Supplementary Online Content

Curtis JR, Back AL, Ford DW, et al. Effect of communication skills training for residents and nurse practitioners on quality of communication with patients with serious illness: a randomized trial. *JAMA*. doi:10.1001/jama.2013.282081.

This supplementary material has been provided by the authors to give readers additional informaiton about their work.

eAppendix. Development and Application of Weights Accounting for Propensity for Survey Inclusion in the Study's Primary Analysis

Although propensity weighting frequently focuses solely on weighting to account for non-response from persons being recruited for participation in an observational study, the focus of our propensity-weighting exercise was somewhat different. We were interested in weighting cases to account for the absence of data in one specific analysis: a clustered regression model of the study's primary end-point (scores on a composite measure reflecting patients' ratings of the communication skills of trainees). The trainees were participants in a randomized trial, with approximately half randomized to receive training in specific communication skills. For each trainee, we requested ratings from several patients whom they had cared for during the pre-intervention period, and ratings from another group of patients whom they had cared for during the post-intervention period. We invited most patients to rate only one trainee; however, a few were asked to rate a second trainee. We distributed 4,614 patient surveys. Of those, 1,866 (40%) were completed and returned; and 1,224 (27%) were included in the primary analysis. A survey could be omitted from the analysis for any of the following reasons:

- The patient didn't return the survey;
- S/he returned a survey, but it was missing information necessary for computation of the primary study outcome;
- S/he returned surveys for two trainees, only one survey of which was selected for use in the analysis (the number of patients who completed multiple surveys was small enough that we opted for a purely hierarchical design for the analysis, with patients clustered under trainees; this required that we use only one survey per patient, excluding the other survey from the analysis);
- The survey was for a trainee who received surveys for only one of the two study time periods (our analysis design required that trainees have at least one patient rating from the pre-intervention period and at least one from the post-intervention period in order to be included in the analysis).

The question of interest in developing the propensity weights was whether the 27% of surveys that were included in the primary analysis were from patients who were significantly different in identifiable ways from the patients to whom the other 73% of the surveys were distributed. If so, we could develop weights to apply to the analysis so that the sample was more nearly representative of the full sample to whom surveys were distributed.

1. Predictor Selection

We first identified 20 variables for which we had complete data for the 4,614 surveys:

- patient's gender
- patient's status on each of 10 criteria, any of which would have conveyed study eligibility:
    - in hospice care
    - documented end-of-life communication or palliative care consultation
    - metastatic cancer
    - advanced COPD
    - restrictive lung disease
    - congestive heart failure with functional deficits
    - end-stage liver disease
    - inpatient aged 80+

    - age-adjusted Charlson comorbidity score of 6+
    - lengthy stay in intensive care
- how many of the 10 criteria the patient met
- recruitment location: Washington or South Carolina
- location of care by the participating trainee: in inpatient or outpatient setting
- trainee type: physician or nurse practitioner
- timing of contact with the trainee: pre-intervention or post-intervention period
- trainee's randomization status: intervention or control group
- trainee's gender
- trainee's racial/ethnic minority status

Although we had considerable data on a few other variables of interest, too many cases lacked the information to make them feasible for use.

2. Reducing the Pool of Predictors

The first step in reducing the predictor pool was to run 20 cross-classified logistic regression models, each with one of the identified variables as the sole predictor of whether the survey was used in the primary analysis. We then removed from the pool all predictors with *P*-values greater than 0.20 in the single-predictor models. This reduced the pool to 14 predictors.

Then beginning with a model with all 14 predictors, we backed predictors out of the model one at a time, at each step removing the predictor with the highest P-value. We repeated this procedure until all remaining predictors had P-values < 0.10. The result was a set of 8 predictors.

Finally, we introduced each of the excluded predictors back into the 8-predictor model to see whether any of their P-values had reduced with the previous predictor exclusions. This step resulted in the reintroduction of one predictor, leaving a final model of 9 predictors. All had P-values < 0.05.

- Seen by trainee in inpatient vs. outpatient setting
- In hospice care
- Advanced COPD
- Restrictive lung disease
- Aged 80+ and receiving inpatient care
- Age-adjusted Charlson comorbidity score = 6+
- Recruited in Washington vs. South Carolina
- Asked to rate a trainee during the pre-intervention vs. post-intervention period
- Asked to rate a trainee in the intervention vs. control group

3. Constructing the Propensity Measure, Based on the Final 9-Predictor Model

Using the intercepts and slopes in the final 9-predictor cross-classified model, we computed for each survey the following three values (where Y = whether the survey was included in the analysis, b_0 = the intercept

from the final model, b_1 to b_9 = slopes for the 9 predictors, and X_1 to X_9 = the values of the 9 predictors on the survey):

- $\text{logit}(Y) = b_0 + b_1X_1 + ... + b_9X_9$;
- the survey's estimated probability of inclusion in the analysis, based on its values on the predictors = $\exp[\text{logit}(Y)]/\{1+\exp[\text{logit}(Y)]\}$;
- propensity weight = 1/ probability of inclusion.

