OPEN

Methodological Challenges and Statistical Approaches in the COMprehensive Post-Acute Stroke Services Study

Matthew A. Psioda, PhD,* Sara B. Jones, PhD,†
James G. Xenakis, PhD,‡ and Ralph B. D'Agostino, PhD§

Background: The COMprehensive Post-Acute Stroke Services study was a cluster-randomized pragmatic trial designed to evaluate a comprehensive care transitions model versus usual care. The data collected during this trial were complex and analysis methodology was required that could simultaneously account for the cluster-randomized design, missing patient-level covariates, outcome non-response, and substantial nonadherence to the intervention.

Objective: The objective of this study was to discuss an array of complementary statistical methods to evaluate treatment effectiveness that appropriately addressed the challenges presented by the complex data arising from this pragmatic trial.

Methods: We utilized multiple imputation combined with inverse probability weighting to account for missing covariate and outcome data in the estimation of intention-to-treat effects (ITT). The ITT estimand reflects the effectiveness of assignment to the COMprehensive Post-Acute Stroke Services intervention compared with usual care (ie, it does not take into account intervention adherence). Per-protocol analyses provide complementary information about the effect of treatment, and therefore are relevant for patients to inform their decision-making. We describe estimation of the complier average causal effect using an instrumental variables approach through 2-stage least squares estimation. For all preplanned analyses, we also discuss additional sensitivity analyses.

Discussion: Pragmatic trials are well suited to inform clinical practice. Care should be taken to proactively identify the appropriate balance between control and pragmatism in trial design. Valid estimation of ITT and per-protocol effects in the presence of complex data requires application of appropriate statistical methods and concerted efforts to ensure high-quality data are collected.

Key Words: causal inference, cluster-randomized trial, pragmatic trial, per-protocol analysis

(*Med Care* 2021;59: S355–S363)

From the *Department of Biostatistics, Collaborative Studies Coordinating Center; †Department of Epidemiology, Gillings School of Global Public Health; ‡Department of Genetics, University of North Carolina, Chapel Hill; and §Department of Biostatistics and Data Science, Division of Public Health Sciences, Wake Forest School of Medicine, Winston-Salem, NC.

Correspondence to: Matthew A. Psioda, PhD, Department of Biostatistics, University of North Carolina at Chapel Hill, 170 Rosenau Hall, CB #7400, Chapel Hill, NC 27599. E-mail: matt_psioda@unc.edu.

In 2010, the Patient-Centered Outcomes Research Institute funded a series of research studies that would spur the development and testing of patient-centered, real-world interventions that could be rapidly translated into clinical care.[1] Despite their increasing utilization, the design and analysis of multicenter pragmatic trials remains complex, particularly when the intervention itself is multifaceted. Interventions frequently span across care settings (eg, acute, post-acute, community), involve multiple interactions with patients at varying timepoints, and require adherence at the provider and patient level. By design, such interventions are often delivered flexibly, allowing for variation in implementation of the intervention across settings. Pragmatic trials also frequently utilize a usual care (UC) comparator that may be heterogenous or share some aspects with the intervention.[2] Given these challenges, there is not yet consensus as to the best way to analyze data from large-scale, complex pragmatic clinical trials.[3] As one moves along the continuum from highly controlled to highly pragmatic trials, there are inherent tradeoffs between data quality (supporting more compelling inferences regarding efficacy) and generalizability (supporting more compelling inferences regarding real-world effectiveness). The COMprehensive Post-Acute Stroke Services (COMPASS) Study was one of the first large-scale pragmatic clinical trials of transitional care in the United States. Our mandate was to balance the competing goals of data quality and pragmatism within the framework of a large cluster-randomized trial. Herein we discuss aspects of the study's design, the analytic challenges we encountered, the statistical approaches we employed to understand the intervention's effectiveness, and considerations for the design and analysis of future studies.

THE COMPREHENSIVE POST-ACUTE STROKE SERVICES STUDY DESIGN

Post-acute care in the United States is often fragmented and ineffective at managing patients' recoveries and transitions to

home.[4,5] Patients may not receive needed rehabilitation therapy, have difficulty with medication management, and are at risk for recurrent stroke, complications and rehospitalizations.[5–13] The COMPASS study was a pragmatic, cluster-randomized trial designed to evaluate the effectiveness of an evidence-based transitional care model compared with hospitals' UC to improve the functional status of stroke and transient ischemic attack patients discharged home.[14] The study investigated the real-world effectiveness and feasibility of implementing comprehensive transitional care management in current practice. The primary outcome for the study was 90-day patient-reported functional status. It was measured by the 16-item Stroke Impact Scale (SIS-16), a well-established semicontinuous measure represented as a percentage between 0 and 100, with higher scores representing better function.[15] The study enrolled nearly 6000 adult stroke and transient ischemic attack patients discharged home from 40 hospital units.[14,16]

Pragmatic Intervention

The COMprehensive Post-Acute Stroke Services transitional care (COMPASS-TC) intervention was designed for scalability and sustainability and was aligned with the Centers for Medicare and Medicaid Services transitional care management reimbursement models.[17] Core elements included a telephone call at 2 days and a face-to-face clinic visit within 7–14 days post-discharge. Standardized clinical assessments were used to generate individualized electronic care (eCare) plans delivered at the clinic visit. The primary goal of COMPASS-TC was to manage patients after discharge through specialized follow-up care including neurological evaluation, referrals to needed services (eg, rehabilitation, community services), and evidence-based coaching on risk factor control. While the intervention established a shared set of processes across hospitals, it was specifically tailored for each hospital, community, and patient. In keeping with the pragmatic nature of the study, the study team did not undertake efforts to increase patient adherence, although it did provide limited financial assistance to support hospitals' efforts at implementation and delivery.

