymcs	pc1	pc2	pc3
curhlth	1							
QMpcs	0.57	1						
QMmcs	0.37	0.02	1					
mypcs	0.58	0.94	0.26	1				
mymcs	0.48	0.31	0.90	0.55	1			
pc1	0.65	0.77	0.59	0.91	0.82	1		
pc2	-0.08	-0.58	0.65	-0.37	0.54	0	1	
pc3	0.30	0.04	0.18	-0.11	0.02	0.01	0	1

Table 11 shows the correlations between eight summary scores (including *curhlth*) and Question 8 (*pnwork*), the question on the degree to which pain is interfering with the respondent's normal activities.

Table 11: Correlations between the summary variables and *pnwork* (pain). Each entry is the correlation of the row variable with *pnwork* at the given measurement period. *Pc4* is included here because, of the first four principal components, it tends to have the largest coefficient, or weight, for *pnwork*.



	baseline	month 4	month 8	month 12	month 24	month 36
curhlth	0.48	0.51	0.57	0.50	0.47	0.53
QMpcs	0.77	0.76	0.75	0.77	0.77	0.77
mypcs	0.75	0.75	0.72	0.73	0.74	0.75
pc1	0.71	0.73	0.70	0.69	0.71	0.73
QMmcs	0.23	0.29	0.26	0.22	0.21	0.28
mymcs	0.46	0.51	0.47	0.45	0.45	0.50
pc2	-0.21	-0.19	-0.19	-0.27	-0.24	-0.22
pc4	-0.41	-0.11	-0.32	-0.48	-0.26	-0.22

How the different summary variables are associated with the question on pain is of special interest since one might expect pain to have a major role in the evaluation of one's physical health. Table 11 shows that *pnwork* is much more highly associated with QM's PCS than with its MCS score. Similarly, *pnwork* is more closely associated with *mypcs* than with *mymcs*. Note that the level of linear association between *pnwork* and *QMpcs*, between *pnwork* and *mypcs*, and between *pnwork* and *pc1* is about the same, all the correlations being over 70%. But *pnwork* has a substantially higher correlation with *mymcs* than with *QMmcs* even though *QMmcs* and *mymcs* are themselves highly correlated (0.90). As noted earlier, a fairly strong correlation between pain and the mental dimension of QoL makes sense because pain, especially chronic pain, wears one down mentally. It is interesting, too, that the correlation between *pnwork* and *QMmcs* is positive when, as we saw earlier in Figure 1, the MCS weights for *pnwork* increase (for the most part) the more pain one has (i.e., the smaller the *pnwork* score). The weight assignment would lead one to expect a negative correlation between the two variables.

Age, Treatment, and Gender as Predictors

We want to know whether throughout the study period the age of the patients, the treatment they are on, or their gender are significant predictors for any of the summary variables. Are the main summary variables similar with respect to these possible predictors? If *QMpcs* and *QMmcs* do a good job tracking, respectively, perceptions of physical and mental health, we would expect to see their variation explained by variables such as age, treatment, and gender to about the same degree that these perceptions are affected, or explained, by them. For the purposes of this analysis, I shall assume that *mypcs* gives us a rough idea of patients' perceptions of their physical health and that *mymcs* gives us a rough idea of patients' perceptions of their mental health.

Longitudinal analysis was used to arrive at the results in Table 12. The response in each case is the difference between the summary variable and its respective baseline value. The table shows whether age, gender, treatment, or their interactions are significant predictors of these responses. For example, to see whether any of these three variables were significant predictors for the response, '*QMpcs – baseline'*, I first determined what the "unconditional growth model" should look like; i.e., I determined if the modeling against time should be linear, quadratic, or cubic. In all cases, only linear modeling was required. I then went through a process of variable selection (backward elimination), starting with the most complex model (see *model.c.QMpcs* in the sample of R code that follows) and used the AIC score as a criterion to choose between models. In the present example, *model.g.QMpcs* has a smaller AIC than either *model.b.QMpcs* (the unconditional growth model) or *model.c.QMpcs* but the AIC difference with *model.b.QMpcs* is only one point and the latter has a six point improvement in BIC score. Also, since *age* is not a significant predictor in *model.g.QMpcs*, we have reason to conclude that *model.b.QMpcs* is the best model among those considered and that none of age, gender, or treatment are significant predictors for the response '*QMpcs – baseline*'.

****

Sample of R code used to get values for Table 12

****

```
## N.B. In what follows, the response QMpcs is really an individual's QMpcs value at a
given measurement period minus that individual's baseline QMpcs score.  When modeling
I am using dataset "delta.dat" rather than "dat"; the "delta" indicates that all the
values in my dataset are the differences from baseline.  All values at t=0 have been
excluded from the dataset so that the models do not suffer from heteroscedasticity.

> iter <- lmeControl(maxIter=200,msMaxIter=200,niterEM=100, opt="optim")

## UNCONDITIONAL GROWTH MODEL:

> model.b.QMpcs <- lme(QMpcs~ month, data=delta.dat, random= ~month | factor(pid),
method="ML",correlation = corCAR1(form= ~month |factor(pid)), na.action=na.omit,
control= iter)
> summary(model.b.QMpcs)

Linear mixed-effects model fit by maximum likelihood
 Data: delta.dat
    AIC   BIC logLik
  34426 34471 -17206

. . .

## Most complex model:

> model.c.QMpcs <- lme(QMpcs~ month*age*trt*gender, data=delta.dat, random= ~month |
factor(pid), method="ML",correlation = corCAR1(form= ~month |factor(pid)),
na.action=na.omit, control= iter)

> summary(model.c.QMpcs)
Linear mixed-effects model fit by maximum likelihood
 Data: delta.dat
    AIC   BIC logLik
  34440 34577 -17199

Random effects:
 Formula: ~month | factor(pid)
 Structure: General positive-definite, Log-Cholesky parametrization
            StdDev Corr
(Intercept) 6.0259 (Intr)
month       0.1333 -0.061
Residual    5.8716

Correlation Structure: Continuous AR(1)
 Formula: ~month | factor(pid)
 Parameter estimate(s):
    Phi
0.20018

Fixed effects: QMpcs ~ month * age * trt * gender
                        Value Std.Error   DF  t-value p-value
(Intercept)           -0.8899    1.8700 3847 -0.47589  0.6342
month                 -0.0357    0.0728 3847 -0.49058  0.6238
age                    0.0176    0.0406 1170  0.43340  0.6648
trt2                  -1.9575    2.7114 1170 -0.72197  0.4705
gender2               -3.5719    3.5215 1170 -1.01433  0.3106
month:age              0.0003    0.0016 3847  0.17274  0.8629
```

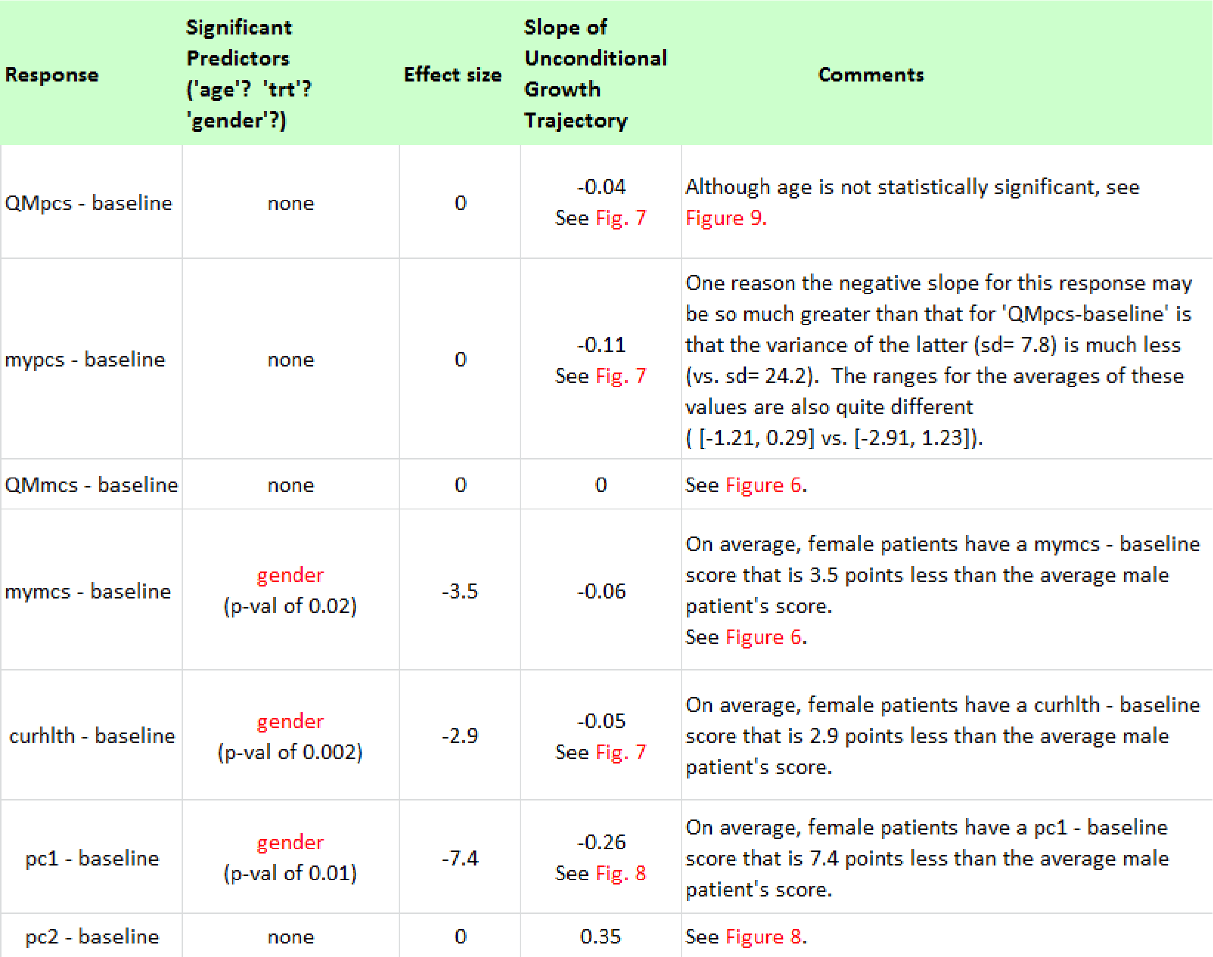
```
month:trt2              0.1000    0.1050 3847  0.95200  0.3412
age:trt2                0.0604    0.0588 1170  1.02751  0.3044
month:gender2           0.0669    0.1409 3847  0.47471  0.6350
age:gender2             0.0797    0.0768 1170  1.03761  0.2997
trt2:gender2            5.5381    4.9739 1170  1.11344  0.2657
month:age:trt2         -0.0026    0.0023 3847 -1.14142  0.2538
month:age:gender2      -0.0017    0.0031 3847 -0.55841  0.5766
month:trt2:gender2     -0.1164    0.1945 3847 -0.59857  0.5495
age:trt2:gender2       -0.1275    0.1099 1170 -1.16010  0.2462
month:age:trt2:gender2  0.0019    0.0043 3847  0.43771  0.6616
...

Number of Observations: 5033
Number of Groups: 1178


## Next best model (among those considered) after the Unconditional Growth Model:

> model.g.QMpcs <- lme(QMpcs~ month + age, data=delta.dat,
random= ~month | factor(pid), method="ML",
correlation = corCAR1(form= ~month |factor(pid)), na.action=na.omit, control= iter)
summary(model.g.QMpcs)

Linear mixed-effects model fit by maximum likelihood
 Data: delta.dat
    AIC   BIC logLik
  34425 34477 -17204

Random effects:
 Formula: ~month | factor(pid)
 Structure: General positive-definite, Log-Cholesky parametrization
            StdDev  Corr
(Intercept) 6.04080 (Intr)
month       0.13488 -0.067
Residual    5.87332

Correlation Structure: Continuous AR(1)
 Formula: ~month | factor(pid)
 Parameter estimate(s):
    Phi
0.20018
Fixed effects: QMpcs ~ month + age
               Value Std.Error   DF t-value p-value
(Intercept) -1.41556   0.99407 3854 -1.4240  0.1545
month       -0.03868   0.00882 3854 -4.3828  0.0000
age          0.03751   0.02156 1176  1.7394  0.0822


. . .
```

