ORIGINAL ARTICLE

Education Outcomes in a Duty-Hour Flexibility Trial in Internal Medicine

S.V. Desai, D.A. Asch, L.M. Bellini, K.H. Chaiyachati, M. Liu, A.L. Sternberg, J. Tonascia, A.M. Yeager, J.M. Asch, J.T. Katz, M. Basner, D.W. Bates, K.Y. Bilimoria, D.F. Dinges, O. Even-Shoshan, D.M. Shade, J.H. Silber, D.S. Small, K.G. Volpp, and J.A. Shea, for the iCOMPARE Research Group*

The authors' full names, academic degrees, and affiliations are listed in the Appendix. Address reprint requests to Dr. Desai at Johns Hopkins University School of Medicine, Rm. 9029, 1830 E. Monument St., Baltimore, MD 21205, or at sanjayvdesai@jhmi.edu.

*A complete list of the members of the iCOMPARE Research Group is provided in the Supplementary Appendix, available at NEJM.org.

This article was published on March 20, 2018, at NEJM.org.

N Engl J Med 2018;378:1494-508.
DOI: 10.1056/NEJMoa1800965

ABSTRACT

BACKGROUND

Concern persists that inflexible duty-hour rules in medical residency programs may adversely affect the training of physicians.

METHODS

We randomly assigned 63 internal medicine residency programs in the United States to be governed by standard duty-hour policies of the 2011 Accreditation Council for Graduate Medical Education (ACGME) or by more flexible policies that did not specify limits on shift length or mandatory time off between shifts. Measures of educational experience included observations of the activities of interns (first-year residents), surveys of trainees (both interns and residents) and faculty, and intern examination scores.

RESULTS

There were no significant between-group differences in the mean percentages of time that interns spent in direct patient care and education nor in trainees' perceptions of an appropriate balance between clinical demands and education (primary outcome for trainee satisfaction with education; response rate, 91%) or in the assessments by program directors and faculty of whether trainees' workload exceeded their capacity (primary outcome for faculty satisfaction with education; response rate, 90%). Another survey of interns (response rate, 49%) revealed that those in flexible programs were more likely to report dissatisfaction with multiple aspects of training, including educational quality (odds ratio, 1.67; 95% confidence interval [CI], 1.02 to 2.73) and overall well-being (odds ratio, 2.47; 95% CI, 1.67 to 3.65). In contrast, directors of flexible programs were less likely to report dissatisfaction with multiple educational processes, including time for bedside teaching (response rate, 98%; odds ratio, 0.13; 95% CI, 0.03 to 0.49). Average scores (percent correct answers) on in-training examinations were 68.9% in flexible programs and 69.4% in standard programs; the difference did not meet the noninferiority margin of 2 percentage points (difference, −0.43; 95% CI, −2.38 to 1.52; P=0.06 for noninferiority).

CONCLUSIONS

There was no significant difference in the proportion of time that medical interns spent on direct patient care and education between programs with standard duty-hour policies and programs with more flexible policies. Interns in flexible programs were less satisfied with their educational experience than were their peers in standard programs, but program directors were more satisfied. (Funded by the National Heart, Lung, and Blood Institute and the ACGME; iCOMPARE ClinicalTrials.gov number, NCT02274818.)

In 2003, the Accreditation Council for Graduate Medical Education (ACGME) established resident duty-hour policies that limited resident workweeks to 80 hours and shifts to 30 hours. Further restrictions that limited shifts to 16 hours for first-year residents (interns) were implemented in 2011.[1] Early evaluations of both policies revealed essentially no differences in patient outcomes after implementation.[2-6]

After the 2011 policies were implemented, program directors reported a reduced quality of training and professional maturation, increased frequency of handoffs of care, and decreased continuity, without improved patient safety or quality of care.[7,8] Observational studies have shown that time spent in direct patient care by interns decreased from 25% in 1994 to 9 to 12% more recently,[9-12] and systematic reviews concluded that stricter duty-hour policies reduced resident education without improving resident well-being.[13,14]

The Individualized Comparative Effectiveness of Models Optimizing Patient Safety and Resident Education (iCOMPARE) trial is a national cluster-randomized trial that compared patient safety, education of trainees (both interns and residents), and intern sleep and alertness in internal medicine residency programs that were regulated by the 2011 duty-hour policies with programs that were regulated by more flexible policies. Flexible policies had no limits on shift length or mandatory time off between inpatient shifts. The outcomes for patient safety and sleep are not yet available. Here, we report the educational experiences of trainees and perceptions of program directors and faculty. We prespecified four hypotheses regarding trainee education: that interns in flexible programs would spend more time involved in direct patient care and in education, that trainees and faculty in flexible programs would report greater satisfaction with their educational experience, and that interns in flexible programs would have noninferior standardized test scores to those in standard programs.

METHODS

TRIAL DESIGN AND PROGRAM SELECTION

From July 1, 2015, to June 30, 2016, we compared a variety of measures in internal medicine residency programs in the United States that were randomly assigned to be regulated by standard 2011 ACGME duty-hour policies or by more flexible policies that did not specify any limits on shift length or mandatory time off between shifts (Table 1). The protocol (available with the full text of this article at NEJM.org) was approved by the institutional review board at the University of Pennsylvania or by the review board at each trial

Table 1. Duty-Hour Policies for Inpatient Rotations in Flexible Programs and Standard Programs.*

Policy	Flexible Programs	Standard Programs
Difference between groups		
Maximum length of shift		
PGY-1	No restriction	Duty-hour periods must not exceed 16 hr
PGY-2 or higher	No restriction	Duty-hour periods must not exceed 24 hr, with an additional 4 hr permitted for transitions in care
Mandatory time off between shifts	No restriction	All residents must have ≥14 hr off after 24 hr of in-house duty and ≥8 hr (and should have ≥10 hr) off after a regular shift
No difference between groups		
Weekly maximum work hr	80 hr	80 hr
Minimum no. of days off	1 day off every 7 days	1 day off every 7 days
Frequency of in-house call	In-house call no more frequent than every third night	In-house call no more frequent than every third night

* Residency programs that were assigned to be governed by flexible policies were allowed to waive limits on maximum shift length and mandatory time off between shifts. In a practical sense, this policy affected only inpatient rotations because outpatient rotations did not include shifts with lengths that would be affected. Flexible programs were provided duty-hour waivers from the Accreditation Council for Graduate Medical Education (ACGME). Time periods were averaged over a 4-week period. PGY denotes postgraduate year.

center. We measured education outcomes using time–motion observations of intern activity at six programs, responses by trainees and faculty on surveys administered by both iCOMPARE investigators and ACGME staff members, and scores on the American College of Physicians Internal Medicine In-Training Examination.