Cluster Randomization

The substantial planning and training efforts required to adapt existing processes of care and to build community resource networks precluded randomization at the patient level. Furthermore, requiring hospital staff to deliver 2 separate interventions would have resulted in logistical challenges, and would likely have interfered with timely delivery of care, led to contamination of treatment groups, and resulted in drift over time had 1 treatment been perceived to be more beneficial. Instead, hospitals were randomized to COMPASS-TC or UC, with randomization stratified into 4 levels according to annual stroke patient discharge volume and stroke center certification status to ensure balance in these key cluster-level characteristics.[14,16] UC sites maintained their UC for stroke patients discharged home and the study team made no effort to change their standards of care.

Outcome Assessment and "Opt-Out" Consent Model

The need to concurrently deliver clinical care while conducting research impacted patient enrollment and the

choice of an informed consent model. We developed an opt-out consent model whereby, before discharge, patients were enrolled and informed of their hospitals' participation in the study.[18] Hospital staff collected contact information and baseline data on medical history, stroke severity and demographics. At 90 days, a survey research laboratory called patients to obtain informed consent and, if given, collect patient-reported outcomes. At least 3 call attempts were made followed by a mailed survey for patients who were not successfully reached by telephone. Participants could opt-out of participation in the outcome survey at any time.

ANALYTIC CHALLENGES AND STATISTICAL METHODOLOGY

While the study design choices had their advantages, such as enhancing the generalizability of results, they simultaneously impacted the type of data collected and the choice of statistical methodologies required to produce unbiased estimates of treatment effectiveness. First, cluster-randomization does not guarantee balance in patient-level covariates. This necessitates adjustment for factors that are strongly associated with study outcomes as well as appropriate handling of missing covariate data. Second, while the opt-out consent facilitated data capture (eg, disease severity, medical history) on all enrolled patients, the lack of baseline consent may have reduced engagement in the study and contributed to outcome nonresponse at 90 days. A third analytic challenge we faced was substantial nonadherence to the intervention by both patients and hospitals. Having been aware that this phenomenon is common in pragmatic trials,[3] our prespecification of per-protocol (PP) analyses was integral to our analysis approach, particularly as these results are of importance to patients when making treatment decisions. Finally, treatment recommendations and patients' associated adherence were not well-characterized in the UC comparator arm. Hospitals assigned to UC typically referred patients to primary care or, in some cases, provided limited hospital-based follow-up. This variation presented challenges in the interpretation of results. Below, we describe the statistical methodologies used to address these key challenges, their assumptions and limitations, and strategies for sensitivity analyses.

Considerations for Analyses Based on Cluster-level Randomization

For cluster-randomized trials, the random assignment of treatment group is to a cluster (ie, hospital unit) even if the intervention is delivered to patients individually. Since randomization is not at the patient level, there is no assurance (and in many cases there should be no expectation) that important patient characteristics will be balanced, as different hospital units might serve fundamentally different populations. Figure 1 illustrates the age distribution for 2 pairs of hospital units that were paired for randomization as well as the overall distribution by treatment arm. This distribution was heterogenous across hospitals resulting in an overall older patient population in the intervention arm. Because age is associated with outcomes of interest (eg, SIS-16), such
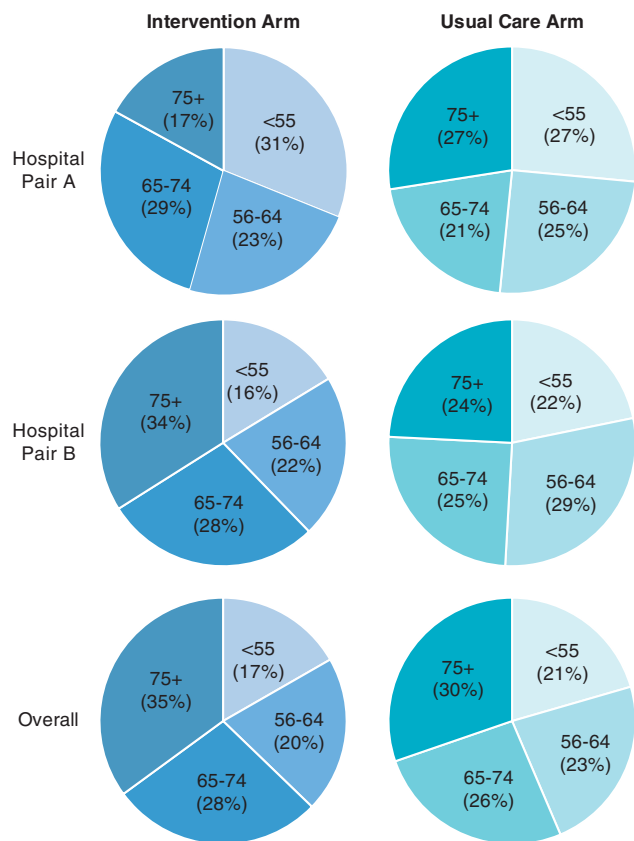
FIGURE 1. Distribution of age. The top two rows display age for 4 individual hospitals representing 2 paired blocks for randomization; the bottom row displays the overall age distribution according to study arm.

imbalances confound treatment effect estimates and must therefore be accounted for in the analysis. Age and other key clinical variables (eg, primary diagnosis, stroke severity) were prespecified for inclusion in these analysis models to minimize patient-level confounding. Moreover, sensitivity

analyses were prespecified to assess the robustness of our inferences to covariate adjustment. These approaches are described in more detail below.