4. Weighted Regression Model of Intervention Effect on the Primary Study Outcome

Finally, we repeated the analysis of the primary outcome, applying the propensity weights. We used the same clustered design and the same surveys as in the unweighted analysis, with only one evaluation per patient included, and evaluations for trainees included only if the trainee had at least one evaluation for the pre-intervention period and one for the post-intervention period. As in the unweighted analysis, this model was based on surveys from 1224 patients, clustered under the 194 trainees they evaluated. Results appear in the table below:

eTable 1. Association of the Intervention With Patient-Assessed QOC Scores, Seven-Predictor Model with Propensity Weights

Predictor	# Patients	# Trainees	b	*P*	95% CI
	1224	194			
Intervention Effect[a]			0.407	0.156	-0.156, 0.970
Randomization Group[b]			-0.425	0.050	-0.850, 0.000
Time Period[c]			-0.008	0.962	-0.354, 0.337
Study Site[d]			0.326	**0.029**	0.033, 0.619
Stratum			--	**0.009**	--
First-Year Resident			0.000	--	--
Second-Year Resident			0.542	**0.010**	0.131, 0.954
Third-Year Resident or Fellow			0.306	0.136	-0.097, 0.709
Nurse Practitioner			0.928	**0.000**	0.537, 1.319

a Interaction term (randomization group times time period)
b 0=control group; 1=intervention group
c 0=pre-intervention period; 1=post-intervention period
d 0=University of Washington; 1=Medical University of South Carolina

As with the unweighted model, this model showed no significant effect of the intervention. Unlike the unweighted model, it failed to demonstrate significantly better communication by third-year residents and fellows than by their first-year counterparts, although the overall difference among the four strata remained statistically significant. Other associations were similar to those in the unweighted model. This model, weighted to make surveys used in the analysis more representative of the full set of surveys requested from patients, shows no evidence that the non-response or the exclusion of surveys from the analysis produced bias in the primary study finding.

eTable 2: Adjusted Patient Questionnaire Return Rates by Patient Characteristics[a]

	Est. % Returned[b]	*P*[c]
General Characteristics		
Gender		0.646
Male	41.6	
Female	42.3	
Race/Ethnicity		**<0.001**
White non-Hispanic	44.0	
Minority Group	37.8	
Setting of Care		**<0.001**
Inpatient	37.4	
Outpatient	66.4	
Time Period		0.286
Trainee's pre-intervention period	40.8	
Trainee's post-intervention period	42.6	
Eligibility Criteria Met by the Patient		
In hospice care		**<0.001**
No	41.8	
Yes	26.7	
Documented Communication about End-of-life Issues		**0.002**
No	41.7	
Yes	34.0	
Metastatic cancer		**0.002**
No	42.2	
Yes	36.1	
Chronic Obstructive Pulmonary Disease		0.970
No	40.5	
Yes	40.7	
Restrictive lung disease		0.271
No	40.5	
Yes	50.6	
Congestive Heart Failure		0.568
No	40.4	
Yes	42.1	
End-stage liver disease		**0.002**
No	41.4	
Yes	31.8	
Age-adjusted Charlson score of 6+		0.563
No	41.3	
Yes	40.2	
Inpatient care at age 80+		**<0.001**
No	42.3	
Yes	31.1	

Length stay in ICU		0.200
No	40.8	
Yes	35.9	
Referred by trainee		0.224
No	40.5	
Yes	52.3	

a Estimated return rates for characteristics listed under the heading "General Characteristics" were based on 4,443 (96.3%) of the 4,614 patient survey records. The remaining 171 records were for patients whose racial/ethnic status was unknown. Estimated rates for characteristics listed under the heading "Eligibility Criteria Met by the patient" were based on the full set of 4,614 requested surveys.

