Annals of Internal Medicine ORIGINAL RESEARCH

Trends in Depressive Symptoms and Associated Factors During Residency, 2007 to 2019

A Repeated Annual Cohort Study

Yu Fang, MSE; Amy S.B. Bohnert, PhD, MHS; Karina Pereira-Lima, PhD; Jennifer Cleary, MS; Elena Frank, PhD; Zhuo Zhao, MS; Walter Dempsey, PhD; and Srijan Sen, MD, PhD

Background: Efforts to address the high depression rates among training physicians have been implemented at various levels of the U.S. medical education system. The cumulative effect of these efforts is unknown.

Objective: To assess how the increase in depressive symptoms with residency has shifted over time and to identify parallel trends in factors that have previously been associated with resident physician depression.

Design: Repeated annual cohort study.

Setting: U.S. health care organizations.

Participants: First-year resident physicians (interns) who started training between 2007 and 2019.

Measurements: Depressive symptoms (9-item Patient Health Questionnaire [PHQ-9]) assessed at baseline and quarterly throughout internship.

Results: Among 16 965 interns, baseline depressive symptoms increased from 2007 to 2019 (PHQ-9 score, 2.3 to 2.9; difference, 0.6 [95% CI, 0.3 to 0.8]). The prevalence of baseline predictors of greater increase in depressive symptoms with internship also increased across cohorts. Despite the higher prevalence of baseline risk factors, the average change in depressive symptoms with internship decreased 24.4% from 2007 to 2019 (change in PHQ-9 score, 4.1 to 3.0; difference, −1.0 [CI, −1.5 to −0.6]). This change across cohorts was greater among women (4.7 to 3.3; difference, −1.4 [CI, −1.9 to −0.9]) than men (3.5 to 2.9; difference, −0.6 [CI, −1.2 to −0.05]) and greater among nonsurgical interns (4.1 to 3.0; difference, −1.1 [CI, −1.6 to −0.6]) than surgical interns (4.0 to 3.2; difference, −0.8 [CI, −1.2 to −0.4]). In parallel to the decrease in depressive symptom change, there were increases in sleep hours, quality of faculty feedback, and use of mental health services and a decrease in work hours across cohorts. The decrease in work hours was greater for nonsurgical than surgical interns. Further, the increase in mental health treatment across cohorts was greater for women than men.

Limitation: Data are observational and subject to biases due to nonrandom sampling, missing data, and unmeasured confounders, limiting causal conclusions.

Conclusion: Although depression during physician training remains high, the average increase in depressive symptoms associated with internship decreased between 2007 and 2019.

Primary Funding Source: National Institute of Mental Health.

Ann Intern Med. doi:10.7326/M21-1594 **Annals.org**
For author, article, and disclosure information, see end of text.
This article was published at Annals.org on 16 November 2021.

The past decade has seen growing recognition and concern around the substantial increase in depressive symptoms during residency training (1, 2). Low well-being among resident physicians is associated with work dissatisfaction (3), career attrition (4), and medical errors (5–8), among other negative outcomes (9–11). Although depression has become more prevalent in the general population, particularly among younger Americans (12, 13), the degree to which depression among persons entering and during residency has changed over this decade is unknown.

Initiatives across organizational levels have sought to improve well-being among residents. National organizations have enacted regulations and provided guidance (2, 14), and institutions and residency programs have implemented measures designed to promote well-being (2, 15–19). How much these initiatives have cumulatively changed intern experiences over time is unclear.

The Intern Health Study is a repeated annual cohort study of stress and depression during residency training; it enrolled cohorts of incoming residents during 2007 to 2019 at the beginning of internship and followed each cohort for 1 year. Here, we analyze trends over time in interns' reports of depressive symptom change with residency training, as well as internship factors that correlate with depressive symptom change.

METHODS

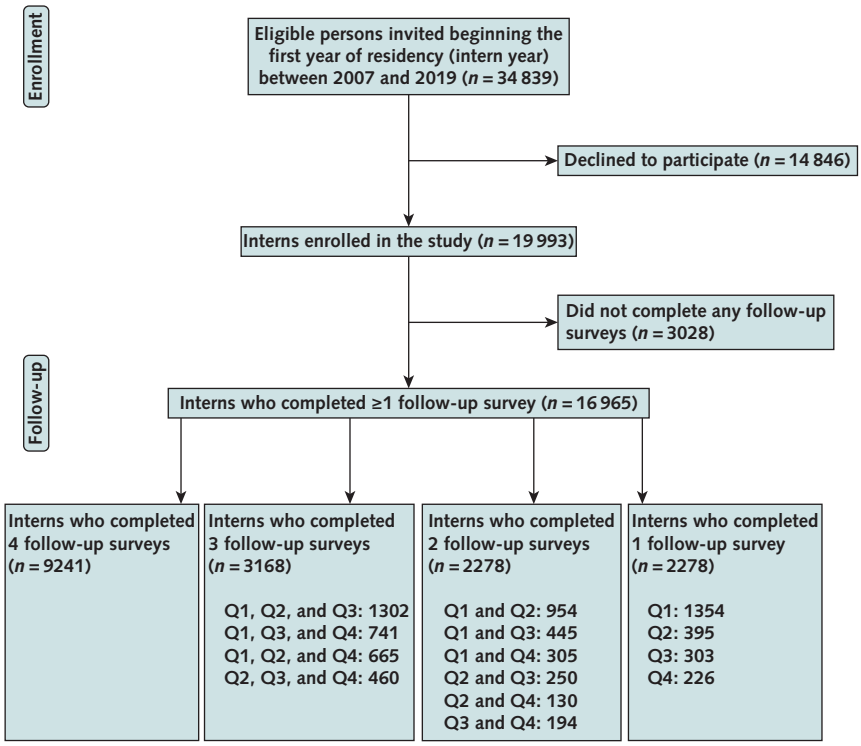
Participants

Between 2007 and 2019, a total of 34 839 incoming interns at U.S. residency institutions were invited via e-mail to participate in the study 2 to 3 months before the start of internship (1 July). In total, 19 993 participants (57.4%) consented to enroll. Figure 1 and the Appendix (available at Annals.org) provide recruitment details. The institutional review board at the University of Michigan approved the study. Participants provided informed consent and received between $50 and $125 in compensation, depending on the cohort year (20, 21).

