From: **Dean Gladish** gladish.dean@gmail.com
Subject: Re: t-test for women and men subsets, promis_first_assessment_data_v1.1
Date: May 17, 2020 at 8:37 PM
To: Christopher Hanes cshanes@umich.edu
Cc: Dean Gladish gladishd@carleton.edu

Hi Chris.

There are many different aspects of the paper which I would look into. We really need both the HUI-3 and PROMIS 29 scores. Is HUI-3 more useful as an outcome measure?

The title is *Using Linear Equating to Map PROMIS Global Health Items and the PROMIS-29 V. 2 Profile Measure to the Health Utilities Index--Mark 3.* This looks like a good paper and it might serve as a guide for **our goal, which is to use our PROMIS scores to model *other* assessments.** Primarily because of their methodology. So, they're mapping PROMIS 29 scores to HUI-3 scores. HUI-3 is a preference-based measure; that is, it comprises a number of domains which patients can use to describe various aspects of their health; patient-reported values are algorithmically converted to an index score. What they're doing is using analyses of their linear regression to identify significant predictors, and they use something called linear equating (transforming *predicted* HUI-3 scores such that they have the same mean and standard deviation of the *observed* HUI-3 scores) in order to avoid the problem of regression to the mean, in which random variables approach the mean with each new measurement. So what this linear equating does is adjust so that the two forms have a comparable mean and standard deviation. So their abstract elaborates on how useful HUI-3 is for "outcome measures in clinical studies, for monitoring health of populations, and for estimating quality-adjusted life years". They emphasize how well their linear regression models explain variance in the HUI-3 preference score (the R^2 value).

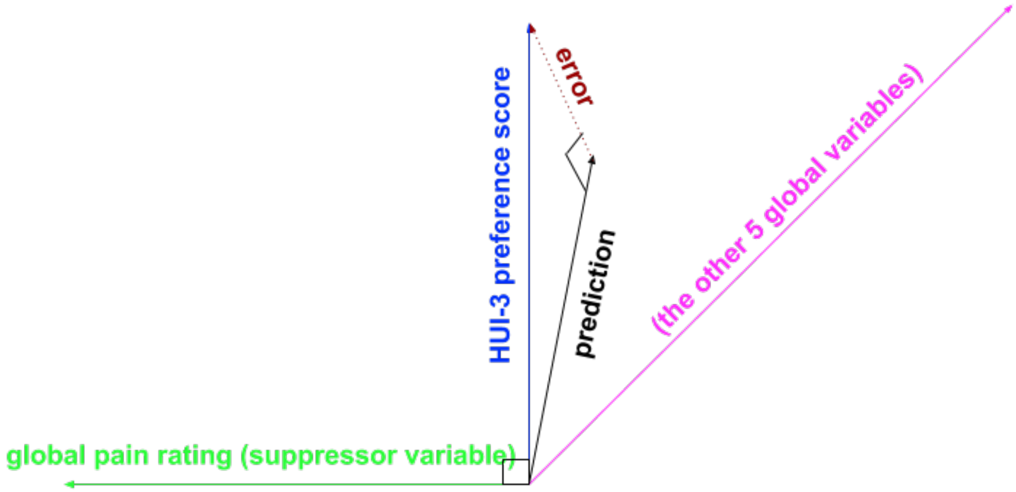
So after giving the specific advantages of HUI-3 as it may be more widely used, and elaborating on their main point which is that all these preference-based measures, health-related quality of life measures (HRQOL, which are useful for monitoring the health of populations, estimating quality-adjusted life years for economic evaluations, and estimating health outcomes for cost-effectiveness) etc., all these measures have different scoring systems. HUI-3 is on a 0 to 1 scale, for example. By October 16, the time of writing, PROMIS measures had not been mapped to anything yet except for EQ-5D-3L. It was uncharted territory, estimating health preference scores based on PROMIS measures was. That is their exigency for writing the paper. Not only that, but the HUI-3 has eight, multi-level attributes. We may not have all the preferences (for HUI-3) available to us, which is why we are following in their footsteps. They followed established recommendations for reporting mapping studies, which are on a checklist.

Now. They start mentioning how HUI-3 scores are based on a multi-attribute utility function. This function is derived using visual analog scale (that is, a survey in which respondents specify their level of agreement to a statement by indicating a position along a continuous line between two endpoints) as well as via using standard gamble elicited preferences (suppose a patient has two alternatives. Alternative A, in which the person will live with a particular health problem with certainty for the remainder of his or her life, and Alternative B, in which a risky treatment allows them to 1. live in a state of optimal health with probability p, and 2. undergo immediate death, with probability (1-p). Our measurement objective is to identify p for which the patient is indifferent between A and B; the health state valuation for this particular health problem of interest is equal to p, and this is the standard gamble value).