We selected programs to meet sample-size requirements for the primary hypothesis that the 30-day patient mortality under a flexible policy would not be inferior to that under the standard policy by more than 1 percentage point. (Mortality outcomes will be assessed with the use of 2015 and 2016 Medicare data, when available.) We included only programs that had at least one affiliated hospital in both the upper half of resident-to-bed ratios and the upper three quartiles of patient volume for diagnoses used to measure mortality. We excluded New York programs because state legislation prevented the necessary duty-hour waivers. Of the 179 eligible programs, 63 volunteered and were randomly assigned to be governed by flexible policies (32 programs) or standard policies (31 programs). Directors were informed about the assignment of their program at the time of randomization.

The decision to be included in the iCOMPARE trial was made by the leadership at each program, not by individual trainees or faculty. Prospective interns could choose which programs they applied to; almost all the programs had undergone randomization by November 2014 and could inform applicants of their duty-hour assignment during recruitment. Each intern who participated in the time–motion observations provided written informed consent. Trainees who responded to surveys that were administered by iCOMPARE investigators were assumed to have provided consent at the time of survey completion.

TIME–MOTION OBSERVATIONS

We conducted time–motion observations of interns to address our hypothesis that trainees in flexible programs would spend more time on direct patient care and education than their colleagues in standard programs. From March through May 2016, we conducted time–motion observations at three flexible and three standard programs, both community and university-based, located in the mid-Atlantic region. Preliminary data[9,10] suggested that the mean (±SD) percentage of time that would be spent in direct patient care in standard programs was 13±4%. We prespecified that the observation of 30 intern shifts on general internal medicine inpatient rotations at each program (180 total) would provide a power of at least 80% to detect an absolute difference of 3 percentage points in the percentage of time spent in direct patient care. Observers were scheduled to cover the entire shift. The types and lengths of shifts that were observed reflected the site-specific distribution of shifts in terms of shift length and overnight schedules. A total of 80 interns (44 in flexible programs and 36 in standard programs) who provided consent were observed for 2173 hours over 194 shifts (1072 hours in 96 shifts in flexible programs and 1101 hours in 98 shifts in standard programs). The mean length of observed shifts was 11.2 hours in both flexible and standard programs, the median length was 10.1 and 11.8 hours, respectively, and the maximum length was 27.8 and 14.5 hours. In the flexible programs, 4.2% of the shifts lasted more than 24 hours.

The observations, which were restricted to shifts beginning on weekdays, were performed by 23 trained observers. During the training of the observers, the median kappa coefficient among pairs of observers was 0.67, and the median agreement was 90%. During the trial, 10% of shifts were simultaneously observed by two observers (median kappa, 0.74; median agreement, 89%). Activity was recorded in milliseconds with the use of custom-built tablet-based software, with choices of direct patient care, education, indirect patient care, handoffs, rounds, or miscellaneous; more than one category could be selected if different types of activities occurred simultaneously.

MEDICAL KNOWLEDGE

We used the score (percent of questions answered correctly) on the American College of Physicians In-Training Examination to address our hypothesis that the medical knowledge acquired by interns in flexible programs would not be inferior to the knowledge of those in standard programs. The American College of Physicians shared deidentified 2015 and 2016 scores for trainees who had provided consent for research. (A total of 88% of all test takers provided such consent in 2016. Not all programs require trainees to take the examination.) In 2016, a total of 1687 trainees (852

of 1228 in flexible programs [69%] and 835 of 1300 in standard programs [64%]) took the examination as second-year residents. A total of 1766 trainees were included in 2015 baseline data (882 in flexible programs and 884 in standard programs).

TRAINEE EXPERIENCES

We used three surveys to address our hypothesis that the trainees' satisfaction with their educational experience would be superior in flexible programs. In accordance with the data-use policy of the ACGME, the iCOMPARE research team specified the analyses to be conducted on ACGME data and provided statistical code, and ACGME researchers completed the analyses and provided summary results. ACGME researchers performed and provided analyses of responses to their 2015 and 2016 annual resident surveys. The primary outcome for this hypothesis about trainee satisfaction with their educational experience was a single statement from the ACGME resident survey: "Major assignments provide an appropriate balance between education and other clinical demands." Potential responses were "never," "rarely," "sometimes," "often," and "very often." Additional ACGME measures were questions in seven content areas. (Details are provided in the Supplementary Appendix, available at NEJM.org.)

For each content area, ACGME dichotomized the 5-level response to each component question into "compliant" or "noncompliant" and pooled the dichotomized responses to provide a content-level dichotomized response. For example, for the question listed above, the options "sometimes," "rarely," and "never" would be considered noncompliant. The response for a content area was noncompliant if the respondent provided a noncompliant response to any of the questions in the content area. The mean rates of response among flexible and standard programs were 91% and 90%, respectively, in 2015 and 91% each in 2016.

Investigators administered an end-of-year survey to all trainees in May 2015 in 55 programs (which served as a baseline survey before the start of the trial) and in May 2016 in 62 programs (end-of-trial survey). The instrument was developed for the FIRST (Flexibility in Duty Hour Requirements for Surgical Trainees) trial[5] in surgery and adapted for internal medicine (Table S1 in the Supplementary Appendix). The 5-level response for each question was dichotomized into a "positive" or "negative" response to parallel the results of the ACGME resident survey. The survey ended with the Maslach Burnout Inventory–Human Services Survey, a 22-item scale assessing emotional exhaustion, depersonalization, and perception of personal accomplishment.[15] The survey response rates among trainees were 58% in flexible programs and 57% in standard programs in 2015 and 46% and 44%, respectively, in 2016. (In 2016, overall response rates were 45% for all trainees and 49% for interns.)