****

End of R code sample

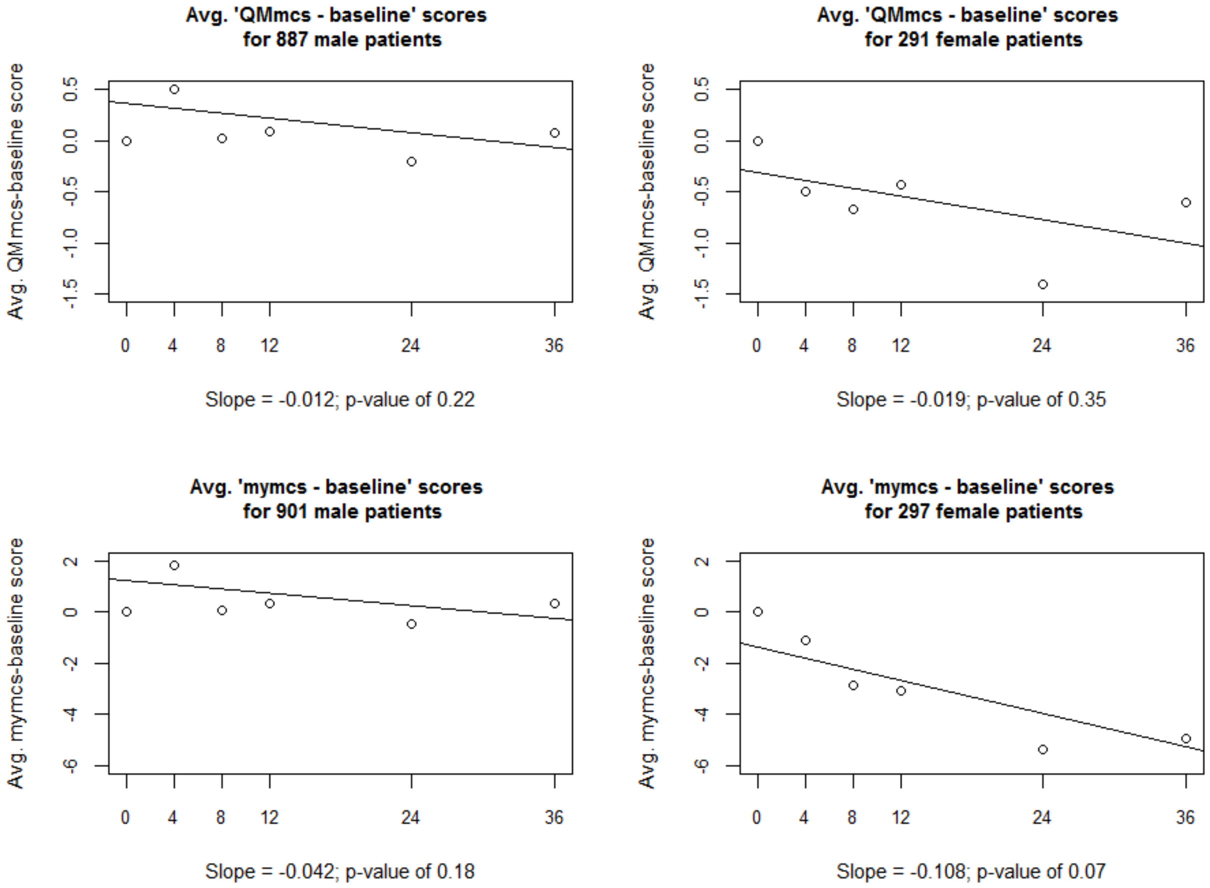
****

Table 12: Whether any of treatment, gender, or age are significant predictors for the main summary variables. Treatment and age are not significant predictors for any of the responses. Gender is a significant predictor for *mymcs*, *curhlth*, and *pc1*.



Response	Significant Predictors ('age'? 'trt'? 'gender'?)	Effect size	Slope of Unconditional Growth Trajectory	Comments
QMpcs - baseline	none	0	-0.04 See Fig. 7	Although age is not statistically significant, see Figure 9.
mypcs - baseline	none	0	-0.11 See Fig. 7	One reason the negative slope for this response may be so much greater than that for 'QMpcs-baseline' is that the variance of the latter (sd= 7.8) is much less (vs. sd= 24.2). The ranges for the averages of these values are also quite different ([-1.21, 0.29] vs. [-2.91, 1.23]).
QMmcs - baseline	none	0	0	See Figure 6.
mymcs - baseline	gender (p-val of 0.02)	-3.5	-0.06	On average, female patients have a mymcs - baseline score that is 3.5 points less than the average male patient's score. See Figure 6.
curhlth - baseline	gender (p-val of 0.002)	-2.9	-0.05 See Fig. 7	On average, female patients have a curhlth - baseline score that is 2.9 points less than the average male patient's score.
pc1 - baseline	gender (p-val of 0.01)	-7.4	-0.26 See Fig. 8	On average, female patients have a pc1 - baseline score that is 7.4 points less than the average male patient's score.
pc2 - baseline	none	0	0.35	See Figure 8.

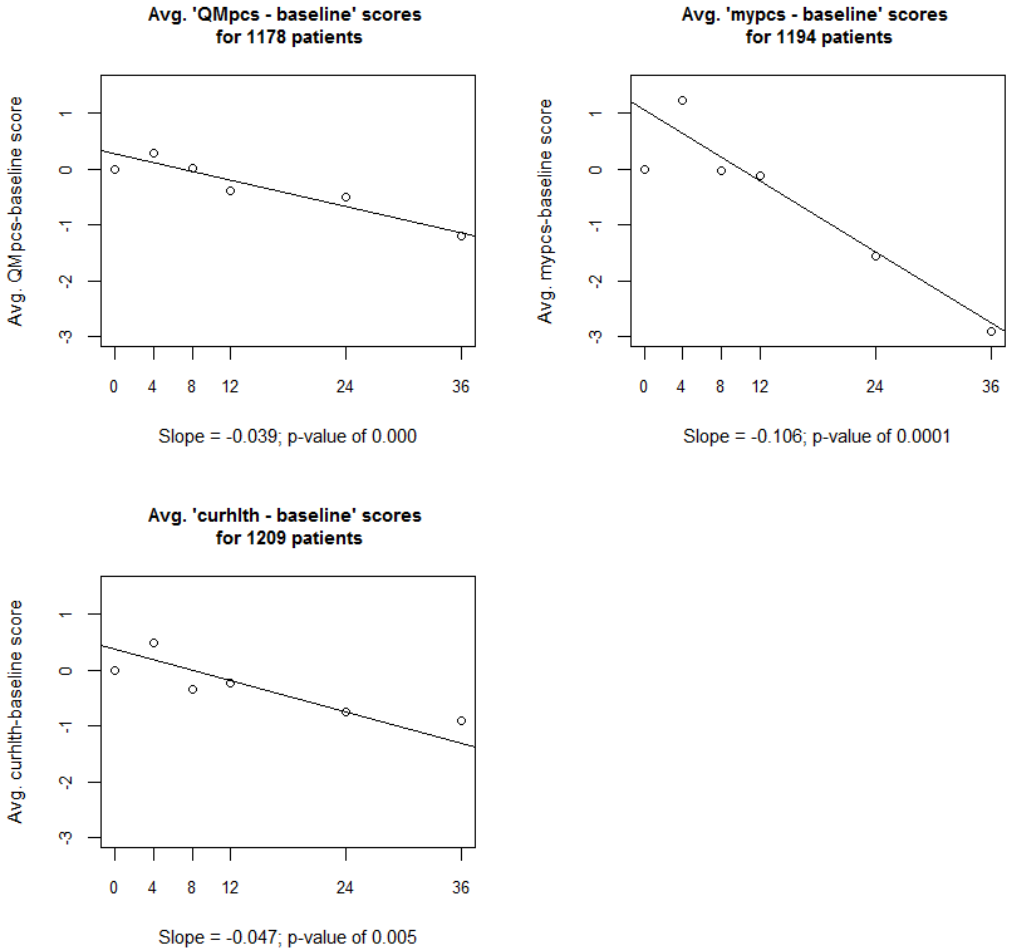
Figure 6: Comparing average differences from baseline for the *QMmcs* and *mymcs* scores, broken down by gender. The points show the average value of each response (e.g., *QMmcs – baseline*) at each measurement period. The lines show the average trajectory when modeling the response against time (using the 'lme' function in R applied to the indicated subset of patients). The modeling does not include the t=0 values.



Figures 6, 7, and 8 illustrate how the average values of each of the responses change over the course of the study period and, for some of the responses, the degree to which these changes differ based on gender.

Referring to Table 12, we see that *age* is not a significant predictor for any of the responses. However, for the variables *QMpcs*, *mypcs*, and *pc1* the change from baseline scores of the older patients (those older than 44) are on average slightly higher than the change from baseline scores of the younger patients (those 44 years old and younger; see Figures 9 and 11 below)[9]—leading one to expect that the rate of decline in these scores will be less for the older patients. Yet in each of the panels of these figures the slopes of the lines are not statistically different. The differences between the older and younger patients' change from baseline scores for *QMpcs* that we see in Figure 9 is also interesting given that at baseline the older patients have an average *QMpcs* score of 46.0 while the younger patients' average score is 50.7. (By contrast, the older patients, on average, have a slightly higher *QMmcs* score at baseline: 44.71 vs. 44.28.)

Figure 7: Comparing average differences from baseline for the *QMpcs*, *mypcs*, and *curhlth* scores. The points show the average value of each response (e.g., *QMpcs – baseline*) at each measurement period. The lines show the average trajectory when modeling the response against time (using the 'lme' function in R applied to the indicated subset of patients and excluding values at t=0).


