iCOMPARE

individualized Comparative Effectiveness Models Optimizing Patient Safety and Resident Education

Statistical Analysis Plan

Version 1

9 October 2015

Table of Contents

1 Introduction

iCOMPARE is a cluster randomized trial comparing the current duty hour standards for interns (**Curr**) with a more flexible duty hour schedule (**Flex**) in at least 58 Internal Medicine training programs. Training programs randomized to **Flex** must schedule intern duty hours in conformance with these 3 rules, each averaged over 4 weeks:

1. Work no more than 80 hours per week
2. Call no more frequent than every 3rd night
3. 1 day off in 7

Current duty hour standards (**Curr**) require that interns work no more than 16h of duty periods in a day.

iCOMPARE has one primary hypothesis:

H1a: 30-day patient mortality under **Flex** will not exceed (will not be inferior to) mortality under **Curr**.

iCOMPARE will test related and complementary secondary hypotheses regarding:

Patient safety and costs:

- **H1b**: 7-day and 30-day hospital readmission rates under **Flex** will not exceed (will not be inferior to) the rates under **Curr**.
- **H1c**: Complication rates, defined by selected AHRQ Patient Safety Indicators, under **Flex** will not exceed (will not be inferior to) complication rates under **Curr**.
- **H1d**: The rate of prolonged length of stay, defined as a stay that exceeds the Hollander-Proschan point or that length of stay for a given condition at which the discharge rate begins to decline, under **Flex** will not exceed (will not be inferior to) the rate of prolonged length of stay under **Curr**.
- **H1e**: Overall costs, as indicated by total Medicare payments, under **Flex** will not exceed (will not be inferior to) overall costs under **Curr**.

Trainee education:

- **H2a**: Interns in **Flex** will spend greater time in direct patient care and education compared to interns in **Curr**.
- **H2b**: Trainees in **Flex** will report greater satisfaction with their educational experience (greater ownership, greater continuity and lower burnout) than trainees in **Curr**.
- **H2c**: Faculty in **Flex** will report greater satisfaction with their clinical teaching experiences and greater perceptions of safety, teamwork and supervision than faculty in **Curr**.
- **H2d**: Standardized test scores for interns in **Flex** will not be less than (inferior to) those for interns in **Curr**.

and Intern sleep and alertness:

- **H3a**: Average daily sleep obtained by interns in **Flex** will not be less than (will not be inferior to) that of interns in **Curr**, as determined by a 14-day period of sleep monitoring using actigraphy and daily sleep diaries.
- **H3b**: Interns in **Flex** will not have (will not be inferior to) greater average subjective sleepiness via Karolinska Sleepiness Score (KSS), or lower average behavioral alertness via psychomotor vigilance test (PVT) than interns in **Curr**, as determined by a 14-day period of morning sleepiness-alertness monitoring.

The iCOMPARE primary outcome (30-day mortality) was chosen to ensure that any policy change in resident duty hours will not result in inferior patient safety. However, additional patient safety measures, as well as costs, education and fatigue management, are critically important considerations which our study addresses. The results of iCOMPARE will help the ACGME in its ongoing deliberations about optimal resident duty hour schedules. Changes in ACGME policies affect every teaching hospital in the United States, and as a consequence, every patient.

2 Data sources

Data on patient outcomes and costs of health care will come from Medicare claims records. These records will be obtained through application to and purchase from the Research Data Assistance Center located at the University of Minnesota School of Public Health (ResDAC; http://www.resdac.org/). All requests for Medicare data proceed through ResDAC. We will obtain claims data from Medicare for calendar years 2013 through 2016.

Data collected directly from trainees and program directors through iCOMPARE surveys, actigraphy, and/or observation will be supplemented with data collected on trainees, program directors and faculty by national organizations such as the ACGME, the American College of Physicians (ACP) and the Association of Program Directors in Internal Medicine (APDIM). The time period of direct data collection by iCOMPARE survey, actigraphy or observation will be May 2015 through June 2016. The data obtained from the ACGME, ACP, and APDIM includes responses to their 2015 and 2016 surveys and ITE scores from fall 2015 and fall 2016.