Investigators administered end-of-shift surveys to all trainees (60 programs) every 2 weeks from September 2015 through April 2016. The surveys reflected trainees' perceptions of their experience with education, ownership, work intensity, and continuity; 72% of interns in flexible programs and 67% of those in standard programs (64% and 61% of all trainees in flexible and standard programs, respectively) participated in at least one survey cycle.

FACULTY EXPERIENCES

To address our hypothesis that faculty in flexible programs would report greater satisfaction with clinical teaching experiences and perceptions of patient safety, teamwork, and supervision, we used ACGME faculty surveys and an iCOMPARE survey of program directors. Using the same process that has been described for ACGME resident surveys, the ACGME provided analyses of responses to their 2015 and 2016 annual faculty surveys. The primary outcome for this hypothesis about faculty satisfaction was a single statement from the ACGME faculty survey: "Residents' clinical workload exceeds their capacity to do the work." Response options were "never," "rarely," "sometimes," "often," and "very often." Additional measures were questions in six content areas, with responses in each component of the content area dichotomized and pooled, as described for the ACGME resident survey. (Details are provided in the Supplementary Appendix.)

Mean rates of survey response among flexible and standard programs were 90% each in 2015 and 91% each in 2016. In addition, iCOMPARE investigators surveyed program directors in May 2015, when 63 programs were surveyed (response rate, 88% in flexible programs and 100% in standard programs, although a data-acquisition er-

ror limited secondary analyses to 19 flexible and 18 standard programs), and in May 2016 (63 programs, with a survey response rate of 100% for flexible programs and 97% for standard programs). The survey instrument mirrored the one used in a previous survey of program directors[16] (Table S2 in the Supplementary Appendix). The 5-level response for each question was dichotomized into "positive" or "negative" to provide parallel reporting to the results of the ACGME faculty survey.

STATISTICAL ANALYSIS

The sample-size calculation of 58 programs (29 per group) was based on the primary hypothesis that the 30-day patient mortality under a flexible policy would not be inferior to that under a standard policy by more than 1 percentage point. The achieved sample size of 63 programs provides a power of more than 80% in the comparison of each of the four prespecified education hypotheses. The hypothesis regarding medical knowledge is a noninferiority hypothesis with a noninferiority margin of 2 percentage points that was chosen by consensus of the investigators. We hypothesized that interns in flexible programs would spend more time in direct patient care and education and would be more satisfied with their education and that faculty in flexible programs would be more satisfied with their teaching experiences, patient safety, teamwork, and supervision than their peers in standard programs.

We used a mixed-effects linear-regression model with a random intercept for each program cluster to determine the between-group difference in outcomes obtained from the time–motion observations; for each activity, we analyzed the mean percentage of the observed shift time that was spent in the activity over all shifts observed for the intern. We used mixed-effects linear or logistic regression with a random intercept for each program cluster to determine the between-group difference in ordinal outcomes obtained from the ACGME surveys of trainees and faculty and the binary outcomes obtained from the end-of-year trainee surveys. We used logistic regression with generalized estimating equations and robust variance estimation to account for the correlations between responses from respondents at the same program to determine the between-group difference in binary outcomes obtained from the ACGME trainee and faculty surveys. Exact logistic regression was used to determine the difference in outcomes obtained from the end-of-year survey of program directors. We used mixed-effects linear regression with a random intercept for each program cluster to determine the between-group difference in outcomes obtained from the end-of-shift surveys; for each question, we analyzed the mean of all ratings provided by the trainee over the survey cycles in which the trainee participated. When program-level data were available for the baseline year, we completed secondary analyses after adjustment for the respondent's trial-year outcome for the baseline year program-level mean outcome.

Here, we report marginal duty-hour group effects; observed effects and variance components from the mixed-effects regression models are provided in the Supplementary Appendix. Each marginal effect is similar to an observed mean or percentage but is derived from the regression model and accounts for correlations between respondents at the same program, averaging across random effects caused by variation in respondent outcomes within programs. Each linear mixed-effects model provides a measure and test of the clustering effect of programs on the outcome (program variance and P value) and a measure of the variability of the respondent's response (error variance). Each logistic mixed-effects model provides a measure and test of the clustering effect of programs on the outcome (random program variance and P value).

Data were analyzed according to the intention-to-treat principle. We assumed that missing responses were missing completely at random and analyzed all available responses. We report P values for the primary outcome measures. For the secondary outcome measures, we report 95% confidence intervals without P values, given the multiplicity of comparisons. Analyses were conducted with the use of SAS and Stata software.

RESULTS

PROGRAM CHARACTERISTICS

The characteristics of the trainees and programs that participated in the trial and those that did not participate did not differ significantly according to group assignment (Table 2). However,

Table 2. Characteristics of the Residency Programs and Trainees.*

Characteristic	Eligible Programs Not Included	Eligible Programs Included		P Value for Not Included vs. Included Programs†	P Value for Flexible vs. Standard Programs‡
		Flexible Programs§	Standard Programs		
Programs					
No. of programs	116	32	31		
Program type — no. (%)				0.07¶	0.48¶
Community	12 (10)	1 (3)	3 (10)		
University	45 (39)	20 (62)	20 (65)		
Both community and university	57 (49)	11 (34)	8 (26)		
Military	2 (2)	0	0		
Geographic region — no. (%)				0.18¶	0.30¶
Northeast	30 (26)	8 (25)	14 (45)		
Midwest	40 (34)	8 (25)	5 (16)		
South	30 (26)	13 (41)	8 (26)		
Mountain or Pacific	16 (14)	3 (9)	4 (13)		
Residents					
Mean no. of residents per program (±SD)	70.0±32.2	94.5±39.1	98.9±45.8	<0.001‖	0.68**
Resident-to-bed ratio	0.46	0.61	0.59	0.004‖	0.82**
No. of residents with available information		3099	3214		
Sex — no./total no. (%)††					0.34¶
Male		683/1299 (53)	622/1207 (52)		
Female		614/1299 (47)	585/1207 (48)		
Other		2/1299 (<1)	0		
Missing data		1800	2007		
Residency year — no./total no. (%)††					0.96¶
PGY-1		800/1844 (43)	767/1788 (43)		
PGY-2		551/1844 (30)	538/1788 (30)		
PGY-3 or higher		493/1844 (27)	483/1788 (27)		
Missing data		1255	1426		