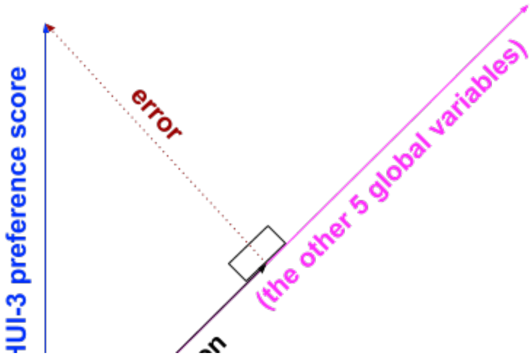
One major difference is that our participants haven't *all* done the HUI-3 test. What incentives do our participants have to complete the surveys? It also looks like they had more demographic information, which I will go into in just a moment. So Chris, what they ended up doing is as described: they took their HUI-3 attribute levels and estimated Spearman correlations between those and their corresponding PROMIS domain scores (the Spearman correlation is just a non-parametric measure of rank correlation, the Pearson correlation coefficient between the rank variables, and they're essentially just finding the correlation between HUI-3 attribute levels and their corresponding PROMIS domain scores). They mention ordinary least squares regression equations - these are in order to minimize the sum of the squares of residuals; that is, to minimize the sum of the squares of the differences between the observed, dependent variable (HUI-3) and the HUI-3 values predicted by the linear function. I also agree with them, it is a little weird that within the PROMIS conventions larger scale scores correspond to more of the concept depicted in the name, that is, higher scores may or may not represent better health depending on the measurement used. An interesting problem they mention is that regression-based prediction results in biased estimates. This is interesting to me because they suggest specifically that linear regression models overpredict low scores and underpredict high scores. So I'm on the bottom of the third page and it looks like they basically just moved outliers to the closest acceptable value. On a boxplot, this resembles moving all the outlier points and just dragging them to 1.5 x the IQR away from the 3rd quartile. They did this and made sure the predicted scores from each of their three regression models were linearly transformed so that they had the same mean and standard deviation as the observed HUI-3 scores. Essentially, they already had observed HUI-3 scores and they used their mean + standard deviation in order to standardize their predicted scores. Are we going to do this kind of linear equating (a linear transformation) for our translated (from PROMIS to HUI-3) scores? '

Another estimate they obtain is of **capitalization on chance**, which just refers to inferred causality which is really only a random act of chance, **product-moment** (think Pearson-distance to line of best fit) and **intraclass correlations** (do values from the same group tend to be similar?) between predicted and observed HUI-3 preference scores in one random half sample and those in the complement random half sample. Basically, they compared correlations in one half to the correlations in the other half. They also have a bunch of demographic information - race, education, age, marital status - and so they can infer that their data is similar to the general population. A big thing for them is accounting for the variance in the HUI-3 score, so looking at the adjusted R^2 value, the proportion of the variance in the HUI-3 scores which is predictable from the PROMIS scores, is going to be essential. And the other thing they look at is **standardized beta** which measures the strength of unique associations; these betas are calculated by taking an independent variable and subtracting its mean and dividing by its standard deviation. A standardized beta coefficient compares the effect of each individual PROMIS score on the HUI-3 preference score. Now, their product-moment is equal to their intraclass correlation in that both values are at the relatively high value of 0.70, which indicates that between-group differences are just as important as across-group differences; that is why I think our demographic stuff (age and gender) is very valuable. There's also a typo on the second paragraph of the fourth page, right before Results. They said that they estimated HUI-3 preference scores by "using on the second paragraph of the fourth page, right before Results. They said that they estimated HUI-3 preference scores by "using ~~used~~" the regression equation, and I think that they were trying to use the present tense which is what we should do as well.