The Medicare claims data will be received, analyzed and stored at Children's Hospital of Philadelphia. Surveys are administered by the CCC located at University of Pennsylvania and responses are collected by the CCC; files will be transferred to the DCC periodically for review and analysis. Sleep and Alertness Substudy data will be collected by Pulsar, transferred to the Sleep and Alertness Substudy group for post collection processing, and then transferred periodically to the DCC for review and analysis. The data received by the DCC will be devoid of personal identifiers such as name, phone number, and email address. Backup files of the database at the DCC will be generated and stored at regular intervals in a secure, off-site location, to permit regeneration of the database in the event that it is destroyed. Freeze dates for data sets created for interim and publication analyses will be documented.

3 Sample size and power

We approached the statistical design by designating the non-inferiority mortality hypothesis H1a as the primary hypothesis for which sample size calculations were based. The PASS 11 software for sample size and power analysis was used to calculate the sample size required for this hypothesis. The PASS 11 software accommodates the complex statistical calculations needed to allow for superiority or non-inferiority hypotheses, and correlations in responses such as those we will see due to clustering on the internal medicine programs.

The 30-day mortality rate in the iCOMPARE target population was estimated to be 11% [11.1% in 2007 and 11.5% in 2008; personal communication from Dr. Silber] and the pooled standard deviation (SD) for the pairs of rate differences was estimated to be 1.5%. The consensus non-inferiority mortality margin among the iCOMPARE investigators was assumed to be 1%. The 30-day mortality outcome measure is defined, for each IM program, as the difference between the 30-day mortality rate in the trial year minus

the 30-day mortality rate in the pre-trial year. This approach permits the use of a simple model (two-sample non-inferiority t-test) for the set of at least N=29 pairs of test year vs. pre-test year differences in each group (**Curr** vs. **Flex**) in annual 30-day mortality rates that obviates the need for complex risk adjustment models, since it adjusts each outcome for secular trends in 30-day mortality as well as in IM program population risk profiles that are likely to cancel out by comparing successive years. We performed the calculations with both 80% and 90% power to gauge any gains in power by recruiting beyond the N=58 IM programs required for 80% power. The results of the calculations for mortality non-inferiority from PASS 11 are as follows, where Type-1 error (alpha) is based on a one-sided t-test as is appropriate for a non-inferiority design.

Power	N1 (Curr) /N2 (Flex)	Non-inferiority Margin (NIM)	Actual Difference (D)	Significance Level (Alpha)	Beta	Standard Deviation 1 (Curr) (SD1)	Standard Deviation 2 (Flex) (SD2)
0.8059	29/29	0.01	0	0.05	0.1941	0.015	0.015
0.9050	40/40	0.01	0	0.05	0.0950	0.015	0.015

Although sample size calculations were based on the mortality outcome, this number of programs will give excellent power for other study hypotheses. The 58 randomized programs are expected to include 4640 internal medicine residents: 1740 interns (approximately 30 interns per program) and approximately 1450 PGY2 trainees (approximately 25 PGY2 per program) and 1450 PGY3 trainees (approximately 25 PGY3 per program). Each program will include one program director (total of 58) and approximately 10 associated faculty (total of about 580 faculty).

The sample size for the Time and Motion Substudy of 6 programs was a practical choice; this Substudy addresses hypothesis H2a. Preliminary data (Block et al, 2013; Fletcher et al 2012) suggest that the mean percent time spent in direct patient care and education in **Curr** will be about 13% (SD = 4%). With 60 interns (10 each from the 6 IM programs), we will be able to detect a 3% difference in the time spent outcome between **Curr** and **Flex** with greater than 80% power.

Hypothesis H3a was designated as the primary hypothesis for the Sleep and Alertness Substudy. For H3a, with 90% power, one-sided Type I error of 0.05, and a non-inferiority margin of 0.5 hours, the required sample size is 290 interns. The proposed sample size is higher: 384 interns (48 at each of 8 programs) for H3a.