b The estimated return rate for each row represents the percentage of questionnaires returned by patients having the characteristic listed on the row, adjusted for differences between patients on remaining characteristics listed under the header. (For example, the estimate for males was adjusted so that the male sample approximated the distribution of the total sample with regard to race/ethnicity, setting of care, and time period. The estimate for patients who were not receiving hospice care was adjusted so that the non-hospice-care sample approximated the distribution of the total sample with regard to the other 10 eligibility criteria.) The estimate was based on the intercept from a cross-classified logistic regression model in which the row predictor was an indicator variable coded 0 for the row category and 1 for its counterpart, with the remaining variables in the predictor group centered on their grand means. This estimate was then converted to a probability, using the formula exp[b0]/(1+exp[b0]), and multiplied by 100 for representation as a percentage.

c *P*-values were based on two multi-predictor cross-classified logistic regression models. All predictors except gender and race/ethnicity were level-1 predictors; gender and race/ethnicity were level-2 row predictors. Values for the general characteristics came from a model that regressed survey return on gender, race/ethnicity, setting of care, and time period. Values for the patient eligibility criteria came from a model that regressed survey return on the 11 criteria.

eTable 3: Adjusted Family Questionnaire Return Rates by Patient Characteristics[a]

	Est. % Returned[b]	*P*[c]
General Characteristics		
Patient Gender		0.358
Male	67.1	
Female	64.5	
Patient Race/Ethnicity		**0.003**
White non-Hispanic	69.0	
Minority Group	59.9	
Setting of Care		0.133
Inpatient	66.9	
Outpatient	60.5	
Time Period		0.136
Trainee's pre-intervention period	63.3	
Trainee's post-intervention period	67.7	
Patient Died during Hospitalization		**<0.001**
No	78.1	
Yes	29.4	
Patient-Eligibility Criteria[c]		
In hospice care		0.493
No	77.7	
Yes	74.2	
Documented Communication about End-of-life Issues		0.058
No	78.2	
Yes	68.7	
Metastatic cancer		0.327
No	76.7	
Yes	80.2	
Chronic Obstructive Pulmonary Disease		0.150
No	76.9	
Yes	85.0	
Restrictive lung disease		0.713
No	77.4	
Yes	82.3	
Congestive Heart Failure		0.515
No	77.8	
Yes	74.7	
End-stage liver disease		0.892
No	77.5	
Yes	76.8	
Age-adjusted Charlson score of 6+		0.762
No	76.8	
Yes	77.8	

Inpatient care at age 80+		0.887
No	77.6	
Yes	77.0	
Length stay in ICU		0.570
No	77.1	
Yes	79.5	
Referred by trainee		0.577
No	77.6	
Yes	69.8	

a Estimated rates for characteristics listed under the heading "General Characteristics" were based on 1,438 (98.4%) of the 1462 surveys for which the patient's racial/ethnic status was known. Estimated rates for characteristics listed under the heading "Patient Eligibility Criteria" group were based on the 1,041 surveys requested from family members of patients who were recruited while the patient was alive. Patient eligibility criteria were not recorded for the remaining 421 surveys, which were requested from families subsequent to their patients' deaths during hospitalization. The overall unadjusted response rate was much higher for the surveys requested from family members of living patients than for those requested from families of decedents (response rates of 77.4% and 30.9%, respectively).

b The estimated return rate for each row represents the percentage of questionnaires returned by family members of patients having the characteristic listed on the row, adjusted for differences between patients on remaining characteristics listed under the header. (For example, the estimate for male patients was adjusted so that the male sample approximated the distribution of the total sample with regard to race/ethnicity, setting of care, time period, and mortality status. The estimate for family members of patients who were not receiving hospice care was adjusted so that the non-hospice-care sample approximated the distribution of the total sample with regard to the other 10 patient eligibility criteria.) The estimate was based on the intercept from a cross-classified logistic regression model in which the row predictor was an indicator variable coded 0 for the row category and 1 for its counterpart, with the remaining variables in the predictor group centered on their grand means. This estimate was then converted to a probability, using the formula exp[b0]/(1+exp[b0]), and multiplied by 100 for representation as a percentage.

c *P*-values were based on two multi-predictor cross-classified logistic regression models. All predictors except patient gender and race/ethnicity were level-1 predictors; gender and race/ethnicity were level-2 row predictors. Values for the general characteristics came from a model that regressed survey return on gender, race/ethnicity, setting of care, time period, and patient's mortality status. Values for the patient eligibility criteria came from a model that regressed survey return on the 11 criteria. P-values associated with all level-1 predictors (i.e., all predictors except patient gender and race) should be interpreted as rough estimates.

eTable 4: Adjusted Clinician-Evaluator Return Rages by Evaluator and Trainee Characteristics

