Special Communication

Development of the Flexibility in Duty Hour Requirements for Surgical Trainees (FIRST) Trial Protocol
A National Cluster-Randomized Trial of Resident Duty Hour Policies

Karl Y. Bilimoria, MD, MS; Jeanette W. Chung, PhD; Larry V. Hedges, PhD; Allison R. Dahlke, MPH; Remi Love, BS; Mark E. Cohen, PhD; John Tarpley, MD; John Mellinger, MD; David M. Mahvi, MD; Rachel R. Kelz, MD, MSCE; Clifford Y. Ko, MD, MS, MSHS; David B. Hoyt, MD; Frank H. Lewis, MD

IMPORTANCE Debate continues regarding whether to further restrict resident duty hour policies, but little high-level evidence is available to guide policy changes.

OBJECTIVE To inform decision making regarding duty hour policies, the Flexibility in Duty Hour Requirements for Surgical Trainees (FIRST) Trial is being conducted to evaluate whether changing resident duty hour policies to permit greater flexibility in work hours affects patient postoperative outcomes, resident education, and resident well-being.

DESIGN, SETTING, AND PARTICIPANTS Pragmatic noninferiority cluster-randomized trial of general surgery residency programs with 2 study arms. Participating in the study are Accreditation Council for Graduate Medical Education (ACGME)–approved US general surgery residency programs (n = 118), their affiliated hospitals (n = 154), surgical residents and program directors, and general surgery patients from July 1, 2014, to June 30, 2015, with additional patient safety outcomes collected through June 30, 2016. The data collection platform for patient outcomes is the American College of Surgeons National Surgical Quality Improvement Program (ACS NSQIP), thus only hospitals participating in the ACS NSQIP were included.

INTERVENTIONS In the usual care arm, programs adhered to current ACGME resident duty hour standards. In the intervention arm, programs were allowed to deviate from current standards regarding maximum shift lengths and minimum time off between shifts through an ACGME waiver.

MAIN OUTCOMES AND MEASURES Death or serious morbidity within 30 days of surgery measured through ACS NSQIP, as well as resident satisfaction and well-being measured through a survey delivered at the time of the 2015 American Board of Surgery in Training Examination (ABSITE).

RESULTS A total of 118 general surgery residency programs and 154 hospitals were enrolled in the FIRST Trial and randomized. Fifty-nine programs (73 hospitals) were randomized to the usual care arm and 59 programs (81 hospitals) were randomized to the intervention arm. Intent-to-treat analysis will be used to estimate the effectiveness of assignment to the intervention arm on patient outcomes, resident education, and resident well-being compared with the usual care arm. Several sensitivity analyses will be performed to determine whether there were differential effects when examining only inpatients, high-risk patients, and emergent/urgent cases.

CONCLUSIONS AND RELEVANCE To our knowledge, the FIRST Trial is the first national randomized clinical trial of duty hour policies. Results of this study may be informative to policymakers and other stakeholders engaged in restructuring graduate medical training to enhance the quality of patient care and resident education.

TRIAL REGISTRATION clinicaltrials.org Identifier: NCT02050789

JAMA Surg. doi:10.1001/jamasurg.2015.4990
Published online December 30, 2015.

Editor's Note

Supplemental content at jamasurgery.com

Author Affiliations: Author affiliations are listed at the end of this article.

Corresponding Author: Karl Y. Bilimoria, MD, MS, Surgical Outcomes and Quality Improvement Center (SOQIC), Department of Surgery, Feinberg School of Medicine, Northwestern University, 633 N St Clair St, Floor 20, Chicago, IL 60611 (k-bilimoria@northwestern.edu).

Resident duty hour requirements in the United States are still evolving as key stakeholders, such as the Accreditation Council for Graduate Medical Education (ACGME), debate how best to (re)structure graduate medical education. Educators struggle to balance the acquisition of clinical judgment, technical skills, care continuity, professionalism, and a commitment to the patient against avoiding fatigue and the related issues of resident well-being and patient safety.[1-5] Many argue that tired residents working long hours may make errors that result in patient harm.[6-10] Conversely, it may be that current duty hour limits interrupt continuity of care and require more frequent handoffs, thus exposing patients to undue risk as the doctors who know their case best are unable to follow them during the critical aspects of their care (eg, stabilization in the intensive care unit and during surgery or reoperation).[1,11,12] Moreover, the duty hour restrictions may limit resident education as they are unable to follow the "natural history" of a particular patient whom they admitted or on whom they performed surgery.[4,13]

In 2003 and 2011, the ACGME placed restrictions on duty hours mandating a maximum 80-hour work week, regulating mandatory time off between shifts, and limiting on-call periods (**Table**).[14-23] However, there is concern that these duty hour restrictions were implemented without a strong evidence base and may result in unintended detrimental effects on patient care and resident training.[17] To our knowledge, no large national randomized trials have addressed this issue, and the available evidence is often challenging to interpret owing to a number of limitations and study heterogeneity.[24] A comprehensive review of the available literature on duty hour reforms, published in 2013 by authors from the ACGME, found mixed results and possibly worse outcomes for surgical patients after duty hour reform.[25]

To address this evidence gap, a 1-year prospective 2-arm cluster-randomized pragmatic noninferiority trial of ACGME-approved residency programs was initiated, the Flexibility in Duty Hour Requirements for Surgical Trainees (FIRST) Trial (clinicaltrials.gov Identifier NCT02050789). The objective of this study is to investigate whether elimination of several ACGME resident duty hour requirements would adversely affect patient outcomes and resident education and well-being. Residency programs were randomized to (1) an intervention arm that allows modification of selected ACGME resident duty hour restrictions or (2) usual care (ie, adherence to all current ACGME restrictions) (Table). The primary outcome will be a composite measure of 30-day risk-adjusted postoperative death or serious morbidity rates. We hypothesize that hospital-level patient safety outcomes in the intervention arm will be no worse than hospital-level outcomes in the usual care arm. In addition, we will evaluate residents' and program directors' perceptions of duty hours, quality and safety, professional training, and well-being. Results from this trial will help inform the ACGME and other stakeholders in efforts to redesign postgraduate training to optimize the quality and safety of patient care, as well as resident education and well-being.