4 Data analysis

4.1 Patient safety hypothesis H1a – 30-day mortality

The primary outcome will be the difference in the pre-trial year and trial year 30-day mortality rates. The SD of the set of paired annual differences (2008 vs. 2007) in 30-day mortality from the preliminary data was equal to 1.5%. The mortality rates were similar across the two years: 11.1% and 11.5% for 2007 and 2008, respectively, consistent with minimal secular trends in mortality. The non-inferiority sample size

calculations described above show high power and low one-sided type-**I** error for the non-inferiority hypothesis with a 1% margin.

The program level data needed for Specific Aim 1 outcome measures (patient safety and costs) will be aggregated into rates or other measures at the program level across two 1-year periods – the rates in the pre-trial year and the rates in the year of the trial. The outcome measure will be the change in these rates from the pre-trial year to the trial year and will be compared by treatment group using the same non-inferiority test proposed in the sample size justifications above.

The model (model A1) is:

$Y_i = \gamma + \beta_1 x_i + \varepsilon_i, \ \ i = 1, \ldots, n_p \ \ where$
$Y_i =$ Outcome measure in IM program i =*Test year rate - Prior year rate*,
$\gamma =$ the intercept in the reference group (Curr),
$\beta_1 =$ the difference in intercepts between Flex and Curr,
$x_i = 1$ if the ith IM program is in **Flex**, 0 if the ith IM program is in **Curr**
$\varepsilon_i =$ i.i.d. random Gaussian errors with mean 0 and variance σ^2
$n_p =$ Number of clusters (IM programs)

Tests of β_1 (or $\beta_1 - nim$, where *nim* is the non-inferiority margin) estimated using linear regression will be used to test this hypothesis, since β_1 is the expected difference in outcome: **Flex** vs. **Curr**. All randomized programs will be included in this model and we expect no missing data.

The primary outcome measure for this cluster-randomized trial is difference in 30-day in-patient mortality rates during the trial year versus the prior, which is 58 rate differences, 1 per randomized program, with a two-sample t-test for non-inferiority patient mortality (difference no greater than 1%), comparing the **Flex** vs. **Curr** schedules. While these rate differences should adjust for most program to program differences in patient and institutional mix, we will also perform an important series of sensitivity analyses for mortality and other safety and costs aims using risk-adjustment modeling methods that we developed(Patel et al 2014) and which we will apply to the iCOMPARE data to confirm the findings of the primary analysis or to add a risk-adjustment caution to the statements of findings.

Briefly, sensitivity analyses will use iCOMPARE program-level risk-adjusted data on mortality and other patient safety and costs by aggregating the patient-level risks derived from our fitted models. We will compute each patient's predicted mortality in the pre-trial period as we did in previous work(Volpp et al, 2007; Volpp et al, 2007; Patel et al, 2014) in order to estimate each patient's predicted mortality in the pre-trial period. We will then average the predicted mortality for all patients in a program in the pre-trial period to obtain program i's expected mortality rate in the pre-trial period, call it RA_{pi}, and average the predicted mortality for all patients in a program in the trial period to obtain what program i's expected mortality rate would have been had these same patients been treated in the pre-trial period, which we call RA_{ti}. Then we update model A1 (at the program level) to include a term for the difference (RA_{ti} - RA_{pi}) so that the coefficient on treatment assignment is our estimate of being assigned to treatment versus not being assigned on the change in the outcome (e.g., mortality rate), adjusting for any changes in the risks of the patients at the program.

The model (Model B) is:

$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i, \ \ i = 1, \ldots, n_p, \ \ where$
$Y_i =$ Mean outcome measure in IM for program i, e.g., difference in mortality in pre − trial and trial period

x_{1i} = 1 if the ith IM program is in Flex, 0 if the ith IM program is in Curr
$x_{2i} = (RA_{ti} - RA_{pi})$
β_0 = Intercept for the IM programs
ε_i = i. i. d. random Gaussian errors with mean 0 and variance σ^2
n_p = Number of clusters (IM programs)

The adjusted non-inferiority tests will be t-tests based using the same non-inferiority margin (1%) and the estimates and standard errors for β_1 from the regression models.