Characteristic	Est. % Returned[b]	*P*[c]
Characteristic		
Evaluator Gender		0.619
Male	57.0	
Female	58.7	
Trainee Gender		0.644
Male	57.5	
Female	58.6	
Evaluator Type		**<0.001**
Physician	64.8	
Nurse or Other	51.8	
Trainee Type		0.059
Physician	57.3	
Nurse Practitioner Student	77.9	
Setting of Interaction		0.867
Inpatient	58.0	
Outpatient	58.8	

a Estimated rates were based on 4,969 survey requests to clinician-evaluators.

b The estimated return rate for each row represents the percentage of questionnaires returned by clinician-evaluators when the evaluator or trainee had the characteristic listed in the row header, adjusted for differences on the remaining characteristics. (For example, the estimate for male clinician-evaluators was adjusted so that the male-evaluator sample approximated the distribution of the total sample with regard to evaluator type, trainee gender and type, and setting of interaction.) Each estimate was based on the intercept from a cross-classified logistic regression model in which the row predictor was an indicator variable coded 0 for the row category and 1 for its counterpart, with the remaining variables in the predictor group centered on their grand means. This estimate was then converted to a probability, using the formula exp[b0]/(1+exp[b0]), and multiplied by 100 for representation as a percentage.

c *P*-values were based on a multi-predictor cross-classified logistic regression model. Setting of interaction between the clinician-evaluator and trainee was a level-1 predictor; evaluator gender and type were level-2 row predictors; trainee gender and type were level-2 column predictors. The model regressed survey return on all five characteristics.

eTable 5. Patient-Assessed Outcomes, Total Sample and Predictor Groups, Mean Values by Time Period[a]

Outcome	Predictor	Pre-Intervention Mean (95% CI)	Post-Intervention Mean (95% CI)	Difference (95% CI)
QOC score	Total Sample	6.2 (6.1 to 6.4)	6.5 (6.3 to 6.7)	0.2 (-0.0 to 0.5)
	Randomization Group			
	Control	6.3 (6.1 to 6.5)	6.4 (6.2 to 6.7)	0.1 (-0.2 to 0.4)
	Intervention	6.1 (5.8 to 6.5)	6.5 (6.2 to 6.8)	0.4 (0.0 to 0.8)
	Site			
	UW	6.0 (5.7 to 6.3)	6.2 (5.9 to 6.5)	0.2 (-0.2 to 0.6)
	MUSC	6.6 (6.3 to 6.8)	6.7 (6.4 to 7.0)	0.1 (-0.2 to 0.5)
	Stratum			
	1st Year Resident	6.1 (5.8 to 6.3)	6.2 (5.9 to 6.4)	0.1 (-0.2 to 0.4)
	2nd Year Resident	7.0 (6.5 to 7.4)	6.7 (6.3 to 7.1)	-0.2 (-0.8 to 0.3)
	3rd Year Resident/Fellow	6.3 (5.9 to 6.8)	7.0 (6.6 to 7.4)	0.7 (0.1 to 1.2)
Overall QOC[b]	Total Sample	9.8 (9.4 to 10.2)	10.0 (9.6 to 10.4)	0.2 (-0.3 to 0.7)
	Randomization Group			
	Control	9.9 (9.5 to 10.4)	10.1 (9.6 to 10.5)	0.2 (-0.5 to 0.8)
	Intervention	9.6 (8.9 to 10.3)	9.9 (9.3 to 10.5)	0.3 (-0.5 to 1.1)
	Site			
	UW	9.3 (8.8 to 9.8)	9.3 (8.8 to 9.7)	-0.0 (-0.6 to 0.6)
	MUSC	10.5 (9.9 to 11.1)	10.8 (10.2 to 11.3)	0.3 (-0.4 to 1.1)
	Stratum			
	1st Year Resident	9.5 (9.1 to 9.9)	9.3 (8.9 to 9.7)	-0.2 (-0.7 to 0.4)
	2nd Year Resident	10.3 (9.2 to 11.3)	11.6 (10.6 to 12.6)	1.3 (0.0 to 2.6)
	3rd Year Resident/Fellow	10.5 (9.3 to 11.8)	10.9 (10.1 to 11.8)	0.4 (-0.9 to 1.7)