See also:

Editorial comment

Figure 1. Study flow diagram.



Q = quarter.

Assessments

One to 3 months before the start of internship, participants completed a baseline survey assessing depressive symptoms via the 9-item Patient Health Questionnaire (PHQ-9). The PHQ-9 is a validated self-report measure of the 9 symptoms of depression, as defined by the *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition* (22). Interns indicated whether, during the previous 2 weeks, the symptom had bothered them "not at all," "several days," "more than half the days," or "nearly every day," yielding a score of 0 to 3 for each of the 9 symptom items and resulting in a total score of 0 to 27. This total score, indicating severity of depression symptoms, is consistent with the validated and recommended use of the measure (23). Total scores of 0 to 4, 5 to 9, 10 to 14, 15 to 19, and 20 to 27 correspond to minimal, mild, moderate, moderately severe, and severe depression, respectively. The diagnostic validity of the PHQ-9 is similar to that of clinician-administered assessments (23, 24). The baseline survey also assessed demographic information, including gender (woman or man), race (White, Asian, or underrepresented minority), specialty (surgical or nonsurgical), and residency institution; neuroticism (scale range, 0 to 56; NEO Five-Factor Inventory) (25); personal history of depression; and early family environment (scale range, 0 to 65; Risky Families Questionnaire) (26).

Quarterly surveys administered at months 3, 6, 9, and 12 of internship year (20) assessed the following variables: 1) PHQ-9 depressive symptoms; 2) total work hours over the past week; 3) average daily sleep hours over the past week; 4) occurrence of stressful life events unrelated to internship during the past 3 months; 5) occurrence of perceived major medical errors (5–8) during the past 3 months; 6) among participants who reported a PHQ-9 score of 10 or higher (PHQ depression) at least once during internship (27), treatment from professional mental health care services in the past 3 months (assessed since 2009 cohort); 7) timely and proper faculty feedback (28) (scale range, 0 to 5; assessed in quarter 4 only); and 8) learning experience in inpatient rotations (28) (scale range, 0 to 5; assessed in quarter 4 only). Participants were given 1 month to complete each quarterly survey.

The depressive symptom change with internship, defined as the difference between the mean of 4 quarterly internship PHQ-9 scores and the baseline PHQ-9 score, was the primary outcome measure of the study.

All surveys were done through Qualtrics, a secured website designed to maintain confidentiality, with participants identified only by identification numbers. The **Appendix** provides detailed definitions of assessed variables.

Missing Data

Participants who completed the baseline survey and at least 1 follow-up survey were included in analyses. Multiple imputation (29) was used to impute missing data from the baseline survey for all enrolled participants and from the follow-up survey for participants who

completed at least 1 follow-up survey. The Appendix provides details on the missing data imputation.

Statistical Analysis

Sample and Attrition Weighting

We generated survey weights to reduce potential bias due to nonrepresentative sampling (poststratification weights [30]) and to account for the differences between participants who completed and those who did not complete follow-up surveys (attrition weights [31]). For the poststratification weighting, the demographic composition (gender, race, and specialty) of the complete set of first-year residents in the United States between 2007 and 2019, provided by the Association of American Medical Colleges, was used as the target population. The Appendix provides details on the survey weights generation.

Trends of Change in Depressive Symptoms With Internship and Its Risk Factors

The primary outcome of the study was the change in depressive symptoms with internship. The secondary outcomes were the internship factors that have previously been found to be associated with the change in depressive symptoms with internship (20, 32–34)–specifically, the means of 4 quarterly assessments for work hours and sleep hours and the occurrence of noninternship stressful life events, self-reported medical errors, and treatment from professional mental health care services, as well as the interns' ratings of faculty feedback and learning experience in inpatient rotations in quarter 4 (Appendix Table 1, available at Annals.org).

To estimate how the average change in depressive symptoms with internship has shifted over time (represented by cohort year), we fitted a natural splines regression model for depressive symptom change with cohort year as the predictor, with robust SEs, on the survey weighted data. The survey data were clustered at the cohort year and residency institution levels (35, 36). The same models were fitted for the secondary outcomes to estimate the trends of the associated internship factors. Next, we applied marginal prediction (37, 38) with these models at the first and last cohort year of the study (2007 and 2019, respectively), with standardization on gender, race, and specialty. We then compared the average predicted depressive symptom change and internship factors in 2007 and 2019 among all participants as well as

Table 1. Baseline and Internship Characteristics of Interns (*n* = 16 965; Weighted *n* = 19 867)

Characteristic	Unweighted	Weighted
Baseline		
Median age (IQR), y	27 (26–28)	27 (26–29)
Gender, *n (%)*		
Men	8193 (48.3)	10 165 (51.2)
Women	8772 (51.7)	9702 (48.8)
Specialty, *n (%)*		
Nonsurgical	13 638 (80.4)	16 086 (81.0)
Surgical	3327 (19.6)	3781 (19.0)
Race, *n (%)*		
White	10 334 (60.9)	11 592 (58.4)
Asian	3645 (21.5)	4471 (22.5)
Underrepresented minority	2986 (17.6)	3804 (19.1)
Depressive symptoms: median PHQ-9 score (IQR)*	2 (0–4)	2 (0–4)
Depression history, *n (%)*		
No	9154 (54.0)	10 772 (54.2)
Yes	7811 (46.0)	9095 (45.8)
Neuroticism: median NEO Five-Factor Inventory score (IQR)†	22 (16–28)	22 (16–28)
Difficult early family environment: median Risky Families Questionnaire score (IQR)‡	10 (6–17)	11 (6–18)
Internship		
Average internship depressive symptoms: median PHQ-9 score (IQR)*	5 (2.8–8.3)	5 (2.8–8.3)
Median self-reported weekly work hours (IQR)	65 (55.5–73.3)	65 (56.3–74)
Median self-reported daily sleep hours (IQR)	6.5 (6–7)	6.5 (6–7)
Self-reported medical error, *n (%)*		
No	10 594 (62.4)	12 313 (62.0)
Yes	6371 (37.6)	7554 (38.0)
Stressful life event, *n (%)*		
No	9096 (53.6)	10 595 (53.3)
Yes	7869 (46.4)	9272 (46.7)
Sought mental health treatment when depressed, *n (%)*		
No	4365 (75.1)	5453 (78.1)
Yes	1447 (24.9)	1531 (21.9)
Timely and proper faculty feedback: median rating (IQR)§	4 (3–4)	4 (3–4)
Learning experience in inpatient rotations: median rating (IQR)§	4 (4–4)	4 (4–4)

IQR = interquartile range; PHQ-9 = 9-item Patient Health Questionnaire.
* Range, 0–27. Scores of 0–4 indicate minimal depression, 5–9 mild depression, 10–14 moderate depression, 15–19 moderately severe depression, and 20–27 severe depression.
† Range, 0–56.
‡ Range, 0–65.
§ Range, 1–5.