4.2 Other patient safety hypotheses – H1b-H1e

The following outcomes are the remaining outcomes for the patient safety and cost hypotheses:

a) Patient safety and costs hypothesis H1b:
- Measure: 7-day and 30-day hospital readmission rates
- Non-inferiority margin: 1%

b) Patient safety and costs hypothesis H1c:
- Measure: complications rates, defined by selected AHRQ Patient Safety Indicators
- Non-inferiority margin: 1%

c) Patient safety and costs hypothesis H1d:
- Measure: The rate of prolonged length of hospital stay
- Non-inferiority margin: 1%

d) Patient safety and costs hypothesis H1e:
- Measure: Overall resources utilized and Medicare payments for patient care
- Non-inferiority margin: 1%

Analyses for H1b-e will use the same approach described for model A1 with either linear, logistic, or Poisson models depending on whether the outcome measure is measured/ordered, a proportion, or a count. The model estimates, 95% CIs, and p-values will be derived using Stata, R or SAS.

4.3 Trainee education hypotheses – H2a-H2d

The following outcomes are the outcomes for the trainee education hypotheses:

a) Education hypothesis H2a:
- Measure: Direct patient care and education measured from Time and Motion Substudy, specifically percent of time spent by the intern in direct patient care
- Minimum important difference is 3% (0.75 SD)

b) Education hypothesis H2b:
- Measure: Trainee satisfaction with their educational experience measured from surveys, primarily the trainee's perception of having an 'appropriate balance for education' on an ordinal scale and is expected to have a mean of 3.7 (SD 0.7) in the **Curr** schedule [31-34]
- Minimum important difference is 0.175

c) Education hypothesis H2c:

- Measure: Faculty satisfaction with their clinical teaching experiences measured from surveys, primarily the faculty ranking on 'residents workload exceeds capacity to do the work' from the ACGME survey measured on an ordinal scale with expected mean in the **Curr** schedule of 4.1 (SD 0.7) [31-34]
- Minimum important difference is 0.175

d) Education hypothesis H2d:

- Measure: Standardized test scores for interns on the In-Training Examination (ITE) measured as the percent correct with expected the mean score in the **Curr** schedule of approximately 65 (SD = 18) [Lisa Bellini, personal communication].
- Non-inferiority margin is 2%

The trainee education analyses will be modeled using the model (Model A2):

$Y_{ij} = \gamma_i + \beta_1 x_{ij} + \varepsilon_{ij},\ \ i = 1, \ldots, n_p;\ \ j = 1, \ldots, n_i, \qquad where$
Y_{ij} = Mean outcome measure in IM for intern (or faculty or director) j in program i ,
x_{ij} = 1 if the ith IM program is in **Flex**, 0 if the ith IM program is in **Curr**
γ_i = i.i.d. random Gaussian intercept for the IM program i with mean β_0 and variance σ_1
β_1 = difference in intercepts in Flex and Curr
ε_{ij} = i.i.d. random Gaussian errors with mean 0 and variance σ^2
n_p = Number of clusters (IM programs)
n_i = Number of interns in program i

Note that γ_i is the random intercept needed to account for clustering. Model A2 is a multilevel mixed effects model that may be estimated using the Stata software mixed command with REML estimates, R (lme4 package) or SAS (PROC MIXED). The hypotheses will be tested using model A2 with either linear, logistic, or Poisson mixed effects models depending on whether the outcome measure is measured/ordered, a proportion, or a count.

4.4 Intern sleep and alertness hypotheses – H3a-H3b

The following outcomes are the outcomes for the intern sleep and alertness hypotheses:

a) Sleep hypothesis H3a:

- Measure: Average daily sleep measured by a 14-day period of sleep monitoring using actigraphy (verified by daily sleep diaries) with expected average sleep in **Curr** of 6.946 hours (SD=1.451 hours) [David Dinges, personal communication].
- Non-inferiority margin is 0.5 hours.

b) Sleep hypothesis H3b:

- Measure: Average subjective sleepiness measured by Karolinska Sleepiness Scale (KSS)
- Non-inferiority margin: 1 unit on KSS Likert scale

The intern sleep and alertness hypotheses will be tested using Model A2 described above.

5 Missing data

The data for the primary outcome will be requested from ResDAC for each randomized program. We expect no missing data for any of the patient safety outcomes.