Study Aims and Hypotheses

Our study aims are 2-fold. First, we will evaluate whether modifying current resident duty hour policies to permit greater flexibility in scheduling (intervention) has any effect on the average hospital-level risk for adverse postoperative outcomes. Our first and primary study hypothesis is that patients undergoing surgery in hospitals affiliated with programs randomized to intervention (flexible duty hours) will be at no greater risk for 30-day postoperative death or serious complications compared with patients undergoing surgery in hospitals affiliated with programs randomized to usual care (hypothesis 1).

Second, we will evaluate whether modifying current resident duty hour policies to allow greater flexibility in resident duty hours will have any effect on the quality of resident education and resident well-being. Our second hypothesis is that residents in programs randomized to the intervention will be no less satisfied with their residency training and experience compared with residents in programs randomized to usual care (hypothesis 2A) and that there will be no difference between residents in programs randomized to the intervention and those in programs randomized to usual care in levels of self-reported well-being (hypothesis 2B).

Table. Overview of ACGME Resident Duty Hour Requirements in Usual Care and Intervention Arms of the FIRST Trial[a]

Requirement	Usual Care Arm	Intervention Arm
80-h Per week (averaged over 4 wk)	No change	No change
1 d Off per week (averaged over 4 wk)	No change	No change
In-house call no more than every third night (averaged over 4 wk)	No change	No change
PGY-1 resident duty periods must not exceed 16 h	No change	Eliminated
PGY-2 residents and above must not work more than 24 h with an additional 4 h for transitions in care	No change	Eliminated
Residents must have 14 h off after 24 h in-house duty and at least 8-10 h off after a regular shift	No change	Eliminated

Abbreviations: ACGME, Accreditation Council for Graduate Medical Education; FIRST, Flexibility in Duty Hour Requirements for Surgical Trainees; PGY, postgraduate year.

[a] For full details regarding the difference in duty hour requirements between the usual care and intervention arms, see the study protocol and redlined version of the ACGME duty hour requirements for intervention arm hospitals at http://www.TheFirstTrial.org.

Methods

Study Design

The FIRST Trial is a 2-arm, cluster-randomized, pragmatic noninferiority trial of ACGME-accredited general surgery residency programs in the United States. Cluster randomization was implemented at the level of residency programs because duty hour policies are designed and implemented at the program level, not at the level of hospitals or individuals. For purposes of external validity, this study was designed as a pragmatic trial where we do not impose a uniform work-hour schedule for residents in either the intervention or usual care study arms—the exact implementation of duty hour policies is left up to the individual programs. Considerable variation exists in the institutional implementation of ACGME duty hour restrictions. It would not be feasible to ensure uniform interpretation and implementation of duty hour restrictions and intervention conditions across programs. Thus, our study resembles the real-world conditions under which resident duty hour policies are implemented.

Institutional Review Board Determination

The FIRST Trial protocol was reviewed by the institutional review board (IRB) office of Northwestern University. The IRB office deemed the FIRST Trial to be nonhuman subjects research and thus waived from further review. This was further confirmed by the principal investigator (K.Y.B.) and the study team in discussions with bioethicists not involved with the study team. Residency programs and hospitals may elect to submit IRB applications to their local IRB and seek any type of determination they deem appropriate.

Study Population

Our study population consists of ACGME-approved general surgery residency programs in the United States as of January 1, 2014, including their affiliated hospitals, surgical residents, and patients therein.

Sample Enumeration

To enumerate our sample for recruitment, we used data from the American Board of Surgery, the ACGME website (https://www.acgme.org/acgmeweb/), and the American Medical Association FREIDA Online database (http://www.ama-assn.org/ama/pub/education-careers/graduate-medical-education/freida-online.page). We identified 252 general surgery residency programs that served as our sampling frame.

Eligibility Criteria

Because the data collection platform for patient-level outcomes was the American College of Surgeons (ACS) National Surgical Quality Improvement Program (NSQIP),[26] general surgery residency programs were eligible to enroll and participate in the FIRST Trial if they were affiliated with 1 or more hospitals enrolled in the ACS NSQIP. Hospitals must have been enrolled in ACS NSQIP as of January 1, 2014 (ie, this ensures that all were in the program and abstracting data for at least 6 months prior to the study start date). Programs were ineligible for the FIRST Trial if they met any of the following study exclusion criteria: the program was located in a state where resident duty hours are regulated by state law (eg, New York); the program was in poor standing with ACGME for work-hour violations (determined by the Surgery Residency Review Committee) or was a new program without a full resident complement; and/or the program's only ACS NSQIP affiliates were children's hospitals, Veterans Administration hospitals, and/or other military hospitals (**Figure 1**).

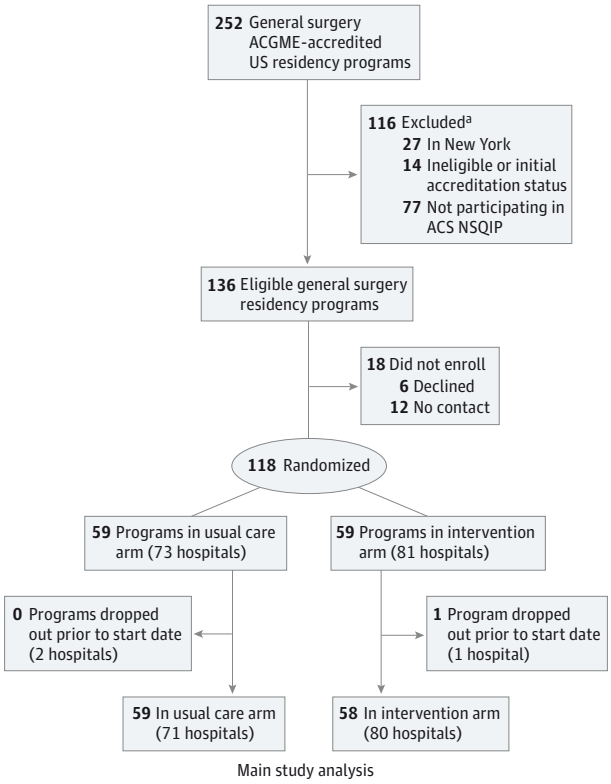
Recruitment

Programs that had at least 1 affiliated hospital that participates in the ACS NSQIP were recruited to participate in the FIRST Trial (eAppendix 1 in the Supplement). Programs enrolling in the FIRST Trial voluntarily consented to 8 terms and requirements to be included (eTable in the Supplement).