Avg. 'QMpcs - baseline' scores for 1178 patients

Slope = -0.039; p-value of 0.000

Avg. 'mypcs - baseline' scores for 1194 patients

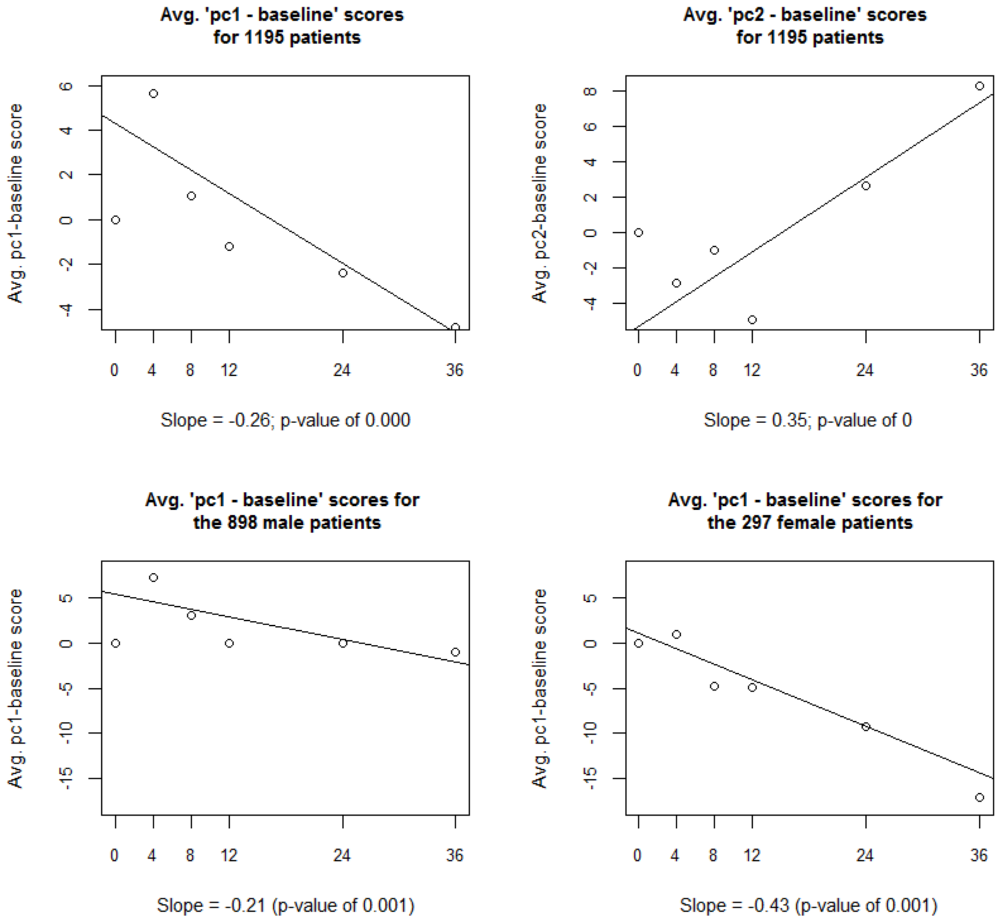
Slope = -0.106; p-value of 0.0001

Avg. 'curhlth - baseline' scores for 1209 patients

Slope = -0.047; p-value of 0.005

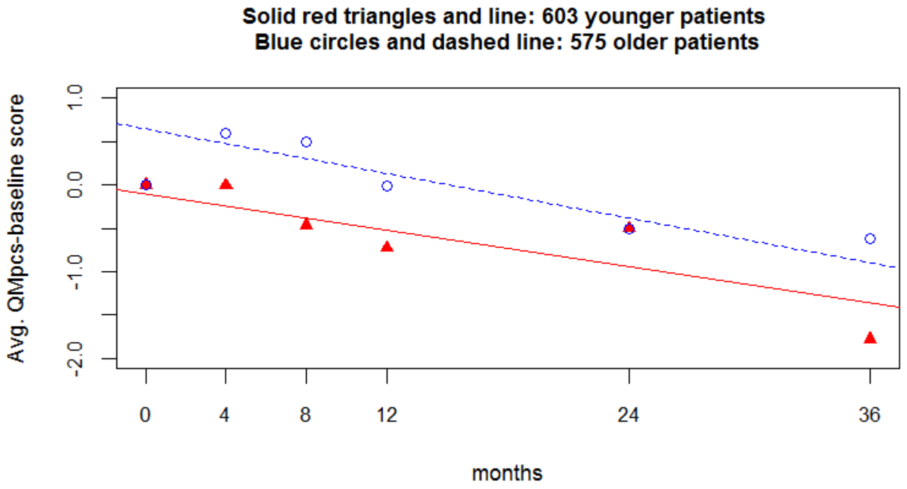
[9] Of course, the difference between the two age groups in change from baseline scores is not statistically significant. I did not investigate whether the average change from baseline scores of the older patients differed from those of the younger patients for the other four variables in Table 12.

Figure 8: Average differences from baseline for the *pc1* and *pc2* scores (top panels); then broken down by gender for the *pc1* scores (bottom panels). Here too the points show the average value of the respective responses at each measurement period. The lines show the average trajectory when modeling the response against time (using the 'lme' function in R applied to the indicated subset of patients and excluding the points at t=0).



Again referring to Table 12, we see that *treatment* is not a significant predictor of any of the responses. On the other hand, *gender* does help us predict the changes from baseline in the *mymcs, curhlth,* and *pc1* scores. In each case the female patients have lower scores, on average. It is surprising, however, to find that there is no *month-gender* interaction for each of the variables when in each case the female patients appear to have a more rapid rate of decline in scores (see the appropriate panels in Figures 6 and 8). For example, we see in Figure 8 that over the course of the study the *pc1* scores are on average changing very little from baseline for the men. The average change from baseline scores for the men at months 24 and 36 are -0.07 and -0.97, respectively. By contrast, the average change from baseline scores for the women at months 24 and 36 are -9.29 and -17.11.

Figure 9: Showing the difference *age* makes to the average *QMpcs – baseline* scores. Here the younger patients are those less than or equal to 44 years old at baseline and the older patients are those older than 44. (The median age of all patients at baseline is 44.) While the slopes of the unconditional growth trajectories are essentially the same, the points show that the older patients, on average, have slightly more positive (or less negative) QMpcs scores relative to their baseline scores than the younger patients. This difference in average scores, however, is not statistically significant.



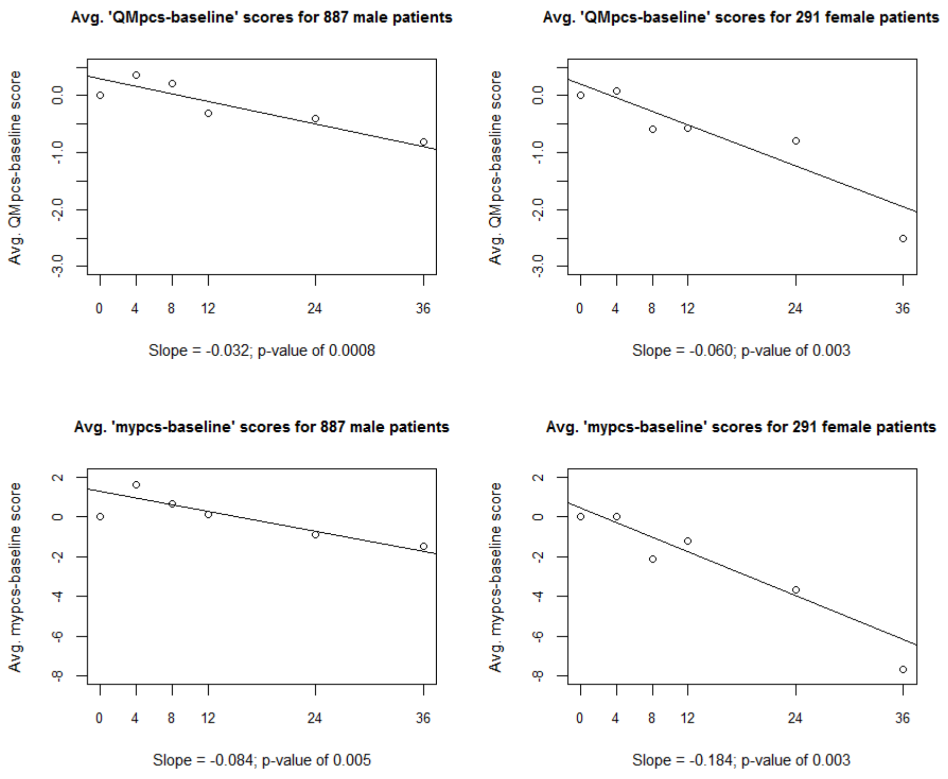
Conclusions from Initial Look at the Data

Regarding the nature of *QMpcs* and *QMmcs*, it is most interesting to see that the correlation between the two (0.02) is so close to 0 in our dataset. While the SF-36 PCS and MCS scores were arrived at using principal components analysis and thus have 0 correlation in the original dataset used for the derivation of the SF-12 PCS and MCS weights, the degree to which the SF-12 summary scores derived from the SMART data retain this aspect of the SF-36 scores is surprising. The very low correlation between *QMpcs* and *QMmcs* tells us that these variables are capturing quite different information from the survey. In one respect this lack of redundancy is a good thing, for it means that we are capturing more information from the survey than might otherwise be the case. On the other hand, the downside is that the relatively sharp distinction fails to reflect a real connection that exists between patients' perceptions of their physical health and patients' perceptions of their mental health. (For comparison, the cross-sectional correlation between *mypcs* and *mymcs* at baseline is 0.58.)

In general, however, none of the observations made so far indicates that *QMpcs* is not giving us a good measure of perceptions of physical health, or that *QMmcs* is not giving us a good measure of perceptions of mental health. *QMpcs* is much more strongly correlated with *mypcs* (0.94) than with *mymcs* (0.31), and *QMmcs* is much more strongly correlated with *mymcs* (0.90) than with *mypcs* (0.26); see Table 10. This suggests that the SF-12 PCS score *is* associated with a distinctly physical dimension of a patient's QoL, and that the SF-12 MCS score *is* associated with a distinctly mental dimension of a patient's QoL. Also, *QMpcs* is correlated with *pnwork* (i.e., pain) very much like *mypcs* is; see Table 11. However, the correlation between *pnwork* and *mymcs* is nearly double that between *pnwork* and *QMmcs*, a discrepancy probably due in part to the strange MCS weight assignment for the *pnwork* question. Still, there is a 90% correlation between the two variables and, as we will see below, *QMmcs* does an excellent job tracking, or predicting, *mymcs*.

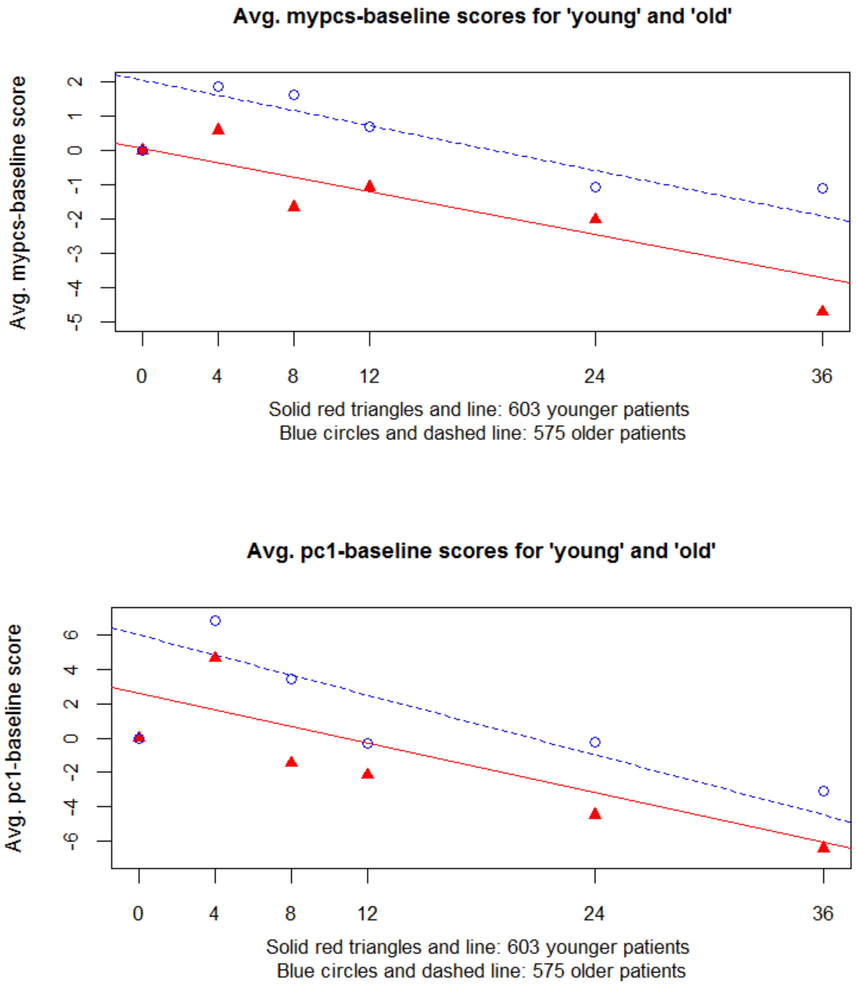
While *age* and *treatment* are not predictors for any of the summary scores, *gender* is a predictor for *mymcs, curhlth,* and *pc1*. The changes from baseline of the *QMpcs* and *mypcs* scores also appear to show a faster rate of decline for the female patients (see Figure 10). But we do not see even this much of a distinction between the male and female patients in terms of the changes in their *QMmcs* scores (see the top two panels in Figure 6). There is only a 0.4 point difference between male and female *QMmcs* scores

Figure 10: Showing the similar effects of *gender* on the average *QMpcs* – baseline and *mypcs* – baseline scores. The points show the average value of the respective responses at each measurement period. The lines show the average trajectory when modeling the response against time (using the 'lme' function in R applied to the indicated subset of patients).



at baseline. Nonetheless, this difference between *QMmcs* and *mymcs* does not prevent the former from being a good predictor of the latter. We will also see in the analysis that follows that '*~QMpcs + QMmcs*' can be an excellent predictor of *curhlth* even though *gender* does not have the effect on either *QMpcs* or *QMmcs* that it has on *curhlth*.