For the survey, sleep and time motion data, we will employ recommended strategies to prevent missing data, based on published research (National Research Council 2010; Mills, et al., 2006; Ross, et al., 1999; Booker, et al., 2011) and previous experience of the Data Coordinating Center.

If data are missing for reasons related to the study, i.e., informative censoring, the treatment effect estimate may be biased and we will perform sensitivity analyses using methods that have been described such as multiple imputation techniques for missing data, best and worst-case scenarios, and use of the drop out event as a study end-point.

6 Intern safety outcomes

We will compare the rates of serious adverse events (SAEs) by duty hour group using negative binomial models. The following events will be considered SAEs: death or hospitalization of an intern, removal of an intern from their schedule or rotation because of mental of physical condition potentially related to their duty hours, motor vehicle accident in which the intern was the driver, needle stick experienced by intern, and other on the job injury to intern.

7 Interim monitoring

iCOMPARE data and safety will be monitored by the Steering Committee and by an independent Data and Safety Monitoring Board (DSMB) as required by NIH guidelines for multicenter trials. The Steering Committee and the DSMB will monitor accumulating safety and performance data. No interim analyses are planned because the intervention phase will be completed before patient mortality data are available. Many of the education outcomes also will not be available until after completion of the intervention phase. However, during the intervention phase, the DSMB may monitor performance data, sleep and alertness outcomes, time and motion outcomes (as available), and reports of safety concerns relating to trainees, faculty or patients as provided by program directors or otherwise brought to the attention of the iCOMPARE investigators. The DSMB is a multidisciplinary group with a written charge, with members appointed by NHLBI. The DSMB will be advisory to the NHLBI. The NHLBI has appointed a multidisciplinary DSMB that will be responsible for the protection of the safety of participants in the trial. Details on the DSMB responsibilities, meetings and reports can be found in the DSMB charter.

References

Block L, Habicht R, Wu AW, Desai SV, Wang K, Silva KN, Niessen T, Oliver N, Feldman L. In the wake of the 2003 and 2011 duty hours regulations, how do internal medicine interns spend their time? J Gen Intern Med 2013;28:1042-7. PMCID:PMC3710392

Booker CL, Harding S, Benzeval M. A systematic review of the effect of retention methods in population-based cohort studies. BMC Public Health 2011, 11:249. PMID: 21504610

Fletcher KE, Visotcky AM, Slagle JM, Tarima S, Weinger MB, Schapira MM. The composition of intern work while on call. J Gen.Intern.Med. 2012 Nov;27(11):1432-7. PMCID:PMC3475836

Mills E, Wilson K, Rachlis B, et al. Barriers to participation in HIV drug trials: a systematic review. Lancet Infectious Diseases 2006, 6(1):32-38. PMID: 16377532

National Research Council. (2010). *The Prevention and Treatment of Missing Data in Clinical Trials.* Panel on Handling Missing Data in Clinical Trials. Committee on National Statistics, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

Patel MS, Volpp KG, Small DS, Hill AS, Even-Shoshan O, Rosenbaum L, Ross RN, Bellini L, Zhu J, Silber JH. Association of the 2011 ACGME resident duty hour reforms with mortality and readmissions among hospitalized Medicare patients. JAMA 2014;in press

Ross S, Grant A, Counsell C, et al. Barriers to participation in randomized controlled trials: a systematic review. Journal of Clinical Epidemiology 1999, 52(12): 1143-1156. PMID: 10580777

Shih W. Problems in dealing with missing data and informative censoring in clinical trials. Curr Control Trials Cardiovasc Med. 2002;3:4

Volpp KG, Rosen AK, Rosenbaum PR, Romano PS, Even-Shoshan O, Wang Y, Bellini L, Behringer T, Silber JH. Mortality among hospitalized Medicare beneficiaries in the first 2 years following ACGME resident duty hour reform. JAMA 2007;298:975-83.

Volpp KG, Rosen AK, Rosenbaum PR, Romano PS, Even-Shoshan O, Canamucio A, Bellini L, Behringer T, Silber JH. Mortality among patients in VA hospitals in the first 2 years following ACGME resident duty hour reform. JAMA 2007;298:984-92.