Randomization

Residency programs were randomly assigned to study arms using a stratified cluster-randomization strategy. The ACS NSQIP data from 2012 and 2013 for enrolled hospitals were used to calculate hospital-level observed rates of 30-day postoperative death/serious morbidity among general surgery patients. For programs with more than 1 ACS NSQIP hospital enrolled in the FIRST Trial, we calculated the program-level rate as the mean of hospital-level rates. Programs were stratified into tertiles according to program-level observed 30-day postoperative death/serious morbidity. Within each stratum, programs were randomly assigned to study arms. The decision to adopt a stratified approach to randomization was predicated on preliminary analyses in which we found that stratifying programs into tertiles based on observed 30-day postoperative death and morbidity could reduce between-hospital variance by more than 40% and reduce between-program variance by approximately 100%, thus improving the statistical efficiency of our study.

Figure 1. FIRST Trial Enrollment Diagram



ACGME indicates Accreditation Council for Graduate Medical Education; ACS, American College of Surgeons; FIRST, Flexibility in Duty Hour Requirements for Surgical Trainees; and NSQIP, National Surgical Quality Improvement Program.

[a] The categories are not mutually exclusive; thus, the number of exclusions for each category do not sum to 116.

Study Sample

A total of 118 general surgery residency programs and 154 hospitals were enrolled in the FIRST Trial and randomized (Figure 1). Fifty-nine programs were randomized to usual care along with 73 hospitals and 59 programs and 81 hospitals were randomized to the intervention. After randomization, 2 hospitals affiliated with programs randomized to usual care withdrew. One program and its hospital affiliate withdrew from the intervention arm because they were located in New York and unsuccessfully petitioned to join the FIRST Trial. Thus, our final sample consisted of 117 general surgery residency programs and a total of 151 hospitals, with 59 programs and 71 hospitals in usual care and 58 programs and 80 hospitals in the intervention arm (Figure 1).

Study Arms

Residency programs and their participating ACS NSQIP hospital affiliates were randomized to 1 of 2 arms: usual care or intervention (Figure 1). A letter and email were sent to each hospital informing them of their assigned study arm.

Usual Care Arm

Residency programs randomized to the usual care arm adhere to all current ACGME resident duty hour regulations as mandated since 2011. There are no changes at all for these programs and hospitals.

Intervention Arm

Residency programs randomized to the intervention arm were subject to all ACGME duty hour regulations for the 2014-2015 academic year with the exception of certain clauses pertaining to maximum work period length, minimum time off between scheduled duty periods, and maximum frequency of in-house on-call hours. The Table details the specific ACGME resident duty hour regulations for programs in the usual care and intervention arms of the FIRST Trial. Three important regulations were unchanged: the 80-hour weekly maximum (averaged over 4 weeks), 1 day off every 7 days (averaged over 4 weeks), and being on call no more than every third night. The guiding principles for these changes were to maximize continuity of patient care, reduce handoffs, and improve resident education by increasing the ability of residents to care for patients they admitted, consulted on, or operated on. Hospitals in the intervention arm were given a 2-year waiver by the ACGME to allow flexibility in duty hour structure (redlined common program requirements are available at http://www.TheFirstTrial.org).

Outcomes and Measures

Patient Safety Outcomes (Hypothesis 1)

Our primary patient safety outcome for testing is the composite measure of death or serious morbidity within 30 days of surgery. This is a composite measure reported to ACS NSQIP hospitals, publicly reported on the Centers for Medicare and Medicaid Services' Hospital Compare website, and endorsed by the National Quality Forum (0697). Secondary patient safety outcomes include 30-day postoperative mortality, serious morbidity, any morbidity, failure to rescue, surgical site infection, pneumonia, renal failure, urinary tract infection, reoperation, sepsis, prolonged length of stay, and readmissions.

These outcomes will be measured according to standard ACS NSQIP procedures. The data collection and waiver for the intervention arm hospitals will continue for 2 years (July 1, 2014, to June 30, 2016) (**Figure 2**). This will allow for future analyses assessing the impact of the intervention in a second year after implementation. It is unclear whether the second year will facilitate further adoption of the intervention arm policies or whether there will be reversion to standard duty hour policies at intervention arm hospitals.

Resident Education Outcomes (Hypothesis 2A)

Our primary resident education outcome is the level of overall satisfaction with their education reported by residents on the 2015 American Board of Surgery in Training Examination (ABSITE) Resident Survey. Secondary outcomes include resident perceptions of patient safety, availability of supervision, work scheduling, patient interaction, clinical responsibility/accountability, and continuity of care. This outcome will be assessed 6 months into the trial period as this is the scheduled timing of the ABSITE, and the survey will be administered at the end of the ABSITE (Figure 2).

Resident Well-Being Outcomes (Hypothesis 2B)

Our primary resident well-being outcome is residents' self-reported overall satisfaction with well-being in the 6 months preceding the ABSITE, as assessed on the ABSITE Resident Survey Addendum. Secondary end points include self-reported levels of rest, personal safety, fatigue, and work/life balance. This outcome will be assessed 6 months into the trial period as this is the scheduled timing of the ABSITE survey (Figure 2).

Data Collection

Patient-Level Data

Detailed validated data on patient characteristics, comorbidities, surgical procedures, and outcomes will be obtained through the ACS NSQIP data platform (**Figure 3**; eAppendix 2 in the Supplement).