Table 2. Temporal Changes in the Annual Average Baseline Risk Factors of Incoming Interns*

Baseline Risk Factor	Expected Value		Difference 2019 vs. 2007
	2007	2019	
Mean depressive symptom score at baseline (95% CI)	2.3 (2.2 to 2.4)	2.9 (2.6 to 3.1)	0.6 (0.3 to 0.8)
Percentage women (95% CI)	47.9 (46.3 to 49.5)	50.0 (49.4 to 50.7)	2.2 (0.4 to 3.9)
Mean neuroticism score (95% CI)	22.7 (21.8 to 23.5)	23.5 (23.1 to 23.8)	0.8 (−0.1 to 1.7)
Percentage with depression history (95% CI)	45.5 (43.7 to 47.2)	48.5 (47.7 to 49.3)	3.1 (1.2 to 5.0)
Mean early family environment score (95% CI)	13.8 (13.3 to 14.3)	13.5 (13.3 to 13.6)	−0.3 (−0.8 to 0.2)

* Expected values of 2007 and 2019 were estimated from survey weighted natural splines regression models based on all data from 2007 to 2019, with the baseline risk factors as the outcome; the cohort year as the predictor; and standardization on gender, race, and specialty.

subgroups stratified by gender (women and men) and specialty (surgical and nonsurgical). We also graphically compared changes across all years for all outcomes between women and men and between surgical and nonsurgical interns to explore effect modification based on stratified marginal predictions. When the trajectory differed between groups, we tested for the presence of interaction in the natural splines regression model. For binary outcomes, we used relative excess risk due to interaction (39) to test for additive interactions.

To understand the degree to which baseline risk factors for change in depressive symptoms evolved over the years, we used the same approach to test for changes across time for the baseline risk factors (baseline depressive symptoms, baseline neuroticism, depression history, and difficult early family environment) (Appendix Table 1) for all participants.

Sensitivity Analyses

The intern year of the 2019 cohort lasted from July 2019 to June 2020, during which time the COVID-19 pandemic outbreak began in the United States. To assess whether the study findings were influenced by the pandemic, we compared the depressive symptom change with internship between the 2007 and 2018 cohorts.

Second, to evaluate the possibility that the trend in depressive symptom change was driven by the trend in baseline PHQ-9 score of incoming interns, we assessed for the presence of a ceiling effect by comparing the skewness (40) of distributions of internship PHQ-9 scores across the cohort years.

Third, the participating residency institutions varied during the years covered by this study (Appendix Figure 1, available at Annals.org). Institution factors likely affect residents' depressive symptoms during internship, and the change in the composition of institutions represented in the sample could have influenced findings. To assess whether patterns observed in the full sample hold in a fixed set of institutions, we restricted our analysis to participants from residency institutions that were represented in every cohort between 2008 and 2019 (called here the "common institution subsample").

Finally, we assessed the trend in depressive symptom change including only participants who finished all 4 follow-up surveys to explore the potential for bias introduced by missing follow-up surveys.

Statistical analyses were done using R, version 4.0.1 (R Foundation). Additional details of the statistical analysis are given in the Appendix.

Role of the Funding Source

The National Institute of Mental Health had no role in the design, conduct, or analysis of the study or the decision to submit the manuscript for publication.

RESULTS

Sample Characteristics

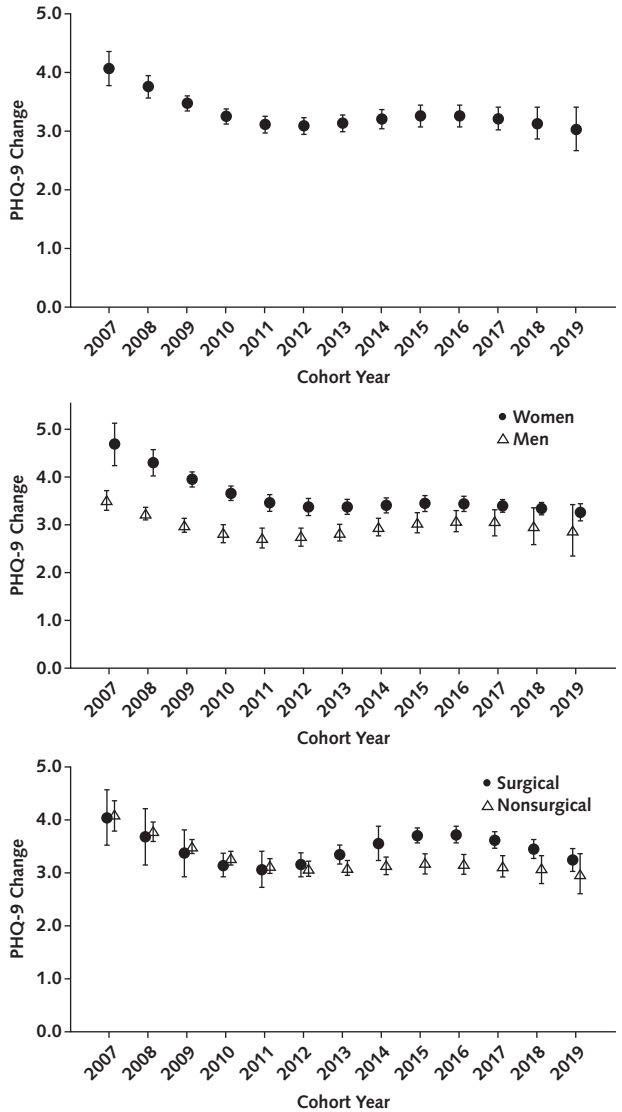
Across 13 cohorts between 2007 and 2019, a total of 16 965 participants (84.9% of enrolled participants; mean age, 27.6 years [SD, 0.04]; 48.8% women; 19.0% surgical interns) from 645 residency institutions completed the baseline survey and at least 1 follow-up survey during the intern year and were included in the analyses (Figure 1 and Table 1; Appendix Tables 2 and 3, available at Annals.org).