So the docs might consider creating a table like Hays et al. have in Table 1, wherein estimates resulting from a regression analysis that have been standardized so that the variances of variables (independent HUI-3 or dependent) are all 1. They can say, *our six global health items account for X amount of variance in HUI-3 preference score.* An interesting thing to note is the possibility of **suppression effects:**


error
HUI-3 preference score
prediction
(the other 5 global variables)
global pain rating (suppressor variable)

As indicated, global pain rating is pretty weak as a predictor itself; this is the meaning of zero-order correlation (the correlation between global pain rating and HUI-3 preference score without controlling for the influence of any other variables) being negative; in fact, the global pain rating hardly correlates with the HUI-3 preference score at all *by itself*; this is demonstrated in the above diagram in which those two vectors are orthogonal. However, global pain rating *suppresses* the error (that is, residuals left by the model if it's not included) of the reduced model. However, it is correlated with the error in prediction:

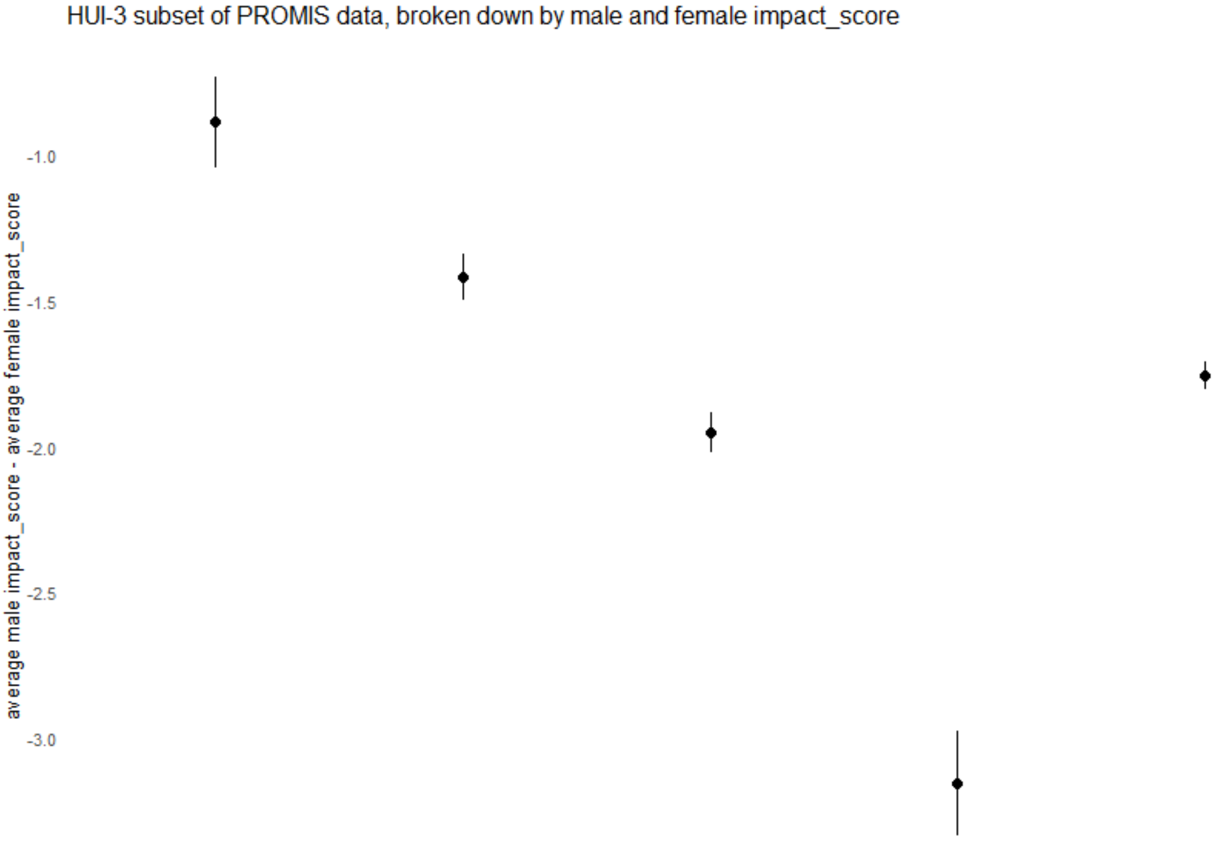


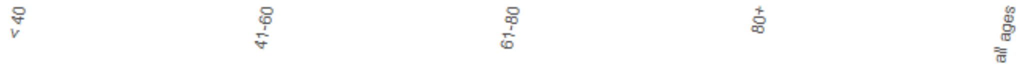
This is just some stuff about how removing a suppressor variable causes the error to increase substantially. I hope all of this can be included in the paper in due time, because this is a topic which will continue to be of importance.