Outcome	Predictor	Pre-Intervention Mean (95% CI)	Post-Intervention Mean (95% CI)	Difference (95% CI)
QEOLC[b]	Total Sample	8.4 (8.2 to 8.6)	8.7 (8.5 to 8.8)	0.3 (0.0 to 0.5)
	Randomization Group			
	Control	8.5 (8.3 to 8.7)	8.7 (8.5 to 8.9)	0.2 (-0.1 to 0.5)
	Intervention	8.2 (7.9 to 8.6)	8.6 (8.3 to 8.9)	0.4 (-0.1 to 0.8)
	Site			
	UW	8.2 (7.9 to 8.4)	8.4 (8.1 to 8.6)	0.2 (-0.2 to 0.5)
	MUSC	8.7 (8.4 to 9.0)	8.9 (8.7 to 9.1)	0.2 (-0.1 to 0.6)
	Stratum			
	1st Year Resident	8.3 (8.0 to 8.5)	8.3 (8.1 to 8.6)	0.1 (-0.2 to 0.4)
	2nd Year Resident	8.5 (8.0 to 9.0)	9.3 (8.9 to 9.7)	0.8 (0.2 to 1.4)
	3rd Year Resident/Fellow	8.8 (8.3 to 9.2)	9.0 (8.6 to 9.3)	0.2 (-0.3 to 0.8)
Depression Score	Total Sample	9.1 (8.5 to 9.7)	9.1 (8.6 to 9.6)	-0.0 (-0.8 to 0.8)
	Randomization Group			
	Control	9.4 (8.7 to 10.1)	8.5 (7.9 to 9.1)	-0.9 (-1.9 to 0.0)
	Intervention	8.6 (7.6 to 9.6)	10.0 (9.1 to 10.8)	1.4 (0.2 to 2.6)
	Site			
	UW	9.0 (8.3 to 9.7)	9.4 (8.7 to 10.1)	0.5 (-0.6 to 1.5)
	MUSC	9.3 (8.3 to 10.2)	8.8 (8.0 to 9.5)	-0.5 (-1.7 to 0.7)
	Stratum			
	1st Year Resident	9.4 (8.7 to 10.0)	9.6 (8.9 to 10.2)	0.2 (-0.8 to 1.1)
	2nd Year Resident	9.1 (7.3 to 11.0)	9.2 (8.1 to 10.3)	0.1 (-2.1 to 2.2)
	3rd Year Resident/Fellow	8.0 (6.7 to 9.2)	7.7 (6.6 to 8.8)	-0.3 (-2.1 to 1.5)
Physical Status	Total Sample	29.9 (29.0 to 30.8)	30.0 (29.2 to 30.9)	0.1 (-1.1 to 1.4)
	Randomization Group			
	Control	29.8 (28.7 to 31.0)	29.5 (28.4 to 30.6)	-0.3 (-1.9 to 1.3)
	Intervention	30.0 (28.6 to 31.3)	30.8 (29.4 to 32.1)	0.8 (-1.1 to 2.8)
	Site			
	UW	29.7 (28.5 to 30.9)	29.1 (28.0 to 30.2)	-0.6 (-2.2 to 0.9)
	MUSC	30.1 (28.9 to 31.3)	31.0 (29.7 to 32.2)	0.8 (-1.1 to 2.8)
	Stratum			
	1st Year Resident	28.9 (27.9 to 29.9)	29.2 (28.2 to 30.2)	0.3 (-1.3 to 1.8)
	2nd Year Resident	32.1 (30.2 to 34.1)	30.5 (28.7 to 32.4)	-1.6 (-4.6 to 1.4)
	3rd Year Resident/Fellow	31.6 (29.8 to 33.4)	32.3 (30.0 to 34.6)	0.7 (-2.0 to 3.4)

Outcome	Predictor	Pre-Intervention Mean (95% CI)	Post-Intervention Mean (95% CI)	Difference (95% CI)
Mental Status	Total Sample	43.3 (42.3 to 44.2)	44.0 (43.0 to 44.9)	0.7 (-0.7 to 2.1)
	Randomization Group			
	Control	43.6 (42.5 to 44.7)	44.2 (43.0 to 45.4)	0.6 (-1.1 to 2.3)
	Intervention	42.8 (41.1 to 44.5)	43.7 (42.1 to 45.2)	0.8 (-1.5 to 3.2)
	Site			
	UW	43.4 (42.0 to 44.8)	43.5 (42.2 to 44.8)	0.1 (-1.8 to 2.0)
	MUSC	43.1 (41.8 to 44.4)	44.4 (43.0 to 45.8)	1.3 (-0.7 to 3.3)
	Stratum			
	1st Year Resident	42.9 (41.7 to 44.1)	42.9 (41.8 to 44.0)	-0.0 (-1.7 to 1.7)
	2nd Year Resident	42.2 (39.8 to 44.7)	44.5 (42.0 to 47.0)	2.2 (-1.6 to 6.1)
	3rd Year Resident/Fellow	45.0 (42.8 to 47.3)	46.7 (44.6 to 48.9)	1.7 (-1.2 to 4.6)