Figure 11: Showing that the difference *age* makes in the *mypcs* – baseline and *pc1* – baseline scores is very similar to what we see in Figure 9 with the *QMpcs* – baseline scores. The younger patients are those less than or equal to 44 years old at baseline and the older patients are those older than 44 at baseline. The lines model the average unconditional growth trajectories for the respective sets of patients. Age is not a significant predictor for either *mypcs* or *pc1*.



Tracking *curhlth*, *mypcs*, and *mymcs*

Aim of this section

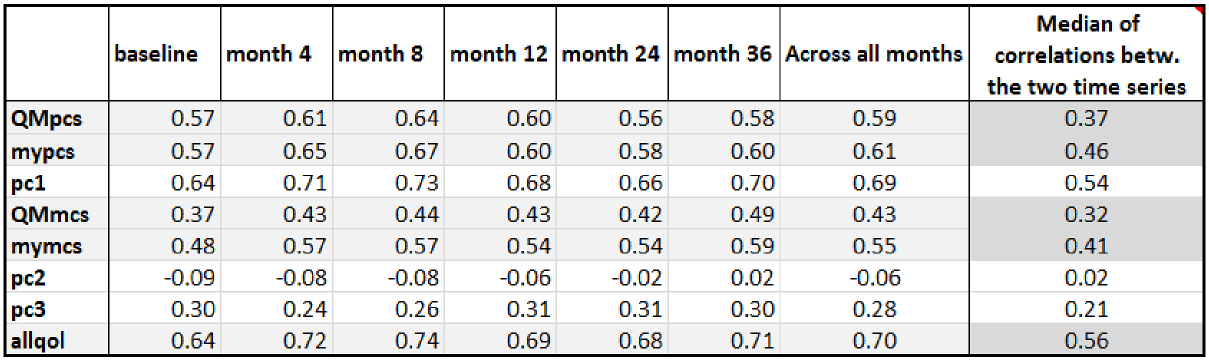
Although not ideal, *curhlth* is the one metric we have from the SMART trial data that we can use as a benchmark for assessing how well the summary scores are tracking patients' perceptions of their overall health. Do *QMpcs* and *QMmcs* do a better job predicting, or tracking, *curhlth* than the combination of *mypcs* and *mymcs*? or than *allqol*? or than *pc1* and *pc2* combined?

We also can get a better understanding of the nature of *QMpcs* (and *QMmcs*) by seeing how well it tracks *mypcs*, and comparing this to how well it tracks *mymcs*. If *QMpcs* does a great job tracking *mypcs* but does not do nearly as well tracking *mymcs*, then we have an additional reason for saying that *QMpcs* reflects perceptions of a distinctly physical dimension of health. We can say something similar for *QMmcs*. The cross-sectional correlations that we saw above suggest that *QMpcs* will in fact be a better predictor of *mypcs* than of *mymcs*, and that *QMmcs* will be a better predictor of *mymcs* than of *mypcs*. The time-series plots and correlations that follow corroborate this.

Results

Table 13 shows the cross-sectional correlations each summary variable has with *curhlth* as well as the medians of the sets of all patient-level time series correlations (one time series being the patient's

Table 13: Correlations with *curhlth* at each measurement period and averaged across all periods. The column on the far right shows the median of the set of patient-level time series correlations between the two variables, each variable being of the form '*variable – baseline*'.



	baseline	month 4	month 8	month 12	month 24	month 36	Across all months	Median of correlations betw. the two time series
QMpcs	0.57	0.61	0.64	0.60	0.56	0.58	0.59	0.37
mypcs	0.57	0.65	0.67	0.60	0.58	0.60	0.61	0.46
pc1	0.64	0.71	0.73	0.68	0.66	0.70	0.69	0.54
QMmcs	0.37	0.43	0.44	0.43	0.42	0.49	0.43	0.32
mymcs	0.48	0.57	0.57	0.54	0.54	0.59	0.55	0.41
pc2	-0.09	-0.08	-0.08	-0.06	-0.02	0.02	-0.06	0.02
pc3	0.30	0.24	0.26	0.31	0.31	0.30	0.28	0.21
allqol	0.64	0.72	0.74	0.69	0.68	0.71	0.70	0.56

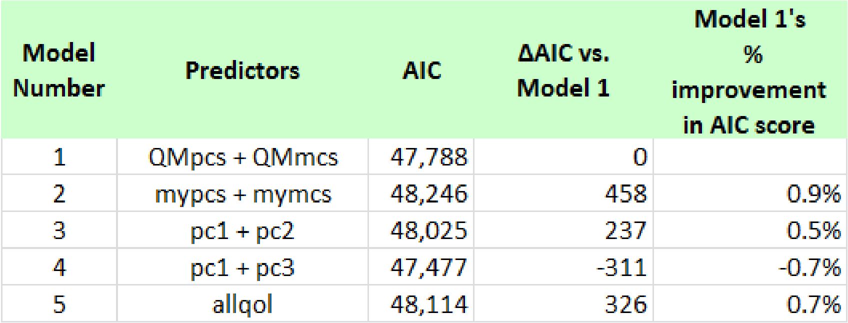
set of *curhlth – baseline* scores, the other being the *variable – baseline* scores for each of the variables on the far left of the table). *Pc2*, we see, has essentially no linear association with *curhlth*. We also see in Table 13 that *allqol* and *pc1* are the most effective at predicting *curhlth*. And once again we see similarities between *QMpcs* and *mypcs* and between *QMmcs* and *mymcs*, although both *mypcs* and *mymcs* are somewhat better at tracking *curhlth* than *QMpcs* and *QMmcs* are, respectively.

Since *QMpcs* and *QMmcs* are supposed to be measuring different dimensions of patients' perceptions of their health, it makes sense to ask how well the two *together* can track, or predict, *curhlth*. Table 14 compares the strength of different pairings of the summary variables for predicting the response, *curhlth – baseline*. We see from the results in Table 14 that '~(*QMpcs* – baseline) + (*QMmcs* – baseline)' is a better predictor than all of the other combinations of predictors, except for '~*pc1* + *pc3*'. Because it is difficult to judge from the AIC scores alone how well QualityMetric's SF-12 summary scores are predicting *curhlth*, I have included Table 15. While the R^2 values in Table 15 do not tell us anything about how well *individuals'* SF-12 PCS and MCS scores predict their changing perceptions of health, they do tell us something about how well the different combinations of terms are capturing the variability in *curhlth* scores. Note the similarities in the "rankings" of the far right columns of the two tables. In each case, '~*pc1* + *pc3*' is outperforming '~*QMpcs* + *QMmcs*', while the latter is outperforming the remaining pairs of terms. Also, in both tables we see that '~*mypcs* + *mymcs*' is performing least well.

Table 14: Comparing how well the summary scores can predict *curhlth – baseline*. Each predictor term is of the form 'variable – baseline score for that variable'. All longitudinal models are run on the same set of 1194 patients. The (abbreviated) model form is always the same in each case:

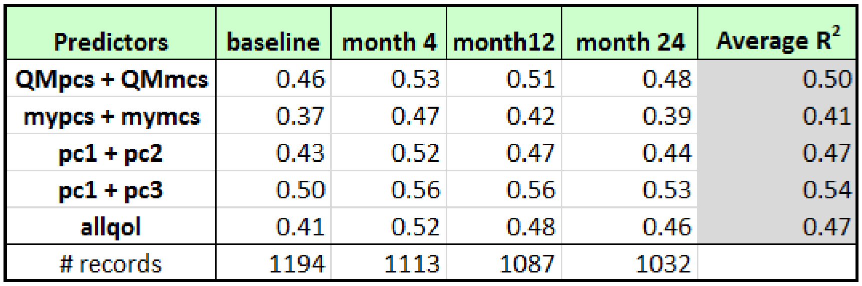
Fixed effects: curhlth ~ month + term1 + term2

Random effects: ~ month + term1 + term2 | factor(pid)



Model Number	Predictors	AIC	ΔAIC vs. Model 1	Model 1's % improvement in AIC score
1	QMpcs + QMmcs	47,788	0	
2	mypcs + mymcs	48,246	458	0.9%
3	pc1 + pc2	48,025	237	0.5%
4	pc1 + pc3	47,477	-311	-0.7%
5	allqol	48,114	326	0.7%

Table 15: Showing R^2 values for OLS regressions of *curhlth* on the given sets of predictors. (Here there is no subtraction of baseline scores from the response or the predictor terms.) The regressions in each measurement period are done on the same set of records, although the number of usable records changes for each period. The R^2 values for predictor *allqol* are calculated from the correlations for *allqol* given in Table 13.



Predictors	baseline	month 4	month12	month 24	Average R^2
QMpcs + QMmcs	0.46	0.53	0.51	0.48	0.50
mypcs + mymcs	0.37	0.47	0.42	0.39	0.41
pc1 + pc2	0.43	0.52	0.47	0.44	0.47
pc1 + pc3	0.50	0.56	0.56	0.53	0.54
allqol	0.41	0.52	0.48	0.46	0.47
# records	1194	1113	1087	1032	

We see from Table 13 that *allqol* and *pc1* are better predictors of *curhlth* than either *QMpcs* or *QMmcs* alone. But Tables 14 and 15 show that together the two QM summary variables outperform both *allqol* and *pc1*. How well do *allqol* and *pc1* track *curhlth*? Figures 12-14 give us a concrete sense of their tracking ability; together *QMpcs* and *QMmcs* are doing at least this well.

Figure 15 illustrates how the averages of the *QMpcs – baseline* values at the different measurement periods track the averages (halved) of the *mypcs – baseline* values. Figure 16 illustrates this tracking at the individual patient level. Similarly, Figure 17 shows the degree to which the averages of the *QMmcs – baseline* values align with the averages of the *mymcs – baseline* values. Figure 18 illustrates this tracking at the level of individuals.

Figures 19-22 give us a concrete sense of how *pc1* compares with *mypcs* and how *pc2* compares with *mymcs*. We get a bit more information about how QM's summaries differ from the first two principal components by seeing how each pair of summaries ({*QMpcs*, *pc1*}, {*QMmcs*, *pc2*}) compares to *mypcs* and *mymcs*.

Table 16: Showing the medians of the different sets of patient-level time series correlations. Each variable is of the form '*variable – baseline*'. The number of patients varies because some patients will have a standard deviation of 0 for a particular variable's time series.

Variable 1	Variable 2	Median of correlations betw. the two time series	Number of patients
QMpcs	mypcs	0.95	923
	mymcs	-0.15	1096
QMmcs	mymcs	0.96	1096
	mypcs	0.00	923
mypcs	mymcs	0.42	904
QMpcs	QMmcs	-0.40	1146
pc1	mypcs	0.91	923
	mymcs	0.86	1096
	QMpcs	0.73	1146
	QMmcs	0.71	1146
pc2	mypcs	-0.54	923
	mymcs	0.71	1096
	QMpcs	-0.75	1146
	QMmcs	0.78	1146

Figure 12: Comparing *pc1*'s and *allqol*'s abilities to track *curhlth*. Each point in the graph is the average of all patients' *variable – baseline* score for the given measurement period. The correlation between the *allqol* averages and the *curhlth* averages is 0.99. For *curhlth* and *pc1*, the correlation is 0.96.

