The NEW ENGLAND JOURNAL of MEDICINE

ESTABLISHED IN 1812 FEBRUARY 25, 2016 VOL. 374 NO. 8

National Cluster-Randomized Trial of Duty-Hour Flexibility in Surgical Training

Karl Y. Bilimoria, M.D., M.S.C.I., Jeanette W. Chung, Ph.D., Larry V. Hedges, Ph.D., Allison R. Dahlke, M.P.H., Remi Love, B.S., Mark E. Cohen, Ph.D., David B. Hoyt, M.D., Anthony D. Yang, M.D., John L. Tarpley, M.D., John D. Mellinger, M.D., David M. Mahvi, M.D., Rachel R. Kelz, M.D., M.S.C.E., Clifford Y. Ko, M.D., M.S.H.S., David D. Odell, M.D., M.M.Sc., Jonah J. Stulberg, M.D., Ph.D., M.P.H., and Frank R. Lewis, M.D.

ABSTRACT

BACKGROUND

Concerns persist regarding the effect of current surgical resident duty-hour policies on patient outcomes, resident education, and resident well-being.

METHODS

We conducted a national, cluster-randomized, pragmatic, noninferiority trial involving 117 general surgery residency programs in the United States (2014–2015 academic year). Programs were randomly assigned to current Accreditation Council for Graduate Medical Education (ACGME) duty-hour policies (standard-policy group) or more flexible policies that waived rules on maximum shift lengths and time off between shifts (flexible-policy group). Outcomes included the 30-day rate of postoperative death or serious complications (primary outcome), other postoperative complications, and resident perceptions and satisfaction regarding their well-being, education, and patient care.

RESULTS

In an analysis of data from 138,691 patients, flexible, less-restrictive duty-hour policies were not associated with an increased rate of death or serious complications (9.1% in the flexible-policy group and 9.0% in the standard-policy group, P=0.92; unadjusted odds ratio for the flexible-policy group, 0.96; 92% confidence interval, 0.87 to 1.06; P=0.44; noninferiority criteria satisfied) or of any secondary postoperative outcomes studied. Among 4330 residents, those in programs assigned to flexible policies did not report significantly greater dissatisfaction with overall education quality (11.0% in the flexible-policy group and 10.7% in the standard-policy group, P=0.86) or well-being (14.9% and 12.0%, respectively; P=0.10). Residents under flexible policies were less likely than those under standard policies to perceive negative effects of duty-hour policies on multiple aspects of patient safety, continuity of care, professionalism, and resident education but were more likely to perceive negative effects on personal activities. There were no significant differences between study groups in resident-reported perception of the effect of fatigue on personal or patient safety. Residents in the flexible-policy group were less likely than those in the standard-policy group to report leaving during an operation (7.0% vs. 13.2%, P<0.001) or handing off active patient issues (32.0% vs. 46.3%, P<0.001).

CONCLUSIONS

As compared with standard duty-hour policies, flexible, less-restrictive duty-hour policies for surgical residents were associated with noninferior patient outcomes and no significant difference in residents' satisfaction with overall well-being and education quality. (FIRST ClinicalTrials.gov number, NCT02050789.)

From the Surgical Outcomes and Quality Improvement Center (SOQIC), Department of Surgery and Center for Healthcare Studies, Feinberg School of Medicine and Northwestern Medicine, Northwestern University (K.Y.B., J.W.C., A.R.D., R.L., A.D.Y., D.M.M., D.D.O., J.J.S.), and the American College of Surgeons (K.Y.B., M.E.C., D.B.H., C.Y.K.), Chicago, the Department of Statistics, Northwestern University, Evanston (L.V.H.), and the Department of Surgery, Southern Illinois University, Springfield (J.D.M.) — all in Illinois; the Department of Surgery, Vanderbilt University, Nashville (J.L.T.); the Department of Surgery and the Center for Surgery and Health Economics, Perelman School of Medicine, University of Pennsylvania (R.R.K.), and the American Board of Surgery (F.R.L.) — both in Philadelphia; and the Department of Surgery, University of California, Los Angeles, School of Medicine, Los Angeles (C.Y.K.). Address reprint requests to Dr. Bilimoria at the Surgical Outcomes and Quality Improvement Center (SOQIC), Department of Surgery, Feinberg School of Medicine and Northwestern Medicine, Northwestern University, 633 N. St. Clair St., 20th Fl., Chicago, IL 60611, or at k-bilimoria@northwestern.edu.

This article was published on February 2, 2016, at NEJM.org.

N Engl J Med 2016;374:713-27.
DOI: 10.1056/NEJMoa1515724

In response to concerns about patient safety and resident well-being, the Accreditation Council for Graduate Medical Education (ACGME) introduced national regulations in 2003 that limited resident duty periods to 80 hours per week, capped overnight shift lengths, and mandated minimum time off between shifts.[1,2] Concerns persisted,[3] and in 2011, the ACGME implemented further restrictions to shorten maximum shift lengths for interns and increase time off after overnight on-call duty for residents.[1,4,5]

A Quick Take is available at NEJM.org

Although most observers agree that some duty-hour regulation was necessary, critics cite a weak evidence base for the 2003 and 2011 reforms.[3,6,7] Several retrospective studies and systematic reviews have questioned whether duty-hour reforms achieved their intended goals of improved patient outcomes, resident education, and resident well-being.[6-18] In surgical settings, most studies have shown no difference or a worsening in patient postoperative outcomes and resident education after duty-hour reforms.[3,6,7,13-15,18-21] However, many studies have suggested that duty-hour reforms resulted in improved well-being and less fatigue among surgical residents.[6]

Although the ACGME reforms were intended to prevent fatigue-related errors in clinical care delivered by residents,[3,5] the restrictions may reduce continuity of care and increase the frequency of handoffs,[3,22-24] which could jeopardize patient safety by forcing residents to leave at critical times and could undermine the goals of surgical training if residents are unable to follow patients through critical aspects of their care.[20,25-29] Evidence from large, prospective, randomized trials to inform duty-hour regulations is currently lacking.[3,6,7] In a widely cited report on resident duty hours, the Institute of Medicine called for additional high-level research to inform policy.[3]

We conducted the Flexibility in Duty Hour Requirements for Surgical Trainees (FIRST) Trial[30-32] to test whether surgical-patient outcomes under flexible, less-restrictive duty-hour policies would be no worse than outcomes under standard ACGME policies. Resident satisfaction and perceptions of patient care, resident education, and resident well-being were also assessed.

METHODS

STUDY DESIGN AND OVERSIGHT

This study was a prospective, cluster-randomized, pragmatic, noninferiority trial comparing standard ACGME duty-hour policies with flexible duty-hour policies.[32] The study was conducted from July 1, 2014, to June 30, 2015.

The initial trial protocol was reviewed by the Northwestern University institutional review board office, which determined the trial to be non–human-subjects research (see the Supplementary Appendix, available with the full text of this article at NEJM.org).[31,32] The authors vouch for the accuracy and completeness of the data and data analyses and for the fidelity of the study and this report to the protocol (available at NEJM.org).

Members of the American Board of Surgery (ABS) and the American College of Surgeons (ACS) staff 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 the decision to submit the manuscript for publication, because the leaders of these organizations are coauthors and collaborators. The ACGME had a role only in the design of the study, insofar as it approved the waiver requirements for the hospitals in the flexible-policy group. The boards of these organizations had no role in the study design and conduct, data analysis, manuscript preparation or review, or the decision to submit the manuscript for publication.