a Mean values were based on the dataset that included only one evaluation per patient. For each outcome, evaluations were included only for trainees who had at least one valid pre-intervention and one valid post-intervention rating. Confidence intervals were estimated with clustered regression models (linear models except where indicated in footnote b) and a restricted maximum likelihood estimator. We computed the means and confidence intervals adjusted for clustering by first selecting the sample indicated by the row header. Then, to estimate the pre-intervention mean and confidence intervals, we ran a model regressing the outcome on an indicator coded 0 (if the evaluation was done by a respondent who interacted with the trainee during the pre-intervention period) or 1 (if the evaluation was done by a respondent who interacted with the trainee during the post-intervention period). The intercept and its confidence interval from this model represented the mean and 95% confidence interval for the pre-intervention period. To compute the confidence values for the post-intervention period, we ran a similar model, but with time period coded in the reverse order (0 for post-intervention evaluations and 1 for pre-intervention evaluations). The confidence interval for the difference between the pre- and post-intervention scores was the confidence interval around the slope for the pre/post indicator in the model with pre-intervention coded 0 and post-intervention coded 1.

b This variable had a strong ceiling effect and was modeled with tobit regression. The means represent the means for a latent variable believed to represent the true value of the rating, rather than the means of the empirically observed variables.

Abbreviations: QOC, Quality of Communication; QEOLC, Quality of End-of-life Care; UW, University of Washington; MUSC, Medical University of South Carolina

eTable 6. Family-Assessed Outcomes, Total Sample and Predictor Groups, Mean Values by Time Period[a]

Outcome	Predictor	Pre-Intervention Mean (95% CI)	Post-Intervention Mean (95% CI)	Difference (95% CI)
QOC score	Total Sample	6.9 (6.5 to 7.2)	7.0 (6.7 to 7.3)	0.1 (-0.3 to 0.6)
	Randomization Group			
	Control	6.6 (6.1 to 7.0)	6.6 (6.3 to 7.0)	0.1 (-0.5 to 0.6)
	Intervention	7.3 (6.9 to 7.7)	7.5 (7.1 to 7.9)	0.2 (-0.5 to 0.8)
	Site			
	UW	6.7 (6.2 to 7.3)	6.8 (6.3 to 7.4)	0.1 (-0.7 to 0.9)
	MUSC	6.9 (6.5 to 7.3)	7.0 (6.7 to 7.4)	0.1 (-0.4 to 0.7)
	Stratum			
	1st Year Resident	6.8 (6.3 to 7.2)	7.1 (6.7 to 7.4)	0.3 (-0.3 to 0.9)
	2nd Year Resident	6.9 (6.3 to 7.6)	6.9 (6.2 to 7.5)	-0.1(-1.0 to 0.9)
	3rd Year Resident/Fellow	7.0 (6.5 to 7.6)	6.9 (6.2 to 7.6)	-0.1 (-1.0 to 0.8)
Overall QOC[b]	Total Sample	9.9 (9.4 to 10.4)	9.7 (9.2 to 10.2)	-0.2 (-0.8 to 0.4)
	Randomization Group			
	Control	9.8 (9.2 to 10.4)	9.6 (9.0 to 10.2)	-0.2 (-1.0 to 0.6)
	Intervention	10.1 (9.3 to 10.9)	9.9 (9.1 to 10.7)	-0.2 (-1.2 to 0.8)
	Site			
	UW	9.1 (8.4 to 9.9)	9.1 (8.4 to 9.8)	-0.1 (-1.1 to 0.9)
	MUSC	10.4 (9.7 to 11.1)	10.1 (9.4 to 10.7)	-0.3 (-1.1 to 0.5)
	Stratum			
	1st Year Resident	9.7 (9.1 to 10.2)	9.5 (9.0 to 10.1)	-0.1 (-0.9 to 0.6)
	2nd Year Resident	10.1 (8.6 to 11.6)	10.0 (8.7 to 11.3)	-0.1 (-1.3 to 1.0)
	3rd Year Resident/Fellow	10.4 (9.0 to 11.7)	10.0 (8.8 to 11.2)	-0.4 (-2.1 to 1.3)
QEOLC[b]	Total Sample	8.6 (8.3 to 8.9)	8.6 (8.3 to 8.8)	-0.0 (-0.4 to 0.3)
	Randomization Group			
	Control	8.5 (8.1 to 8.9)	8.5 (8.1 to 8.8)	-0.1 (-0.6 to 0.4)
	Intervention	8.7 (8.3 to 9.1)	8.7 (8.3 to 9.1)	-0.0 (-0.6 to 0.5)
	Site			
	UW	8.3 (7.8 to 8.8)	8.4 (7.8 to 8.9)	0.1 (-0.7 to 0.8)
	MUSC	8.8 (8.4 to 9.1)	8.7 (8.4 to 9.0)	-0.1 (-0.6 to 0.3)
	Stratum			
	1st Year Resident	8.6 (8.2 to 8.9)	8.5 (8.2 to 8.9)	-0.0 (-0.5 to 0.5)
	2nd Year Resident	8.4 (7.7 to 9.1)	8.5 (8.0 to 9.0)	0.1 (-0.7 to 0.8)
	3rd Year Resident/Fellow	9.0 (8.1 to 9.9)	8.7 (8.1 to 9.3)	-0.3 (-1.2 to 0.7)