* Percentages may not total 100 because of rounding.
† The P value is for the difference between eligible programs that did not volunteer to participate in the trial and those that volunteered to participate and underwent randomization to be governed by flexible duty-hour policies or standard duty-hour policies.
‡ The P value is for the comparison between flexible programs and standard programs that were included in the study.
§ Residency programs that were assigned to be governed by flexible policies were allowed to waive limits on maximum shift length and mandatory time off between shifts.
¶ The P value was calculated with the use of a two-tailed chi-square test.
‖ The P value was calculated with the use of one-way analysis of variance.
** The P value was calculated with the use of a Student's t-test.
†† These data were collected from iCOMPARE surveys of trainees. The question regarding sex was not asked in seven programs owing to original agreements with institutional review boards. In addition, many trainees did not complete the survey.

Table 3. Percentage of Observed Shift Time Spent on Activity by Interns.*

Activity	Flexible Programs	Standard Programs	Difference (95% CI)	P Value†
	mean percentage of shift time		*percentage points*	
Primary outcomes				
Direct patient care‡	13.0	11.8	1.2 (–0.7 to 3.1)	0.21
Education§	7.3	7.3	0.0 (–5.9 to 5.9)	>0.99
Secondary outcomes				
Indirect patient care¶	67.9	63.7	4.2 (–6.7 to 15.1)	
Handoffs	2.7	4.0	–1.3 (–3.8 to 1.1)	
Rounds	22.4	19.0	3.4 (–7.6 to 14.5)	
Miscellaneous‖	5.6	9.7	–4.2 (–7.8 to –0.6)	

* Each intern who was observed is included in each model. The quantity analyzed was the intern's mean percentage of shift time observed in the activity overall. If an activity was not observed during a shift, the time spent on the activity was 0 minutes. In six flexible programs, 44 interns were observed for 1072 hours during 96 shifts. In six standard programs, 36 interns were observed for 1101 hours during 98 shifts. The sum of the percentages may exceed 100% because more than one activity could occur simultaneously.

† The mean percentages (marginal means) and differences were estimated from a mixed-effects linear-regression model. Observed mean percentages and random-effect variance components are provided in Table S4 in the Supplementary Appendix, which shows P<0.05 for between-program variance on all activities except direct patient care and miscellaneous. This finding reflects a large variation across all programs, independent of assignment to flexible or standard policies.

‡ Direct patient care includes evaluation and in-person communication with patients or their families.

§ Educational activities include teaching or being taught (including teaching rounds), educational conferences, and reading about medicine.

¶ Indirect patient care includes activities such as interacting with the electronic chart, viewing imaging, attending rounds not involving the patient directly, and discussing care with a consultant.

‖ Miscellaneous activities include eating and sleeping.

participating programs were larger and more likely to be university-based and had higher resident-to-bed ratios than eligible programs that chose not to participate. All the programs that were assigned to the flexible group made use of the flexible rules and implemented them in an average of three rotations. At any one time, about two thirds of the interns in those hospitals were on a rotation with flexible hours, and the average maximum shift length across different rotations was about 24 hours (Table S3 in the Supplementary Appendix).

TIME–MOTION OBSERVATIONS

There was no significant between-group difference in the mean percentage of observed shift time that was spent on direct patient care, with 13.0% in the flexible programs and 11.8% in the standard programs (difference in the mixed-effects model, 1.2 percentage points; 95% confidence interval [CI], –0.7 to 3.1; P=0.21), or in the mean percentage of time spent on education. The time spent on other activities was similar in the two groups (Table 3, and Table S4 in the Supplementary Appendix). The variance among the programs in each group was large and significant for most activities, which indicated a large variation in how interns spent their time across all the programs in the trial.

MEDICAL KNOWLEDGE

Average scores (percent correct answers) on in-training exams were 68.9% in the flexible programs and 69.4% in the standard programs. The between-group difference in examination scores did not meet the margin of 2 percentage points for noninferiority of the flexible programs, as compared with the standard programs (difference in the mixed-effects linear-regression model, –0.43; 95% CI, –2.38 to 1.52; P=0.06 for noninferiority). However, after adjustment of the 2016 scores for the 2015 mean program score, noninferiority was confirmed (difference, 0.64; 95% CI, –0.56 to 1.84; P<0.001 for noninferiority) (Table S5 in the Supplementary Appendix). This analysis was not prespecified in the protocol but was planned before any data were analyzed. The variance among the programs in each group was significant for this outcome, which indicated a large variation in scores across programs.

TRAINEE EXPERIENCES

There was no significant between-group difference in the responses to the prespecified primary question on the ACGME survey regarding appropriate balance for education. The mean score on scale of 1 (never) to 5 (very often) was 3.87 for the flexible programs and 3.85 for the standard programs (difference, 0.02; 95% CI, –0.12 to 0.17; P=0.74). Trainees in each group had similar responses in the seven content areas provided by the ACGME. Analyses after adjustment of the 2016 responses for the 2015 mean program response provided results that were consistent with the unadjusted analyses (Table S6 in the Supplementary Appendix).