While cluster-level randomization ensures the theoretical independence of cluster-level characteristics and treatment assignment, in practice, unless the number of clusters is very large, imbalances may remain, suggesting these characteristics should still be controlled for. To this end, our analysis models included a 4-level variable for randomization strata to increase the precision of effect estimates. This stratification variable was defined as the cross classification of each hospital's designation as a primary or comprehensive stroke center (either vs. neither) and stroke volume (high vs. low).

Intention-to-Treat Analysis

For the sake of brevity, we focus our discussion of the prespecified intention-to-treat (ITT) analyses on the primary outcome (SIS-16). Our prespecified analysis methods attempted to address 3 main challenges enumerated above: (1) confounding due to imbalance in patient-level characteristics that are associated with outcomes; (2) missing values for key patient-level characteristics; and (3) selection bias associated with outcome nonresponse. The primary analysis of SIS-16 scores was based on a weighted linear mixed model (LMM) that included a hospital-specific random effect (ie, random intercept), the 4-level randomization stratification variable, a cluster-level indicator for randomization arm, and patient-level covariates (ie, primary diagnosis, age, race, National Institutes of Health Stroke Scale score). Further sensitivity analyses, discussed below, included additional patient-level covariates (Table 1).

Accounting for Missing Covariate Data Using Multiple Imputation

Efforts taken during the trial successfully minimized missing data; however, key patient characteristics still suffered from varying degrees of missingness, ranging from <1% to 6.5% (see footnote to Table 1). For example, documentation of whether the patient was in need of

TABLE 1. Intention-to-treat and Per-protocol Effect Estimates Obtained From Prespecified and Sensitivity Analyses

Analysis Type	Method	Effect Estimate	95% CI	*P*
Intention-to-treat	Complete case, restricted to participants with observed outcome and covariate data	0.95	−1.60 to 3.50	0.4479
	Multiple imputation, prespecified adjustment set*	0.42	−1.91 to 2.76	0.7223
	Inverse probability weighting+multiple imputation, pre-specified adjustment set	0.61	−1.74 to 2.97	0.6098
	Inverse probability weighting+multiple imputation, additional covariate adjustment using model selection procedure[†]	0.29	−2.12 to 2.69	0.8152
Per-protocol (estimation of the complier average causal effect)	Instrumental variables[‡]	1.11	−2.90 to 5.12	0.5867
	Compliance mixture model	1.34	−2.55 to 5.23	0.5002

Values in parentheses below indicate degree of missing data for these covariates.

*Prespecified covariates included: age, race (<1%), National Institutes of Health Stroke Scale score (2.8%), diagnosis (stroke, TIA), and randomization stratum. Values in parentheses indicate degree of missing data for these covariates.

[†]Additional covariates considered in model selection were: insurance status (2.1%), need for rehabilitation (6.5%), history of stroke, history of TIA, history of cardiovascular disease, referral for home health (6.5%), presence of primary care provider, rural residence (<1%), history of depression, history of hypertension, history of smoking.

[‡]Primary, prespecified analyses.

CI indicates confidence interval; TIA, transient ischemic attack.

rehabilitation was missing for ~6.5% of patients. At least 1 of the National Institutes of Health Stroke Scale score, race, and need of rehabilitation characteristics was missing for ~9.2% of patients; this is a degree of missingness that could render complete case analysis biased to a non-negligible degree. Moreover, in complete case analysis, changing the covariate adjustment sets for sensitivity analyses results in different subsets of complete cases, making comparisons of results across models more difficult.

We used multiple imputation by chained equations (MICE) to impute missing values of 6 covariates that we planned to explore in the sensitivity analyses.[19] MICE requires specifying a regression model for each sometimes-missing variable such that some of the covariates included may themselves be sometimes-missing variables. The method assumes that the missing data are missing at random. Because the tenability of this assumptions relies on correctly specifying the imputation models, it becomes more plausible as the number of covariates included in them increases.[19,20] This militates in favor of including "all but the kitchen sink" in each imputation model, although this presents practical challenges, especially when missing covariates are categorical in nature (eg, prediction models can perform poorly when there is insufficient information to estimate the model parameters). We attempted to balance these competing concerns using a conservative model selection procedure to first identify the variables to be included in the imputation models. Specifically, we used an ad hoc Hot Deck imputation procedure to fill in missing covariate values using the observed values from complete cases.[21] For each case that had 1 or more missing covariates, we created 10 imputed observations, each of which was given one-tenth the weight of a case with fully observed data. Each imputation model (be it a linear, logistic, or generalized logistic regression) was then fit to the Hot Deck imputed dataset, using a large number of covariates. Using backward selection with a relatively liberal *P*-value threshold, we identified the subset of covariates to include in each imputation model; selection proceeded until the effects of all remaining covariates had associated unadjusted *P* values of <0.1.

