SUPPLEMENTAL MATERIALS

Supplemental Methods

Internal validity tests:

Attribute dominance

No respondent from the online-panel or the DUHS sample always dominated on one of the four attributes (i.e., years in NYHA class II, years in NYHA class III, risk of death within 30 days, and risk of complications leading to additional 2 days in hospital) that had natural ordering. That is, no respondent selected the alternative with the better level of one specific attribute in all 8 choice questions.

Across-set monotonicity

Across-set monotonicity tests can occur naturally in an experimental design, especially for designs with an opt-out (i.e. no-device) alternative with fixed attribute levels. This is the case for our study. Below is an excerpt of how the test works from Johnson et al. 2019 (1).

> Suppose there are two choice sets, 1 and 2. A1 and B1 designate attribute-level profiles for alternatives A and B in set 1. A2 and B2 designate attribute level profiles in set 2. The alternative C profile is common to both choice sets.
>
> Choice set 1: A1 ~ B1 ~ C
>
> Choice set 2: A2 ~ B2 ~ C
>
> Suppose alternative B is chosen in choice set 1, indicating that B1 is preferred to both A1 and C. If the attributes in B are naturally ordered and B1 is inferior to B2 for every attribute, then choosing C in choice set 2 fails a logic test; if B1 is preferred to C in choice set 1, then B2 must be even more attractive than C in choice set 2. Alternatively, if C is chosen in choice set 2, then logic prohibits choosing B1 in choice set 1 because C was shown to be preferred to the better levels of B2.

In both the web panel and DUHS samples, most respondents (80%) passed the across-set monotonicity tests that were embedded in the design. About 20% of the respondents failed at least one test, of which

[1] Johnson FR, Yang JC, Reed SD. The Internal Validity of Discrete Choice Experiment Data: A Testing Tool for Quantitative Assessments. Value Health. 2019;22(2):157-160.

most (about 87%) had chosen the 'no-device' option at least once in the series of choice questions. Supplemental table I provides the detailed results by data source.

Scope test

Scope tests are performed to evaluate whether respondents appear to be sensitive to the magnitude of shown across interval-scale attribute levels. Study participants in the web panel and DUHS panels were randomized to a low-risk or high-risk arm. For participants in the low-risk arm, the highest risk of 30-day mortality was 10%. For participants in the high-risk arm, instead of 10%, they saw 15% risks as the highest mortality risk level. Both groups saw the same risk levels for the 0%, 2% and 5% levels. The scope test had two components: one assessed whether respondents in both the low-risk and high-risk arms had similar preferences for overlapping risk levels (i.e., 0%, 2%, and 5%) and the other assessed whether respondents viewing 15% risk levels had more negative preferences than respondents viewing 10% risk levels as a means to evaluate whether they were responding to the magnitude of the maximum risk shown or possibly recoding risks in an ordinal fashion.

To analyze the scope test data, we estimated separate parameters for each risk level by arm using a dummy-coded version of the mortality risk attribute levels (i.e. omitting the preference estimate for 0%). We then interacted a binary "arm" indicator with each level of the remaining risk levels as interaction terms in the mixed-logit model with the following model specification:

$$V = \alpha_{opt-in} + \left(\beta_{Risk_{2\%}} + \beta_{Arm*Risk_{2\%}} * Arm\right) * Risk_{2\%} + \left(\beta_{Risk_{5\%}} + \beta_{Arm*Risk_{5\%}} * Arm\right) * Risk_{5\%} + \left(\beta_{Risk_{max}} + \beta_{Arm*Risk_{max}} * Arm\right) * Risk_{max} + \sum \beta_x * X + \varepsilon$$

where V represents the utility associated with a specific combination of attribute levels, "arm" is equal to 1 if the respondent evaluated the high-risk arm and 0 if the respondent evaluated the low-risk arm. The 0% risk is the omitted category in the dummy-coded attributes. Riskmax represents a single indicator variable for 10% mortality risk in the low-risk arm and 15% mortality risk in the high-risk arm. $\sum \beta_x * X + \varepsilon$ captured the effect of the remaining study attributes and the modelling error.

If respondents evaluated the actual numeric risk levels, preference weights for overlapping levels (i.e., 0%, 2%, and 5%) would not be statistically significantly different between arms. To evaluate this, we conducted a scale-adjusted chi-square test of joint significance on $\beta_{Arm*Risk_{2\%}} = 0$ and $\beta_{Arm*Risk_{5\%}} = 0$. The result of this test (p=0.22 for online panel, p=0.22 for DUHS, and p=0.16 for overall) indicated

preference weights for overlapping mortality risk levels (i.e., 2% and 5%) was similar between the two arms in each sample and the combined sample.

Since the preference weights for 2% and 5% risk were not statistically different across arms, then we specified an additional model in which we forced overlapping levels (2% and 5%) to have equal preference weights across arms. To assess whether respondents were sensitive to the magnitude of the maximum mortality risk shown, we assessed whether the importance of increasing the risk from 0% to 15% is statistically significantly larger than the importance of increasing the risk from 0% to 10%. To evaluate this, we conducted a one-tailed test of significance on $\beta_{Arm*Risk_{max}}$ or a test that evaluates the probability that $\beta_{Arm*Risk_{max}}$ is strictly positive. Failing to reject the null hypothesis that $\beta_{Arm*Risk_{max}}$=0 will indicate the data are consistent with attribute-level re-coding.