PARTICIPANTS

The study population comprised all 252 ACGME-accredited general surgery residency programs in the United States in 2014 and, by extension, residents in those programs, hospitals with which they were affiliated, and general surgery patients at those hospitals (Fig. S1 in the Supplementary Appendix). Because the ACS National Surgical Quality Improvement Program (ACS NSQIP)[33] was the intended platform for patient data collection, program eligibility required affiliation with at least one hospital in ACS NSQIP (77 programs were therefore excluded).[31,32] Programs located in New York were excluded because resident duty hours there are regulated by state law (27 programs were excluded).[5] Programs were also excluded if they were new or in poor standing with the ACGME (12 programs were excluded).

RANDOMIZATION

A total of 118 general surgery residency programs (87% of the 136 eligible programs) and 154 affiliated hospitals were enrolled in the

Table 1. Duty-Hour Requirements and Adherence Rates According to Study Group.*

Requirement Category	Standard-Policy Group		Flexible-Policy Group	
	Standard ACGME Policies	Adherent Programs†	Policies‡	Adherent Programs†
		no. (%)		*no. (%)*
Maximum shift length	PGY 1 (interns): Duty periods may not exceed 16 hr	59 (100)	PGY 1 (interns): Duty periods can exceed 16 hr	58 (100)
	PGY 2–5 (residents): Duty periods may not exceed 28 hr (24 hr plus 4 hr for transition)	59 (100)	PGY 2–5 (residents): Duty periods can exceed 28 hr (24 hr plus 4 hr for transition)	49 (84)
Minimum time off between shifts	Residents must have ≥8 hr off between shifts but should have 10 hr off between shifts	59 (100)	Residents are not required to have ≥8–10 hr off between shifts	47 (81)
	Residents must have ≥14 hr off after 24 hr of continuous duty	57 (97)	Residents are not required to have ≥14 hr off after 24 hr of continuous duty	51 (88)
Maximum work hr/wk	Residents must not work >80 hr/wk, averaged over 4 wk§	—	Residents must not work >80 hr/wk, averaged over 4 wk§	—
Mandatory time free of duty	Residents must have 1 in every 7 days off from all educational and clinical duties, averaged over 4 wk§	—	Residents must have 1 in every 7 days off from all educational and clinical duties, averaged over 4 wk§	—
Frequency of on-call duty	Residents must not be on call more frequently than every third night§	—	Residents must not be on call more frequently than every third night§	—

* ACGME denotes Accreditation Council for Graduate Medical Education, and PGY postgraduate year.
† Program adherence was defined by residency program directors regarding which policies were followed at their institution during the trial period (100% response rate).
‡ Residency programs assigned to the flexible-policy group were allowed to waive four ACGME duty-hour requirements concerning maximum shift length and minimum time off between shifts.
§ These ACGME duty-hour requirements remained the same in both study groups.

FIRST Trial. Programs were stratified into three strata on the basis of the rates in 2012 and 2013 of a composite measure of death or serious complications.[31,32,34-38] Programs and their hospital affiliates were then randomly assigned as clusters within strata to one of two study groups.[39] Programs assigned to the standard-policy group were to continue adhering to existing ACGME duty-hour policies (Table 1). Programs assigned to the flexible-policy (intervention) group were required to adhere to ACGME duty-hour requirements of limiting work to 80 hours per week, 1 day off in 7 days, and on-call duty no more frequently than every third night, but they were granted a waiver by the ACGME to waive four duty-hour requirements (from the 2003 and 2011 reforms) concerning maximum shift length and minimum time off between shifts (to facilitate continuity of care) (Table 1).[40,41] Residents were not specifically kept from knowing their study-group assignment.

DATA COLLECTION

Patient-level data on patient characteristics, coexisting conditions, operative details, and surgical outcomes were obtained for general surgery cases from ACS NSQIP, a validated system developed in the 1990s for collection of high-quality clinical data to measure surgical outcomes; the system has been described extensively elsewhere.[33,35,42] Data on patients 18 years of age or older are collected in ACS NSQIP for most surgical specialties, excluding trauma and transplantation surgery, by trained, certified, and audited data abstractors at each site.[42] The abstractors ascertain patient outcomes by examining the medical record, discussing with treating physicians, and contacting patients directly when needed. The ACS NSQIP data abstractors were not specifically informed of the study-group assignments.

Data on resident outcomes were collected in collaboration with the ABS, which administered a close-ended (i.e., multiple-choice) resident survey

at the end of the January 2015 ABS In-Training Examination (ABSITE)[43] to all surgical resident examinees in the United States (Table S7 in the Supplementary Appendix). The ABSITE is a computer-based multiple-choice examination given annually in January to assess resident knowledge and management of surgical problems. Survey items were adapted from previously published surveys, pretested with residents through cognitive interviews, and iteratively revised.[32]

MEASURES

Our primary patient outcome was based on the ACS NSQIP composite outcome measure of the 30-day rate of postoperative death or serious complications, which is based on a National Quality Forum–endorsed metric (NQF#0697).[36,37] Serious complications include stroke, myocardial infarction, cardiac arrest with cardiopulmonary resuscitation, pulmonary embolism, ventilation for more than 48 hours, acute renal failure, bleeding requiring transfusion of more than 4 units, sepsis or septic shock, organ-space surgical-site infection, or wound dehiscence. Secondary outcomes included the following 10 other ACS NSQIP outcome measures: 30-day rate of postoperative death, serious complications, any complication, failure to rescue (i.e., death in a patient who had a serious complication), pneumonia, renal failure, unplanned reoperation, sepsis, surgical-site infection, and urinary tract infection[34,38] (Table S6 in the Supplementary Appendix).

Primary resident outcome measures were specified before trial initiation and included resident-reported level of satisfaction (very dissatisfied, dissatisfied, neutral, satisfied, or very satisfied) with overall quality of resident education and overall well-being. Secondary resident outcomes included residents' perceptions and satisfaction regarding the effect of 2014–2015 institutional duty-hour policies on aspects of patient care, residency training, and personal well-being; how often fatigue affected personal safety and patient safety; and how often in the past month residents had breaks in continuity of care and education because of duty-hour policies (Table S7 in the Supplementary Appendix).

STATISTICAL ANALYSIS

Using general surgery data from hospitals in 2012 before the trial began, we calculated a baseline raw rate of death or serious complications of 9.94%. Our rationale for the noninferiority design has been described previously.[32] A noninferiority margin was specified before trial initiation as an absolute difference of 1.25 percentage points (13% relative difference, which corresponds to a noninferiority margin odds ratio of 1.15) on the basis of examination of the empirical distribution of hospital-level 30-day rates of death or serious complications, intracluster correlations, and power calculations.[31,32] Using a noninferiority margin of an absolute difference of 1.25 percentage points in 30-day rates of postoperative death or serious complications, we estimated that minimum sample sizes of 90 programs (45 per group) with an average of 1.1 hospitals per program and an average of 950 patients per hospital would be necessary to obtain 80% power at an alpha level of 0.05 (see the Supplementary Appendix).

Data analyses were performed at Northwestern University. Because one midpoint interim analysis was performed for data and safety monitoring purposes, the level of statistical significance for our final analyses of only patient outcomes was adjusted to 0.04 in order to maintain an overall significance level for the entire trial of 0.05.[31,32,44] In the context of a hypothesis of no difference in outcomes across study groups, correction for multiple comparisons was not a conservative approach for reducing the false discovery rate; thus, we report non–Bonferroni-corrected P values for all estimates. Bonferroni adjustment of P values for patient outcomes entails lowering the value from 0.04 to 0.004 (adjustment for 11 tests), whereas adjustment of P values for resident outcomes entails lowering the value from 0.05 to 0.0015 (adjustment for 34 tests).

We assessed how well randomization balanced observable characteristics of programs, hospitals, patients, and residents between the flexible-policy and standard-policy groups by comparing differences in means and frequencies using Student's t-tests and chi-square tests with cluster-corrected P values. Program characteristics were obtained from the ABS, and hospital-level characteristics were obtained from the American Hospital Association (AHA) annual survey.