Outcome	Predictor	Pre-Intervention Mean (95% CI)	Post-Intervention Mean (95% CI)	Difference (95% CI)
Depression Score	Total Sample	6.3 (5.5 to 7.1)	6.5 (5.7 to 7.3)	0.2 (-0.8 to 1.2)
	Randomization Group			
	Control	6.2 (5.3 to 7.1)	6.0 (5.1 to 6.8)	-0.3 (-1.5 to 0.9)
	Intervention	6.5 (5.0 to 8.0)	7.3 (6.0 to 8.7)	0.8 (-1.0 to 2.6)
	Site			
	UW	6.0 (4.7 to 7.2)	7.5 (6.1 to 8.9)	1.5 (-0.4 to 3.4)
	MUSC	6.5 (5.5 to 7.5)	6.1 (5.2 to 6.9)	-0.5 (-1.6 to 0.7)
	Stratum			
	1st Year Resident	6.5 (5.5 to 7.5)	6.3 (5.5 to 7.2)	-0.2 (-1.5 to 1.1)
	2nd Year Resident	6.9 (5.5 to 8.2)	7.1 (5.7 to 8.4)	0.2 (-1.8 to 2.2)
	3rd Year Resident/Fellow	5.3 (3.6 to 7.1)	6.4 (4.0 to 8.9)	1.1 (-1.4 to 3.6)

a Mean values were based on the dataset that included only one evaluation per family member. For each outcome, evaluations were included only for trainees who had at least one valid pre-intervention and one valid post-intervention rating. Confidence intervals were estimated with clustered regression models (linear models except where indicated in footnote b) and a restricted maximum likelihood estimator. We computed the means and confidence intervals adjusted for clustering by first selecting the sample indicated by the row header. Then, to estimate the pre-intervention mean and confidence intervals, we ran a model regressing the outcome on an indicator coded 0 (if the evaluation was done by a respondent who interacted with the trainee during the pre-intervention period) or 1 (if the evaluation was done by a respondent who interacted with the trainee during the post-intervention period). The intercept and its confidence interval from this model represented the mean and 95% confidence interval for the pre-intervention period. To compute the confidence values for the post-intervention period, we ran a similar model, but with time period coded in the reverse order (0 for post-intervention evaluations and 1 for pre-intervention evaluations). The confidence interval for the difference between the pre- and post-intervention scores was the confidence interval around the slope for the pre/post indicator in the model with pre-intervention coded 0 and post-intervention coded 1.

b This variable had a strong ceiling effect and was modeled with tobit regression. The means represent the means for a latent variable believed to represent the true value of the rating, rather than the means of the empirically observed variables.

Abbreviations: QOC, Quality of Communication ; QEOLC, Quality of End-of-life Care ; UW, University of Washington; MUSC, Medical University of South Carolina

eTable 7. Clinician-Assessed Outcomes, Total Sample and Predictor Groups, Mean Values by Time Period[a]