Interns in the two groups often differed in

Table 4. Percentage of Interns Who Reported That Flexible or Standard Duty-Hour Rules Had a Negative Effect on Their Training and Personal Experiences.*

Survey Item with Negative Effect	Flexible Programs	Standard Programs	Odds Ratio (95% CI)†
	percentage of interns		
Safety of patient care	14	6	2.71 (1.65–4.48)
Continuity of care	11	26	0.35 (0.22–0.57)
Ability to acquire clinical skills	10	10	1.01 (0.62–1.63)
Intern autonomy	6	8	0.68 (0.40–1.14)
Availability for urgent patient care encounters	7	10	0.68 (0.44–1.05)
Availability for elective patient care encounters	15	11	1.47 (0.94–2.30)
Ability to attend required educational conferences	22	10	2.61 (1.77–3.85)
Relationship between interns and residents	7	3	2.21 (1.20–4.06)
Time for teaching of medical students	22	17	1.44 (1.02–2.04)
Need for performing work related to patient care outside the hospital	24	25	0.94 (0.65–1.36)
Ability to participate in research	26	11	2.88 (2.01–4.12)
Professionalism	8	3	3.26 (1.72–6.16)
Job satisfaction	21	6	4.32 (2.72–6.85)
Satisfaction with career choice	18	5	4.26 (2.52–7.20)
Intern morale	24	4	8.14 (4.65–14.26)
Time with family and friends	33	7	6.11 (3.76–9.91)
Time for hobbies and outside interests	31	8	5.20 (3.17–8.54)
Health	29	7	5.53 (3.32–9.20)
Ability to acquire clinical reasoning skills	9	6	1.47 (0.87–2.51)
Pace of intern's workday	19	17	1.14 (0.76–1.71)
Intern's overall well-being	26	6	5.27 (3.22–8.64)

* Data are from end-of-year surveys of interns in 30 flexible programs (with 638 interns) and 31 standard programs (with 608 interns). Two of the flexible programs did not provide survey data for inclusion in this analysis. Interns were asked their perceptions of 21 aspects of their trainee experiences in relation to duty-hour rules; response choices were a negative effect, no effect, or positive effect. For each item, the response choice was dichotomized into a binary response of negative effect versus no effect or positive effect. In the flexible programs, 622 of 1228 interns (51%) answered every survey question, and 16 (1%) answered one or more but not all questions. In the standard programs, 594 of 1300 interns (46%) answered every survey question, and 14 (1%) answered one or more but not all questions.

† The percentages (marginal means) and odds ratios were estimated from a mixed-effects logistic-regression model. Observed percentages and random program variances are provided in Table S7 in the Supplementary Appendix, which shows $P<0.05$ for between-program variance on many outcomes. This finding reflects a large variation across all programs, independent of assignment to flexible or standard programs.

responses to iCOMPARE end-of-year surveys (Tables 4 and 5, and Table S7 in the Supplementary Appendix). Interns in flexible programs were more likely than those in standard programs to report dissatisfaction with the overall quality of education (odds ratio, 1.67; 95% CI, 1.02 to 2.73), with overall well-being (odds ratio, 2.47; 95% CI, 1.67 to 3.65), and with the effect of the program on their personal lives (e.g., time with family and friends) (odds ratio, 6.11; 95% CI, 3.76 to 9.91). In contrast, interns in flexible programs were less likely to report dissatisfaction with the number of admissions that they were able to complete (odds ratio, 0.48; 95% CI, 0.27 to 0.85) and were less likely to perceive negative effects of duty-hour rules on the continuity of care (odds ratio, 0.35; 95% CI, 0.22 to 0.57).

Reports of burnout were high in each group.

Table 5. Measures of Intern Satisfaction with Education and Assessment of Fatigue, Patient Safety, and Continuity of Patient Care.*

Survey Item	Flexible Programs	Standard Programs	Odds Ratio (95% CI)†
	percentage of interns		
Interns reporting dissatisfaction with:‡			
Overall quality of resident education	13	8	1.67 (1.02–2.73)
Overall well-being	30	15	2.47 (1.67–3.65)
Patient safety	6	4	1.40 (0.83–2.36)
Continuity of care	5	7	0.80 (0.46–1.41)
Quality and ease of handoffs and transitions in care	6	7	0.89 (0.54–1.46)
Duty-hour regulations of the program	13	5	2.78 (1.69–4.57)
Work hours and scheduling	21	11	2.21 (1.45–3.37)
Time for rest	34	17	2.43 (1.62–3.63)
Level of attending supervision	3	2	1.45 (0.69–3.03)
Ability to follow the clinical care of patients admitted by the intern	5	6	0.83 (0.49–1.39)
No. of admissions handled entirely by intern	5	9	0.48 (0.27–0.85)
Interns perceiving that their fatigue:§			
Almost always or often affected their personal safety	16	8	2.01 (1.28–3.14)
Almost always or often affected patient safety	12	7	1.64 (1.06–2.52)
Interns reporting at least 1 problematic occurrence during their most recent month on an internal medicine rotation¶			
Left during an encounter with a patient because of duty-hour limits	5	5	0.98 (0.56–1.73)
Missed an encounter with a patient because of duty-hour limits	18	15	1.29 (0.80–2.09)
Handed off an issue involving active patient care because of duty-hour limits	28	32	0.81 (0.53–1.22)
Left or missed educational conference during a scheduled shift because of duty-hour limits	31	27	1.22 (0.87–1.71)
Worked more than 16 hr continuously in-house	58	31	3.02 (1.79–5.10)
Had less than 8 hr off between daily shifts	30	32	0.90 (0.56–1.44)

* Data are from end-of-year surveys of interns, with the same numbers as described in Table 4. Interns were asked to score 19 aspects of their trainee experiences in relation to satisfaction with education and their assessments of fatigue, patient safety, and continuity of care. For each item, the response choices were dichotomized into a binary response, as indicated for each question theme.

† The percentages (marginal means) and odds ratios were estimated from a mixed-effects logistic-regression model. Observed percentages and random program variances are provided in Table S7 in the Supplementary Appendix, which shows $P<0.05$ for between-program variance on many outcomes. This finding reflects a large variation across all programs, independent of assignment to flexible or standard programs.

‡ In the section on dissatisfaction, responses were "very dissatisfied" or "dissatisfied" versus "neutral," "satisfied," or "very satisfied."

§ In the section on fatigue, responses were "always" or "often" versus "sometimes," "rarely," or "never."

¶ In the section on problematic occurrences, responses were one or more occurrences in the past month versus no occurrence.