Using an intention-to-treat approach, we modeled the association between patient outcomes and study-group assignment using three-level hierarchical logistic-regression models with program-level and hospital-level random intercepts and controls for program-level strata of 2013 rates of postoperative death or serious

complications (i.e., performance in the previous year was used as the stratifying variable in randomization).[31,32] These analyses are referred to in the results as "unadjusted" and were the primary prespecified analyses. Given a noninferiority design with a 0.04 alpha level, 92% confidence intervals [$100 \times (1-2\alpha)$] were used on the basis of a "two one-sided tests" (TOST) approach.[45,46] A significant odds ratio of less than 1.00 favored flexible policies over standard policies. Noninferiority was assessed by comparison of the odds ratio and 92% confidence interval with the noninferiority margin expressed as an odds ratio. An outcome was deemed to be noninferior if the point estimate and upper boundary of the 92% confidence interval were less than the prespecified noninferiority margin odds ratio of 1.15. Analyses were also performed that adjusted for any residual differences in patient demographic characteristics, coexisting conditions, and procedural case mix[35,38] (referred to in the results as "adjusted" analyses). For secondary patient outcomes, the noninferiority margin was defined in a manner analogous to that for the primary outcome as a 13% relative difference in rates (Table S8 in the Supplementary Appendix).

Numerous additional prespecified analyses were conducted to examine the sensitivity of our results with respect to minor variations in modeling or estimation approaches (e.g., conditional and population-averaged estimates). We performed prespecified subgroup analyses of primary patient outcomes to test for significant interactions between study-group assignment and subgroups defined according to type of surgery (emergency vs. elective), risk of death or serious complications (highest quartile vs. lower three quartiles of patients), and surgical setting (inpatient vs. outpatient).[34]

The association between resident outcomes and study-group assignment was modeled with the use of two-level hierarchical logistic regression with program-level random intercepts and controls for program-level strata of 2013 rates of postoperative death or serious complications (i.e., the stratifying variable in randomization).[31,32] A noninferiority margin for assessing resident outcomes was not specified; thus, we used two-tailed tests and standard 95% confidence intervals.

Prespecified sensitivity analyses were performed to examine the robustness of our results with respect to alternative modeling approaches for resident outcomes (e.g., hierarchical ordered and multinomial logistic-regression models and conditional and population-averaged estimates) and the inclusion of additional program-level covariates. Prespecified subgroup analyses tested for significant interactions between study-group assignment and subgroups defined according to resident sex, postgraduate year, program geographic region, and program type (academic, community, or military).

Because implementation and enforcement of study-group conditions were at the discretion of program directors (i.e., flexible-policy programs were not required to eliminate all four policies waived by the ACGME), a separate survey of residency program directors in the FIRST Trial was conducted in June and July 2015 to collect data on program-level adherence to study-group conditions (i.e., policy changes enacted). Standard-policy programs were defined as adherent if their duty-hour policies had zero departures from the four ACGME duty-hour requirements regarding minimum time off between shifts and maximum shift length (Table 1).[31,32] Flexible-policy programs were defined as adherent if they instituted at least one of these four allowed policy changes. Three types of analyses were undertaken to explore the influence of adherence on our main results: a per-protocol analysis (limited to adherent programs), an as-treated analysis (which assessed actual exposure to policy change), and analysis of local average treatment effects with the use of instrumental variables, with study-group assignment serving as an instrumental variable for actual exposure to policy change (see the Supplementary Appendix).[31,32] No data were collected regarding on-call schedules, duty-hour logs, sleep, midlevel providers, handoff protocols, or adherence to policies that remained unchanged across the two study groups (e.g., 80-hour workweek).

Analyses were conducted with the use of Stata statistical software, release 13 (StataCorp).[47] Details of our methods have been described previously[30-32] and can also be found in the Supplementary Appendix and study protocol.

RESULTS

STUDY SAMPLE

Participating programs had more residents per year, a lower proportion of international medical graduates, and higher board-examination scores than nonparticipating programs (Tables S31 and

Table 2. Characteristics of the Residency Programs, Hospitals, Patients, and Residents According to Study Group.*

Characteristic	Total	Standard-Policy Group	Flexible-Policy Group	P Value
Residency programs				
No. of programs	117	59	58	
Program type — no. (%)				
Academic	70 (60)	37 (63)	33 (57)	0.81†
Community	45 (38)	21 (36)	24 (41)	
Military	2 (2)	1 (2)	1 (2)	
Geographic region — no. (%)				
Northeast	34 (29)	14 (24)	20 (34)	0.07†
Southeast	26 (22)	16 (27)	10 (17)	
Midwest	33 (28)	18 (31)	15 (26)	
Southwest	11 (9)	8 (14)	3 (5)	
West	13 (11)	3 (5)	10 (17)	
No. of chief residents per program‡	4.7±2.2	4.8±2.2	4.6±2.2	0.59§
Proportion of international medical graduates	0.17±0.17	0.16±0.17	0.19±0.17	0.34§
First-attempt pass rate on qualifying board examination, 2009–2013	88.6±8.0	89.0±8.5	88.4±7.5	0.65
First-attempt pass rate on certifying board examination, 2009–2014	83.9±10.7	84.9±10.4	83.0±11.0	0.27
Hospitals				
No. of hospitals¶	148	70	78	
Total bed capacity	578±287	598±290	560±285	0.42‖
Total surgical volume	23,387±16,089	23,239±15,480	23,519±16,716	0.92‖
Nurse-to-bed ratio	2.56±0.90	2.54±0.82	2.58±0.97	0.73‖
Resident-to-bed ratio	0.39±0.26	0.40±0.26	0.38±0.27	0.71‖
CMS case-mix index**	1.90±0.24	1.87±0.27	1.93±0.21	0.15‖
30-Day rate of postoperative death or serious complications in previous year, 2013	8.95±3.36	9.16±3.76	8.76±2.95	0.47‖
Patients				
No. of patients	138,691	65,849	72,842	
Age — yr	54.3±16.4	53.9±16.4	54.7±16.4	0.23‖
Nonwhite race — no. (%)††	30,848 (22.2)	13,784 (20.9)	17,064 (23.4)	0.92‡‡
ASA classification score — no. (%)§§				
1	10,233 (7.4)	4,866 (7.4)	5,367 (7.4)	0.19¶¶
2	61,491 (44.3)	29,262 (44.4)	32,229 (44.2)	
3	59,958 (43.2)	28,399 (43.1)	31,559 (43.3)	
4 or 5	7,009 (5.1)	3,322 (5.0)	3,687 (5.1)	
Emergency surgery — no. (%)	15,433 (11.1)	7,706 (11.7)	7,727 (10.6)	0.80‡‡
Inpatient surgery — no. (%)	82,698 (59.6)	39,451 (59.9)	43,247 (59.4)	0.83§
Diabetes requiring medication — no. (%)	20,743 (15.0)	10,067 (15.3)	10,676 (14.7)	0.26‡‡
BMI classification — no. (%)‖‖				
Normal weight	35,187 (25.4)	16,327 (24.8)	18,860 (25.9)	<0.001¶¶
Underweight	2,754 (2.0)	1,259 (1.9)	1,495 (2.1)	
Overweight	40,990 (29.6)	19,221 (29.2)	21,769 (29.9)	
Class I obesity	27,483 (19.8)	13,052 (19.8)	14,431 (19.8)	
Class II obesity	14,822 (10.7)	7,162 (10.9)	7,660 (10.5)	
Class III obesity	17,455 (12.6)	8,828 (13.4)	8,627 (11.8)	

Table 2. (Continued.)