Trends of Change in Depressive Symptoms With Internship and Its Risk Factors

Mean depressive symptom score at baseline, measured before the start of internship, increased from 2.3 in 2007 to 2.9 in 2019 (difference, 0.6 [95% CI, 0.3 to 0.8]) (Table 2; Appendix Figure 2, available at Annals.org). Among the baseline factors previously established as predictors of PHQ-9 score increase during residency training, 2 factors changed significantly across cohorts, both in the direction of predicting greater increases in depressive symptoms with internship. These were the proportion of women (47.9% to 50.0%; difference, 2.2 percentage points [CI, 0.4 to 3.9 percentage points]) and the proportion with a history of depression (45.5% to 48.5%; difference, 3.1 percentage points [CI, 1.2 to 5.0 percentage points]).

Based on expected values generated from marginal predictions standardized for intern gender, race, and specialty, the magnitude of the change in PHQ-9 depressive symptoms with internship decreased from 4.1 in 2007 to 3.0 in 2019 (24.4% decrease; difference, −1.0 [CI, −1.5 to −0.6]) (Figure 2 [*top*] and Table 3). The change across cohorts was not consistent over time: The rate of decrease was greater during the 2007-to-2011 period than afterward (Figure 2 [*top*]; Appendix Table 4, available at Annals.org). Four internship factors also changed significantly across cohorts, all in the direction of predicting less increase in depressive symptoms with internship (Table 3): Weekly work hours decreased by 11.3% (69.0 to 61.2 hours; difference, −7.7 hours [CI, −9.8 to −5.6 hours]), daily sleep hours increased by 4.8% (6.2 to 6.5 hours; difference, 0.2 hour [CI, 0.1 to 0.4 hour]), interns' rating of the timeliness and quality of faculty feedback

Figure 2. The weighted, model-based, expected changes in depressive symptoms with internship across cohorts: full sample.



The natural splines fitting of annual average change in PHQ-9 depressive symptoms with internship between 2007 and 2019 cohorts among all participants (*top*), stratified by gender (*middle*), and stratified by specialty (*bottom*). Error bars represent 95% CIs. PHQ-9 = 9-item Patient Health Questionnaire.

increased by 8.0% (3.2 to 3.4; difference, 0.3 [CI, 0.2 to 0.3]), and the prevalence of participants with depression getting professional mental health care treatment increased by 165.7% (14.3% to 38.0%; difference, 23.7 percentage points [CI, 21.2 to 26.1 percentage points]). Of note, the use of most forms of mental health services increased: institutional employee assistance program (3.0% to 14.4%; difference, 11.3 percentage points [CI, 9.4 to 13.3 percentage points]), general practitioners (3.9% to 15.3%; difference, 11.4 percentage points [CI, 7.8 to 15.0 percentage points]), and therapists (7.7% to 27.0%; difference, 19.3 percentage points [CI, 18.5 to 20.1 percentage points]). The exception to this trend was psychiatric hospitalizations. The occurrence of stressful life events unrelated to internship, the occurrence of medical errors, and the interns' evaluation of learning experience in inpatient rotations did not change significantly.

Comparisons Between Gender and Specialty Groups

Although the decrease in PHQ-9 depressive symptoms change with internship over time was significant for both genders (women: 4.7 to 3.3; difference, −1.4 [CI, −1.9 to −0.9]; men: 3.5 to 2.9; difference, −0.6 [CI, −1.2 to −0.05]) (**Figure 2** [*middle*] and **Table 3**), the decrease across cohorts in the depression symptom change with internship was significantly greater among women than men. Further, seeking of mental health treatment increased more across cohorts for women than men (women: 13.0% to 43.6%; difference, 30.5 percentage points [CI, 24.2 to 36.9 percentage points]; men: 15.9% to 31.1%; difference, 15.2 percentage points [CI, 9.6 to 20.8 percentage points]) (**Table 3**; **Appendix Figure 3** [*E*], available at Annals.org).

Although both specialty groups had a significant decrease in PHQ-9 depressive symptom change with internship over time (nonsurgical: 4.1 to 3.0; difference, −1.1 [CI, −1.6 to −0.6]; surgical: 4.0 to 3.2; difference, −0.8 [CI, −1.2 to −0.4]) (**Figure 2** [*bottom*] and **Table 3**), the change associated with internship across cohorts was greater among nonsurgical interns than surgical interns. Among the internship factors, there was a significantly greater decrease in work hours across cohorts for nonsurgical interns than surgical interns (nonsurgical: 67.9 to 58.7 hours; difference, −9.2 hours [CI, −11.1 to −7.4 hours]; surgical: 73.6 to 71.6 hours; difference, −2.0 hours [CI, −7.9 to 3.9 hours]) (**Table 3**; **Appendix Figure 4** [*A*], available at Annals.org).

Sensitivity Analyses

In the sensitivity analysis assessing whether the result was influenced by the COVID-19 pandemic, the change in depressive symptom increase associated with internship between the 2007 and 2018 cohorts was −0.9 (CI, −1.3 to −0.5), similar to that found between the 2007 and 2019 cohorts.

We assessed the skewness of the distribution of internship PHQ-9 symptom scores across cohorts and found no evidence for the presence of a ceiling effect (change in skewness per year = 0.0022 [CI, −0.034 to 0.038]; $P = 0.89$).