Table 6. Interns' Scores on the Maslach Burnout Inventory.*

Subscale and Scoring	Flexible Programs	Standard Programs	Odds Ratio (95% CI)
Emotional exhaustion subscale (scored 0–54)			
Observed score	25.9±11.7	24.7±12.0	
High or moderate score (≥17) (% of interns)†	79	72	1.43 (0.96–2.13)
Depersonalization subscale (scored 0–30)			
Observed score	11.9±6.8	11.3±6.9	
High or moderate score (≥7) (% of interns)†	75	72	1.18 (0.81–1.71)
Personal accomplishment subscale (scored 0–48)			
Observed score	33.5±8.4	34.2±8.1	
Low or moderate score (0–38) (% of interns)†	71	69	1.12 (0.84–1.49)

* Plus–minus values are means ±SD. Data are for 30 flexible programs (594 interns) and for 31 standard programs (563 interns). Two of the flexible programs did not provide survey data for inclusion in this analysis. The Maslach Burnout Inventory–Human Services version is a 22-item scale that assesses how respondents in the human services or helping professionals view their job and the persons with whom they work closely.[15] The respondent rates how often each item (statement) is true from 0 (never) to 6 (every day). Three subscales are scored for rating emotional exhaustion (9 items scored from 0 to 54), depersonalization (5 items scored from 0 to 30), and personal accomplishment (8 items scored from 0 to 48); on all three subscales, higher scores indicate a greater level of response. In the flexible groups, 594 of 1228 interns (48%) completed all the items; in the standard programs, 563 of 1300 interns (43%) completed all the items.

† The percentages (marginal means) and odds ratios were estimated from a mixed-effects logistic-regression model. Observed percentages and random program variances are provided in Table S8 in the Supplementary Appendix, which shows $P<0.05$ for between-program variance on all subscales except personal accomplishment. This finding reflects a large variation across all programs, independent of assignment to flexible or standard programs.

The interns in each group had a similar likelihood of having high or moderate scores on the Maslach Burnout Inventory subscale for emotional exhaustion (79% in flexible programs and 72% in standard programs; odds ratio in mixed-effects logistic-regression model, 1.43; 95% CI, 0.96 to 2.13), high or moderate scores on the depersonalization subscale (75% and 72%, respectively; odds ratio, 1.18; 95% CI, 0.81 to 1.71), and low or moderate scores on the personal accomplishment subscale (71% and 69%, respectively; odds ratio, 1.12; 95% CI, 0.84 to 1.49) (Table 6, and Table S8 in the Supplementary Appendix). These prevalences were slightly higher than those of emergency medicine trainees[17] and slightly lower than those of surgical trainees.[18] On the subscales of emotional exhaustion and depersonalization, the mean scores were higher than those in the general population,[15] and the prevalences are much higher than those reported in many other professions, such as university faculty.[19] However, there is variability in the cutoffs that have been used to determine burnout across studies.[20]

Analyses of responses from all trainees (Table S9 in the Supplementary Appendix) and analyses after adjustment of the 2016 survey responses for the 2015 mean program responses (Tables S7, S8, and S9 in the Supplementary Appendix) provided results that were consistent with those in the unadjusted analyses. In more than 11,000 end-of-shift surveys, interns in each group were similar in their perceptions of their experience with respect to education, sense of ownership, work intensity, and continuity over the previous 24 hours, findings that were consistent with analyses from all the trainees (Table S10 in the Supplementary Appendix).

FACULTY EXPERIENCES

There was no significant between-group difference in faculty responses to the prespecified primary question on the ACGME survey asking whether residents' workload exceeded their capacity to do the work, which ranged from 1 (very often) to 5 (never) (mean score, 4.22 in the flexible programs and 4.18 in the standard pro-

grams; difference, 0.04; 95% CI, −0.04 to 0.12; P=0.46). There also was no significant between-group difference in faculty responses for the six content areas provided by the ACGME. Analyses after adjustment of the 2016 responses for the 2015 mean program responses provided results

Table 7. Survey Responses of Program Directors.*

Survey Item	Flexible Programs	Standard Programs	Odds Ratio (95% CI)
	percentage of program directors		
Dissatisfaction with learning environment			
Ownership of patient care			
Interns	0	23	0.08 (0.00–0.57)
Residents	0	10	0.23 (0.00–2.22)
Ability of interns to manage the patients they admit	0	13	0.16 (0.00–1.36)
Morale			
Interns	3	27	0.09 (0.00–0.77)
Residents	9	20	0.42 (0.06–2.22)
Time for trainees to reflect	25	57	0.26 (0.07–0.84)
Effectiveness in performing clinical duties			
Interns	0	10	0.23 (0.00–2.22)
Residents	0	3	0.94 (0.00–36.6)
Ability of attending physician to provide real-time feedback on patient care activities			
Interns	3	63	0.02 (0.00–0.15)
Residents	3	43	0.04 (0.00–0.34)
Frequency of handoffs	6	67	0.04 (0.00–0.19)
Quality of handoffs	13	40	0.22 (0.04–0.87)
Ability of residents to work in interprofessional teams	3	3	0.94 (0.01–75.9)
Dissatisfaction with workload			
Type of staff member			
Faculty	29	40	0.62 (0.18–2.02)
Residents	10	33	0.22 (0.03–1.00)
Interns	6	37	0.12 (0.01–0.66)
Program director	29	43	0.54 (0.16–1.74)
Opportunity for residents to transition care when fatigued	13	14	0.93 (0.15–5.55)
Ability of trainees to perform necessary work during the scheduled duty period	10	55	0.09 (0.01–0.40)
Reliance on residents to provide clinical service	29	31	0.91 (0.26–3.18)
Dissatisfaction with educational opportunities			
Adequacy of time for bedside teaching			
Interns	13	55	0.13 (0.03–0.49)
Residents	13	52	0.14 (0.03–0.56)
Ability to attend conferences while on inpatient rotations			
Interns	19	55	0.20 (0.05–0.70)
Residents	17	38	0.33 (0.08–1.27)

Table 7. (Continued.)