Characteristic	Total	Standard-Policy Group	Flexible-Policy Group	P Value
COPD — no. (%)	5,318 (3.8)	2,579 (3.9)	2,739 (3.8)	0.52‡‡
Renal failure — no. (%)	632 (0.5)	305 (0.5)	327 (0.4)	0.93‡‡
Functional status of partially or totally dependent — no. (%)	2,648 (1.9)	1,276 (1.9)	1,372 (1.9)	0.40‡‡
Preoperative SIRS, sepsis, or septic shock — no. (%)	10,983 (7.9)	5,188 (7.9)	5,795 (8.0)	0.92‡‡
Residents				
No. of residents	4330	2220	2110	
Sex — no. (%)				
Female	1,737 (40.1)	866 (39.0)	871 (41.3)	0.23†
Male	2,593 (59.9)	1,354 (61.0)	1,239 (58.7)	
Postgraduate year — no. (%)				
1	1,156 (26.7)	616 (27.7)	540 (25.6)	0.57†
2	1,081 (25.0)	554 (25.0)	527 (25.0)	
3	872 (20.1)	438 (19.7)	434 (20.6)	
4	628 (14.5)	313 (14.1)	315 (14.9)	
5	593 (13.7)	299 (13.5)	294 (13.9)	
Resident type — no. (%)				
Categorical	3,699 (85.4)	1,874 (84.4)	1,825 (86.5)	0.73†
Preliminary	621 (14.3)	340 (15.3)	281 (13.3)	
Other	10 (0.2)	6 (0.3)	4 (0.2)	

* Plus–minus values are means ±SD. COPD denotes chronic obstructive pulmonary disease, and SIRS the systemic inflammatory response syndrome.
† The P value was calculated with the use of a two-tailed chi-square test.
‡ Chief residents are fifth-year residents who are eligible to take the American Board of Surgery qualifying examination (written boards).
§ The P value was calculated with the use of Student's t-test.
¶ Two hospitals were not included in the final patient-level analysis owing to data-availability issues. Two pairs of hospitals reported hospital data to the American Hospital Association jointly, and thus each pair was treated as a single hospital-level unit in these analyses.
‖ The P value was calculated with the use of hierarchical linear regression with program intercepts.
** The Centers for Medicare and Medicaid Services (CMS) case-mix index represents the average diagnosis-related group (DRG) relative weight for that hospital, with higher values indicating that the hospital provides care for sicker patients.
†† Race was determined on the basis of clinical records by American College of Surgeons National Surgical Quality Improvement Program abstractors at each site.
‡‡ The P value was calculated with the use of hierarchical logistic regression with program intercepts.
§§ An American Society of Anesthesiologists (ASA) classification score of 1 indicates a normal healthy patient, 2 a patient with mild systemic disease, 3 a patient with severe systemic disease, 4 a patient with severe systemic disease that is a constant threat to life, and 5 a moribund patient who is not expected to survive without the operation.
¶¶ The P value was calculated with the use of hierarchical multinomial logistic regression with program intercepts.
‖‖ The body-mass index (BMI) is the weight in kilograms divided by the square of the height in meters. BMI classifications are as follows: underweight, less than 18.5; normal weight, 18.5 to 24.9; overweight, 25.0 to 29.9; class I obesity, 30.0 to 34.9; class II obesity, 35.0 to 39.9; and class III obesity, 40.0 or more.

S32 in the Supplementary Appendix). Our study sample included 117 ACGME-accredited general surgery residency programs and 151 affiliated hospitals (Fig. S1 in the Supplementary Appendix), because 1 program and 3 hospitals dropped out after randomization but before the trial start date. Of these, 59 programs and their affiliated 71 hospitals were assigned to the standard-policy group, and 58 programs and their 80 affiliated hospitals were assigned to the flexible-policy group. The study groups were well balanced with respect to a broad range of program, hospital, patient, and resident characteristics (Table 2).

PATIENT OUTCOMES

Owing to issues with respect to the availability of final data, 2 hospitals were dropped from the final analysis, resulting in the loss of 2 programs

in our sample for patient outcomes only (Fig. S1 in the Supplementary Appendix). Two pairs of hospitals, each pair from the same residency program, reported data under the same AHA identification number, so each pair was treated as a single unit for analysis. Thus, patient outcome analyses included 115 programs (58 in the standard-policy group and 57 in the flexible-policy group) and 148 hospitals (70 in the standard-policy group and 78 in the flexible-policy group), which contributed data on 138,691 general surgery patients (65,849 in the standard-policy group and 72,842 in the flexible-policy group).

The rate of death or serious complications did not differ significantly between study groups (9.1% in the flexible-policy group and 9.0% in the standard-policy group, P=0.92). Figure 1 presents both unadjusted and adjusted (for patient characteristics) odds ratios comparing the association between study-group assignment and patient outcomes. The risk of death or serious complications did not differ significantly between patients who underwent surgery in hospitals affiliated with programs assigned to flexible, less-restrictive duty-hour policies and those who underwent surgery in standard-policy hospitals (unadjusted odds ratio for the flexible-policy group, 0.96; 92% confidence interval [CI], 0.87 to 1.06; P=0.44; adjusted odds ratio, 0.96; 92% CI, 0.90 to 1.04; P=0.38) (Fig. 1). The upper boundaries of the 92% confidence interval from both unadjusted and adjusted models were greater than 1.00 but less than the noninferiority margin odds ratio of 1.15; thus, flexible policies were deemed to be noninferior to standard policies with respect to death or serious complications.

With respect to secondary outcomes, flexible policies were noninferior to standard policies with respect to serious complications, any complication, unplanned reoperation, sepsis, surgical-site infection, and urinary tract infection in unadjusted and adjusted models (Fig. 1). The results were inconclusive for 30-day mortality in the unadjusted analysis, but the noninferiority criterion was met in the adjusted analysis. There was no significant difference between study groups with respect to failure to rescue and renal failure, but the upper boundary of the 92% confidence interval exceeded the margin; therefore, noninferiority was not established for these outcomes. The upper boundaries of the 92% confidence intervals from unadjusted and adjusted analyses of the 30-day rate of postoperative pneumonia coincided exactly with the noninferiority margin.

There were no significant subgroup effects for death or serious complications according to type of surgery (emergency vs. elective), risk of death or serious complications (highest quartile vs. lower three quartiles of patients), or surgical setting (inpatient vs. outpatient) (Table S13 in the Supplementary Appendix). All results were robust with respect to variations in modeling specifications and the inclusion of additional covariates for patients, hospitals, or both. The results were qualitatively similar for conditional and population-averaged estimates (Table S34 in the Supplementary Appendix).

RESIDENT OUTCOMES

ABSITE survey data were obtained for a total of 4330 general surgery residents who were undergoing training in 117 FIRST Trial programs (2110 residents in the flexible-policy group and 2220 in the standard-policy group). Response rates varied across survey items, ranging from 84 to 87% for the outcomes examined (Tables S25 through S30 in the Supplementary Appendix).

With respect to the two primary resident outcomes, residents in flexible-policy programs were not significantly more likely than those in standard-policy programs to be dissatisfied (very dissatisfied or dissatisfied vs. neutral, satisfied, or very satisfied) with overall education quality (11.0% in the flexible-policy group and 10.7% in the standard-policy group, P=0.86; odds ratio for the flexible-policy group, 1.08; 95% CI, 0.77 to 1.52; P=0.64) or overall well-being (14.9% and 12.0%, respectively; P=0.10; odds ratio, 1.31; 95% CI, 0.99 to 1.74; P=0.06) (Table 3).