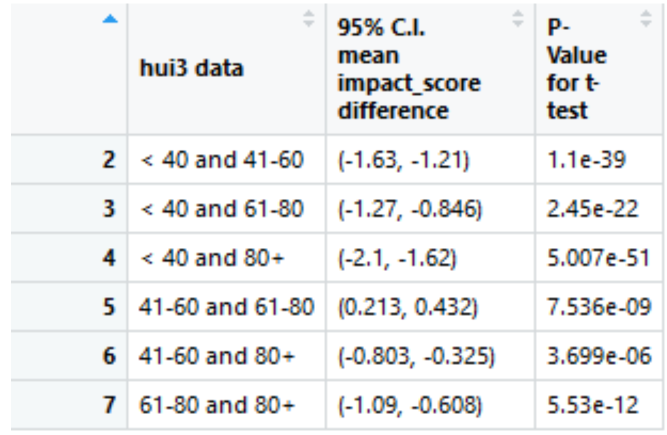
-Dean

On Fri, May 15, 2020 at 4:44 PM Dean Gladish <gladish.dean@gmail.com> wrote:

> Hi Chris,
>
> It's great that the HUI-3 data includes patient_id, because this was essential to linking with the PROMIS data.

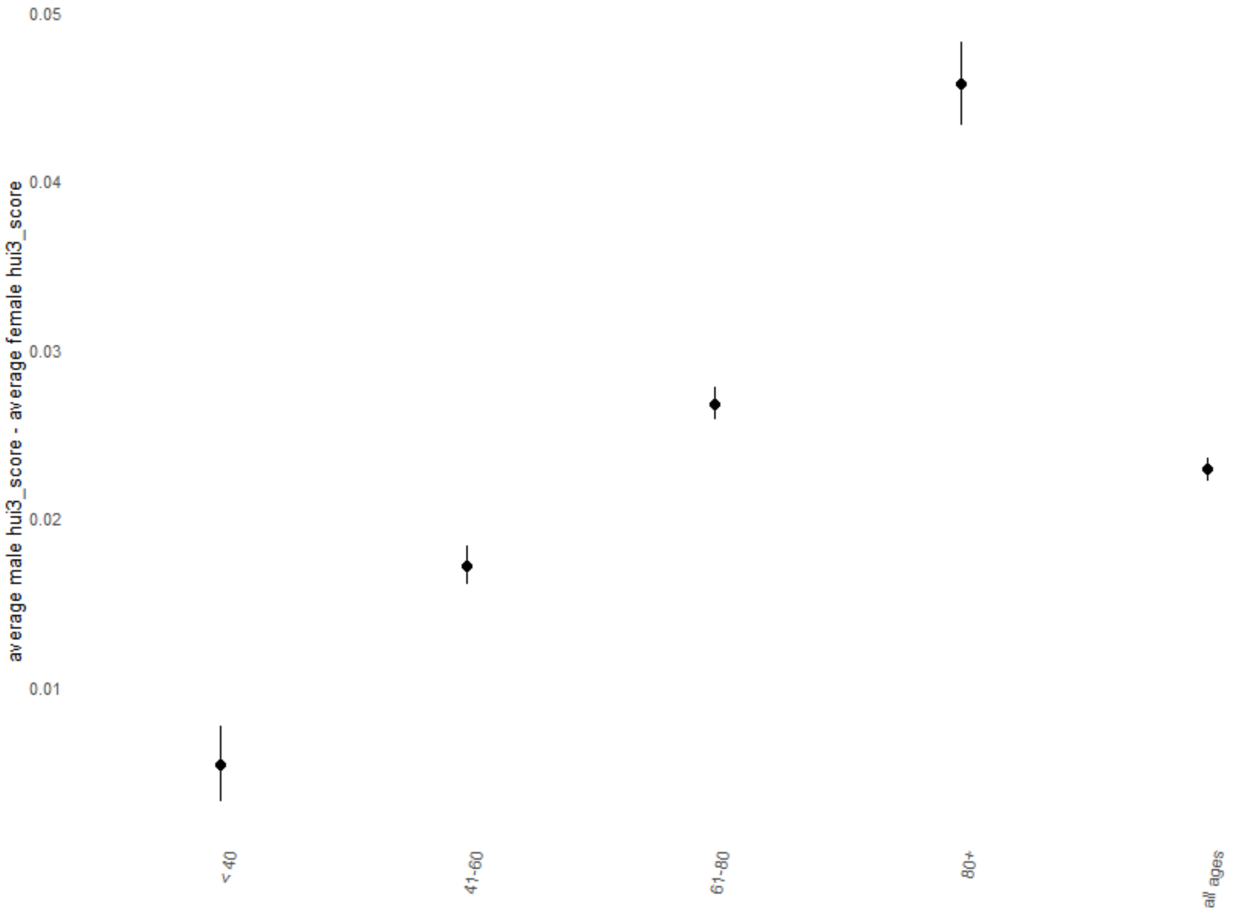

HUI-3 subset of PROMIS data, broken down by male and female impact_score

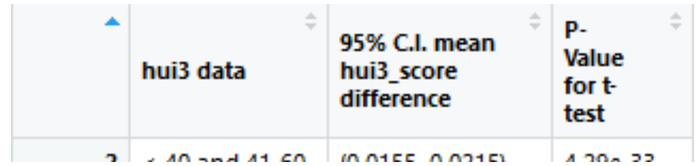




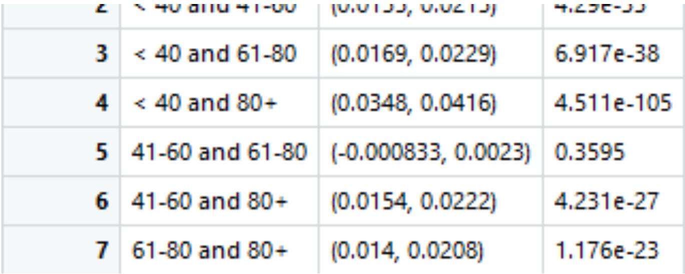
	hui3 data	95% C.I. mean impact_score difference	P-Value for t-test
2	< 40 and 41-60	(-1.63, -1.21)	1.1e-39
3	< 40 and 61-80	(-1.27, -0.846)	2.45e-22
4	< 40 and 80+	(-2.1, -1.62)	5.007e-51