Outcome	Predictor	Pre-Intervention Mean (95% CI)	Post-Intervention Mean (95% CI)	Difference (95% CI)
QOC score	Total Sample	7.5 (7.3 to 7.6)	7.6 (7.5 to 7.8)	0.2 (0.0 to 0.3)
	Randomization Group			
	Control	7.5 (7.3 to 7.6)	7.5 (7.4 to 7.7)	0.1 (-0.1 to 0.3)
	Intervention	7.5 (7.3 to 7.7)	7.7 (7.6 to 7.9)	0.3 (0.1 to 0.5)
	Site			
	UW	7.4 (7.2 to 7.5)	7.6 (7.4 to 7.7)	0.2 (0.1 to 0.4)
	MUSC	7.6 (7.4 to 7.9)	7.7 (7.5 to 7.9)	0.1 (-0.2 to 0.3)
	Stratum			
	1st Year Resident	7.4 (7.2 to 7.5)	7.5 (7.4 to 7.7)	0.2 (0.0 to 0.3)
	2nd Year Resident	7.9 (7.5 to 8.3)	7.7 (7.4 to 8.0)	-0.2 (-0.6 to 0.3)
	3rd Year Resident/Fellow	7.9(7.5 to 8.2)	8.0 (7.7 to 8.3)	0.1 (-0.3 to 0.5)
Overall QOC	Total Sample	7.6 (7.4 to 7.7)	7.7 (7.6 to 7.9)	0.2 (0.0 to 0.3)
	Randomization Group			
	Control	7.6 (7.4 to 7.8)	7.7 (7.5 to 7.9)	0.1 (-0.1 to 0.3)
	Intervention	7.5 (7.3 to 7.7)	7.8 (7.6 to 8.0)	0.2 (0.0 to 0.5)
	Site			
	UW	7.5 (7.3 to 7.7)	7.7 (7.5 to 7.9)	0.2 (0.1 to 0.4)
	MUSC	7.8 (7.5 to 8.0)	7.8 (7.6 to 8.0)	0.0 (-0.2 to 0.3)
	Stratum			
	1st Year Resident	7.5 (7.3 to 7.6)	7.6 (7.5 to 7.8)	0.1 (0.0 to 0.3)
	2nd Year Resident	8.0 (7.6 to 8.5)	7.8 (7.5 to 8.2)	-0.2 (-0.7 to 0.3)
	3rd Year Resident/Fellow	7.8 (7.5 to 8.2)	8.1 (7.8 to 8.4)	0.2 (-0.2 to 0.7)
QEOLC	Total Sample	7.6 (7.4 to 7.7)	7.7 (7.6 to 7.9)	0.1 (0.0 to 0.3)
	Randomization Group			
	Control	7.6 (7.4 to 7.8)	7.6 (7.5 to 7.8)	0.1 (-0.1 to 0.2)
	Intervention	7.5 (7.3 to 7.7)	7.8 (7.6 to 8.0)	0.3 (0.1 to 0.5)
	Site			
	UW	7.5 (7.3 to 7.7)	7.7 (7.5 to 7.8)	0.2 (0.0 to 0.3)
	MUSC	7.7 (7.5 to 7.9)	7.8 (7.6 to 8.0)	0.1 (-0.1 to 0.3)
	Stratum			
	1st Year Resident	7.5 (7.3 to 7.6)	7.6 (7.4 to 7.7)	0.1 (0.0 to 0.3)
	2nd Year Resident	8.0 (7.6 to 8.3)	7.8 (7.5 to 8.2)	-0.1 (0.0 to 0.3)
	3rd Year Resident/Fellow	8.0 (7.7 to 8.3)	8.1 (7.8 to 8.4)	0.1 (-0.3 to 0.4)

a Mean values were based on multiple evaluations per clinician evaluator. For each outcome, evaluations were included only for trainees who had at least one valid pre-intervention rating and one valid post-intervention rating. Confidence intervals were estimated with cross-classified hierarchical linear regression models and full maximum likelihood estimation. We computed the means and confidence intervals adjusted for clustering by first selecting the sample indicated by the row header. Then, to estimate the pre-intervention mean and confidence intervals, we ran a model regressing the outcome on an indicator coded 0 (if the evaluation was done by a respondent who interacted with the trainee during the pre-intervention period) or 1 (if the evaluation was done by a respondent who interacted with the trainee during the post-intervention period). The intercept and its confidence interval from this model represented the mean and 95% confidence interval for the pre-intervention period. To compute the confidence values for the post-intervention period, we ran a similar model, but with time period coded in the reverse order (0 for post-intervention evaluations and 1 for pre-intervention evaluations). The confidence interval for the difference between the pre- and post-intervention scores was the confidence interval around the slope for the pre/post indicator in the model with pre-intervention coded 0 and post-intervention coded 1. Because confidence intervals were not provided by the software used for cross-classified models, we hand-computed all confidence intervals, using the standard errors provided by the software, and basing the confidence intervals on a t-statistic, which is the statistic used to test for statistical significance in these models.

Abbreviations: QOC, Quality of Communication questionnaire; QEOLC, Quality of End-of-life Care questionnaire; UW, University of Washington; MUSC, Medical University of South Carolina