Flexible-policy residents were significantly less likely than standard-policy residents to be dissatisfied with continuity of care (odds ratio, 0.44; 95% CI, 0.32 to 0.60; P<0.001) and with the quality and ease of handoffs and transitions in care (odds ratio, 0.69; 95% CI, 0.52 to 0.92; P=0.01) but were more likely to be dissatisfied with time for rest (odds ratio, 1.41; 95% CI, 1.06 to 1.89; P=0.02) (Table 3). There was no significant difference between study groups regarding resident satisfaction with the duty-hour regulations of their program (odds ratio, 0.99; 95% CI, 0.71 to 1.40; P=0.97).

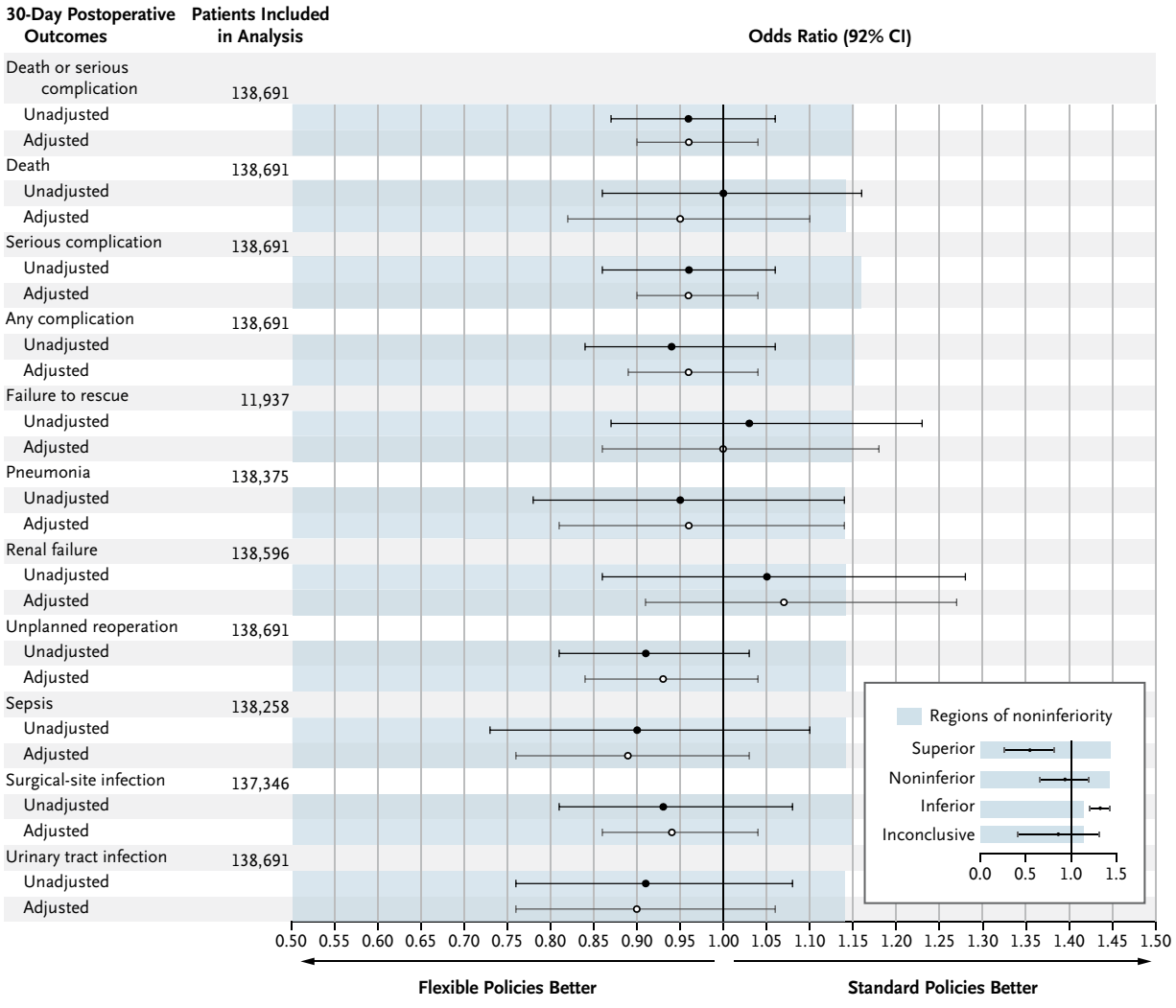

Figure 1. Comparison of Postoperative Outcomes between Flexible, Less-Restrictive Duty-Hour Policies and Standard Policies.

In all regressions, 115 programs and 148 hospitals were included. Solid black circles indicate the unadjusted effect of assignment to the flexible-policy group (vs. the standard-policy group). Open circles indicate the adjusted effect of assignment to the flexible-policy group (vs. the standard-policy group), expressed as an odds ratio from similar models that also adjusted for patient characteristics. Estimates reported are conditional estimates (not population-averaged effects) that were obtained from three-level hierarchical mixed-effects logistic-regression models. In these models, outcomes were regressed on assignment to the flexible-policy group (vs. the standard-policy group) with controls for program-level strata of 30-day rates of postoperative death or serious complications in 2013 (variable used in randomization) with program- and hospital-level random intercepts. To account for interim analysis, the alpha level was adjusted to 0.04 for the final analysis (alpha level for the overall trial, 0.05). Given a noninferiority design with a 0.04 alpha level, 92% confidence intervals [$100\times(1-2\alpha)$] were used on the basis of a "two one-sided tests" (TOST) approach. Thus, error bars indicate 92% confidence intervals, and shaded blue regions represent the area of noninferiority for each outcome. Flexible policies were considered to be noninferior to standard policies if the estimated odds ratio (circle) and upper boundary of the 92% confidence interval are contained within the shaded region; inferior to standard policies if the estimated odds ratio and lower boundary of the 92% confidence interval are both to the right, outside the shaded region for an outcome; and superior to standard policies if the estimated odds ratio and upper boundary of the 92% confidence interval are both within the shaded region and below 1.00 (see inset). If the estimated odds ratio is within the shaded region but the upper boundary of the 92% confidence interval extends outside the region, the results were considered to be inconclusive. The number of patients per outcome differs because patients were excluded from the analysis if the condition was preexisting at the time of surgery. The number of patients is reduced for failure to rescue (i.e., death in a patient who had a serious complication), because only patients who had a serious complication were included in the analysis.

Table 3. Resident-Reported Satisfaction and Perceptions of Well-Being, Education, and Patient Safety.*