5	41-60 and 61-80	(0.213, 0.432)	7.536e-09
6	41-60 and 80+	(-0.803, -0.325)	3.699e-06
7	61-80 and 80+	(-1.09, -0.608)	5.53e-12

HUI-3 subset of PROMIS data, broken down by male and female hui3_score





	hui3 data	95% C.I. mean hui3_score difference	P-Value for t-test
2	< 40 and 41-60	(0.0155, 0.0215)	4.29e-33



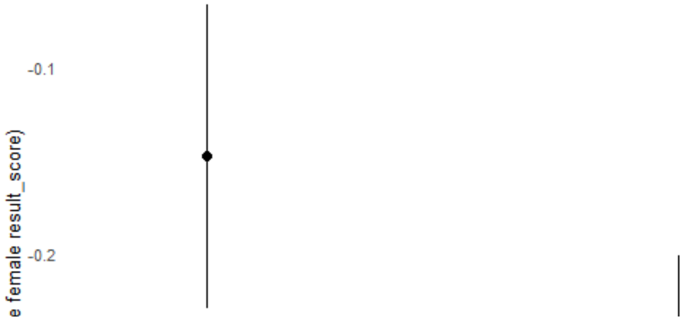
2	< 40 and 41-60	(0.0155, 0.0215)	4.29e-33
3	< 40 and 61-80	(0.0169, 0.0229)	6.917e-38
4	< 40 and 80+	(0.0348, 0.0416)	4.511e-105
5	41-60 and 61-80	(-0.000833, 0.0023)	0.3595
6	41-60 and 80+	(0.0154, 0.0222)	4.231e-27
7	61-80 and 80+	(0.014, 0.0208)	1.176e-23