While a detailed explanation and theoretical justification of MICE is beyond the scope of this paper, it is an iterative procedure, with a Bayesian motivation. MICE uses Markov Chain Monte Carlo methods to draw samples (ie, predictions) for the missing variables to create so-called *full* datasets where no covariates have missing values. The use of MICE has been shown to perform well compared with other commonly used methods, although there is potential for improving performance using machine learning approaches, especially when the number of covariates is large or their relationships are highly complex.[22] Regardless of imputation method, the procedure yields some number, M, of full datasets (we specified M = 100). Each complete dataset is then analyzed, and the results are combined using standard techniques to account for the uncertainty associated with the unknown values of the missing covariates.[23]

When data are missing for reasons related to the (unobserved) missing values, they are considered *missing not at random*, in which case MICE cannot fully address the issues presented by missingness. In that case, investigators might

consider conducting sensitivity analyses to assess the robustness of their findings to the imputation method. Increasingly, software exists to support such efforts. For example, The MI Procedure in SAS (which we used to implement MICE) allows the user to employ a pattern-mixture approach to impute missing data under plausible models for why they are missing (eg, patients with worse outcomes are more likely to have them missing). Such investigations are particularly important if outcome ascertainment rates differ by group. If inferences from such sensitivity analyses differ with primary analyses, this must be taken into account when describing the results of the study.

Minimizing Selection Bias From Outcome Nonresponse

The primary outcome, SIS-16, was observed for 59% of patients. Thus, there was significant cause for concern that analyses performed using observed outcomes without further adjustment would be biased. There are several strategies to correct for selection bias due to outcome nonresponse. The 2 most common choices are multiple imputation (MI) of both the missing covariate data and outcomes and inverse probability weighting (IPW). Both approaches are valid when the missing data are missing at random, which in the context of the COMPASS study outcomes data means patients' baseline characteristics are sufficient to predict their outcome values. We chose to employ IPW to account for outcome nonresponse, a decision made on practical grounds.

One key factor that influenced our decision to use IPW over MI to account for missing outcomes was that we observed a moderate ceiling effect during an early blinded examination of the distribution of SIS-16 scores (~24% of participants scored at the maximum value). A ceiling (or, analogously, a floor) effect refers to the mounding of responses at the upper (lower) boundary of an instrument's range. As a consequence, no transformation can be applied to make the distribution approximately normally distributed (ie, to remove the ceiling effect). In response, our team conducted large-scale simulation studies to assess the robustness of inference using the linear model in this setting.[24] While we found that inferences were indeed robust, we noted that MI would require accurate prediction of the missing outcome values (subject to the ceiling effect) which could not be achieved by assuming a normal prediction model or using other prediction models available in standard software. This led us to choose IPW rather than imputation methods to deal with challenges presented by outcome nonresponse.

Briefly, we used a logistic regression model to estimate the probability (propensity) that a patient with given characteristics would provide an outcome (ie, respond to the 90 d outcome assessment). The outcome model was then fit to the observed data, weighting patients by the inverse of their propensities. Heuristically, selection bias due to outcome nonresponse is accounted for by upweighting data from patients who are less likely to respond. For a more rigorous exposition of IPW and its assumptions, we refer the reader to the excellent text by Hernán and Robins.[25]

Using IPWs estimated from logistic random effects models can yield biased estimation when the number of observations

within clusters is small.[26] This is precisely the setting in the COMPASS study, which employs models with hospital-specific random intercepts. As a remedy, we used conditional logistic regression, which is more robust to the presence of small clusters. We estimated IPWs separately for each study arm using a common set of covariates as predictors.[26]

A final complication was that some of the covariates used to estimate the IPWs exhibited missingness. To address this issue, for each imputed dataset, the IPW analysis procedure described above was performed, and the parameter estimates were combined across imputations. Simultaneously accounting for these sources of bias presents a somewhat daunting computational challenge, especially in cases where the number of endpoints examined in the trial is large, since the multistep analysis process must be repeated for each endpoint.

Sensitivity Analysis to Assess Robustness of Prespecified Analyses

The methods described above to address covariate missingness, outcome nonresponse, and confounding due to imbalance in patient characteristics are subject to criticism to the degree that their underlying assumptions are violated. We felt it important to present results of alternative analyses that would allow critical readers to assess the robustness of the approaches taken. To this end, we performed 3 sensitivity analyses of the primary endpoint, including: (1) a complete case analysis which made no use of the sophisticated statistical machinery described above; (2) an analysis that did not incorporate IPWs to gauge their influence on the analysis; and (3) an analysis that included an expanded covariate adjustment set to assess whether confounding was robustly controlled for in the analysis models based on the prespecified covariates. Results of these analyses are presented in Table 1 and additional details are available in a supplement to the primary results publication.[27]

Per-Protocol Analyses

Increasingly, supplementing ITT analyses with PP analyses is viewed as important in comparative effectiveness research in pragmatic trials.[28] PP analysis was an integral part of our analysis plan given its relevance to patient stakeholders and the subsequent substantial nonadherence to treatment that we observed in the COMPASS study. Some hospitals did not consistently implement the intervention during the trial because of limited resources including staff and competing demands. Others never achieved true buy-in by those implementing the intervention.[29] Patients also experienced barriers to receipt of treatment including costs, transportation challenges, and difficulty balancing multiple follow-up visits. As a result, only 35% of patients assigned to treatment received the core elements of the COMPASS-TC intervention (ie, an eCare plan within 30 d of hospital discharge). There was considerable heterogeneity in adherence across hospitals, ranging from 6% to 70% (Fig. 2). ITT estimates in the presence of nonadherence do not represent the biological effect of the intervention but rather a mixture of the effects in compliers and noncompliers. As such, the ITT estimate may not be indicative of the experience in a setting where
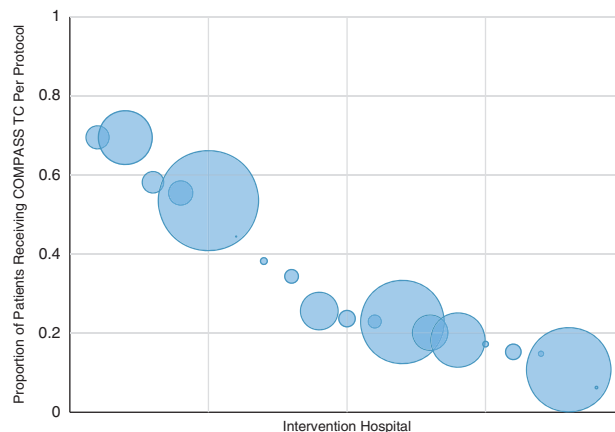