Outcome	Standard-Policy Group	Flexible-Policy Group	P Value†	Odds Ratio for Flexible-Policy Group (95% CI)‡	P Value
	no./total no. (%)				
Primary outcomes					
Dissatisfaction with overall quality of resident education§	200/1874 (10.7)	194/1768 (11.0)	0.86	1.08 (0.77–1.52)	0.64
Dissatisfaction with overall well-being§	226/1876 (12.0)	263/1769 (14.9)	0.10	1.31 (0.99–1.74)	0.06
Secondary outcomes					
Dissatisfaction§					
With patient safety	77/1875 (4.1)	62/1770 (3.5)	0.48	0.85 (0.55–1.31)	0.46
With continuity of care	188/1876 (10.0)	83/1769 (4.7)	<0.001	0.44 (0.32–0.60)	<0.001
With quality and ease of handoffs and transitions in care	190/1873 (10.1)	124/1766 (7.0)	0.009	0.69 (0.52–0.92)	0.01
With duty-hour regulations of the program	161/1876 (8.6)	144/1768 (8.1)	0.74	0.99 (0.71–1.40)	0.97
With work hours and scheduling	236/1874 (12.6)	214/1767 (12.1)	0.76	0.95 (0.71–1.27)	0.72
With time for rest	280/1875 (14.9)	329/1768 (18.6)	0.08	1.41 (1.06–1.89)	0.02
Perception of negative effect of institutional duty hours¶					
On patient safety	491/1891 (26.0)	223/1782 (12.5)	<0.001	0.40 (0.32–0.51)	<0.001
On continuity of care	1053/1892 (55.7)	339/1786 (19.0)	<0.001	0.16 (0.12–0.21)	<0.001
On clinical-skills acquisition	688/1888 (36.4)	232/1777 (13.1)	<0.001	0.24 (0.19–0.31)	<0.001
On operative-skills acquisition	928/1885 (49.2)	337/1781 (18.9)	<0.001	0.22 (0.17–0.27)	<0.001
On resident autonomy	663/1888 (35.1)	232/1782 (13.0)	<0.001	0.26 (0.20–0.34)	<0.001
On operative volume	915/1887 (48.5)	330/1778 (18.6)	<0.001	0.22 (0.17–0.28)	<0.001
On availability for urgent cases	845/1890 (44.7)	266/1783 (14.9)	<0.001	0.20 (0.16–0.25)	<0.001
On availability for elective cases	651/1889 (34.5)	264/1781 (14.8)	<0.001	0.30 (0.24–0.39)	<0.001
On attendance at educational conferences	431/1886 (22.9)	218/1780 (12.2)	<0.001	0.47 (0.36–0.62)	<0.001
On relationship between interns and residents	488/1892 (25.8)	199/1782 (11.2)	<0.001	0.38 (0.29–0.49)	<0.001
On time for teaching medical students	523/1888 (27.7)	262/1781 (14.7)	<0.001	0.45 (0.37–0.56)	<0.001
On case preparation away from hospital	176/1887 (9.3)	427/1781 (24.0)	<0.001	3.37 (2.54–4.47)	<0.001
On participation in research	172/1888 (9.1)	373/1780 (21.0)	<0.001	2.81 (2.12–3.73)	<0.001
On professionalism	240/1891 (12.7)	148/1780 (8.3)	0.002	0.65 (0.49–0.87)	0.003
On job satisfaction	262/1888 (13.9)	226/1782 (12.7)	0.43	0.94 (0.73–1.23)	0.67
On satisfaction with career choice	172/1887 (9.1)	164/1777 (9.2)	0.92	1.03 (0.79–1.33)	0.84
On morale	301/1892 (15.9)	294/1782 (16.5)	0.73	1.09 (0.85–1.40)	0.51

On time with family and friends	168/1888 (8.9)	441/1779 (24.8)	<0.001	3.66 (2.70–4.97)	<0.001
On time for extracurricular activities	172/1886 (9.1)	458/1779 (25.7)	<0.001	3.81 (2.84–5.11)	<0.001
On rest	178/1887 (9.4)	470/1781 (26.4)	<0.001	3.85 (2.88–5.15)	<0.001
On health	128/1883 (6.8)	326/1778 (18.3)	<0.001	3.22 (2.37–4.36)	<0.001
Fatigue always or often affects personal safety‖	175/1878 (9.3)	188/1774 (10.6)	0.26	1.15 (0.91–1.47)	0.25
Fatigue always or often affects patient safety‖	118/1878 (6.3)	133/1774 (7.5)	0.17	1.18 (0.91–1.53)	0.21
Occurrence during past month owing to duty-hour regulations**					
Left during an operation	256/1944 (13.2)	128/1821 (7.0)	<0.001	0.46 (0.32–0.65)	<0.001
Missed an operation	817/1944 (42.0)	544/1821 (29.9)	<0.001	0.56 (0.45–0.69)	<0.001
Handed off an active patient issue	901/1944 (46.3)	583/1821 (32.0)	<0.001	0.53 (0.45–0.63)	<0.001

* Denominators represent the number of respondents per survey item in the trial sample of residents. Response rates varied across survey items, ranging from 84 to 87%. When the Bonferroni correction was applied to the 34 resident outcomes assessed, the level of significance was adjusted from 0.05 to 0.0015, and the differences between the study groups were no longer significant for three outcomes: time for rest, quality and ease of handoffs and transitions in care, and professionalism.

† Cluster-corrected P values were calculated by means of a chi-square test of association between study-group assignment and dichotomized resident outcome.

‡ Odds ratios and 95% confidence intervals (CI) and two-tailed P values were calculated by means of two-level hierarchical logistic regression with program-level random intercepts. Models assessed the association between outcomes and study-group assignment, with adjustment for program-level strata based on 30-day rates of postoperative death or serious complications in 2013 (stratifying variable for randomization). Significant odds ratios of less than 1.00 favor flexible policies over standard policies. Significant odds ratios of more than 1.00 favor standard policies over flexible policies.

§ The numerator represents the number of residents who reported being "very dissatisfied" or "dissatisfied" versus "neutral," "satisfied," or "very satisfied."

¶ The numerator represents the number of residents who perceived a "negative effect" of 2014–2015 institutional duty hours versus "no effect" or a "positive effect."

‖ The numerator represents the number of residents who reported that fatigue "always" or "often" affects personal safety or patient safety versus "sometimes," "rarely," or "never."

** The numerator represents the number of residents who reported one or more occurrences in the past month versus no occurrence.

Flexible-policy residents were significantly less likely than standard-policy residents to perceive a negative effect (vs. a positive effect or no effect) of institutional duty-hour policies on patient safety, continuity of care, clinical-skills acquisition, operative-skills acquisition, autonomy, operative volume, availability for elective and urgent cases, conference attendance, time for teaching medical students, the relationship between interns and residents, and professionalism (all odds ratios <1.00, P<0.001 for all comparisons except P=0.003 for professionalism) (Table 3). However, flexible-policy residents were more likely to perceive negative effects of duty-hour policies on resident outcomes that depended on time away from the hospital, such as case preparation after work, research participation, time with family and friends, time for extracurricular activities, rest, and health (all odds ratios >1.00, P<0.001 for all comparisons). Nonetheless, there were no significant differences between study groups regarding the perceived effects of duty hours on job satisfaction, satisfaction with career choice, or morale (Table 3). Study group was also not associated with resident-reported frequency at which fatigue affected either patient safety or personal safety (Table 3).

In analyses of breaks in continuity of care, flexible-policy residents were significantly less likely than standard-policy residents to leave during an operation (7.0% vs. 13.2%, P<0.001; odds ratio, 0.46; 95% CI, 0.32 to 0.65; P<0.001), miss an operation (29.9% vs. 42.0%, P<0.001; odds ratio, 0.56; 95% CI, 0.45 to 0.69; P<0.001), or hand off an active patient care issue (32.0% vs. 46.3%, P<0.001; odds ratio, 0.53; 95% CI, 0.45 to 0.63; P<0.001) at least once in the past month (Table 3).

When correction for multiple comparisons was applied, the differences in three resident outcomes were no longer significant between the standard-policy and flexible-policy groups: resident satisfaction with time for rest, perception of the quality and ease of handoffs and transitions in care, and perception of professionalism (P>0.0015 with correction for multiple comparisons, for all comparisons). There were no significant differences between the standard-policy and flexible-policy groups in our resident-reported primary outcomes in subgroups defined according to resident sex, program geographic region, or program type (Table S19 in the Sup-

plementary Appendix). There were also no significant differences between the standard-policy and flexible-policy groups in the primary outcomes in the subgroup defined according to postgraduate year (first vs. second and third vs. fourth and fifth) (Table S19 in the Supplementary Appendix). All results were robust with respect to minor variations in modeling specifications. The results were qualitatively similar for conditional and population-averaged estimates (Table S35 in the Supplementary Appendix).