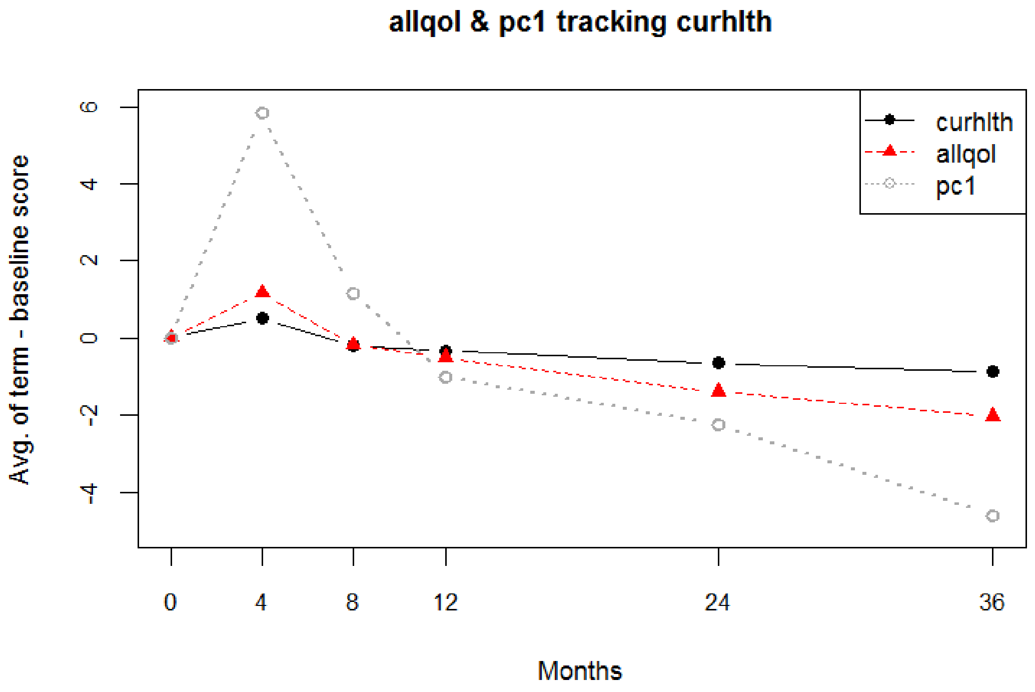


Figure 13: Relationship of *allqol* and *curhlth* at the individual patient level. A sample of 9 patients. The median of the set of all patient-level time series correlations is 0.56 (see Table 13).

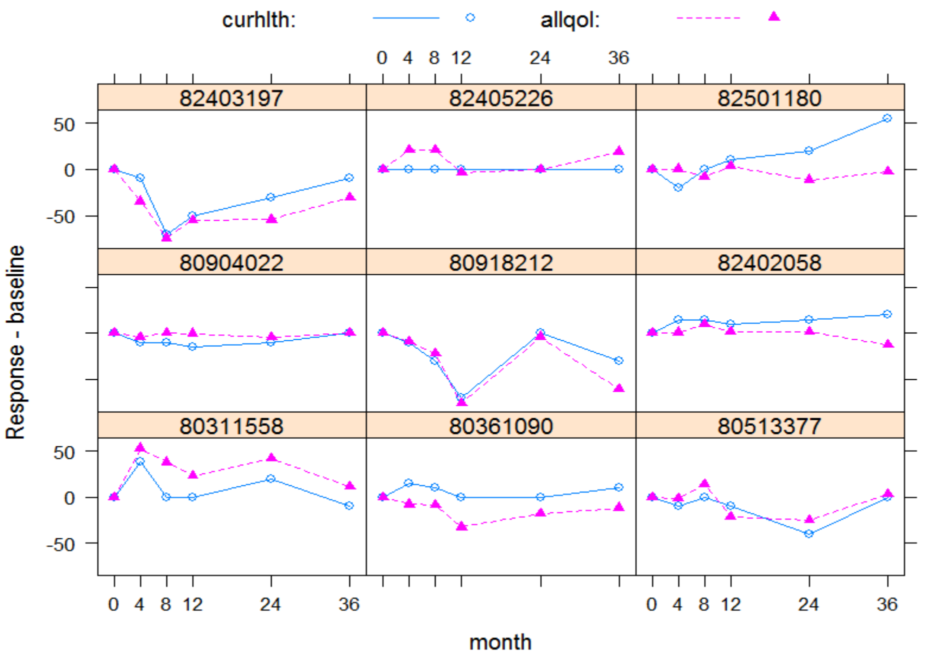


Figure 14: Relationship of *pc1* and *curhlth* at the individual patient level. A sample of 9 patients. The *pc1 – baseline* values have been halved in order to get them closer to the range of the *allqol* values.

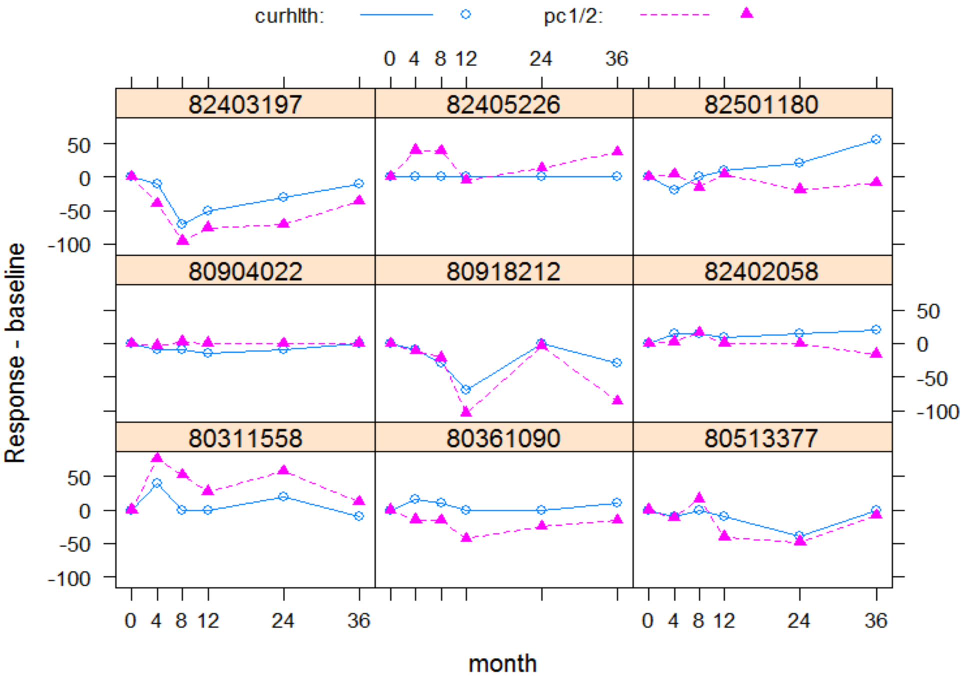


Figure 15: Relationship of *QMpcs* and *mypcs*, with the *mypcs* values divided by 2 for better comparison. Correlation between the averages shown is 0.97.

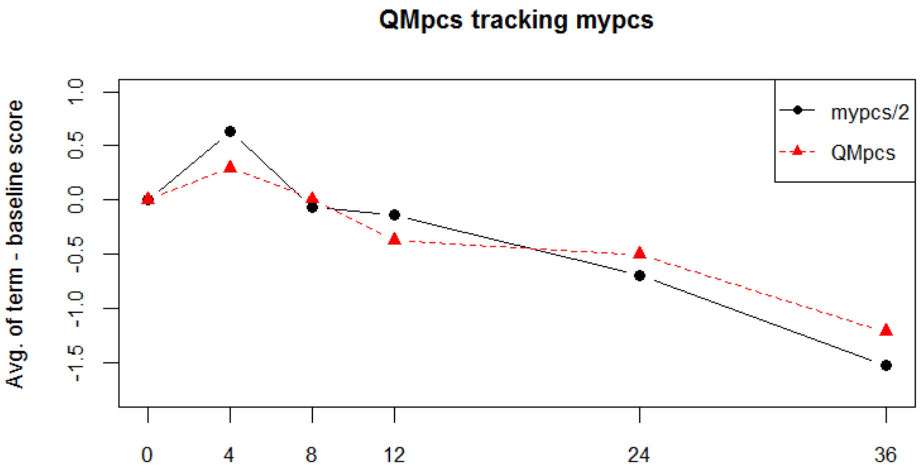


Figure 16: Relationship of *QMpcs* and *mypcs* at the individual patient level. A sample of 9 patients. The average cross-sectional correlation is 0.90. The median of the set of patient-level time series correlations is 0.95.

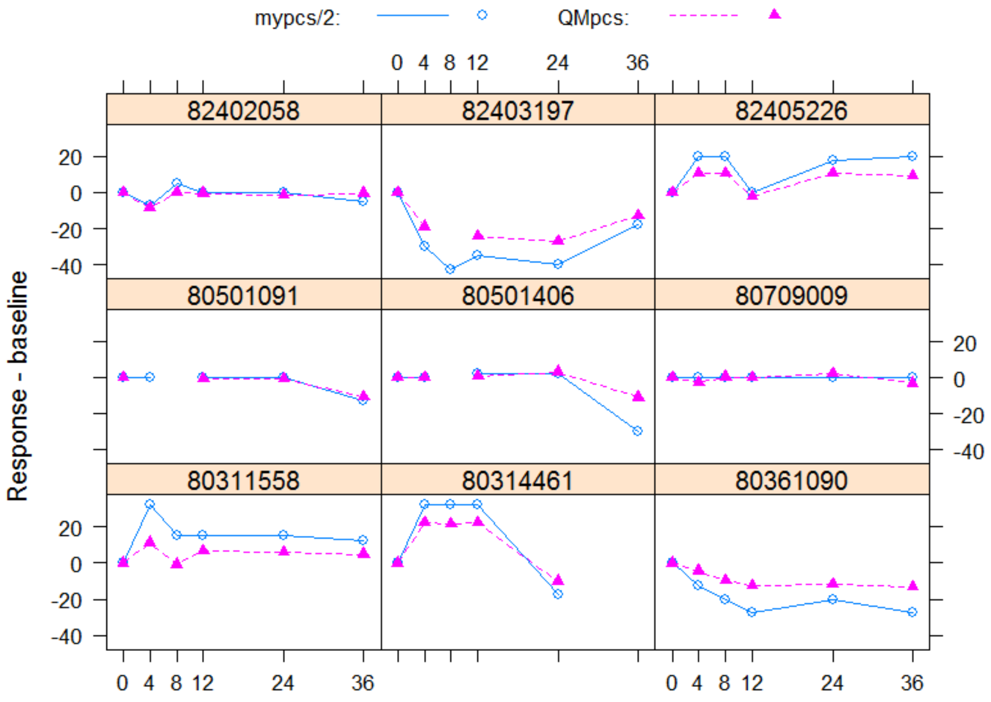


Figure 17: Relationship of *QMmcs* and *mymcs*, with the *mymcs* values divided by 2 for better comparison. Correlation between the averages shown is 0.92.

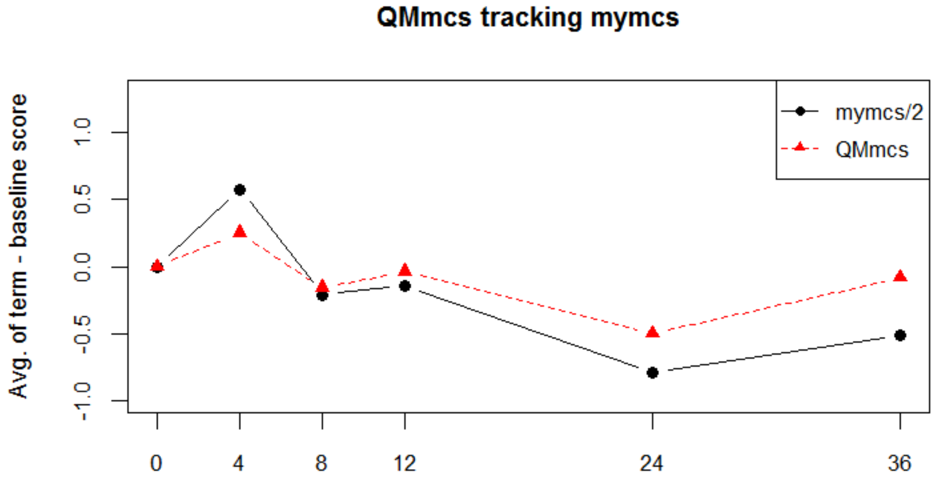


Figure 18: Relationship of *QMmcs* and *mymcs* at the individual patient level. A sample of 9 patients. The average cross-sectional correlation is 0.89. The median of the set of patient-level time series correlations is 0.96.