A total of 24 residency institutions were included in all cohorts from 2008 to 2019. The 5519 participants (weighted mean age, 27.6 years [SD, 0.04]; 49.2% women) (**Appendix Table 5**, available at Annals.org) who attended these institutions composed the "common institution subsample." As in the full sample, baseline depressive symptoms increased significantly (2.4 to 2.8; difference, 0.4 [CI, 0.2 to 0.7]) (**Appendix Figure 5**, available at Annals.org) across cohorts. Similarly, the magnitude of the increase in depressive symptoms with internship declined significantly from 2008 to 2019 in this subsample (4.0 to 3.3; difference, −0.7 [CI, −1.0 to −0.3]) (**Appendix Figure 6**, available at

Table 3. Trends of Annual Average Depressive Symptom Change With Internship and Internship Factors From 2007 to 2019*

Depressive Symptom Change and Internship Factor	Expected Value		Difference 2019 vs. 2007†
	2007	2019	
Mean change in depressive symptom score (95% CI)			
All interns	4.1 (3.8 to 4.4)	3.0 (2.7 to 3.4)	−1.0 (−1.5 to −0.6)
Men	3.5 (3.3 to 3.7)	2.9 (2.3 to 3.4)	−0.6 (−1.2 to −0.05)
Women	4.7 (4.2 to 5.1)	3.3 (3.1 to 3.4)	−1.4 (−1.9 to −0.9)
Surgical interns	4.0 (3.7 to 4.4)	3.2 (3.0 to 3.5)	−0.8 (−1.2 to −0.4)
Nonsurgical interns	4.1 (3.7 to 4.4)	3.0 (2.6 to 3.4)	−1.1 (−1.6 to −0.6)
Mean weekly work hours (95% CI)			
All interns	69.0 (67.9 to 70.1)	61.2 (59.5 to 63.1)	−7.7 (−9.8 to −5.6)
Men	69.1 (67.8 to 70.4)	62.1 (60.1 to 64.1)	−7.0 (−9.4 to −4.6)
Women	69.1 (67.9 to 70.2)	60.5 (59.2 to 61.9)	−8.5 (−10.3 to −6.7)
Surgical interns	73.6 (68.4 to 78.9)	71.6 (69.0 to 74.3)	−2.0 (−7.9 to 3.9)
Nonsurgical interns	67.9 (66.8 to 69.0)	58.7 (57.2 to 60.2)	−9.2 (−11.1 to −7.4)
Mean daily sleep hours (95% CI)			
All interns	6.2 (6.2 to 6.3)	6.5 (6.4 to 6.6)	0.2 (0.1 to 0.4)
Men	6.2 (6.2 to 6.3)	6.4 (6.3 to 6.5)	0.2 (0.04 to 0.3)
Women	6.3 (6.2 to 6.3)	6.6 (6.5 to 6.7)	0.3 (0.2 to 0.4)
Surgical interns	6.0 (5.8 to 6.3)	6.1 (6.0 to 6.2)	0.1 (−0.1 to 0.4)
Nonsurgical interns	6.3 (6.3 to 6.3)	6.6 (6.4 to 6.7)	0.3 (0.1 to 0.4)
Percentage of interns with stressful life events during internship (95% CI)			
All interns	44.3 (37.2 to 51.3)	44.9 (43.7 to 46.0)	0.6 (−11.1 to 12.3)
Men	43.1 (36.2 to 50.1)	43.9 (42.2 to 45.6)	0.7 (−6.4 to 8.0)
Women	46.2 (38.3 to 54.1)	44.8 (42.5 to 47.1)	−1.4 (−9.6 to 6.9)
Surgical interns	49.3 (47.0 to 51.7)	42.7 (42.0 to 43.3)	−6.7 (−9.1 to −4.2)
Nonsurgical interns	43.2 (34.0 to 52.5)	45.4 (43.7 to 47.0)	2.1 (−7.3 to 11.6)
Percentage of interns reporting medical errors (95% CI)			
All interns	39.1 (34.7 to 43.5)	40.7 (35.0 to 46.4)	1.6 (−10.1 to 13.3)
Men	39.7 (35.3 to 44.2)	39.9 (33.9 to 46.0)	0.2 (−7.3 to 7.7)
Women	39.2 (34.2 to 44.2)	41.1 (35.7 to 46.4)	1.9 (−5.5 to 9.2)
Surgical interns	37.5 (32.1 to 42.9)	42.4 (34.8 to 50.0)	4.9 (−4.4 to 14.2)
Nonsurgical interns	39.7 (35.3 to 44.2)	40.0 (34.8 to 45.3)	0.3 (−6.6 to 7.1)
Percentage of depressed interns receiving mental health treatment (95% CI)†			
All interns	14.3 (12.2 to 16.4)	38.0 (36.7 to 39.3)	23.7 (21.2 to 26.1)
Men	15.9 (12.4 to 19.4)	31.1 (26.7 to 35.5)	15.2 (9.6 to 20.8)
Women	13.0 (7.1 to 19.0)	43.6 (41.3 to 45.8)	30.5 (24.2 to 36.9)
Surgical interns	20.5 (17.5 to 23.4)	32.7 (31.8 to 33.6)	12.2 (9.1 to 15.3)
Nonsurgical interns	12.9 (10.9 to 14.9)	39.2 (37.3 to 41.2)	26.3 (23.5 to 29.1)
Mean faculty feedback (95% CI)			
All interns	3.2 (3.1 to 3.2)	3.4 (3.4 to 3.5)	0.3 (0.2 to 0.3)
Men	3.2 (3.1 to 3.3)	3.4 (3.4 to 3.5)	0.2 (0.1 to 0.3)
Women	3.1 (3.1 to 3.2)	3.4 (3.4 to 3.5)	0.3 (0.2 to 0.4)
Surgical interns	3.0 (2.9 to 3.1)	3.3 (3.2 to 3.3)	0.2 (0.1 to 0.3)
Nonsurgical interns	3.2 (3.2 to 3.3)	3.5 (3.4 to 3.5)	0.3 (0.2 to 0.3)
Mean inpatient rotation learning experience (95% CI)			
All interns	3.9 (3.8 to 4.0)	3.9 (3.9 to 3.9)	0.03 (−0.07 to 0.1)
Men	3.9 (3.8 to 4.1)	3.9 (3.8 to 3.9)	−0.1 (−0.2 to 0.1)
Women	3.8 (3.8 to 3.9)	4.0 (4.0 to 4.0)	0.1 (0.1 to 0.2)
Surgical interns	3.8 (3.7 to 3.9)	4.0 (3.9 to 4.0)	0.1 (0 to 0.2)
Nonsurgical interns	3.9 (3.8 to 4.0)	3.9 (3.9 to 3.9)	−0.01 (−0.1 to 0.1)

* Expected values for 2007 and 2019 were estimated from survey weighted natural splines regression models based on all data from 2007 to 2019, with depression symptom change or the internship factors as the outcome; the cohort year as the predictor; and standardization on gender, race, and specialty.
† Receipt of mental health treatment has been assessed since the 2009 cohort, so comparisons for this outcome are for 2019 vs. 2009.