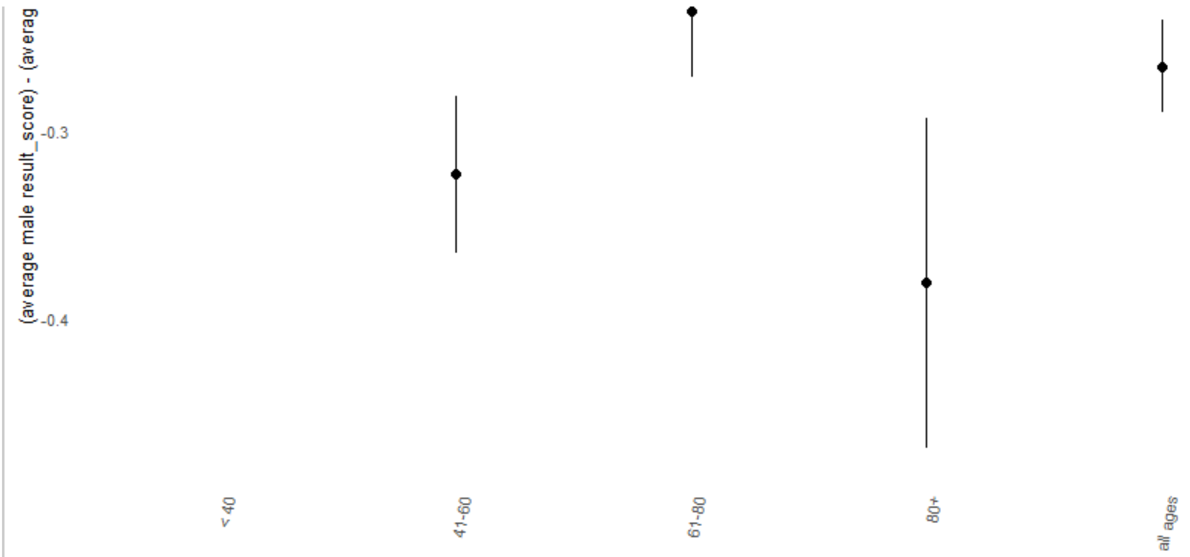
I am simply discovering all kinds of new stuff working for you. It's like we are back in Korea.

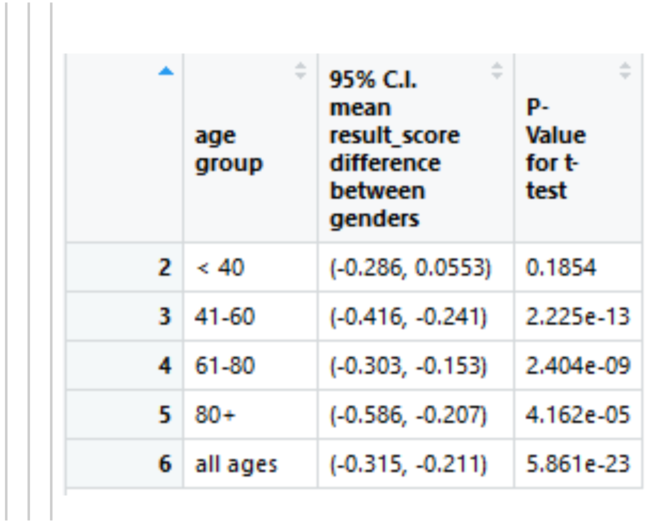
Sincerely,
Dean Gladish

On Thu, May 14, 2020 at 11:44 PM Dean Gladish <gladish.dean@gmail.com> wrote:

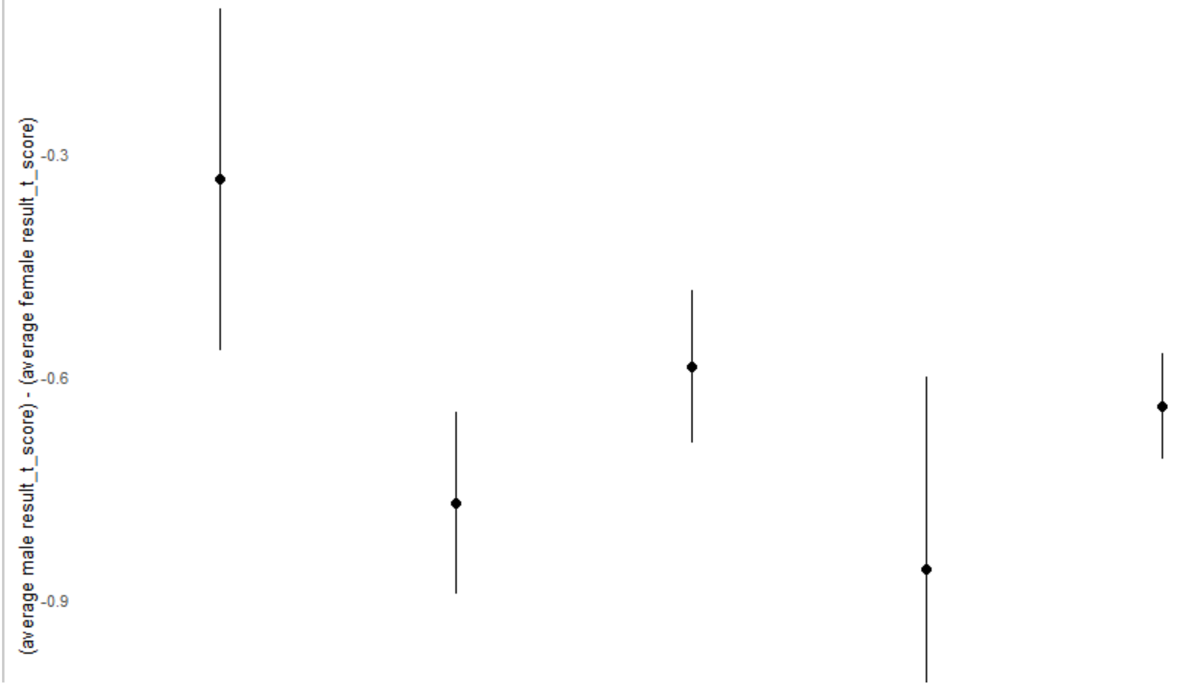
> Hi Chris,
>
> That is really nice that the docs are writing a paper. It's not a problem, I'll do the Welch's t-test for impact score and the hui-3 data. The prediction thing is really cool. Thanks for sending me that paper.
>
> -Dean
>
> On Thu, May 14, 2020 at 11:40 PM Christopher Hanes <cshanes@umich.edu> wrote:
>
>> Hey Dean :)
>>
>> Hopefully, this is the last request. Honestly a bit annoyed with these docs, their process for writing this paper is frustrating for me.
>>
>> To give a little background information for this request, HUI-3 is another widely used assessment that is analogous to PROMIS 29. We have a paper that made a linear regression model that predicts HUI-3 scores using PROMIS 29 scores.
>>
>> Anyways, two more things that Jason is requesting:
>>
>> 1. Welch's t-test numbers comparing impact score between the different age groups and male vs female
>> 2. Welch's t-test numbers comparing hui-3 score between the different age groups and male vs female
>>
>> So basically the same thing you did before but using just impact score and HUI-3. I attached the HUI-3 data. When doing male vs female, it doesn't need to be broken down by age group.
>>
>> I attached the HUI-3 data, please let me know if you need any clarification.
>>
>> Thanks a lot Dean, we couldn't have done this without your help. I think we may have a more immediate project where we model other assessments using PROMIS 29 scores like ZCQ and ODI. This is the paper we used for HUI-3 to PROMIS 29 translation for reference: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5026900/
>>
>> On Tue, May 12, 2020 at 3:01 PM Dean Gladish <gladishd@carleton.edu> wrote:
>>
>>> Hi Chris,
>>>
>>> It looks like when I ran the t-test comparison comparing men vs. women subsets, there are some statistically significant differences and the actual value of the differences suggests that women have slightly higher scores and that the value of the difference diminishes with age; furthermore, result_score and its derivative seem to be staggered while the impact_score demonstrates exponential decrease and yes, there is some difference.