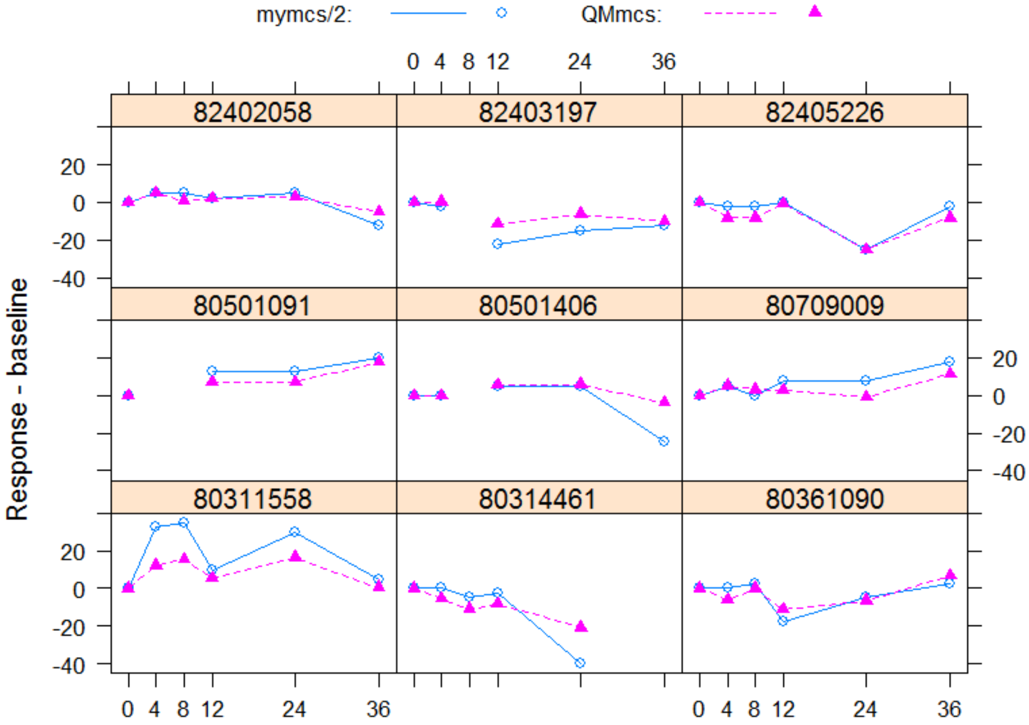


Figure 19: Relationship of *pc1* and *mypcs*, with the *pc1* values halved. Correlation between these averages is 0.94.

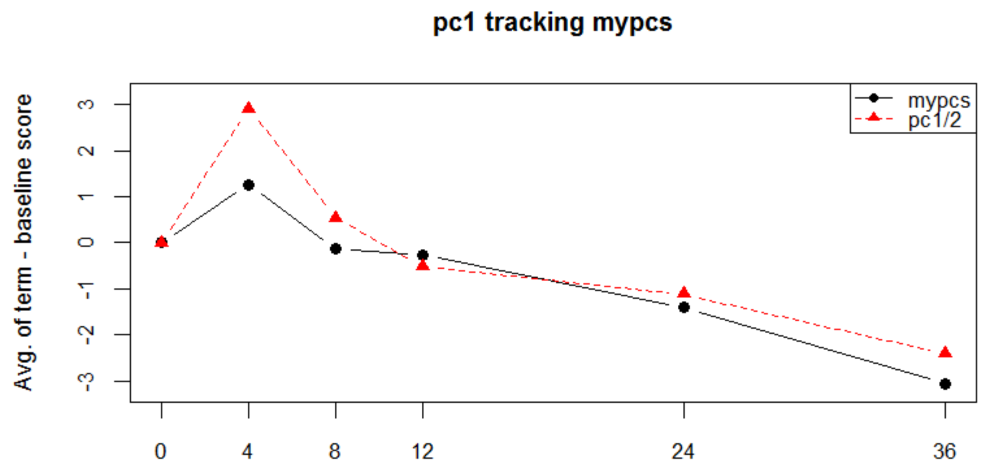


Figure 20: Relationship of *pc1* and *mypcs* at the individual patient level. A sample of 9 patients. The average cross-sectional correlation is 0.84. The median of the set of patient-level time series correlations is 0.91.

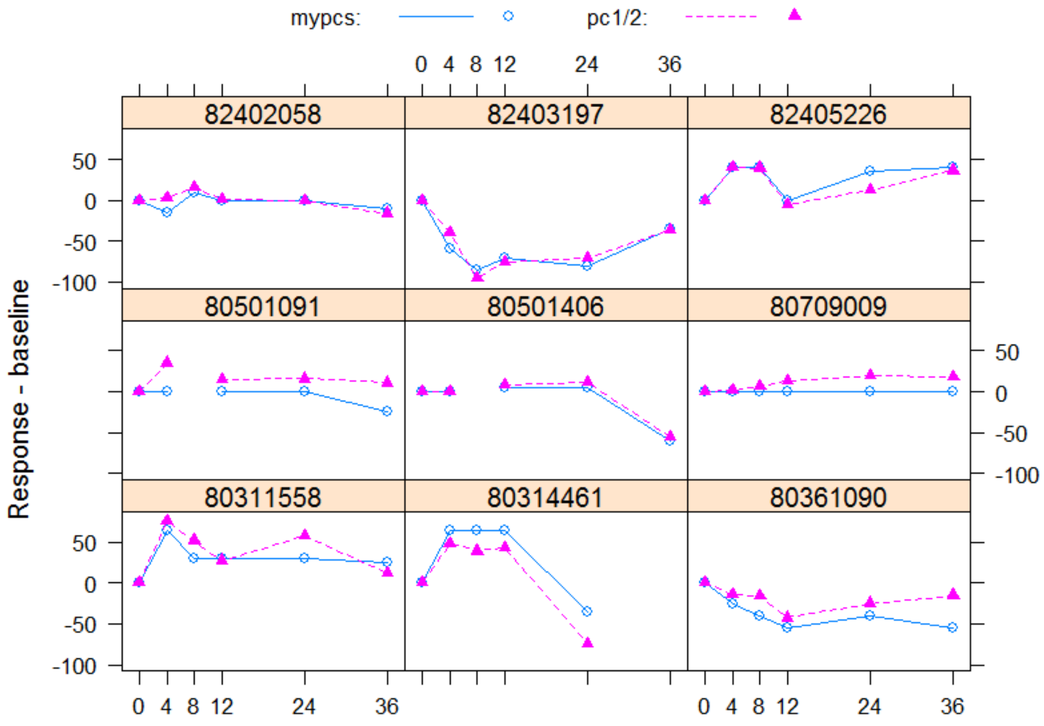


Figure 21: Relationship of *pc2* and *mymcs*, with the *pc2* values divided by 4. Correlation between the averages is -0.63.

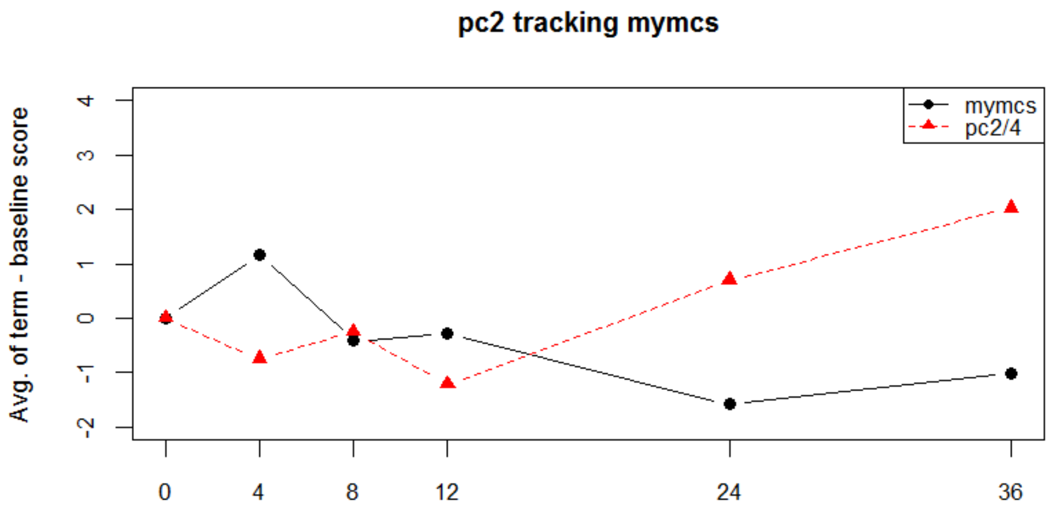
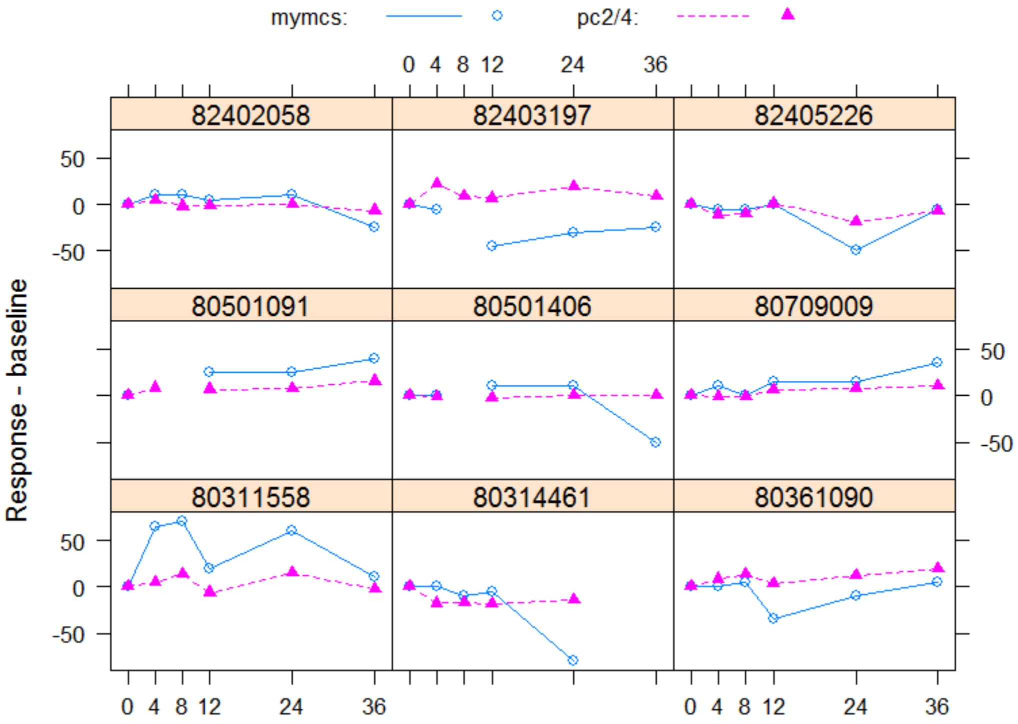


Figure 22: Relationship of *pc2* and *mymcs* at the individual patient level. A sample of 9 patients. The average cross-sectional correlation is 0.68. The median of the set of patient-level time series correlations is 0.68.



Conclusions for this section

One question of interest in this investigation has been whether QualityMetric's SF-12 summary scores do a better job telling us about patients' perceptions of their health and a better job measuring the changes in these perceptions than a simple summary like *allqol*, which is just the average of the numeric scores for each question on the SF-12 (see Figure 1 for the numeric scores). Similarly, *mypcs* and *mymcs* are just simple averages of the numeric scores of certain sets of questions on the SF-12, whereas the construction of *QMpcs* and *QMmcs* is far more complicated. The plots in this section give us a sense of what we gain for the price of that complexity.

The results in Table 14 tell us that *QMpcs* and *QMmcs* together are better in predicting changes in patients' perceptions of their overall health than *allqol* alone, or than *mypcs* and *mymcs* together, or than *pc1* + *pc2*. Only *pc1* + *pc3* is doing a better job than the QM summaries. While Table 14 summarizes how well the different summary variables track or predict the changes over time in patients' perceptions of their overall health, Table 15 tells us how well the pairs of predictors are explaining the variability, at each measurement period, of the *curhlth* scores. The results in Table 15 seem to corroborate the results of Table 14; we see that the QM summaries are doing a better job explaining the variability in *curhlth* than all the other pairs of predictors, except for the *pc1* + *pc3* combination.

The top half of Table 16 provides additional evidence for saying that *QMpcs* reflects perceptions of a distinctly physical dimension of health and for saying that *QMmcs* reflects perceptions of a distinctly mental dimension of health. Figures 15-18 provide graphical complements to this part of Table 16. *QMpcs* does an excellent job predicting changes from baseline in the *mypcs* values, and *QMmcs* does an excellent job predicting changes from baseline in the *mymcs* values. But *QMpcs* and *QMmcs* have almost no ability to predict changes from baseline for *mymcs* and *mypcs*, respectively.