Annals.org). Four internship factors changed significantly across cohorts in the subsample: average weekly work hours decreased (68.6 to 61.7 hours; difference, −6.9 hours [CI, −10.2 to −3.6 hours]), sleep hours increased (6.3 to 6.5 hours; difference, 0.2 hour [CI, 0.1 to 0.4 hour]), treatment among participants with depression increased (14.2% to 39.6%; difference, 25.4 percentage points [CI, 17.4 to 33.4 percentage points]), and rating of faculty feedback increased (3.2 to 3.4; difference, 0.2 [CI, 0.04 to 0.2]).

Among participants who completed all 4 follow-up surveys (n = 9241 [54.5% of the full sample]) (**Appendix Table 6**, available at Annals.org), the depressive

symptom change with internship from 2007 to 2019 was 3.8 to 2.9, a difference of −0.9 (CI, −1.5 to −0.3), similar to that in the full sample.

Discussion

In this repeated annual cohort study comparing physicians during residency training between 2007 and 2019, we found that the degree of increase in depressive symptoms with internship has decreased substantially over time, suggesting that the negative effect of medical training on mental health may have lessened in recent years. Of note, this occurred despite recent intern cohorts having higher baseline risk for depressive symptom increases with internship, suggesting that changes in the residency experience over the years may have been important drivers of the observed reduction in depressive symptom change with internship.

Over the past decade, the prevalence of depression among younger Americans in the general population has been growing (13). In particular, the proportion of U.S. college students who screened positive for depression increased by 20% between 2009 and 2017 (41). Consistent with this national trend, depressive symptoms and depression history among training physicians entering their intern year became more prevalent between 2007 and 2019.

Among internship factors previously associated with the change in depressive symptoms during internship, we identified 4 that changed significantly from 2007 to 2019: decreased work hours, increased quality of faculty feedback, longer total sleep time, and increased mental health treatment. Thus, these 4 factors merit further examination as potential drivers of the decline in depressive symptom change observed in this study. Work hours have long been established as an important factor related to the well-being of training physicians and the general population (42, 43). Consistent with other work, we found that the trend of a smaller average increase in depressive symptoms with internship across cohorts was concurrent with a decrease of 7 to 8 work hours per week. Future studies are needed to assess the extent to which further reduction in work hours will drive more mental health improvements among training physicians. Our findings similarly confirm the established association between short sleep and depression and support further efforts to increase sleep duration in training physicians. We also identified substantial changes in the use of mental health services: Interns with depression were almost 3 times more likely to receive formal treatment in 2019 (38.0%) than in 2007 (14.3%). Overall, because more than half of interns with depression still do not receive treatment in 2019, efforts to increase access to care and reduce stigma around treatment could further improve intern mental health.

For program directors and institutions, these findings suggest that efforts to reduce workload, improve faculty feedback, and decrease barriers to care have had meaningful effects on trainee mental health. The changing profile of incoming physicians toward greater risk for increased depression during training underscores the need to continue these efforts. Given the established association between well-being of physicians and quality of care for patients, future studies should examine whether this trend toward improved mental health translates into improved patient outcomes (5, 7, 44).

Of note, the improvement in mental health was not experienced equally by all interns. Specifically, declines in the degree of depressive symptom increase with internship were significantly greater for nonsurgical interns than surgical interns. This difference in depressive symptom trajectory corresponds with a difference in work hours trajectory, with work hours decreasing significantly more among nonsurgical interns. These results underscore the importance of developing reforms that achieve work hours reduction among training surgeons without harming education. Further evaluating cultural differences between specialties may also elucidate the causes of this disparity (45).

The trend of change in depressive symptoms with internship also differed by gender. We found that the decline over time in the development of internship depressive symptoms was significantly greater for women than men. In parallel to this finding, women increased their use of mental health services during the intern year at twice the rate men did. This trend is in line with gender differences in seeking of mental health care in the general population (46–48). These findings highlight the importance of considering gender in addressing barriers to mental health care seeking among training physicians. The increasing proportion of women entering the physician workforce (49) and changes in the culture of medicine over this 13-year period may also have contributed to the observed decreasing gender difference in depression over time.

This study has several limitations. First, because it was an observational study with convenience sampling, we could not make definitive conclusions about the causal mechanisms driving changes in depressive symptoms over time. Of note, there may have been unmeasured individual, institutional, and societal factors that changed over time and affected the mental health of residents. Second, missing data and study attrition may have biased the results, although the risk for this bias is attenuated, but not completely addressed, by imputation and attrition weighting. Of note, the sensitivity analysis using the subsample with data for all follow-up assessments also did not indicate a high risk for this bias. Third, the number of residents and institutions enrolled in the study has expanded over the years, complicating the comparison of trend estimates between years. However, the sensitivity analysis using the common institution subsample showed that the trends were similar in a fixed set of institutions. Fourth, the sample is restricted to first-year residents; mental health trends may differ among other groups of trainees and physicians. Last, although the data on mental health treatment showed a clear trend over time, they were not collected in the first 2 cohorts of the study.

In summary, across 13 annual cohorts of the Intern Health Study, we found that the increase in depressive symptoms with internship decreased substantially over successive cohorts. These findings represent important

progress in reducing the depression associated with residency training. Despite this progress, depressive symptoms still increase substantially with internship for many trainees. Our findings on the identified internship factors associated with reduced change in depressive symptoms, such as lower work hours, and the specific subpopulations who have had less improvement over time can help guide next steps to further improve mental health among residents.