ADHERENCE ANALYSES

Overall program-level adherence to study-group conditions was 98% (97% in the standard-policy group and 100% in the flexible-policy group) (Tables S4 and S5 in the Supplementary Appendix). Thus, results of per-protocol, as-treated, and instrumental-variables analyses were highly consistent with intention-to-treat results for patient and resident outcomes (Tables S14 through S18 and S20 through S24 in the Supplementary Appendix). The number of policies waived at an institution was not associated with death or serious complications, nor were there any significant effects of waiving specific policies on death or serious complications.

DISCUSSION

This national, prospective, randomized trial showed that flexible, less-restrictive duty-hour policies for surgical residents were noninferior to standard ACGME duty-hour policies with respect to our primary patient outcome of the 30-day rate of postoperative death or serious complications. There was also no significant difference between the standard-policy and flexible-policy groups with respect to residents' satisfaction regarding their overall well-being and education.

Our finding of noninferior patient outcomes under flexible, less-restrictive duty-hour policies as compared with standard duty-hour policies for most postoperative outcomes examined is consistent with the results of previous studies.[6,7,13-15,18-21] Moreover, there were no significant differences between the standard-policy and flexible-policy groups in outcomes for subgroups that may be more sensitive to differences in duty-hour policies,[8,48] including high-risk patients, inpatient surgeries, and emergency cases. Thus, these findings suggest that flexible duty-hour policies appear to be safe for patient care.

Previous surveys showed that residents were concerned about the negative effect of duty-hour policies on patient care and resident education; however, most generally did show improvements in residents' quality of life and well-being.[6,16,19] Similarly, we found that residents in programs with flexible duty-hour policies (as compared with current ACGME duty-hour restrictions) noted numerous benefits with respect to nearly all aspects of patient safety, continuity of care, surgical training, and professionalism. However, residents reported that less-restrictive duty-hour policies had a negative effect on time with family and friends, time for extracurricular activities, rest, and health. Importantly, although there was a trend favoring standard policies with respect to outcomes related to perceptions of personal time, residents' satisfaction with overall well-being did not differ significantly between study groups. Flexible-policy residents did not report less satisfaction with their overall resident education, and they did not perceive that fatigue affected their personal safety or patient safety. There was also no significant difference in satisfaction with duty-hour policies between the study groups. These results suggest that residents found that flexible duty-hour policies improved multiple aspects of patient care and resident education without an appreciable difference in their personal safety, but these benefits came with the recognition that the flexible policies affected time for personal activities and certain aspects of well-being.

Patient care and resident education can be compromised by interruptions in continuity of care (i.e., handoffs)[22-24]; thus, another important finding in our study was that residents in the flexible-policy group were about half as likely to leave or miss an operation or hand off an active patient care issue than were those in the standard-policy group. This suggests that the flexible, less-restrictive duty hours had their intended effect of improving continuity of care, as further reflected in the residents' perceptions of benefit with respect to continuity and patient safety in the intervention group.

Several limitations should be acknowledged. First, our study was limited to programs affiliated with ACS NSQIP hospitals, so the findings

may not be generalizable to programs not represented in ACS NSQIP. Second, we focused on general surgery, and although our results may be relevant to other surgical disciplines, they may not be generalizable to nonsurgical specialties. However, given differences in training requirements and previous evidence of differential effects of resident duty hours between surgery and internal medicine,[7] it may be reasonable to have specialty-specific duty-hour requirements. Third, we conducted this study for a full academic year, but we cannot extrapolate the ways in which flexible duty-hour policies might affect the training and experience of an entire cohort of surgical residents over multiple years. In addition, the resident survey was conducted halfway through the trial during standard dates for ABSITE administration. Although this eased data-collection logistics and increased response rates, measures of residents' perceptions and experiences could vary over a longer exposure period before survey administration (i.e., residents' perceptions over time under flexible policies could improve as they become accustomed to the policies or could worsen if the effects are cumulatively strenuous). Similarly, patient outcomes could improve or worsen with more time under flexible duty-hour policies. Fourth, our patient outcomes were limited to those captured in ACS NSQIP, so there may be other outcomes that would be more sensitive to resident duty-hour policies. Although postoperative complications are the ultimate outcomes that must be assessed for any change in surgical duty-hour policies, we were unable to collect data on medication errors and other potentially resident-sensitive outcomes. Given that ACS NSQIP already performs data-quality checks and audits (see the Supplementary Appendix), no additional data-quality checks were performed by the study team. Fifth, although there was no significant difference between the standard-policy and flexible-policy groups in residents' report that fatigue affected personal safety, we did not specifically collect data on needle sticks and car accidents, because these are notoriously challenging outcomes to capture in surveys.

Finally, adherence to assigned study-group policies was evaluated on the basis of a survey of program directors and the policy changes implemented at that program. Although that does not reflect resident-level adherence, the intention-to-treat analysis is the policy-relevant test: programs are given the flexibility to change policies, and outcomes reflect real-world implementation conditions, irrespective of the level of adherence (i.e., whether they change no policies, one policy, or all four policies).

In conclusion, flexible duty-hour policies for surgical residents were noninferior to current ACGME duty-hour policies with respect to patient outcomes. Residents' satisfaction regarding their overall well-being and education quality was similar in the flexible-policy and standard-policy groups.

The results and conclusions in this article are the authors' own and do not represent the views of organizations providing support or otherwise involved.

Results of the FIRST Trial were presented at the 11th Annual Academic Surgical Congress, February 2–4, 2016, in Jacksonville, Florida.

Supported equally by the American Board of Surgery (ABS), the American College of Surgeons (ACS), and the Accreditation Council for Graduate Medical Education (ACGME).

Disclosure forms provided by the authors are available with the full text of this article at NEJM.org.

We thank those who have contributed to the administration and execution of the trial: Jonathan Fryer, M.D., Anne Grace, Ph.D., Julie K. Johnson, Ph.D., Lindsey J. Kreutzer, M.P.H., Shari Meyerson, M.D., Emily S. Pavey, M.A., Sean Perry, J.D., Christopher M. Quinn, M.S., Alfred Rademaker, Ph.D., and Ravi Rajaram, M.D. (Northwestern University); Judy Shea, Ph.D. (University of Pennsylvania); Sameera Ali, M.P.H., Amy Hart, B.S., Emma Malloy, B.A., Brian Matel, B.A., Craig Miller, B.S.E.E., Lynn Modla, M.S., Ajit Sachdeva, M.D., and Lynn Zhou, Ph.D. (ACS); James Hebert, M.D. (University of Vermont); Michael Englesbe, M.D., M.P.H., and Paul Gauger, M.D. (University of Michigan); Christine V. Kinnier, M.D. (Massachusetts General Hospital); Joseph Cofer, M.D. (University of Tennessee, Chattanooga); Mitchell Posner, M.D. (University of Chicago); Eugene Foley, M.D. (University of Wisconsin); Thomas Louis, Ph.D. (Johns Hopkins); Thomas Biester, M.S., and Andrew Jones, Ph.D. (ABS); Rebecca Miller, M.S., Thomas Nasca, M.D., and John Potts, M.D. (ACGME); Margaret M. Class (Defense Health Agency); all the surgeon champions and surgical clinical reviewers at the 151 participating ACS National Surgical Quality Improvement Program hospitals; and all the program directors and program coordinators at the 117 participating general surgery residency programs (see the Supplementary Appendix).