Survey Item	Flexible Programs	Standard Programs	Odds Ratio (95% CI)
	percentage of program directors		
Ability to participate in teaching rounds by attending physicians			
Interns	7	31	0.16 (0.02–0.91)
Residents	7	17	0.35 (0.03–2.37)
Ability to attend meetings of patients' families			
Interns	7	17	0.35 (0.03–2.37)
Residents	3	7	0.47 (0.01–9.54)
Balance of service versus education			
Interns	17	34	0.39 (0.09–1.49)
Residents	17	24	0.63 (0.14–2.71)
Elective rotation time for house staff	17	21	0.74 (0.15–3.36)
Time for house staff to do research	13	28	0.41 (0.08–1.79)
Time for house staff to engage in medical-student education or quality improvement	30	17	2.03 (0.51–9.01)
Amount of time that house staff need to spend on night rotations	3	28	0.09 (0.00–0.79)
Dissatisfaction with program administration and organization			
Financial support			
For nonteaching services	27	57	0.28 (0.08–0.93)
For hiring allied health professionals (e.g., nurse practitioners) for clinical care delivery	50	72	0.39 (0.11–1.27)
For hiring hospitalists or additional faculty members for clinical care delivery	43	66	0.41 (0.12–1.30)
Relationship of residency program with hospital administration	3	17	0.17 (0.00–1.67)
Program director morale	7	25	0.22 (0.02–1.31)
Effort of tracking duty hours	33	55	0.41 (0.12–1.31)
Dissatisfaction with patient outcomes			
Continuity of care for patients	7	52	0.07 (0.01–0.36)
Safety of patients	0	17	0.13 (0.00–0.98)
Graduates' preparedness for practice after residency	0	14	0.17 (0.00–1.40)

* Data are from end-of-year surveys of directors of 32 flexible programs and 30 standard programs. Program directors were asked to score 43 aspects of the educational environment of their internal medicine training program; response choices were "very dissatisfied," "dissatisfied," "neutral," "satisfied," and "very satisfied." For each item, these response choices were dichotomized into a binary response of "very dissatisfied" or "dissatisfied" versus "other." In the flexible programs, 30 of 32 directors (94%) answered every survey question, and 2 (6%) answered at least one question but not all. In the standard programs, 26 of 30 program directors (87%) answered every survey question, and 4 (13%) answered at least one question but not all.

that were consistent with those in the unadjusted analyses (Table S11 in the Supplementary Appendix).

In contrast, in responses to items on the iCOMPARE end-of-year survey of program directors, directors in flexible programs were less likely than those in standard programs to report dissatisfaction with many elements of training, including aspects of the learning environment, such as intern ownership of patient care, intern

morale, time for trainees to reflect, ability of attending physician to provide real-time feedback, frequency of handoffs, intern workload, educational opportunities (e.g., time for teaching at bedside), and continuity of care for patients (Table 7). Analyses after adjustment of the 2016 responses for the 2015 responses were limited to the 37 programs for which 2015 responses were available and provided results that were similar to those in the unadjusted analysis (Table S12 in the Supplementary Appendix).

DISCUSSION

In this cluster-randomized trial, we compared the standard 2011 ACGME duty-hour policies for medical trainees with more flexible policies mandating no limits on shift length or mandatory time off between shifts. Outcomes with respect to patient mortality and intern sleep and alertness, which are important in interpreting the complete results of the trial, are not reported here.

The design of the trial called for evaluating four primary educational hypotheses. We found no significant between-group differences in how interns who were exposed to the different duty-hour rules spent their time. The primary unadjusted analysis of difference in knowledge acquired by interns did not confirm the noninferiority of the flexible programs. In a secondary analysis that was planned before any data had been analyzed but was not specified in the protocol, in which we adjusted the 2016 in-training examination scores for 2015 mean program scores, the criterion for noninferiority was met. In addition, there was no significant between-group difference in the primary outcomes of trainee satisfaction with the balance between education and work or in faculty satisfaction with trainee workload relative to their capacity. However, in the flexible programs, interns were substantially less satisfied with many aspects of the learning environment and directors were more satisfied with the educational processes than their peers in standard programs.

The similarities in the clinical and educational activities of residents in programs governed by different duty-hour rules suggest that programs can adapt their activities to preserve exposure to such activities. Indeed, the end-of-shift survey data from more than 11,000 shifts and the ACGME survey data from thousands of trainees and hundreds of faculty showed no significant between-group differences in any of the educational domains. Even so, directors of standard programs were more likely than their peers in flexible programs to be dissatisfied with educational processes (e.g., opportunities for bedside teaching or providing interns with real-time feedback), a finding that was consistent with reported opinions.[7,8,16] In contrast, interns in flexible programs were more likely to be dissatisfied with their overall educational experience, especially its effect on their personal lives, a finding that was consistent with the results among surgical trainees in the FIRST trial.[5] This discrepancy may reflect different perspectives of educators who are focused on teaching processes and learners who are focused on overall life experiences during training.

The trial has several limitations. Because the overall response rate for the iCOMPARE trainee survey was only 45%, we cannot be sure that survey respondents reflect the eligible pool. Response rates were in the ranges that have been reported in studies of graduate medical education.[21] Our analyses reflect a comparison between two duty-hour policies, as implemented by internal medicine residency programs, and we have no reason to postulate an interaction between trial group and systematic nonresponse, but we also have no way to exclude such bias. Although it is a large trial, statistical resolution required that we exclude smaller programs. The time–motion substudy was further restricted to six programs in the mid-Atlantic states. Although we have no reason to believe intern time distributions differed according to program type, we included an increased number of large-sized and university-based programs, and other programs may have different effects from flexible rules. We did not measure the actual numbers of hours that were worked, but interns in the two groups were limited to the same average total number of hours worked per week. The trial included only internal medicine programs, so the results may not apply to other specialties. Trainees and program directors were all aware of their program assignments, so survey responses could reflect preconceptions about the effect of the assignments. In addition, there was a large variation across programs in how interns spent their time and in

trainee responses on many of the iCOMPARE survey items, as indicated by the estimates of variance among the programs in each group, a factor that limits the reliability and interpretability of comparisons between the two duty-hour groups. Accounting for this variation in the model helps with the reliability and interpretability of the comparisons.