The findings revealed that we could not reject the null hypothesis (p=0.90 for online panel, p=0.18 for DUHS, and p=0.71 for pooled sample); thus, the importance of increasing the risk from 0% to 15% was not statistically significantly larger than the importance of increasing the risk from 0% to 10%. Failing the second scope test does not necessarily indicate that participants did not pay attention to the maximum risk level. Instead, it could suggest that devices presented in the study did not provide enough improvement in effectiveness for respondents to take on more risk, that incremental increases in risk were less important as risk increased as suggested by prospect theory, or that the study sample size was not adequately powered to detect the difference.

We also tested whether the slope between 0% and 10% and the slope between 0% and 15% were the same. The results indicated the 2 slopes were statistically significantly different from each other (p=0.002 for the online panel, p=0.024 for DUHS, and p=0.001 for overall.) The slope between 0% and 10% was steeper than that between 0% and 15%, meaning that, for a given benefit, the implied MARs using the slope between 0% and 10% are smaller than those using the slope between 0% and 15%. Therefore, to have the most conservative estimates, we applied the slope between 0% and 10% for estimating MARs of death.

Supplemental Tables

Table I. Performance on across-set monotonicity tests

Description	Web panel (n=500)	DUHS Sample (n=126)	Pooled Sample (N=626)
Number of across-set monotonicity tests	12,373	3,111	15,484
Number of failures	352	86	438
Respondents with 0 failures	403 (80.6%)	103 (81.7%)	506 (80.8%)
Respondents with 1 failure	37 (7.4%)	7 (5.6%)	44 (7.0%)
Respondents with >1 failure	60 (12.0%)	16 (12.7%)	76 (12.1%)

DUHS, Duke University Health System

Table II. Scale-adjusted log-odds preference weights from pooled sample random-parameters model

	Web Panel (n=489)				DUHS (n=124)			
Attributes and levels	**Coefficient**	**Standard error**	**Z-score**	***p*-value**	**Coefficient**	**Standard error**	**Z-score**	***p*-value**
Physical functioning								
1-year gain in NYHA class II	3.222	0.592	5.45	0.000	3.465	0.607	5.71	0.000
1-year gain in NYHA class III	2.466	0.469	5.26	0.000	2.598	0.490	5.31	0.000
30-day mortality risk (vs 0%)								
2%	-0.810	0.246	-3.30	0.001	-0.892	0.240	-3.71	0.000
5%	-2.188	0.449	-4.88	0.000	-2.298	0.429	-5.36	0.000
10%	-4.520	0.869	-5.20	0.000	-4.866	0.876	-5.56	0.000
15%	-5.171	0.955	-5.41	0.000	-5.503	0.953	-5.77	0.000
In-hospital complication risk (vs 0%)								
5%	-0.980	0.272	-3.60	0.000	-0.916	0.255	-3.59	0.000
15%	-2.312	0.460	-5.02	0.000	-2.301	0.440	-5.23	0.000
40%	-4.195	0.768	-5.46	0.000	-4.241	0.735	-5.77	0.000
Remote device adjustment (vs no)	0.419	0.192	2.18	0.029	0.403	0.185	2.17	0.030
Device (vs no device)	1.499	0.518	2.89	0.004	4.759	0.961	4.95	0.000

DUHS, Duke University Health System; NYHA, New York Heart Association

Table III. Characteristics of Study Samples and Heart Failure Patients in NHANES

Characteristic	Web Panel (N = 500)	DUHS (N = 126)	NHANES * (N=182)
Male, %	52%	47%	48%
Age in years, mean (SD)	64	66	66
Race, %			
White	89%	66%	
Black	8%	32%	24%
American Indian/Alaskan native	3%	3%	
Asian	2%	0	
Other	1%	0	
Education, %*			
High school or less	19%	23%	28%
Some college but no degree	28%	19%	
Associate degree/tech school	16%	19%	
4-year degree (+/- some grad studies)	23%	20%	
Graduate or professional degree	14%	19%	
Comorbidities, %			
Diabetes	37%	33%	40%
Hypertension	65%	76%	82%
Obesity	29%	38%	49%

* Komanduri S, Jadhao Y, Guduru SS, Cheriyath P, Wert Y. Prevalence and risk factors of heart failure in the USA: NHANES 2013 - 2014 epidemiological follow-up study. J Community Hosp Intern Med Perspect. 2017;7(1):15-20. PMCID: PMC5463661.

Table IV: Predicted choice probabilities for two hypothetical device profiles

	Device profiles				Predicted choice probabilities (95% CIs)	
	Improvement in functioning	Risk of death	Risk of complications	Remote monitoring	DUHS	Web panel
Device 1	1 year in NYHA III	2%	15%	no	98.5% (91.8%, 99.8%)	69.9% (50.1%, 85.7%)
Device 2	1 year in NYHA II	5%	15%	no	97.4% (88.0%, 99.5%)	55.5% (34.2%, 75.1%)