FIGURE 2. Proportion of patients receiving COMprehensive Post-Acute Stroke Services transitional care (COMPASS-TC) per-protocol by hospital. Circles represent the 19 hospitals that adopted the intervention and the area represents the total number of enrolled patients. Values on the *y*-axis represent the proportion of patients who received COMPASS-TC per-protocol (an electronic care plan within 30 d of index discharge).

treatment is widely and successfully adopted. For these reasons, we prespecified PP analyses to better understand the effect of treatment *when received*. We endeavored to estimate the conditional (ie, covariate adjusted) *complier-average causal effect* (CACE) of the intervention, which is based on Rubin compliance principal strata.[30] A rigorous discussion of compliance principal strata is beyond the scope of this paper. Briefly, under this framework, 4 subpopulations are relevant with respect to treatment receipt: (1) always-takers: those patients who would receive an intervention regardless of treatment assignment; (2) never-takers: those who would never receive an intervention; (3) compliers: those patients who would follow COMPASS-TC protocols or those of UC (whichever they were offered); and (4) defiers: those who would receive the opposite care assigned to the hospital unit in which they were enrolled. Our analyses were based on a set of identifiability assumptions, discussed below. One is the unverifiable but plausible assumption that there are no defiers in the study population. In our case, because patients treated at UC sites had no opportunity to receive the COMPASS intervention, there were no defiers or always takers in the population. Therefore, our population consisted of compliers and never-takers so that a noncomplier is essentially a synonym for never-taker. PP analyses attempted to estimate the COMPASS intervention effect among compliers. We can infer that patients who received the COMPASS-TC were compliers. However, as we did not closely monitor care recommendations and adherence in the UC arm, compliance to UC was unobserved, presenting a challenge for estimation of the CACE.

Complier Average Causal Effect Estimation Using an Instrumental Variables Approach

Our primary strategy for estimation of the CACE was an instrumental variables (IV) approach.[31] A rigorous discussion of IV is beyond the scope of this paper, for which we
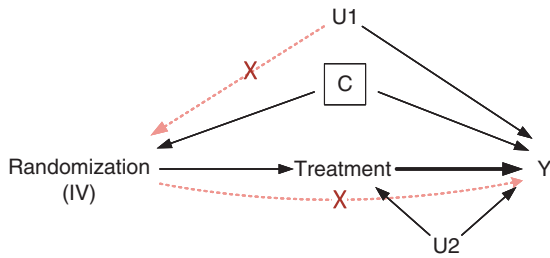
FIGURE 3. Directed acyclic graph representing the COMprehensive Post-Acute Stroke Services cluster-randomized trial where randomization assignment is the instrumental variable (IV) and Y represents the outcome of interest. U1 and U2 represent unmeasured confounders of the randomization-outcome and treatment-outcome relationships, respectively. C represents measured confounders that were adjusted for in the analysis. Direct causal effects are represented by solid arrows. Instrumental conditions require that (1) the IV is associated with treatment; (2) there is no effect of IV on Y except through treatment; and (3) there is no common cause of IV and Y. The dotted arrows are included to illustrate the effects assumed to be absent under conditions 2 and 3. Of note, the solid arrow between C and IV is included to represent observed associations present after randomization, although these are not causal effects.

guide the reader to Hernán and Robins (see Section 16.1 and Technical Point 16.1 therein for more rigorous exposition on this topic). Although fundamental to the approach, identifying a valid IV can be challenging. Briefly: (1) the instrument must be associated with treatment receipt; (2) the instrument must not have a causal effect on the outcome of interest except through treatment receipt; and (3) the instrument and outcome must not have a common cause (that is not accounted for in the analysis). Figure 3 graphically depicts the causal relationships assumed for the preplanned IV analysis for the COMPASS study.

One of the key benefits of IV analysis is that, if the analyst adequately adjusts for the effects of the measured confounders (C) of the relationship between study arm and outcome, and there are no unmeasured confounders (U1), inference on the effect of treatment receipt on the outcome is valid, *even if there are unmeasured confounders of that relationship* (U2). The statistical analysis plan for the COMPASS study provides a thorough description of the rationale for using randomized study arm (a cluster-level variable) as an instrument. Briefly, the fact that criterion 1 is met is obvious. For criterion 2, the research team felt that the majority of the intervention's effectiveness came through attending a specialized clinic visit and receiving an individualized eCare plan. Thus, simply being enrolled at a hospital that was randomized to provide the COMPASS-TC was not viewed as having the potential to provide a meaningful degree of efficacy and thus criterion 2 holds, if only approximately. Criterion 3 requires appropriate adjustment for prognostic characteristics that were imbalanced across study arms, which we attempted to achieve using strategies described above.