The ACGME duty-hour rules aim to protect patients and trainees at the same time that they ensure a workforce capable of caring for future patients. Despite concerns of program directors that ACGME duty-hour rules diminish the training and professional development of physicians, we found little evidence to support variations in activities and knowledge acquisition within the ranges tested in this trial. However, the satisfaction of interns and program directors were meaningfully different in programs with flexible policies. Given that trainees in flexible programs in both medicine and surgery perceived more burdens on their personal lives, program directors may see a need to better understand their dissatisfaction and develop mechanisms to address these effects.

Supported by grants (U01HL125388, to Dr. Asch; and U01HL126088, to Dr. Tonascia) from the National Heart, Lung, and Blood Institute; and by a grant from the Accreditation Council for Graduate Medical Education to Dr. Shea.

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

We thank the participating program directors; Amanda K. Bertram, M.S., of Johns Hopkins University; Kelsey A. Gangemi, M.P.H., of the University of Pennsylvania; James T. Richardson, M.P.H., Lauren Byrne, M.P.H., Rebecca Miller, M.S., and Thomas Nasca, M.D., of the Accreditation Council for Graduate Medical Education; and Margaret Wells, B.A., and Linda J. Harris, A.A.S., of the American College of Physicians.

APPENDIX

The authors' full names and academic degrees are as follows: Sanjay V. Desai, M.D., David A. Asch, M.D., M.B.A., Lisa M. Bellini, M.D., Krisda H. Chaiyachati, M.D., M.P.H., M.S.H.P., Manqing Liu, M.H.S., Alice L. Sternberg, Sc.M., James Tonascia, Ph.D., Alyssa M. Yeager, M.D., Jeremy M. Asch, B.S., Joel T. Katz, M.D., Mathias Basner, M.D., Ph.D., David W. Bates, M.D., Karl Y. Bilimoria, M.D., David F. Dinges, Ph.D., Orit Even-Shoshan, M.S., David M. Shade, J.D., Jeffrey H. Silber, M.D., Ph.D., Dylan S. Small, Ph.D., Kevin G. Volpp, M.D., Ph.D., and Judy A. Shea, Ph.D.

From the Departments of Medicine (S.V.D.), Epidemiology (A.L.S., J.T., D.M.S.), and Biostatistics (J.T.), Johns Hopkins University, Baltimore; the Departments of Medicine (D.A.A., L.M.B., K.H.C., M.L., A.M.Y., J.M.A., J.A.S.), Psychiatry (M.B., D.F.D.), and Medical Ethics and Policy (K.G.V.) and the Department of Statistics, the Wharton School (D.S.S.), University of Pennsylvania, the Corporal Michael J. Crescenz Veterans Affairs Medical Center (D.A.A., K.H.C., K.G.V.), and the Department of Pediatrics, Children's Hospital of Philadelphia (O.E.-S., J.H.S.) — all in Philadelphia; the Department of Medicine, Brigham and Women's Hospital, Boston (J.T.K., D.W.B.); and the Department of Surgery and Center for Healthcare Studies, Northwestern University, Chicago (K.Y.B.).

REFERENCES

1. 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.
2. 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.
3. 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.
4. 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.
5. Bilimoria KY, Chung JW, Hedges LV, et al. National cluster-randomized trial of duty-hour flexibility in surgical training. N Engl J Med 2016;374:713-27.
6. 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.
7. Drolet BC, Khokhar MT, Fischer SA. The 2011 duty-hour requirements — a survey of residency program directors. N Engl J Med 2013;368:694-7.
8. Garg M, Drolet BC, Tammaro D, Fischer SA. Resident duty hours: a survey of internal medicine program directors. J Gen Intern Med 2014;29:1349-54.
9. Block L, Habicht R, Wu AW, et al. In the wake of the 2003 and 2011 duty hours regulations, how do internal medicine interns spend their time? J Gen Intern Med 2013;28:1042-7.
10. Fletcher KE, Visotcky AM, Slagle JM, Tarima S, Weinger MB, Schapira MM. The composition of intern work while on call. J Gen Intern Med 2012;27:1432-7.
11. Mamykina L, Vawdrey DK, Hripcsak G. How do residents spend their shift time? A time and motion study with a particular focus on the use of computers. Acad Med 2016;91:827-32.
12. Guarisco S, Oddone E, Simel D. Time analysis of a general medicine service: results from a random work sampling study. J Gen Intern Med 1994;9:272-7.
13. Bolster L, Rourke L. The effect of restricting residents' duty hours on patient safety, resident well-being, and resident education: an updated systematic review. J Grad Med Educ 2015;7:349-63.
14. Lin H, Lin E, Auditore S, Fanning J. A narrative review of high-quality literature on the effects of resident duty hours reforms. Acad Med 2016;91:140-50.
15. Maslach C, Jackson SE, Leiter MP. Maslach burnout inventory manual. 3rd ed. Palo Alto, CA: Consulting Psychologists Press, 1996.
16. Shea JA, Willett LL, Borman KR, et al. Anticipated consequences of the 2011 duty hours standards: views of internal medicine and surgery program directors. Acad Med 2012;87:895-903.
17. Kimo Takayesu J, Ramoska EA, Clark TR, et al. Factors associated with burnout during emergency medicine residency. Acad Emerg Med 2014;21:1031-5.
18. Elmore LC, Jeffe DB, Jin L, Awad MM,

Turnbull IR. National survey of burnout among US general surgery residents. J Am Coll Surg 2016;223:440-51.
19. Lackritz JR. Exploring burnout among university faculty: incidence, performance, and demographic issues. Teach Teach Edu 2004;20:713-29 (https://www.sciencedirect.com/science/article/pii/S0742051X04000848).
20. Eckleberry-Hunt J, Kirkpatrick H, Barbera T. The problems with burnout research. Acad Med 2018;93:367-70.
21. Cunningham CT, Quan H, Hemmelgarn B, et al. Exploring physician specialist response rates to Web-based surveys. BMC Med Res Methodol 2015;15:32.