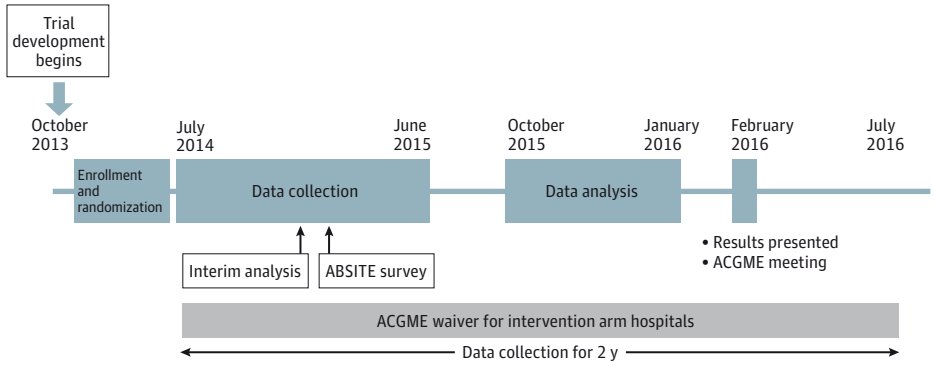
Resident-Level Data

As part of collaboration with the American Board of Surgery (ABS), the ABS agreed to attach a Resident Survey Addendum to the January 2015 ABSITE. The Resident Survey Addendum was a brief close-ended survey designed to collect information on resident perceptions of the quality of their education, workload, quality of life/well-being, and patient safety. Although this survey addendum was designed by FIRST Trial investigators, the survey was administered as part of the ABSITE using existing test administration protocols and infrastructure to all residents in the country irrespective of whether the resident is at a FIRST Trial–participating program. The full survey will be available at http://www.TheFirstTrial.org.

The survey was developed based on prior surveys of resident duty hour policies by a multidisciplinary group.[5,27-32] These existing surveys were adapted to meet the needs of the FIRST Trial: (1) assess resident satisfaction with their educational experience, (2) assess resident well-being, and (3) assess whether residents were adherent to the duty hour policies at their institution. The survey differed slightly for postgraduate year 1 residents vs postgraduate years 2 through 5 residents. Research residents were excluded from having to take the survey. The survey was piloted and iteratively revised based on feedback from postgraduate years 1 through 5 residents at programs in the usual care and intervention arms. The piloting was done in 3 rounds. First, the survey was piloted through semistructured interviews with residents using a "think-aloud" approach. Then, the survey was given to a larger number of residents at multiple institutions (both intervention and usual care), and feedback was solicited to assess the residents' ability to understand the questions easily. Finally, additional residents completed the survey to ensure that it could be completed in a timely fashion.

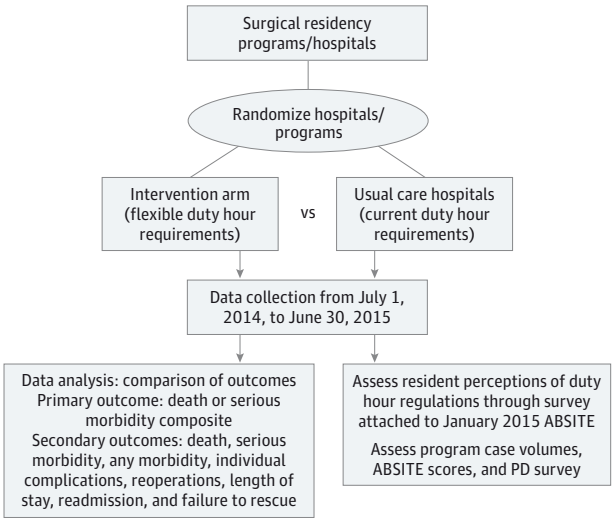
The ABSITE data collection and data processing were solely under the purview of the ABS. The ABS will give the study team access to a deidentified data set containing coded private information and responses to items in the Resident Survey Addendum. This

Figure 2. FIRST Trial Timeline



ABSITE indicates American Board of Surgery in Training Examination; ACGME, Accreditation Council for Graduate Medical Education.

Figure 3. Data Collection Plan



ABSITE indicates American Board of Surgery in Training Examination; PD, program director.

deidentified data set will include no direct personal identifiers (eg, resident names). The data set will include analytic identifiers for individual respondents and an analytic institutional identifier to permit appropriate statistical methods for analyzing the data to account for clustering of data.

Adherence Measures

Data used to measure adherence to the study arm will be taken from the Program Director Survey. The brief close-ended survey allows the program director of each general surgery residency program to designate which ACGME resident duty hour requirements were changed, if any, during the FIRST Trial waiver period (Table). Additionally, the survey asks the reason for which changes were made and the program director's satisfaction level with these changes. The survey was created from prior surveys on resident duty hours for program directors,[3,33] piloted, and iteratively revised in the manner described for the ABSITE Resident Survey. The IRB at Northwestern University deemed the Program Director's Survey to be nonhuman participants research.

Statistical Significance

The overall level of statistical significance for the study will be set at $P < .05$. This level of statistical significance will be adjusted in analyses and reporting to account for higher overall study type I error rates caused by (1) group sequential analysis due to analyses undertaken at interim and at study completion and (2) multiple comparisons in final analyses. We will use the method of Lan and DeMets[34] to set the significance levels for the interim analyses and final analyses to account for the fact that we will be analyzing the data twice. Given that analyses will be conducted at midpoint and at completion with 1-sided hypothesis testing ($\alpha = .05$), overall power set at $\beta = 80\%$; and using an O'Brien-Fleming-type α-spending function, we calculate that the significance-level boundaries for 1-sided tests to be an interim $P < .006$ for interim analyses and final $P < .04$ for final analyses. This approach sets a high bar for early stopping and keeps the final significance-level boundaries closer to that which was intended.