It is important to note that, in the context of the COMPASS study, having an instrument that meets criteria 1–3 above is not sufficient for the IV estimator to adequately

target the CACE. Further assumptions are necessary. Specifically, one must assume that patients enrolled in UC hospitals who were compliers (an unobservable characteristic) did not have the option to receive specialized post-acute care akin to COMPASS-TC (ie, compliance is defined with respect to specialized post-acute care, which was assumed to be unavailable in the UC setting). In addition, one must further assume that intervention arm patients who do not receive the eCare plan (ie, noncompliers) received minimal specialized post-acute care, consistent with the level of intervention provided in UC settings to all patients. These assumptions are admittedly strong, but not unreasonable in our opinion, for several reasons. First, before randomization, hospitals did not provide comprehensive post-acute care fully consistent with COMPASS-TC.[32] Second, a key reason for noncompliance to COMPASS-TC was preference for alternative follow-up, for example, primary care, consistent with that offered in UC.[29]

Two-stage least squares estimation was performed for IV analysis and robust standard errors were computed to account for the cluster-level heterogeneity of patient outcomes.[31,33] Briefly, the first stage regression model regresses treatment receipt on hospital-level and patient-level characteristics using intervention arm data. In the second stage regression, the outcome is regressed on the same set of characteristics except that treatment receipt is replaced by the estimated probability of receipt from the first stage regression (and set to zero for UC patients). Use of IVs offers no protection from selection bias related to outcome ascertainment.[34] Thus, the same methods for employing IPW coupled with MI described above were also incorporated into the IV analysis. This has been shown to improve the quality of IV estimation in the presence of selection bias.[35]

Sensitivity Analysis: Complier Average Causal Effect Estimation Using a Complier-Outcome Mixture Model

A key limitation of IV analysis described above is sensitivity to underlying assumptions. When its assumptions are met, IV analysis protects against unmeasured confounding of the effect of treatment on the outcome of interest. The property of accounting for unmeasured confounding is unique to the IV approach. However, previous work has shown that weak instruments (ie, instruments weakly associated with intervention receipt) can lead to IV-based CACE estimates that are biased and that the bias does not disappear even under large sample sizes.[31] Ultimately, given the low rate of treatment receipt, randomized study arm may be viewed as a relatively weak instrument, and such bias was a concern to the research team. Moreover, as described above, the IV analysis assumes that there was essentially no specialized post-acute care provided by UC hospitals, an assumption that realistically can be viewed to hold only approximately.

Given the potential for biases in the IV analysis stemming from assumptions that are only approximately met, we conducted a sensitivity analysis designed to be more robust than the IV approach *to the assumptions described above* using a compliance-outcome mixture model. Unlike IV analysis with a valid instrument, that approach remains subject to bias associated with unmeasured confounding.

The compliance-outcome mixture model[36] assumes that patients can be viewed as either compliers, who by definition will adhere to specialized post-acute care treatment recommendations, or noncompliers. Unlike the IV analysis, the approach does not require one to assume that specialized post-acute care is not provided by UC but does assume that: (1) outcomes for noncompliers are the same, on average, for patients enrolled in both intervention and UC hospital units after appropriate covariate adjustment; (2) compliers will comply with specialized post-acute care recommendations, regardless of whether they correspond to those from the COMPASS intervention or those from UC; and (3) the effect of specialized post-acute care provided as a part of UC is the same across UC hospital units.

This analysis is based on a joint-model for the primary endpoint and patient-level compliance status. The primary outcome model includes a treatment effect for compliers in the UC arm and for compliers in the intervention arm but otherwise uses the same LMM framework described above. From this model, one is able to directly estimate the CACE while accounting for the fact that UC hospitals may provide specialized post-acute care that provides some degree of efficacy. For the second part of the model, compliance status is modeled via logistic regression. Both the LMM and logistic model incorporated covariates associated with the primary outcome and compliance, respectively, to address potential confounding. Since compliance status is not observed for UC patients, this variable is integrated out of the likelihood, resulting in a mixture of LMMs for the control group. Additional details and results of these analyses are available in the supplement to the primary paper.[27]

DISCUSSION

Large-scale pragmatic trials such as the COMPASS study are designed to generate evidence on the comparative effectiveness of interventions as delivered in real-world settings. The complexity of these trials necessitates utilization of statistical methods appropriately suited for analysis of the complex data that arise, often due to the less controlled (ie, pragmatic) aspects of the trial's design. To the degree possible, these challenges should be identified *a priori* and addressed in a prespecified statistical analysis plan.[3] However, not all challenges can be anticipated, and analysis plans should be flexible and incorporate sensitivity analyses. In this paper, we have used the COMPASS study as an example to describe the application of statistical methods to analyze clustered data that were subject to missingness with respect to key baseline characteristics, outcome nonresponse, and nonadherence to the intervention.

On the basis of our experience with the COMPASS study, we advocate that researchers proactively consider the most appropriate balance between control in trial design and pragmatism, which should be guided by the study goals.[37] For example, we observed that the generalizability gained from an opt-out consent model (as compared with a patient-randomized design) might result in lower patient engagement and therefore play a role, among other factors (eg, telephone-based assessment), in increasing outcome nonresponse to levels that present significant analytic challenges. At a minimum, collection of factors associated with outcome ascertainment is important to facilitate use of methods to address selection bias (ie, imputation, IPW). We successfully applied such methods, having captured extensive baseline data and key prognostic factors on all patients enrolled. However, as a result of not closely following the UC group due to the pragmatic nature of our trial, we had limited data on time-varying confounders between the index hospital discharge and 90-day outcome assessment, limiting the statistical methods at our disposal. In contrast, had we more closely followed UC (and intervention) patients, we would have had better data for this purpose but a trial that was less pragmatic. Continued efforts to more seamlessly integrate clinical care and research, as in learning health systems, may enhance patient engagement in research and allow for the routine capture of patient-centered outcomes and other important data more completely. Increased use of technology may also offer tools to lower the patient burden of participation in research studies.