It is very interesting to see in Table 16 that while the median for the set of patient-level correlations between the *mypcs* and *mymcs* time series is 0.42, the median for the set of correlations between the *QMpcs* and *QMmcs* times series is nearly the same in absolute value but the opposite in sign (-0.40). This reaffirms what we have already observed: that the relationship between QM's summaries is somewhat counterintuitive given what we know about the connection between physical and mental health, a connection that is reflected in the positive correlation between *mypcs* and *mymcs*.

Overall, though, I see no evidence for thinking that *QMpcs* is a poor representation of patients' perceptions of physical health, or that *QMmcs* is a poor summary for patients' perceptions of mental health. Together the two variables do an excellent job predicting changes in patients' perceptions of their overall health.

The Principal Components

Aim of this section

Pc1 and *pc2* are of special interest as benchmarks because they capture the most variability in the data that any two standardized linear combinations of the variables (*hlth*, *physfunc*, *physrole*, etc.) can capture. *QMpcs* and *QMmcs* are also linear combinations of variables associated with the SF-12. So it has made sense to consider whether QualityMetric's summaries do as good a job as *pc1* and *pc2* in measuring changes in health across time. We have seen that they in fact do a better job than the principal components.

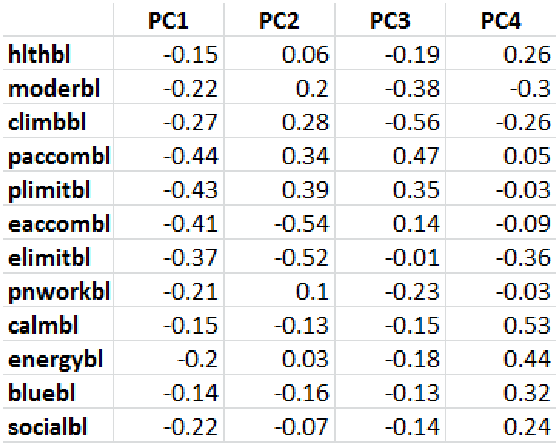
However, since *QMpcs* and *QMmcs* are derived from the SF-36 PCS and MCS, which in turn are constructed using principal components analysis (PCA), it is worth seeing if we can learn more about the nature of *QMpcs* and *QMmcs* by looking a bit more at *pc1 – pc4*. The aim of this section, then, is simply to provide a little more information about the principal components and make a few more comparisons and contrasts with QualityMetric's summary scores.

PCA details

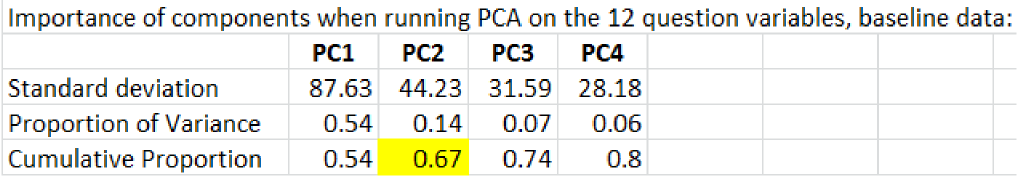
Figure 1 above shows the weights used in the construction of *QMpcs* and *QMmcs*. For each summary score there are 47 weights. *Pc1* and *pc2*, however, are a reduction of 8, rather than 47, variables down to 2. This makes it difficult to draw comparisons between the sets of weights. I didn't work with the 47 indicator variables in Figure 1 because the PCA on them resulted in the first two principal components capturing only around 27% of the variability in the data. Competitive summaries will need to capture more of the variability than this. When using the 12 variables associated with the 12 questions, the proportion of variance captured by pc1 and pc2 is around 67% for the baseline data (Table 17). If we run PCA on the 8 domain variables, we do a little better: the proportion of variance captured is around 71% (Table 18).

We can see from Table 18 that the variables with the largest weights for pc1 are *physrole* (an average of Questions 4 and 5 which deal with physical role limitations) and *emorole* (an average of Questions 6 and 7 which deal with role limitations due to emotional problems such as feeling depressed or anxious).[15] The next heaviest loadings for pc1 are associated with *physfunc*, *pnwork*, *energy*, and *social*. *Pc1* for the baseline data is representing the physical dimension of QoL only slightly more strongly than the mental dimension. Notice that the signs on the weights for all of the pc1 terms are the same. The variables with the largest weights for *pc2* are again *emorole* and *physrole*, but with the former having a significantly larger weight and opposite sign. In fact, variables *emorole, menhlth*, and *social* stand out in *pc2*, each having a sign opposite of the remaining five variables. *Pc2* is thus related to the physical and mental dimensions of QoL in a way that *pc1* is not.

Table 17: Showing the weights for the first four principal components, baseline data only, for the 12 variables associated with the 12 SF-12 questions, along with the cumulative proportion of variance.



	PC1	PC2	PC3	PC4
hlthbl	-0.15	0.06	-0.19	0.26
moderbl	-0.22	0.2	-0.38	-0.3
climbbl	-0.27	0.28	-0.56	-0.26
paccombl	-0.44	0.34	0.47	0.05
plimitbl	-0.43	0.39	0.35	-0.03
eaccombl	-0.41	-0.54	0.14	-0.09
elimitbl	-0.37	-0.52	-0.01	-0.36
pnworkbl	-0.21	0.1	-0.23	-0.03
calmbl	-0.15	-0.13	-0.15	0.53
energybl	-0.2	0.03	-0.18	0.44
bluebl	-0.14	-0.16	-0.13	0.32
socialbl	-0.22	-0.07	-0.14	0.24

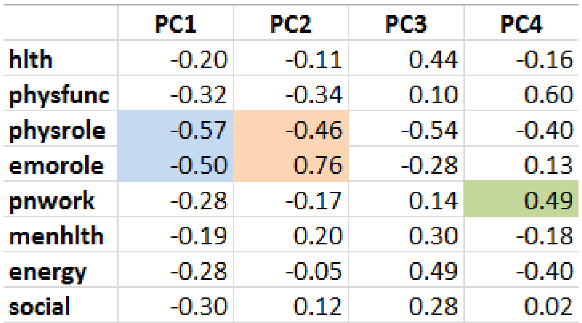


Importance of components when running PCA on the 12 question variables, baseline data:

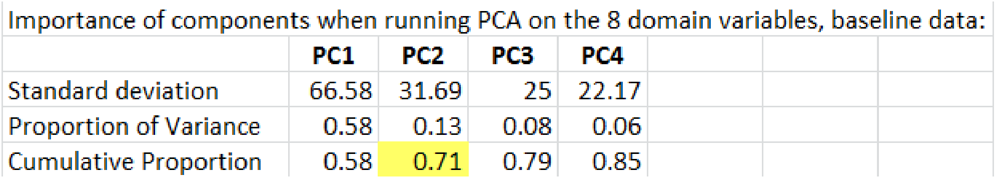
	PC1	PC2	PC3	PC4
Standard deviation	87.63	44.23	31.59	28.18
Proportion of Variance	0.54	0.14	0.07	0.06
Cumulative Proportion	0.54	0.67	0.74	0.8

[15] The set of weights for each principal component are unique up to a multiple of -1.

Table 18: Showing the weights for the first four principal components, baseline data only, for the 8 domain variables, along with the cumulative proportion of variance.



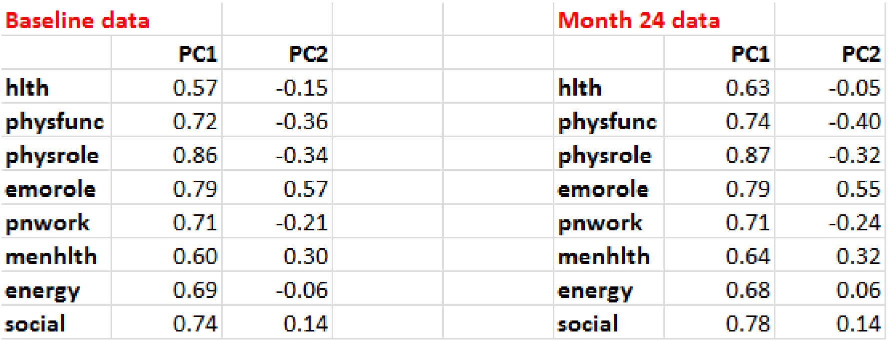
	PC1	PC2	PC3	PC4
hlth	-0.20	-0.11	0.44	-0.16
physfunc	-0.32	-0.34	0.10	0.60
physrole	-0.57	-0.46	-0.54	-0.40
emorole	-0.50	0.76	-0.28	0.13
pnwork	-0.28	-0.17	0.14	0.49
menhlth	-0.19	0.20	0.30	-0.18
energy	-0.28	-0.05	0.49	-0.40
social	-0.30	0.12	0.28	0.02



Importance of components when running PCA on the 8 domain variables, baseline data:

	PC1	PC2	PC3	PC4
Standard deviation	66.58	31.69	25	22.17
Proportion of Variance	0.58	0.13	0.08	0.06
Cumulative Proportion	0.58	0.71	0.79	0.85

Table 19: Showing the correlations of *pc1* and *pc2* with the 8 domain variables for the baseline and month 24 data.



Baseline data				Month 24 data		
	PC1	PC2			PC1	PC2
hlth	0.57	-0.15		**hlth**	0.63	-0.05
physfunc	0.72	-0.36		**physfunc**	0.74	-0.40
physrole	0.86	-0.34		**physrole**	0.87	-0.32
emorole	0.79	0.57		**emorole**	0.79	0.55
pnwork	0.71	-0.21		**pnwork**	0.71	-0.24
menhlth	0.60	0.30		**menhlth**	0.64	0.32
energy	0.69	-0.06		**energy**	0.68	0.06
social	0.74	0.14		**social**	0.78	0.14

In Table 19 we see the correlations of *pc1* and *pc2* with the 8 domain variables when the principal components analysis is done on the baseline data and on the month 24 data. Figure 23 provides graphical representations of these correlation structures. Figure 23 shows us that Question 1 (*hlth*), the alternative form of Question 13 (*curhlth*), is "represented" least well by the first two principal components. Also worth noting in Table 19 is the fact that *pc2* has positive correlation with the mental health domains represented here by *emorole* and *menhlth*, but a negative correlation with the physical health domains (*physfunc*, *physrole*, and *pnwork*). The sign differences on the correlations in Table 19 reflect the sign differences on the weights of *pc1* and *pc2* that we see in Table 18.

In Figure 23 notice how the three physical health domains (2, 3, 5), which are involved in the construction of *mypcs*, are clustered together and how the two mental health domains (4, 6), which are involved in the construction of *mymcs*, are relatively close together and in a different quadrant. Figure 23 gives us the best sense of how *pc1* and *pc2* work together to capture the variability of the different domain variables.

Figure 23: Showing how *pc1* and *pc2* are correlated with the 8 domain variables. A number outside of the inner circle tells us that more than 75% of the variation in the associated variable is captured by *pc1* and *pc2* combined.

Legend: 1= *hlth*; 2= *physfunc*; 3= *physrole*; 4= *emorole*; 5= *pnwork*; 6= *menhlth*; 7= *energy*; 8= *social*.