Statistical Methods

Intent-to-Treat Analyses (Primary Analyses)

Our primary analyses to evaluate all hypotheses will follow an intent-to-treat (ITT) approach, meaning that all available data from all study participants from both study arms will be used, regardless of the level of adherence to duty hour prescriptions in study arms. An ITT approach is consistent with real-world implementation conditions in which there is program-level heterogeneity in adherence and adoption of duty hour policies and is consistent with the pragmatic nature of this trial. For evaluation of a policy change such as this one in the FIRST Trial, an ITT approach is relevant given that the question is whether the patient outcomes and resident outcomes are no worse when programs are allowed the flexibility to change their duty hour policies. Sensitivity analyses will include per-protocol analyses in which we control for program-level adherence to prescribed duty hour regulations.

Evaluation of Hypothesis 1

To evaluate the effect of study arm assignment on postoperative outcomes, we will estimate 3-level logistic regression models (patients nested within hospitals nested within residency programs) with random hospital-level and program-level intercepts that regress postoperative outcomes on a single covariate indicating the study arm assignment of the hospital in which a patient

underwent surgery and fixed effects indicating program strata. Given successful randomization, the coefficient on study arm assignment represents the ITT estimate of the effect of being hospitalized in an environment with flexible duty hours on patient outcomes.

Models will also be estimated with adjustment for hospital and patient characteristics, and sensitivity analyses will explore alternative specifications. Noninferiority of intervention (vs usual care) will be assessed by comparing the exponentiated coefficient for Intervention and its 92% CI against the noninferiority margin expressed as an odds ratio.

We will follow the same modeling strategy described here to evaluate secondary patient outcomes, including 30-day postoperative mortality, serious morbidity, any morbidity, failure to rescue, surgical site infection, pneumonia, renal failure, urinary tract infection, reoperation, sepsis, readmissions, and length of stay.

Planned subgroup analyses will be conducted to investigate whether there were any differential effects of intervention for high-risk patients, inpatients only, and emergent/urgent cases. Subgroup analyses will be executed by including an interaction term between study arm assignment and grouping variables and will be carried out for primary end points only.[35-37]

Evaluation of Hypotheses 2A and 2B

To evaluate the effect of assignment to flexible duty hours on resident satisfaction with educational quality (hypothesis 2A) and well-being (hypothesis 2B), we will estimate hierarchical-ordered logistic regression models with program-level random intercepts that regress each outcome on a single covariate indicating the study arm assignment of the program a resident was enrolled in and fixed effects indicating program strata. Hierarchical-ordered logistic regression will be used because resident outcomes were measured using a 5-point Likert scale. Given successful randomization, the exponentiated coefficients on study arm assignment in these analyses represent the unadjusted ITT estimate of the multiplicative effect of assignment to intervention (vs usual care [reference category]) on the odds of reporting a higher level of satisfaction. We will explore alternative model specifications to assess the robustness of our estimates, including nonhierarchical-ordered logistic regression, generalized-ordered logistic regression, and multinomial logistic regression. All models will also be reestimated with the inclusion of resident and program characteristics (resident sex, postgraduate year, program type [academic, community, or military], and geographic region).

This modeling strategy will be applied to evaluate secondary resident outcomes, including residents' satisfaction with specific aspects of their educational and clinical experience, as well as their perception of the effect of their institutional duty hour policies on dimensions of their educational and clinical experience and quality of life.

Planned subgroup analyses will explore whether there was any differential effect of intervention by resident sex, postgraduate level, geographic region, or program type. As before, subgroup analyses will be carried out by including an interaction term between study arm assignment and grouping variables. Subgroup analyses will only be conducted on the 2 primary resident end points in this study (satisfaction with educational quality and satisfaction with own well-being).

Addressing Study Adherence (Additional Analyses)

The ITT analysis uses the full study sample to estimate the effectiveness of study arm assignment on study outcomes (ie, the effect of permitting residency programs to modify ACGME duty hour restrictions). However, it does not take into account whether programs randomized to the intervention actually implemented changes to duty hour policies. Particularly in pragmatic trials such as this one, the difference between effectiveness (ITT) and efficacy may be great, and there may be additional interest in understanding whether there were differential outcomes among study participants that actually implemented duty hour changes. Thus, we will also perform per-protocol, population average treatment effect, and local average treatment effect analyses (eAppendix 3 in the Supplement).

Design Justifications

Cluster Randomization

Cluster randomization was implemented because duty hour policies and schedules are implemented and enforced at the program level (eAppendix 4 in the Supplement).

Noninferiority Design

We chose to design this study as a noninferiority trial where the burden of proof is on demonstrating that modified resident duty hours result in patient safety and resident outcomes that are no worse than current ACGME standards. An equivalence trial would require demonstrating that modified duty hour policies result in patient and resident outcomes that are no better and no worse than existing standards. A superiority trial would require demonstrating that modified duty hour policies result in better patient and resident outcomes compared with existing standards. The rationale underlying current ACGME policies is to mitigate resident fatigue and fatigue-related threats to patient safety but do so at the expense of continuity of care and training that may also threaten the quality of patient care and resident education. By the same token, our intervention promotes continuity of care and training at the expense of potentially longer work hours during a duty period. Given that results of extant studies on the effects of ACGME duty hour reform remain somewhat equivocal as to whether patient and resident outcomes are better postreform, we could not justify framing our hypothesis as one of superiority. Our choice to adopt a noninferiority frame over an equivalence frame was guided by considerations of sample size efficiency given the fixed number of clusters available for recruitment and enrollment. Nonetheless, we will explore these alternate evaluation approaches post hoc.

Sample Size/Power

Using pilot data on general surgery discharges from the ACS NSQIP 2012 and 2013, we estimated baseline rates of 30-day postoperative death or serious complications at 9.94% in 2012. Examining the empirical distribution of hospital-level observed/raw rates of 30-day postoperative death or serious morbidity, we defined our noninferiority margin to be an absolute difference of 1.25 percentage points. Level-2 and Level-3 variances were estimated from variance components analyses models using pilot data that included stratification of residency programs by tertiles of the previous year's observed rates of death or serious morbidity, as well as a program-

level measure of the average rate of death or serious morbidity across affiliated hospitals within a residency program. Level-2 (hospital) and Level-3 (program) variances were estimated to be 0.062 and 2.44e-12, respectively.