Cluster randomization is frequently used in pragmatic trials when interventions are necessarily system-level due to the significant reorganization of care they require. Stratified randomization improves balance on system-level confounders. This technique was successfully used in the COMPASS study, which achieved good balance in a variety of cluster-level characteristics, including the 2 guaranteed by stratification. Where relevant, run-in periods can be used to identify or randomize only hospitals or systems that have the capacity to implement a complex intervention. However, there is again a tradeoff regarding generalizability, but this may be worthwhile to avoid significant implementation challenges that may preclude meaningful interpretation of study results.

Stepped wedge designs may offer advantages to parallel designs (ie, improved power in certain situations, hospitals contribute to data on both comparators) but may require more investment of time by study staff for site training, especially if individual clusters (eg, hospitals) are rolled into the intervention phase at different times.[38] Such an approach would also necessarily lengthen the trial. However, it may have the additional benefit of temporarily embedding research staff in the real-word research settings to provide direct support in the actual health care setting for a more tailored transition phase.

Implementation of complex care transitions models is challenging for many hospitals. We have previously reported on the substantial efforts required to recruit hospitals and to implement COMPASS-TC.[16,29] Critical to successful TC delivery is a commitment to implementing the care model as new standard practice and having sufficient capacity to do so.[39] In the COMPASS study, research staff provided ongoing, extensive support for training and implementation. Providing financial support may be necessary for future studies, given the high degree of competing demands for health care providers and systems. Conducting research, even if part of clinical care, requires additional investment on the part of hospital staff to screen, notify, and enroll patients and collect study-related data. Incentives may be particularly important to engage hospitals with limited capacity and to generate high-quality data that achieve the best return on investment in pragmatic trials.[14,29,39]

By their very nature, pragmatic trials are subject to greater nonadherence than highly controlled trials.[2,40] Ensuring the correct balance between maximizing adherence and delivering care in a real-world setting during the design phase is important. Pragmatic trials should plan to provide results with a variety of stakeholders in mind.[28] ITT effects (meaningful to policy makers) and PP effects (meaningful to patients) provide complementary information. PP estimates, in particular, require careful thought and rigorous collection of data. With complex interventions such as care transitions, the treatment is not delivered fully at the time of randomization but rather involves ongoing chronic disease management and multiple points of patient-provider interaction. It is critical to carefully plan for the collection of patient-level and cluster-level prognostic factors associated with the time-varying nature of treatment receipt and adherence to facilitate estimation of causal PP effects. Furthermore, the degree of heterogeneity across systems may impact effectiveness of interventions. For example, the quality and availability of resources in a given setting may impact effectiveness. Teasing out these aspects is challenging and requires planned evaluations of treatment effect heterogeneity to provide information about who is most likely to benefit and in what settings various interventions may have the most impact.

In summary, some lessons learned from our experience with pragmatic research are the importance of collecting post-randomization prognostic information and careful measurement of the UC comparator. Maximizing site fidelity should be carefully considered, and higher fidelity may be achieved through use of organizational readiness surveys, run-in periods, financial and staffing support, or alternative study designs (eg, a stepped wedge). Balancing the potential benefits and consequences of these and other possible design choices is necessarily study specific and should be closely tied to the study objectives. Our goal herein is to highlight the universal importance of pre-specifying appropriate analyses in light of the complex data that arise from pragmatic studies, providing examples based on our experience in the COMPASS study.

Pragmatic trials are an increasingly important aspect of research design and are well suited to inform clinical practice. However, deciding which pragmatic design elements to include in a study should not be taken lightly, as such trials almost always entail tradeoffs. The potential benefits of pragmatic trials can often only be realized with the concurrent application of valid (and sometimes quite complex) statistical methods. In this paper, we have shared our experience tackling statistical issues attendant to pragmatic trials (eg, the presence of missing data and nonadherence to treatment) in the context of the COMPASS study to help inform future work.