From Michigan Neuroscience Institute, University of Michigan, Ann Arbor, Michigan (Y.F., E.F., Z.Z.); Departments of Anesthesiology and Epidemiology, University of Michigan, and VA Center for Clinical Management Research, Ann Arbor, Michigan (A.S.B.); Department of Psychiatry, University of Michigan, Ann Arbor, Michigan, and Department of Psychiatry, Federal University of São Paulo, São Paulo, Brazil (K.P.); Michigan Neuroscience Institute and Department of Psychology, University of Michigan, Ann Arbor, Michigan (J.C.); Department of Biostatistics and Institute of Social Research, University of Michigan, Ann Arbor, Michigan (W.D.); and Michigan Neuroscience Institute and Department of Psychiatry, University of Michigan, Ann Arbor, Michigan (S.S.).

Acknowledgment: The authors thank the training physicians for taking part in this study.

Grant Support: By grant R01MH101459 from the National Institute of Mental Health. Dr. Pereira-Lima was supported by a research fellowship grant 2018/21480-4 from São Paulo Research Foundation. Ms. Cleary was supported by grant T32HD007109 from the National Institutes of Health.

Disclosures: Authors have reported no disclosures of interest. Forms can be viewed at www.acponline.org/authors/icmje/ConflictOfInterestForms.do?msNum=M21-1594.

Reproducible Research Statement: *Study protocol*: Further details are available from Dr. Sen (e-mail, srijan@umich.com). *Statistical code*: Statistical code supporting the analysis will be made available in GitHub 18 months after 2019 cohort completion. *Data set*: A deidentified data set supporting the results in this analysis will be made available in ICPSR (www.icpsr.umich.edu) 18 months after 2019 cohort completion.

Corresponding Author: Srijan Sen, MD, PhD, Michigan Neuroscience Institute, University of Michigan, 205 Zina Pitcher Place, Ann Arbor, MI 48109; e-mail, srijan@umich.edu.

Author contributions are available at Annals.org.

References

1. **Mata DA, Ramos MA, Bansal N, et al.** Prevalence of depression and depressive symptoms among resident physicians: a systematic review and meta-analysis. JAMA. 2015;314:2373-83. [PMID: 26647259] doi:10.1001/jama.2015.15845

2. **National Academies of Sciences, Engineering, and Medicine.** Taking Action Against Clinician Burnout: A Systems Approach to Professional Well-Being. National Academies Pr; 2019. doi:10.17226/25521

3. **Becker JL, Milad MP, Klock SC.** Burnout, depression, and career satisfaction: cross-sectional study of obstetrics and gynecology residents. Am J Obstet Gynecol. 2006;195:1444-9. [PMID: 17074551]

4. **Williams ES, Konrad TR, Scheckler WE, et al.** Understanding physicians' intentions to withdraw from practice: the role of job satisfaction, job stress, mental and physical health. 2001. Health Care Manage Rev. 2010;35:105-15. [PMID: 20234217] doi:10.1097/01.HMR.0000304509.58297.6f

5. **Pereira-Lima K, Mata DA, Loureiro SR, et al.** Association between physician depressive symptoms and medical errors: a systematic review and meta-analysis. JAMA Netw Open. 2019;2:e1916097. [PMID: 31774520] doi:10.1001/jamanetworkopen.2019.16097

6. **Fahrenkopf AM, Sectish TC, Barger LK, et al.** Rates of medication errors among depressed and burnt out residents: prospective cohort study. BMJ. 2008;336:488-91. [PMID: 18258931] doi:10.1136/bmj.39469.763218.BE

7. **West CP, Huschka MM, Novotny PJ, et al.** Association of perceived medical errors with resident distress and empathy: a prospective longitudinal study. JAMA. 2006;296:1071-8. [PMID: 16954486]

8. **West CP, Tan AD, Habermann TM, et al.** Association of resident fatigue and distress with perceived medical errors. JAMA. 2009;302:1294-300. [PMID: 19773564] doi:10.1001/jama.2009.1389

9. **Tyssen R, Vaglum P, Grønvold NT, et al.** Suicidal ideation among medical students and young physicians: a nationwide and prospective study of prevalence and predictors. J Affect Disord. 2001;64:69-79. [PMID: 11292521]

10. **West CP, Tan AD, Shanafelt TD.** Association of resident fatigue and distress with occupational blood and body fluid exposures and motor vehicle incidents. Mayo Clin Proc. 2012;87:1138-44. [PMID: 23218084] doi:10.1016/j.mayocp.2012.07.021

11. **de Oliveira GS Jr, Chang R, Fitzgerald PC, et al.** The prevalence of burnout and depression and their association with adherence to safety and practice standards: a survey of United States anesthesiology trainees. Anesth Analg. 2013;117:182-93. [PMID: 23687232] doi:10.1213/ANE.0b013e3182917da9

12. **Global Burden of Disease Study 2013 Collaborators.** Global, regional, and national incidence, prevalence, and years lived with disability for 301 acute and chronic diseases and injuries in 188 countries, 1990-2013: a systematic analysis for the Global Burden of Disease Study 2013. Lancet. 2015;386:743-800. [PMID: 26063472] doi:10.1016/S0140-6736(15)60692-4

13. **Twenge JM, Cooper AB, Joiner TE, et al.** Age, period, and cohort trends in mood disorder indicators and suicide-related outcomes in a nationally representative dataset, 2005-2017. J Abnorm Psychol. 2019;128:185-199. [PMID: 30869927] doi:10.1037/abn0000410

14. **Accreditation Council for Graduate Medical Education.** ACGME Common Program Requirements. 2017. Accessed athttps://acgme.org/Portals/0/PFAssets/ProgramRequirements/CPRs_2017-07-01.pdf on 21 September 2021.