Given these assumptions and using methods developed by Borenstein and Hedges,[38] we estimated that this study would have at least 80% power to detect a 1.25-percentage-point absolute difference in rates of death or serious morbidity across study arms with at least 90 residency programs (45 programs per arm), an average of 1.1 hospitals per program, and an average of 950 patients per hospital. This is a conservative calculation as the number of programs and cases is larger; thus, we should be able to detect a smaller difference in the primary outcome measure. However, the 1.25-percentage-point inferiority margin was the smallest margin consistent with the investigators' clinical opinion and from the literature that could be detected with adequate power (80%) given baseline rates and potential cluster and sample sizes.

Study Organization and Institutional Assurances

Program directors at each participating residency program will complete a survey as a way to ascertain adherence to study arm assignment. A data safety monitoring board (DSMB) will independently review the study progress to ensure data rigor and patient safety. Full details on study organization and assurances are available in eAppendix 5 in the Supplement.

Discussion

The FIRST Trial is a 1-year, 2-arm, cluster-randomized, pragmatic noninferiority trial that is designed to compare patient and resident outcomes in residency programs and their affiliated hospitals randomized to modified duty hour policies that permit greater flexibility in work-hour scheduling as opposed to a group of programs and hospitals randomized to usual care (ie, existing policies).

The FIRST Trial was able to enroll 117 of 135 (87%) eligible programs, and if we exclude the military hospitals that did not receive approval from the Department of Defense in sufficient time to enroll, we would have a 95% enrollment rate. There were 4 programs that refused to enroll in the FIRST Trial because their residents were not supportive or the program director did not feel that duty hour requirements needed to be adjusted. Nonetheless, the participation rate clearly demonstrates the importance of the issue to surgical programs in the United States and allows for generalizable results on which to base policy decisions.

Importantly, the FIRST trial was designed to examine outcomes beyond simply postoperative outcomes. While ensuring the safety of an alternate duty hour policy is critical, it must be combined with the effects on resident well-being and education. When allowing flexibility in duty hour requirements, it is critical to understand whether residents felt they were better able to care for patients, maintain their well-being, and obtain a better surgical education.

A common concern around the FIRST Trial is whether the hospitals in the intervention arm adhered to the flexible duty hours permitted and allowed their residents to leverage all of the duty hour requirements that were eliminated. From a policy standpoint, this is irrelevant because the objective is to ensure that allowing the flexibility of the intervention arm is no worse than usual care. Whether programs elect to enact all of the changes or some depends on their local contextual factors and underlies the rationale for a pragmatic trial where it is not feasible or reasonable to exactly specify and ensure adherence to all of the changes in duty hour policies. Nonetheless, we will query program directors about what policy changes were made and will perform per-protocol analyses to address this question, but these are entirely separate analyses and should not be the basis for policy decisions surrounding this issue.

Potential Contributions of the Study

To our knowledge, the FIRST Trial is the first prospectively designed, national randomized trial to be conducted for the purposes of evaluating the effects of resident duty hour policies on patient and resident outcomes. A strength of our study design is the reliance on existing validated data collection infrastructure for obtaining data on patient and resident outcomes (NSQIP and ABSITE). The results of our study may be informative to policymakers and other stakeholders engaged in restructuring graduate medical training to enhance the quality of patient care and resident education. The analogous trial in internal medicine, the iCOMPARE Trial, began on July 1, 2015, and will use a similar study design as the FIRST Trial.

Limitations

Several potential limitations should be acknowledged. First, hospital participants had to be enrolled in NSQIP to be eligible. Approximately 70% of the 252 general surgery residency programs accredited by the ACGME were affiliated with at least 1 hospital that was enrolled in the NSQIP. Because of the nontrivial participation costs in the NSQIP, NSQIP hospitals may differ systematically from non-NSQIP hospitals. These participating hospitals may have more resources and ability to adapt to changes in duty hours given their resources. We have planned analyses to assess the generalizability of our findings by assessing differences between participating and nonparticipating sites.[39]

Second, our study concerns the evaluation of resident outcomes using a survey that is administered 6 months into the trial rather than at the end of the trial. It is unclear whether resident perceptions of training and well-being assessed at 6 months would serve as a reliable measure of such outcomes at 12 months. We may be able to repeat the survey at the 18-month mark of the study alongside the ABSITE in January 2016. Similarly, a third limitation of the FIRST Trial may be its limited duration (1 year). Ideally, we would follow a cohort of general surgery residents from intern through chief year and compare whether there were any differences across study arms in resident perception of education and the acquisition of professional skills. However, this would take upwards of 7 years and would likely not be policy relevant at the conclusion. Conducting the trial over 1 year allows for some potentially actionable data in a shorter time frame. Moreover, we are continuing the data analysis for a second year, and we will be able to determine whether any changes occur over time (ie, increased adoption of the flexibility or reversion to pre–FIRST Trial standard ACGME duty hour requirements).

Fourth, it was impossible to blind participants in this study. Because participants were aware of their study arm assignment, and possibly aware of the nature of the study, we cannot rule out research bias, such as the Hawthorne Effect, as a possible explanation for our findings. Such bias cannot be addressed by instrumen-

tal variables analysis of local average treatment effect. However, blinding a randomized trial of duty hour policy (or any policy) is impossible.

Finally, as a pragmatic trial, this study is not designed or intended to test for causal relations between specific work hour reforms or combinations of reforms and resident fatigue and patient outcomes. Rather, the design is meant to assess the results of a policy change where programs are permitted to make duty hours more flexible.