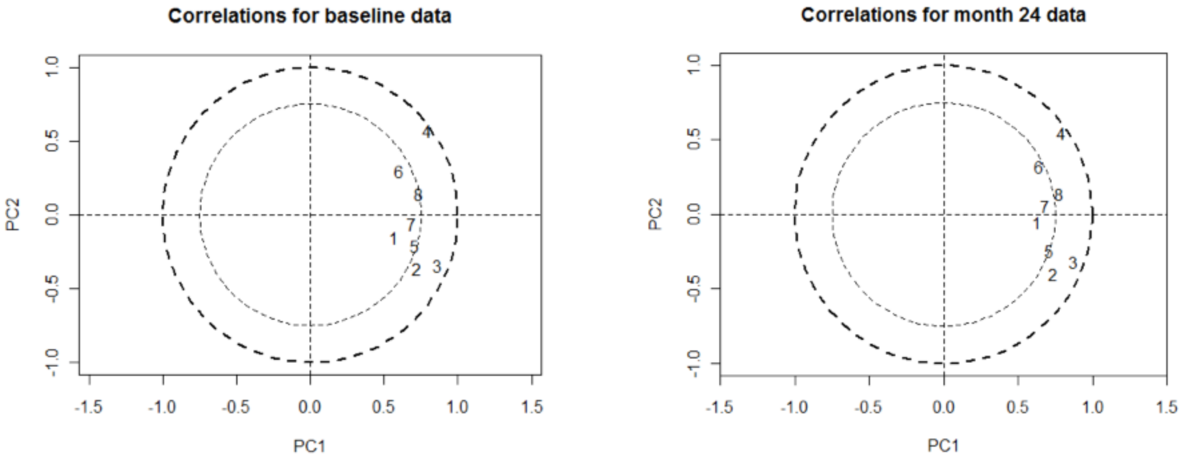
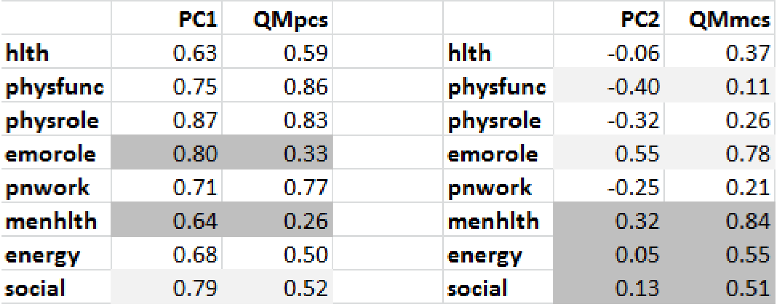


Table 20: Showing the correlations between *pc1*, *pc2* and the 8 domain variables, and between *QMpcs*, *QMmcs* and the 8 domain variables. For the month 24 data only.

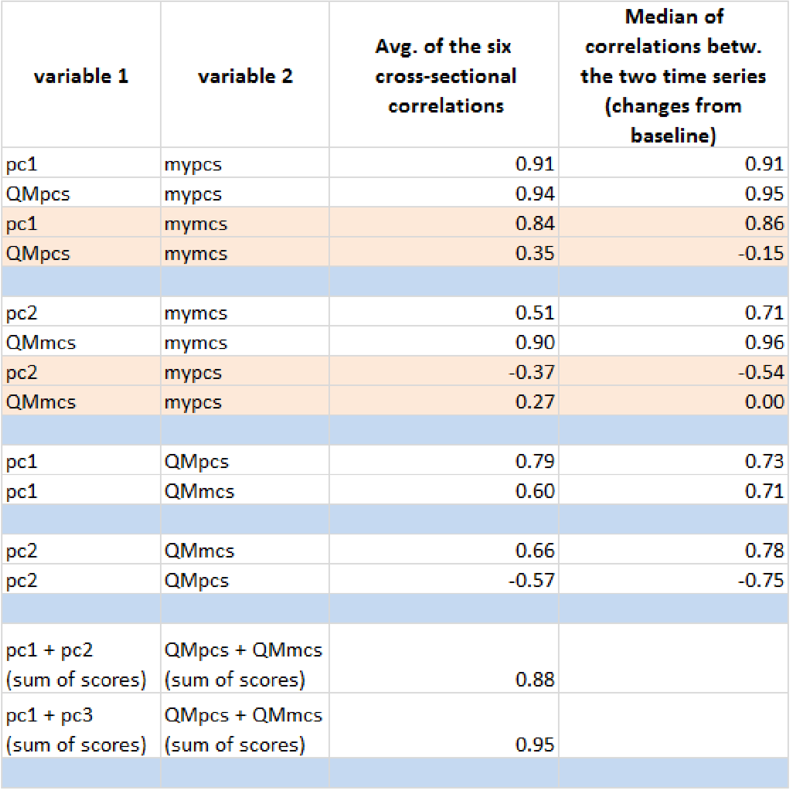


	PC1	QMpcs			PC2	QMmcs
hlth	0.63	0.59		hlth	-0.06	0.37
physfunc	0.75	0.86		physfunc	-0.40	0.11
physrole	0.87	0.83		physrole	-0.32	0.26
emorole	0.80	0.33		emorole	0.55	0.78
pnwork	0.71	0.77		pnwork	-0.25	0.21
menhlth	0.64	0.26		menhlth	0.32	0.84
energy	0.68	0.50		energy	0.05	0.55
social	0.79	0.52		social	0.13	0.51

Table 20 allows us to see exactly how *pc1* differs from *QMpcs* in terms of correlations with the 8 domain variables, and how *pc2* differs from *QMmcs* in terms of these correlations. The rows filled in dark grey highlight the sharpest distinctions, while the rows filled in light grey highlight relatively strong differences. We see that *pc1* places far more weight than *QMpcs* does on *emorole, menhlth,* and *social*. And we see that *pc2* places far less weight than *QMmcs* does on *emorole, menhlth, social,* and *energy* while also placing far greater weight on *physfunc*.

The differences in ability to distinguish the physical and the mental components that we see between these pairs of variables in Table 20 is quantified in another way in Table 21. Looking at the left column of numbers in Table 21 first, we see that *pc1*'s correlation with *mypcs* is close to its correlation with *mymcs*—a difference of only 7 percentage points. Whereas for *QMpcs* this difference is about 60 percentage points. Similarly, the difference in absolute magnitude between *pc2*'s correlation with *mymcs* and its correlation with *mypcs* is 14 points, whereas this difference for *QMmcs* is again close to 60 points. These differences in discriminating ability are even sharper in the far right column.

Table 21: Showing cross-sectional and patient-level time series correlations between the different summary variables in order to point out differences between *pc1* and *QMpcs* and between *pc2* and *QMmcs*. The last two rows may or may not tell us something important about the nature of QM's summary scores.



variable 1	variable 2	Avg. of the six cross-sectional correlations	Median of correlations betw. the two time series (changes from baseline)
pc1	mypcs	0.91	0.91
QMpcs	mypcs	0.94	0.95
pc1	mymcs	0.84	0.86
QMpcs	mymcs	0.35	-0.15
pc2	mymcs	0.51	0.71
QMmcs	mymcs	0.90	0.96
pc2	mypcs	-0.37	-0.54
QMmcs	mypcs	0.27	0.00
pc1	QMpcs	0.79	0.73
pc1	QMmcs	0.60	0.71
pc2	QMmcs	0.66	0.78
pc2	QMpcs	-0.57	-0.75
pc1 + pc2 (sum of scores)	QMpcs + QMmcs (sum of scores)	0.88	
pc1 + pc3 (sum of scores)	QMpcs + QMmcs (sum of scores)	0.95	

PCA Observations

We have seen that *QMpcs* has rough similarities to *pc1* and that *QMmcs* has rough similarities to *pc2*. A major difference between the two pairs of variables, observed in the conclusions on tracking, is that neither *pc1* nor *pc2* make a very good distinction between perceptions of physical health and perceptions of mental health. Table 20 shows one way of quantifying the differences between these pairs of variables and the degree to which *pc1* and *pc2* are not doing as good a job discriminating the mental and physical as *QMpcs* and *QMmcs*. This inability to make the distinction that the QM summaries are making is also evident in Table 21.

One wonders then how the SF-36 PCS and MCS scores which were derived through PCA came to have the weights they did. Presumably the SF-36 PCS derives from a first principal component and the SF-36 MCS derives from a second principal component. Was the data treated in a certain way before the PCA was run in order to have the two summary scores do such a good job distinguishing between perceptions of physical and mental health? And how were the SF-12 PCS and MCS derived from the SF-36 summaries? We are told by Ware and Kosinski (see Figure 2) that the three domains which correlate most highly with the physical component of the SF-36 are "Physical Functioning, Role-Physical, and Bodily Pain". In Table 20 we see that while *physfunc*, *physrole,* and *pnwork* all correlate highly with *pc1*, *emorole* and *social* correlate more highly with *pc1* than does *physfunc*. Further, *menhlth* and *energy* also have relatively high correlations with *pc1*. Ware and Kosinski further observe that "the mental health component [of the SF-36] correlates most highly with the Mental Health, Role-Emotional, and Social Functioning [domains]". But in Table 20 we see that *physfunc* has a higher correlation with *pc2* than does *menhlth* (taking the absolute values of the correlations), and *physrole* is as strongly correlated with *pc2* as *menhlth* is.

Other observations from Table 21:

- *QMpcs* is more highly correlated with *mypcs* than *pc1* is
- *QMmcs* is more highly correlated with *mymcs* than *pc2* is
- *QMpcs* does a slightly better job than *pc1* tracking changes from baseline in *mypcs*
- *QMmcs* does a much better job than *pc2* tracking changes from baseline in *mymcs*

Conclusion

The impetus for this investigation was the observation that the weights used in the computation of the SF-12 PCS and MCS scores are hard to interpret. Indeed, for several questions on the SF-12 the weights appear counter-intuitive: Why are the weight assignments for some of the questions non-monotonic when higher scores are always supposed to reflect more positive perceptions of health? Why is it the case that the more pain one has, the higher one's MCS score? Or the more downhearted and blue one is, the higher one's PCS score? These questions, in turn, led us to ask whether the QM summaries are properly measuring better and worse perceptions of health, including whether *QMpcs* is measuring a distinctly physical dimension of patients' perceptions of QoL, and whether *QMmcs* is measuring a distinctly mental dimension.

For this investigation the only benchmark of overall health that I had to work with that wasn't used in the computation of the QM summaries was *curhlth*. But even this was not an objective, independent measure of patients' health. It was another score the patients assigned to their health each time they took the SF-12. We saw that the average cross-sectional correlation between *curhlth* and *hlth* (another measure of overall health) was only 0.67 (Table 7), indicating that the variability in subjective assessments of one's health is quite large, even between the beginning of a survey and its end—what might be a difference in time of only three minutes or so. I also made use of *mypcs, mymcs, pc1,* and *pc2* as both benchmarks and alternative candidate summaries.

Despite these limitations, the results of the analysis for this investigation are encouraging. The QM summaries outperformed all of the candidate summaries in terms of tracking the changes in *curhlth*, and they "outperformed" the principal components in terms of distinguishing between the mental and physical dimensions of patients' perceptions of their health.[10] The construction of the QM summaries may be complex, but that complexity has a payoff. We may not be entirely at ease with the lack of interpretability of the weights used in the computation of the QM summaries, but we should feel much more confident that the summaries are doing the work we want them to do.

[10] While *~pc1 + pc3* outperformed *~QMpcs + QMmcs* in tracking changes in *curhlth*, the former cannot be used to compare QoL measurements from clinical trials using the SF-36. We should also keep in mind that we have only one set of weights for *QMpcs* and one set of weights for *QMmcs*. We use the same weights, in other words, at every measurement period and with every dataset. This is not the case with the pc1 through pc4 that I used in my analysis. So it is not quite correct to say that *~pc1 + pc3* outperformed *~QMpcs + QMmcs* in tracking changes in *curhlth*.

References

1. Ware JE Jr., Kosinski M., Keller SD. 1996. "A 12-item short-form health survey: construction of scales and preliminary test of reliability and validity." *Med Care* 34: 220-233.

2. Han C., Pulling C., Telke S., Huppler Hullsiek K. 2002. "Assessing the utility of five domains in SF-12 Health Status Questionnaire in an AIDS clinical trial." *AIDS* 16: 431-439.

3. Burman W., Grund B., et. al. 2008. "The Impact of Episodic CD4 Cell Count-Guided Antiretroviral Therapy on Quality of Life". *Journal of Acquired Immune Deficiency Syndrome* 47: 185-193.

4. Ware JE Jr., Kosinski M, Turner-Bowker DM, Gandek B. 2005. "How to score version 2 of the SF-12 health survey (with a supplement documenting version 1)". Lincoln, RI: QualityMetric, Inc.; and Boston, MA: Health Assessment Lab.

5. Ware JE Jr., Kosinski, M. 2005 (8th printing). "SF-36 Physical and Mental Health Summary Scales: A Manual for Users of Version 1, Second Edition". Lincoln, RI: QualityMetric, Inc.

6. CPCRA Form 65A-BAS-1 Version 1 November 2001. SMART Quality of Life and Healthcare Utilization Substudy.