REFERENCES

1. Accreditation Council for Graduate Medical Education. Resident duty hours in the learning and working environment: comparison of 2003 and 2011 standards (https://www.acgme.org/acgmeweb/Portals/0/PDFs/dh-ComparisonTable2003v2011.pdf).
2. Philibert I, Friedmann P, Williams WT. New requirements for resident duty hours. JAMA 2002;288:1112-4.
3. Institute of Medicine. Resident duty hours: enhancing sleep, supervision, and safety. Washington, DC: National Academies Press, 2008.
4. Nasca TJ, Day SH, Amis ES Jr. The

new recommendations on duty hours from the ACGME Task Force. N Engl J Med 2010;363(2):e3.
5. Accreditation Council for Graduate Medical Education. The ACGME 2011 duty hour standards: enhancing quality of care, supervision, and resident professional development (http://www.acgme.org/acgmeweb/Portals/0/PDFs/jgme-monograph[1].pdf).
6. 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;259:1041-53.
7. 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;64:467-83.
8. Rajaram R, Chung JW, Jones AT, et al. Association of the 2011 ACGME resident duty hour reform with general surgery patient outcomes and with resident examination performance. JAMA 2014;312:2374-84.
9. Rajaram R, Chung JW, Cohen ME, et al. Association of the 2011 ACGME resident duty hour reform with postoperative patient outcomes in surgical specialties. J Am Coll Surg 2015;221:748-57.
10. Shetty KD, Bhattacharya J. Changes in hospital mortality associated with residency work-hour regulations. Ann Intern Med 2007;147:73-80.
11. Volpp KG, Rosen AK, Rosenbaum PR, et al. Mortality among patients in VA hospitals in the first 2 years following ACGME resident duty hour reform. JAMA 2007;298:984-92.
12. Volpp KG, Rosen AK, Rosenbaum PR, et al. Mortality among hospitalized Medicare beneficiaries in the first 2 years following ACGME resident duty hour reform. JAMA 2007;298:975-83.
13. Browne JA, Cook C, Olson SA, Bolognesi MP. Resident duty-hour reform associated with increased morbidity following hip fracture. J Bone Joint Surg Am 2009;91:2079-85.
14. Poulose BK, Ray WA, Arbogast PG, et al. Resident work hour limits and patient safety. Ann Surg 2005;241:847-56.
15. Rosen AK, Loveland SA, Romano PS, et al. Effects of resident duty hour reform on surgical and procedural patient safety indicators among hospitalized Veterans Health Administration and Medicare patients. Med Care 2009;47:723-31.
16. 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.
17. Rajaram R, Sadaat L, Chung JW, Dahlke AR, Yang AD, Bilimoria KY. Impact of the 2011 ACGME resident duty hour reform on hospital processes of care and patient experience. BMJ Qual Saf 2015 December 30 (Epub ahead of print).
18. Patel MS, Volpp KG, Small DS, et al. Association of the 2011 ACGME resident duty hour reforms with mortality and readmissions among hospitalized Medicare patients. JAMA 2014;312:2364-73.
19. Antiel RM, Reed DA, Van Arendonk KJ, et al. Effects of duty hour restrictions on core competencies, education, quality of life, and burnout among general surgery interns. JAMA Surg 2013;148:448-55.
20. Mattar SG, Alseidi AA, Jones DB, et al. General surgery residency inadequately prepares trainees for fellowship: results of a survey of fellowship program directors. Ann Surg 2013;258:440-9.
21. Bailit JL, Blanchard MH. The effect of house staff working hours on the quality of obstetric and gynecologic care. Obstet Gynecol 2004;103:613-6.
22. Desai SV, Feldman L, Brown L, et al. Effect of the 2011 vs 2003 duty hour regulation-compliant models on sleep duration, trainee education, and continuity of patient care among internal medicine house staff: a randomized trial. JAMA Intern Med 2013;173:649-55.
23. Drolet BC, Spalluto LB, Fischer SA. Residents' perspectives on ACGME regulation of supervision and duty hours — a national survey. N Engl J Med 2010;363(23):e34.
24. Horwitz LI, Krumholz HM, Green ML, Huot SJ. Transfers of patient care between house staff on internal medicine wards: a national survey. Arch Intern Med 2006;166:1173-7.
25. 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;86:185-91.
26. 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.
27. Barden CB, Specht MC, McCarter MD, Daly JM, Fahey TJ III. Effects of limited work hours on surgical training. J Am Coll Surg 2002;195:531-8.
28. Coverdill JE, Adrales GL, Finlay W, et al. How surgical faculty and residents assess the first year of the Accreditation Council for Graduate Medical Education duty-hour restrictions: results of a multi-institutional study. Am J Surg 2006;191:11-6.
29. Coverdill JE, Carbonell AM, Cogbill TH, et al. Professional values, value conflicts, and assessments of the duty-hour restrictions after six years: a multi-institutional study of surgical faculty and residents. Am J Surg 2011;201:16-23.
30. Flexibility in Duty Hour Requirements for Surgical Trainees Trial: the "FIRST Trial" (http://www.thefirsttrial.org).
31. The Flexibility in Duty Hour Requirements for Surgical Trainees Trial study protocol and statistical analysis plan (http://www.thefirsttrial.org/Documents/FIRSTTRIAL_StatisticalAnalysisPlan_Updated03NOV2015.pdf).
32. Bilimoria KY, Chung JW, Hedges LV, et al. Development of the Flexibility in Duty Hour Requirements for Surgical Trainees (FIRST) trial protocol: a national cluster-randomized trial of resident duty hour policies. JAMA Surg 2015 December 30 (Epub ahead of print).
33. American College of Surgeons National Surgical Quality Improvement Program (ACS NSQIP) (http://site.acsnsqip.org).
34. ACS NSQIP variables and definitions. In: ACS NSQIP operations manual. Chicago: American College of Surgeons. 2013.
35. Cohen ME, Ko CY, Bilimoria KY, et al. Optimizing ACS NSQIP modeling for evaluation of surgical quality and risk: patient risk adjustment, procedure mix adjustment, shrinkage adjustment, and surgical focus. J Am Coll Surg 2013;217(2):336-46.e1.
36. Merkow RP, Hall BL, Cohen ME, et al. Validity and feasibility of the American College of Surgeons colectomy composite outcome quality measure. Ann Surg 2013;257:483-9.
37. National Quality Forum home page (http://www.qualityforum.org/Home.aspx).
38. American College of Surgeons National Surgical Quality Improvement Program semiannual report: January 2015. Chicago: American College of Surgeons.
39. Kernan WN, Viscoli CM, Makuch RW, Brass LM, Horwitz RI. Stratified randomization for clinical trials. J Clin Epidemiol 1999;52:19-26.
40. Bilimoria KY, Hoyt DB, Lewis F. Making the case for investigating flexibility in duty hour limits for surgical residents. JAMA Surg 2015;150:503-4.
41. ACGME duty hour requirements for FIRST Trial intervention arm programs (http://www.thefirsttrial.org/Documents/Redlined%20Duty%20Hour%20Requirements%20for%20Intervention%20Arm%20Hospitals.pdf).
42. Shiloach M, Frencher SK Jr, Steeger JE, et al. Toward robust information: data quality and inter-rater reliability in the American College of Surgeons National Surgical Quality Improvement Program. J Am Coll Surg 2010;210:6-16.
43. The American Board of Surgery In-Training Exam (ABSITE) (http://www.absurgery.org/default.jsp?examoffered_gs).
44. Lan KKG, Demets DL. Discrete sequential boundaries for clinical trials. Biometrika 1983;70:659-63.
45. Piaggio G, Elbourne DR, Pocock SJ,

Evans SJ, Altman DG. Reporting of noninferiority and equivalence randomized trials: extension of the CONSORT 2010 statement. JAMA 2012;308:2594-604.
46. Walker E, Nowacki AS. Understanding equivalence and noninferiority testing. J Gen Intern Med 2011;26:192-6.
47. StataCorp. Stata statistical software: release 13. College Station, TX: StataCorp, 2015.
48. Volpp KG, Rosen AK, Rosenbaum PR, et al. Did duty hour reform lead to better outcomes among the highest risk patients? J Gen Intern Med 2009;24:1149-55.