Conclusions

Study results concerning patient outcomes and resident outcomes will both need to be considered to truly understand the impact of allowing flexibility in duty hour requirements. Many challenges remain in determining how to best train our future surgeons, but the FIRST Trial should provide some guidance on how to structure future duty hour policies for surgical trainees.

ARTICLE INFORMATION

Accepted for Publication: October 26, 2015.

Published Online: December 30, 2015. doi:10.1001/jamasurg.2015.4990.

Author Affiliations: Surgical Outcomes and Quality Improvement Center (SOQIC), Department of Surgery, Feinberg School of Medicine, Northwestern University, Chicago, Illinois (Bilimoria, Chung, Dahlke, Love, Mahvi); Center for Healthcare Studies in the Institute for Public Health and Medicine, Feinberg School of Medicine, Northwestern University, Chicago, Illinois (Bilimoria); American College of Surgeons, Chicago, Illinois (Bilimoria, Cohen, Ko, Hoyt); Department of Biostatics, Northwestern University, Evanston, Illinois (Hedges); Department of Surgery, Vanderbilt University, Nashville, Tennessee (Tarpley); Department of Surgery, Southern Illinois University, Springfield (Mellinger); Center for Surgery and Health Economics, Department of Surgery, Perelman School of Medicine, University of Pennsylvania, Philadelphia (Kelz); American Board of Surgery, Philadelphia, Pennsylvania (Lewis).

Author Contributions: Drs Bilimoria and Chung had full access to all of the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis.
Study concept and design: Bilimoria, Chung, Hedges, Love, Mellinger, Mahvi, Kelz, Ko, Hoyt, Lewis.
Acquisition, analysis, or interpretation of data: Bilimoria, Chung, Dahlke, Cohen, Tarpley, Mellinger.
Drafting of the manuscript: Bilimoria, Chung, Hedges, Dahlke.
Critical revision of the manuscript for important intellectual content: Bilimoria, Chung, Love, Cohen, Tarpley, Mellinger, Mahvi, Kelz, Ko, Hoyt, Lewis.
Statistical analysis: Bilimoria, Chung, Hedges, Cohen.
Obtained funding: Bilimoria, Mahvi, Lewis.
Administrative, technical, or material support: Dahlke, Love, Mellinger, Lewis.
Study supervision: Bilimoria, Ko, Hoyt, Lewis.

Conflict of Interest Disclosures: Dr Bilimoria reported support from the American Board of Surgery, American College of Surgeons, Accreditation Council for Graduate Medical Education, and Health Care Services Corp. Dr Bilimoria has received honoraria from hospitals and professional societies for clinical care and quality improvement research presentations. No other disclosures were reported.

Funding/Support: The FIRST Trial is funded by the American Board of Surgery, the American College of Surgeons, and the Accreditation Council for Graduate Medical Education. Dr Bilimoria has received support from the National Institutes of Health, Agency for Healthcare Research and Quality, National Comprehensive Cancer Network, American Cancer Society, Northwestern University, the Robert H. Lurie Comprehensive Cancer Center, and Northwestern Memorial Hospital.

Role of the Funder/Sponsor: The American Board of Surgery and the American College of Surgeons had a role in the design and conduct of the study; collection, management, and interpretation of the data; preparation, review, and approval of the manuscript; and decision to submit the manuscript for publication, as leaders from these organizations are collaborators and coauthors. The Accreditation Council for Graduate Medical Education had a role in the design of the study, as it approved the waiver requirements for the intervention arm hospitals.

Additional Contributions: We acknowledge those who have contributed to the administration and execution of the FIRST Trial: Anthony Yang, MD, Ravi Rajaram, MD, MS, Emily S. Pavey, MA, Sean Perry, JD, and Anne Grace, JD (Northwestern University); Judy Shea, PhD (University of Pennsylvania); Ajit Sachdeva, MD, Sameera Ali, MPH, Emma Malloy, BA, Lynn Zhou, PhD, and Amy Hart Sachs, BS (American College of Surgeons); James Hebert, MD (University of Vermont); Paul Gauger, MD (University of Michigan); Christine V. Kinnier, MD, MS (Massachusetts General Hospital); Joseph Cofer, MD (University of Tennessee); Thomas Biester, PhD, and Andrew Jones, PhD (American Board of Surgery); John Potts, MD (Accreditation Council for Graduate Medical Education); all of the 152 American College of Surgeons National Surgical Quality Improvement Program Surgeon Champions and Surgical Clinical Reviewers; and all of the program directors and program coordinators at the 117 participating general surgery residency programs. They did not receive compensation for their contributions.

REFERENCES

1. Okie S. An elusive balance: residents' work hours and the continuity of care. *N Engl J Med*. 2007;356(26):2665-2667.

2. Van Eaton EG, Horvath KD, Pellegrini CA. Professionalism and the shift mentality: how to reconcile patient ownership with limited work hours. *Arch Surg*. 2005;140(3):230-235.

3. Antiel RM, Thompson SM, Hafferty FW, et al. Duty hour recommendations and implications for meeting the ACGME core competencies: views of residency directors. *Mayo Clin Proc*. 2011;86(3):185-191.

4. Kort KC, Pavone LA, Jensen E, Haque E, Newman N, Kittur D. Resident perceptions of the impact of work-hour restrictions on health care delivery and surgical education: time for transformational change. *Surgery*. 2004;136(4):861-871.

5. Landrigan CP, Fahrenkopf AM, Lewin D, et al. Effects of the accreditation council for graduate medical education duty hour limits on sleep, work hours, and safety. *Pediatrics*. 2008;122(2):250-258.

6. Landrigan CP, Rothschild JM, Cronin JW, et al. Effect of reducing interns' work hours on serious medical errors in intensive care units. *N Engl J Med*. 2004;351(18):1838-1848.

7. Weinger MB, Ancoli-Israel S. Sleep deprivation and clinical performance. *JAMA*. 2002;287(8):955-957.

8. Lockley SW, Cronin JW, Evans EE, et al; Harvard Work Hours, Health and Safety Group. Effect of reducing interns' weekly work hours on sleep and attentional failures. *N Engl J Med*. 2004;351(18):1829-1837.