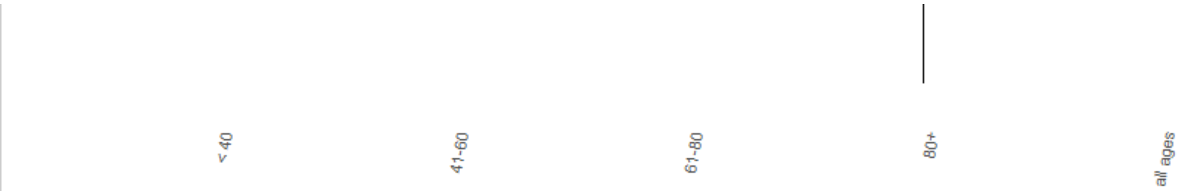




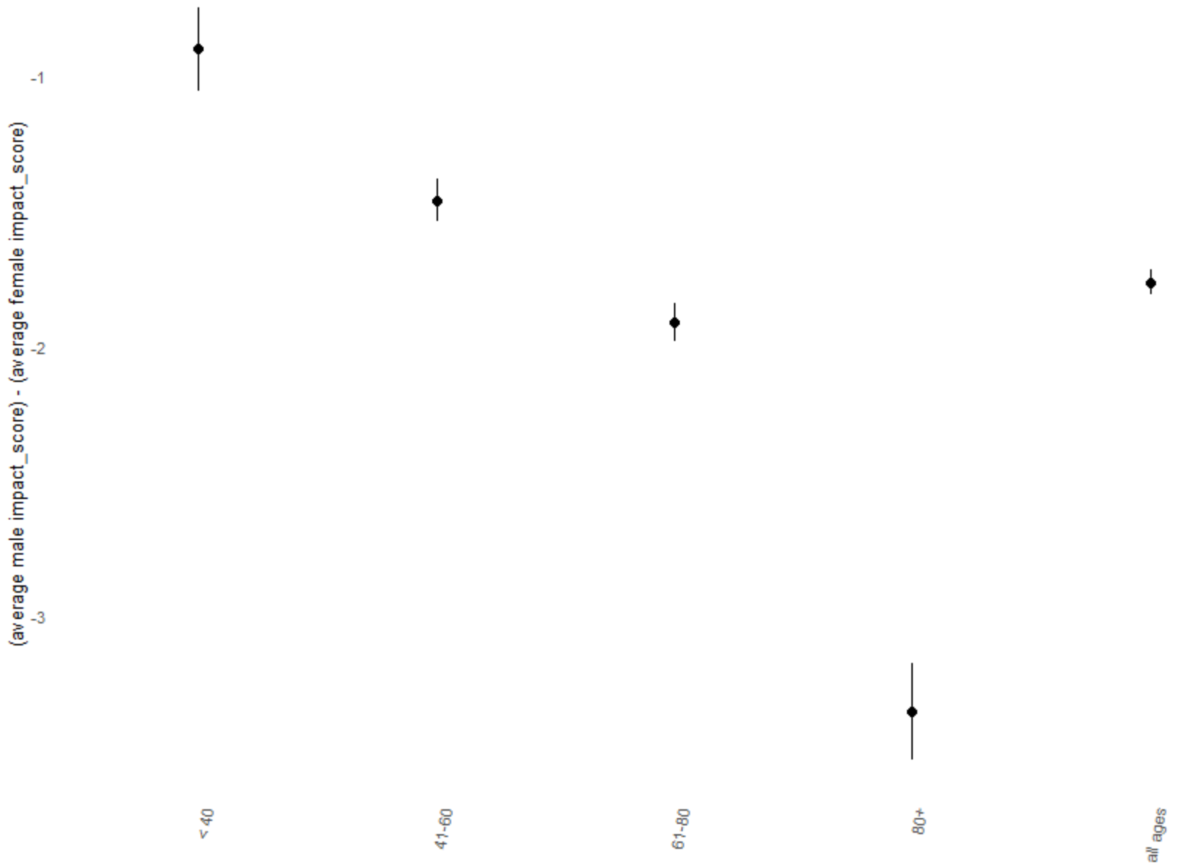


	age group	95% C.I. mean result_score difference between genders	P-Value for t-test
2	< 40	(-0.286, 0.0553)	0.1854
3	41-60	(-0.416, -0.241)	2.225e-13
4	61-80	(-0.303, -0.153)	2.404e-09
5	80+	(-0.586, -0.207)	4.162e-05
6	all ages	(-0.315, -0.211)	5.861e-23

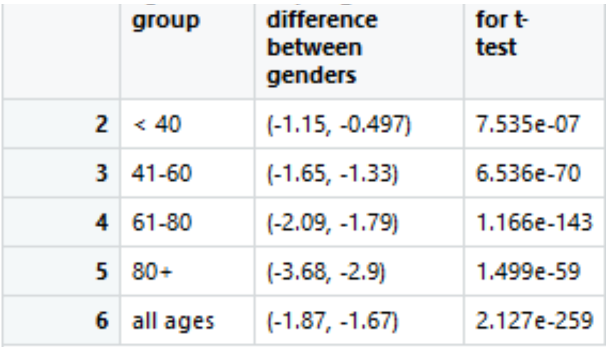




	age group	95% C.I. mean result_t_score difference between genders	P- Value for t-test
2	< 40	(-0.863, 0.162)	0.1803
3	41-60	(-1.11, -0.555)	3.881e-09
4	61-80	(-0.85, -0.385)	1.99e-07
5	80+	(-1.32, -0.0881)	0.02509
6	all ages	(-0.761, -0.435)	5.856e-13



	age	95% C.I. mean impact_score	P- Value



	group	difference between genders	for t-test
2	< 40	(-1.15, -0.497)	7.535e-07
3	41-60	(-1.65, -1.33)	6.536e-70
4	61-80	(-2.09, -1.79)	1.166e-143
5	80+	(-3.68, -2.9)	1.499e-59
6	all ages	(-1.87, -1.67)	2.127e-259

The plots show the difference in means between men and women, with standard deviation of the difference between means drawn through each point. This will help provide some clarity for the t-tests.

-Dean