REFERENCES

1. Patsopoulos NA. A pragmatic view on pragmatic trials. *Dialogues Clin Neurosci.* 2011;13:217–224.
2. Ford I, Norrie J. Pragmatic Trials. *N Engl J Med.* 2016;375:454–463.
3. Weinfurt KP, Hernandez AF, Coronado GD, et al. Pragmatic clinical trials embedded in healthcare systems: generalizable lessons from the NIH collaboratory. *BMC Med Res Methodol.* 2017;17:144.
4. Broderick JP, Abir M. Transitions of care for stroke patients: opportunities to improve outcomes. *Circ Cardiovasc Qual Outcomes.* 2015;8(suppl 3):S190–S192.
5. Adeoye O, Nyström KV, Yavagal DR, et al. Recommendations for the establishment of stroke systems of care: a 2019 update. *Stroke.* 2019;50: e187–e210.
6. Prvu Bettger J, McCoy L, Smith EE, et al. Contemporary trends and predictors of postacute service use and routine discharge home after stroke. *J Am Heart Assoc.* 2015;4:e001038.
7. Bushnell CD, Zimmer LO, Pan W, et al. Persistence with stroke prevention medications 3 months after hospitalization. *Arch Neurol.* 2010;67:1456–1463.
8. White CL, Pergola PE, Szychowski JM, et al. Blood pressure after recent stroke: baseline findings from the secondary prevention of small subcortical strokes trial. *Am J Hypertens.* 2013;26:1114–1122.
9. Fini NA, Holland AE, Keating J, et al. How physically active are people following stroke? Systematic review and quantitative synthesis. *Phys Ther.* 2017;97:707–717.
10. Forster A, Young J. Incidence and consequences of falls due to stroke: a systematic inquiry. *BMJ.* 1995;311:83–86.
11. Benjamin EJ, Blaha MJ, Chiuve SE, et al. Heart disease and stroke statistics-2017 update: a report from the American Heart Association. *Circulation.* 2017;135:e146–e603.
12. Ovbiagele B, Goldstein LB, Higashida RT, et al. Forecasting the future of stroke in the United States: a policy statement from the American Heart Association and American Stroke Association. *Stroke.* 2013;44: 2361–2375.
13. Dhamoon MS, Longstreth WT Jr, Bartz TM, et al. Disability trajectories before and after stroke and myocardial infarction: the Cardiovascular Health Study. *JAMA Neurol.* 2017;74:1439–1445.
14. Duncan PW, Bushnell CD, Rosamond WD, et al. The Comprehensive Post-Acute Stroke Services (COMPASS) study: design and methods for a cluster-randomized pragmatic trial. *BMC Neurol.* 2017;17:133.
15. Duncan PW, Lai SM, Bode RK, et al. Stroke Impact Scale-16: a brief assessment of physical function. *Neurology.* 2003;60:291–296.
16. Johnson AM, Jones SB, Duncan PW, et al. Hospital recruitment for a pragmatic cluster-randomized clinical trial: lessons learned from the COMPASS study. *Trials.* 2018;19:74.
17. Bushnell CD, Duncan PW, Lycan SL, et al. A person-centered approach to poststroke care: the COMprehensive Post-Acute Stroke Services Model. *J Am Geriatr Soc.* 2018;66:1025–1030.
18. Andrews JE, Moore JB, Weinberg RB, et al. Ensuring respect for persons in COMPASS: a cluster randomised pragmatic clinical trial. *J Med Ethics.* 2018;44:560–566.
19. Van Buuren S. Multiple imputation of discrete and continuous data by fully conditional specification. *Stat Methods Med Res.* 2007;16:219–242.
20. Schafer JL. *Analysis of Incomplete Multivariate Data*. New York, NY: Chapman and Hall/CRC; 1997.
21. Andridge RR, Little RJ. A review of hot deck imputation for survey nonresponse. *Int Stat Rev.* 2010;78:40–64.
22. Ibrahim JG, Chen M-H, Lipsitz SR, et al. Missing-data methods for generalized linear models. *J Am Stat Assoc.* 2005;100:332–346.
23. Rubin DB. *Multiple Imputation for Nonresponse in Surveys*. New York, NY: Wiley; 1987.
24. DeBarmore BM, Schilsky SR, Psioda MA, et al. Abstract P351: Evaluation of analytic approaches to address ceiling effect in patient-reported functional status after stroke. *Circulation.* 2018;137: AP351.
25. Hernán MA, Robins J. *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC; 2020.
26. Skinner CJ, D'Arrigo J. Inverse probability weighting for clustered nonresponse. *Biometrika.* 2011;98:953–966.
27. Duncan PW, Bushnell CD, Jones SB, et al. Randomized pragmatic trial of stroke transitional care: the COMPASS study. *Circ Cardiovasc Qual Outcomes.* 2020;13:e006285.
28. Hernán MA, Robins JM. Per-protocol analyses of pragmatic trials. *N Engl J Med.* 2017;377:1391–1398.
29. Gesell SB, Bushnell CD, Jones SB, et al. Implementation of a billable transitional care model for stroke patients: the COMPASS study. *BMC Health Serv Res.* 2019;19:978.
30. Frangakis CE, Rubin DB. Principal stratification in causal inference. *Biometrics.* 2002;58:21–29.

31. Baiocchi M, Cheng J, Small DS. Instrumental variable methods for causal inference. *Stat Med*. 2014;33:2297–2340.

32. Bettger JP, Jones SB, Kucharska-Newton AM, et al. Meeting Medicare requirements for transitional care: do stroke care and policy align? *Neurology*. 2019;92:427–434.

33. White H. Instrumental variables regression with independent observations. *Econometrica*. 1982;50:483–499.

34. Hughes RA, Davies NM, Davey Smith G, et al. Selection bias when estimating average treatment effects using one-sample instrumental variables analysis. *Epidemiology*. 2019;30:350–357.

35. Canan C, Lesko C, Lau B. Instrumental variable analyses and selection bias. *Epidemiology*. 2017;28:396–398.

36. Jo B, Asparouhov T, Muthén BO, et al. Cluster randomized trials with treatment noncompliance. *Psychol Methods*. 2008;13:1–18.

37. Loudon K, Treweek S, Sullivan F, et al. The PRECIS-2 tool: designing trials that are fit for purpose. *BMJ*. 2015;350:h2147.

38. Hemming K, Haines TP, Chilton PJ, et al. The stepped wedge cluster randomised trial: rationale, design, analysis, and reporting. *BMJ*. 2015; 350:h391.

39. Lutz BJ, Reimold AE, Coleman SW, et al. Implementation of a Transitional Care Model for Stroke: perspectives from frontline clinicians, administrators, and COMPASS-TC Implementation staff. *Gerontologist*. 2020;60:1071–1084.

40. Ware JH, Hamel MB. Pragmatic trials–guides to better patient care? *N Engl J Med*. 2011;364:1685–1687.