Appendix F. Analytic Methods

Analyses for Aim 1: The intervention's effect on occurrence of goals-of-care discussions at the target visit and on patients' ratings of the quality of clinician–patient communication at that visit. We assessed the intervention's effect on the occurrence of communication about end-of-life care with 2 binary outcomes, each analyzed with data from 2 samples. One outcome used patients' reports on the 2-week follow-up questionnaire of whether a goals-of-care discussion had occurred at the target visit; the other used the EHR of whether the target visit had included any of the following: discussion of advance care planning, end-of-life treatment preferences, POLST, patient prognosis, hospice, or a referral to palliative care services. We analyzed each of the 2 outcomes with data from all patients and also with data from the subsample of patients who had not objected on the baseline questionnaire to a future goals-of-care discussion with their participating clinician. All 4 models were based on probit regression and estimated with weighted least squares with mean and variance adjustment, with standard errors corrected for clustering of patients under clinicians, and with covariate adjustment for any variables that confounded the association between randomization group and the outcome. The models of patient-reported discussions included automatic adjustment for the patient's report on the baseline questionnaire of whether a goals-of-care discussion had occurred with the participating clinician before the patient's study enrollment.

We assessed the intervention's effect on the quality of communication with patients' ratings of their clinician's quality of communication at the target visit, using 7 end-of-life-specific items drawn from the (QOC instrument and collected 2 weeks after the target visit. The 0 to 10 ratings were recoded to 1 to 11; a code of 0 was imputed when the patient indicated that the clinician had not attended to the communication aspect measured by the item. Large floor effects on all the variables led us to define them as censored from below and to use clustered Tobit regression models for the analyses. Preliminary analyses indicated that both a construct measured with the full set of 7 items and a construct measured with a subset of 4 of the items that were judged to be most supported by the intervention showed nonsignificant misfit to the observed data when measurement invariance was imposed between groups (control and intervention) and over time (baseline and 2-week follow-up). A judgment of nonsignificant misfit was based on a chi-square test of fit with $P \geq 0.05$. We elected to use the 4-indicator construct for measuring overall quality of communication, basing the analysis on patients who had complete data for all 4 indicators at both time points. After confirming model fit with the 2-group model, we used a single-group model, with randomization group as the predictor of interest and automatic adjustment for the baseline construct, to estimate the coefficient for randomization group; there were no additional confounders. In addition to these analyses, we examined each of the 7 QOC items as individual outcomes, again using clustered Tobit

regression models, adjusting each model for the baseline measure of the outcome and for any other variables that acted as confounders.

Analyses for Aim 2: The intervention's effect on improvement in goal-concordant care 3 months after the target visit and on increased use of palliative care services during the 6-month follow-up period. We tested the intervention's effect on concordance between care desired and care received 3 months after the target visit with clustered probit regression models, using a binary outcome based on responses from 2 groups of patients: (1) patients with a stated preference (life extension or comfort care) at 3 months and adequate information to assess goal-concordant care at baseline (a variable that was automatically included as a covariate in the analyses), and (2) patients meeting those same criteria and whose goal of care was stable between the 3-month follow-up assessment and the last previous assessment (at 2-week follow-up if a questionnaire was completed at that time; otherwise, at baseline). We deemed care concordant at an assessment point if the preference was comfort and current medical care was focused on comfort, or if both the preference and current focus were life extension. If preference and current focus did not match or the patient was uncertain about either preference or current focus, we deemed care discordant. The model with each patient group was automatically adjusted for the patient's preference at 3 months and for whether goal-concordant-care variable existed at baseline; additional adjustment was made for any variables that confounded the association between randomization group and goal-concordant care at follow-up.

We tested the intervention's effect on EHR documentation of referrals of the patient to palliative care services during the 6-month follow-up period with clustered probit regression models adjusted for confounders and using data from 2 samples: all patients and patients deemed most likely to benefit (ie, those whose previous assessment had indicated a preference for comfort care and a desire for future goals-of-care discussion with their enrolled clinician). In addition to testing an outcome that measured whether a referral had occurred during any visit (outpatient or inpatient), we also looked at referrals from each visit type separately.

We had planned to test the intervention's effect on reducing unwanted life-sustaining therapies over the 6-month follow-up period using clustered probit regression models, adjusted for confounders, and using a dichotomous outcome based on the EHR. This outcome would have indicated whether the patient received any of 3 targeted life-sustaining therapies (admission to the ICU, receipt of CPR, or receipt of mechanical ventilation), and the sample was to include patients whose preferences at the assessment point closest to death indicated that they

wanted care focused on comfort. However, only 40 patients died during the follow-up period—a sample too small for the intended analysis.

Analysis for Aim 3: The intervention's effect on patients' and family members' anxiety and depression 3 and 6 months after the target visit. We assessed anxiety with items from the GAD-7 questionnaire and depression with items from the PHQ-8 questionnaire. Preliminary confirmatory factor analyses had established that neither set of items measured a unidimensional construct in our patient sample. Further analyses, using exploratory factor analysis in a confirmatory factor analysis framework located a 2-indicator depression construct (PHQ items 1 and 2) and a 2-indicator anxiety construct (GAD items 1 and 3), each of which showed nonsignificant misfit to the observed data when measurement invariance was imposed between groups (intervention and control) and over time (baseline, 3 months, 6 months). The 2-indicator PHQ construct is parallel to an abbreviated PHQ-2 score that is used as a standard depression screening tool. We based the CFA and E/CFA analyses on clustered probit regression models with the indicators defined as ordered categorical variables. As with the earlier 2-group latent variable analysis of quality of care variables, we followed our preliminary 2-group models of the 2-indicator anxiety and depression constructs with single-group models in which the randomization group was the predictor of interest, automatic adjustment was made for the baseline level on the construct, and covariate adjustment was made for any confounders of the association between the randomization group and the outcome. We ran these single-group models for the outcome measured at each of the 2 follow-up points: 3 months and 6 months. Finally, although our data showed significant misfit to models using the full set of indicators for each construct, we ran analyses using the standard composite scores (PHQ-8 and GAD-7) as outcomes for comparison with other studies. As with the latent variable outcomes, we ran models of the standard PHQ and GAD scores with outcomes measured at each of the follow-up points. The PHQ-8 score was modeled with robust clustered linear regression, estimated with restricted maximum likelihood. Because of a strong floor effect in this sample, the GAD-7 score was defined as censored from below and modeled with clustered Tobit regression.

We had planned to conduct similar analyses to test the 8-item PHQ scale score and the 7-item GAD scale score for unidimensionality and between-group measurement invariance, using data provided by family members, and to substitute latent constructs measured with fewer indicators, should those tests show significant misfit to the data. However, the family member sample that provided follow-up data was too small to allow these analyses. Therefore, we limited our analyses of family data to the standard 8-item PHQ scale score and the 7-item GAD scale score without evidence of model fit, using the same techniques as we used for analyses of the same outcomes in the patient sample.