9. Barger LK, Ayas NT, Cade BE, et al. Impact of extended-duration shifts on medical errors, adverse events, and attentional failures. *PLoS Med*. 2006;3(12):e487.

10. Eastridge BJ, Hamilton EC, O'Keefe GE, et al. Effect of sleep deprivation on the performance of simulated laparoscopic surgical skill. *Am J Surg*. 2003;186(2):169-174.

11. Fischer JE. Continuity of care: a casualty of the 80-hour work week. *Acad Med*. 2004;79(5):381-383.

12. Fletcher KE, Saint S, Mangrulkar RS. Balancing continuity of care with residents' limited work hours: defining the implications. *Acad Med*. 2005;80(1):39-43.

13. Antiel RM, Van Arendonk KJ, Reed DA, et al. Surgical training, duty-hour restrictions, and implications for meeting the Accreditation Council for Graduate Medical Education core competencies: views of surgical interns compared with program directors. *Arch Surg*. 2012;147(6):536-541.

14. Accreditation Council for Graduate Medical Education (ACGME). ACGME duty hours. https://www.acgme.org/acgmeweb/tabid/271/GraduateMedicalEducation/DutyHours.aspx. Accessed September 1, 2015.

15. Public Citizen. Petition requesting medical residents work hour limits. http://www.citizen.org/Page.aspx?pid=614. Published 2001. Accessed September 1, 2015.

16. Institute of Medicine. *Resident Duty Hours: Enhancing Sleep, Supervision, and Safety*. Washington, DC: The National Academies Press; 2008.

17. Bilimoria KY, Hoyt DB, Lewis F. Making the case for investigating flexibility in duty hour limits for surgical residents. *JAMA Surg*. 2015;150(6):503-504.

18. Patient and Physician Safety and Protection Act of 2005. HR 1228, 109th Leg. https://www.govtrack.us/congress/bills/109/hr1228. Published 2005. Accessed September 1, 2015.

19. Patient and Physician Safety and Protection Act of 2001. HR 3236, 107th Leg. https://www.govtrack.us/congress/bills/107/hr3236. Published 2001. Accessed September 1, 2015.

20. Patient and Physician Safety and Protection Act of 2003. S 952, 108th Leg. https://www.govtrack.us/congress/bills/108/s952. Published 2003. Accessed September 1, 2015.

21. Nasca TJ, Day SH, Amis ES Jr; ACGME Duty Hour Task Force. The new recommendations on duty hours from the ACGME Task Force. *N Engl J Med*. 2010;363(2):e3.

22. SEIU Healthcare Committee of Interns and Residents. http://www.cirseiu.org/. Accessed September 1, 2015.

23. European Commission. Working conditions: working time directive. http://ec.europa.eu/social/main.jsp?catId=706&langId=en&intPageId=205. Accessed September 1, 2015.

24. Ahmed N, Devitt KS, Keshet I, et al. A systematic review of the effects of resident duty hour restrictions in surgery: impact on resident wellness, training, and patient outcomes. *Ann Surg*. 2014;259(6):1041-1053.

25. Philibert I, Nasca T, Brigham T, Shapiro J. Duty-hour limits and patient care and resident outcomes: can high-quality studies offer insight into complex relationships? *Annu Rev Med*. 2013;64:467-483.

26. American College of Surgeons. American College of Surgeons National Surgical Quality Improvement Program. https://www.facs.org/quality-programs/acs-nsqip. Accessed September 1, 2015.

27. Landrigan CP, Barger LK, Cade BE, Ayas NT, Czeisler CA. Interns' compliance with accreditation council for graduate medical education work-hour limits. *JAMA*. 2006;296(9):1063-1070.

28. Barden CB, Specht MC, McCarter MD, Daly JM, Fahey TJ III. Effects of limited work hours on surgical training. *J Am Coll Surg*. 2002;195(4):531-538.

29. Stamp T, Termuhlen P, Miller S, et al. Before and after resident work hour limitations: an objective assessment of the well-being of surgical residents. *Curr Surg*. 2005;62(1):117-121.

30. Vidyarthi AR, Auerbach AD, Wachter RM, Katz PP. The impact of duty hours on resident self reports of errors. *J Gen Intern Med*. 2007;22(2):205-209.

31. Drolet BC, Sangisetty S, Tracy TF, Cioffi WG. Surgical residents' perceptions of 2011 Accreditation Council for Graduate Medical Education duty hour regulations. *JAMA Surg*. 2013;148(5):427-433.

32. Drolet BC, Christopher DA, Fischer SA. Residents' response to duty-hour regulations: a follow-up national survey. *N Engl J Med*. 2012;366(24):e35.

33. Antiel RM, Thompson SM, Reed DA, et al. ACGME duty-hour recommendations: a national survey of residency program directors. *N Engl J Med*. 2010;363(8):e12.

34. Lan KKG, Demets DL. Discrete sequential boundaries for clinical trials. *Biometrika*. 1983;70(3):659-663. doi:10.2307/2336502.

35. Assmann SF, Pocock SJ, Enos LE, Kasten LE. Subgroup analysis and other (mis)uses of baseline data in clinical trials. *Lancet*. 2000;355(9209):1064-1069.

36. Rothwell PM. Treating individuals 2: subgroup analysis in randomised controlled trials: importance, indications, and interpretation. *Lancet*. 2005;365(9454):176-186.

37. Wang R, Lagakos SW, Ware JH, Hunter DJ, Drazen JM. Statistics in medicine: reporting of subgroup analyses in clinical trials. *N Engl J Med*. 2007;357(21):2189-2194.

38. Borenstein M, Hedges L. *CRT Power Manual*. Englewood, NJ: Biostat Inc; 2012.

39. Hedges L. Scale-up in education: generalizability of treatment effects: psychometrics and education. In: McDonald BSS-K, ed. *Ideas in Principle*. Vol 1. Lanham, MD: Rowman and Littlefield Publishers; 2006:55-78.