15. **Aggarwal R, Deutsch JK, Medina J, et al.** Resident wellness: an intervention to decrease burnout and increase resiliency and happiness. MedEdPORTAL. 2017;13:10651. [PMID: 30800852] doi:10.15766/mep_2374-8265.10651

16. **Cheston CC, Sox CM, Michelson CD, et al.** MINdI: mindfulness instruction for new interns. MedEdPORTAL. 2020;16:10933. [PMID: 32754632] doi:10.15766/mep_2374-8265.10933

17. **Fischer J, Alpert A, Rao P.** Promoting intern resilience: individual chief wellness check-ins. MedEdPORTAL. 2019;15:10848. [PMID: 31921994] doi:10.15766/mep_2374-8265.10848

18. **Salles A, Liebert CA, Esquivel M, et al.** Perceived value of a program to promote surgical resident well-being. J Surg Educ. 2017;74:921-927. [PMID: 28457875] doi:10.1016/j.jsurg.2017.04.006

19. **Jennings ML, Slavin SJ.** Resident wellness matters: optimizing resident education and wellness through the learning environment. Acad Med. 2015;90:1246-50. [PMID: 26177527] doi:10.1097/ACM.0000000000000842

20. **Sen S, Kranzler HR, Krystal JH, et al.** A prospective cohort study investigating factors associated with depression during medical internship. Arch Gen Psychiatry. 2010;67:557-65. [PMID: 20368500] doi:10.1001/archgenpsychiatry.2010.41
21. **Guille C, Zhao Z, Krystal J, et al.** Web-based cognitive behavioral therapy intervention for the prevention of suicidal ideation in medical interns: a randomized clinical trial. Mo Med. 2016;113:19. [PMID: 30228438]
22. **American Psychiatric Association.** Diagnostic and Statistical Manual of Mental Disorders. 5th ed. American Psychiatric Assoc; 2013.
23. **Kroenke K, Spitzer RL, Williams JB.** The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med. 2001;16:606-13. [PMID: 11556941]
24. **Spitzer RL, Kroenke K, Williams JB.** Validation and utility of a self-report version of PRIME-MD: the PHQ primary care study. JAMA. 1999;282:1737-44. [PMID: 10568646]
25. **Costa PT Jr, McCrae RR..** Stability and change in personality assessment: the revised NEO Personality Inventory in the year 2000. J Pers Assess. 1997;68:86-94. [PMID: 9018844]
26. **Taylor SE, Way BM, Welch WT, et al.** Early family environment, current adversity, the serotonin transporter promoter polymorphism, and depressive symptomatology. Biol Psychiatry. 2006;60:671-6. [PMID: 16934775]
27. **Levis B, Benedetti A, Thombs BD; DEPRESsion Screening Data (DEPRESSD) Collaboration.** Accuracy of Patient Health Questionnaire-9 (PHQ-9) for screening to detect major depression: individual participant data meta-analysis. BMJ. 2019;365:l1476. [PMID: 30967483] doi:10.1136/bmj.l1476
28. **Seelig CB, DuPre CT, Adelman HM.** Development and validation of a scaled questionnaire for evaluation of residency programs. South Med J. 1995;88:745-50. [PMID: 7597480]
29. **Little RJA, Rubin DB.** Statistical Analysis With Missing Data. 2nd ed. J Wiley; 2002. doi:10.1002/9781119013563
30. **DeBell M.** Best practices for creating survey weights. In: Vannette DL, Krosnick JA, eds. The Palgrave Handbook of Survey Research. 2018;159-62. doi:10.1007/978-3-319-54395-6_21
31. **Griffin BA, Ridgeway G, Morral AM, et al.** Toolkit for Weighting and Analysis of Nonequivalent Groups (TWANG). RAND Corporation; 2014. Accessed atwww.rand.org/statistics/twang.html on 10 June 2021.
32. **Hollon SD, Thase ME, Markowitz JC.** Treatment and prevention of depression. Psychol Sci Public Interest. 2002;3:39-77. [PMID: 26151569] doi:10.1111/1529-1006.00008
33. **Dobson KS, Hollon SD, Dimidjian S, et al.** Which treatment approach is most effective in the prevention of relapse and recurrence of major depression? PsycEXTRA Dataset. 2008. doi:10.1037/e558102009-008
34. **Pereira-Lima K, Gupta RR, Guille C, et al.** Residency program factors associated with depressive symptoms in internal medicine interns: a prospective cohort study. Acad Med. 2019;94:869-875. [PMID: 30570500] doi:10.1097/ACM.0000000000002567
35. **Korn EL, Graubard BI.** Analysis of Health Surveys. J Wiley; 1999. doi:10.1002/9781118032619
36. **Lumley T.** Analysis of complex survey samples. J Stat Softw. 2004;9:1-19. doi:10.18637/jss.v009.i08
37. **Graubard BI, Korn EL.** Predictive margins with survey data. Biometrics. 1999;55:652-9. [PMID: 11318229]
38. **Witt M, Spagnola K.** Using predictive marginals to produce standardized estimates. 2009. Accessed atwww.semanticscholar.org/paper/Using-Predictive-Marginals-to-Produce-Standardized-Witt-Spagnola/9030646d9a8102a674ddbff065aac5831e81670d on 29 June 2021.
39. **Mathur MB, VanderWeele TJ.** R function for additive interaction measures [Letter]. Epidemiology. 2018;29:e5-e6. [PMID: 28901974] doi:10.1097/EDE.0000000000000752
40. **Koedel C, Betts J.** Value added to what? How a ceiling in the testing instrument influences value-added estimation. Educ Finance Policy. 2010;5:54-81. doi:10.1162/edfp.2009.5.1.5104
41. **Lipson SK, Lattie EG, Eisenberg D.** Increased rates of mental health service utilization by U.S. college students: 10-year population-level trends (2007-2017). Psychiatr Serv. 2019;70:60-63. [PMID: 30394183] doi:10.1176/appi.ps.201800332
42. **Valko RJ, Clayton PJ.** Depression in the internship. Dis Nerv Syst. 1975;36:26-9. [PMID: 1109883]
43. **Pega F, Náfrádi B, Momen NC, et al; Technical Advisory Group.** Global, regional, and national burdens of ischemic heart disease and stroke attributable to exposure to long working hours for 194 countries, 2000-2016: a systematic analysis from the WHO/ILO Joint Estimates of the Work-related Burden of Disease and Injury. Environ Int. 2021;154:106595. [PMID: 34011457] doi:10.1016/j.envint.2021.106595
44. **Vigo D, Thornicroft G, Atun R.** Estimating the true global burden of mental illness. Lancet Psychiatry. 2016;3:171-8. [PMID: 26851330] doi:10.1016/S2215-0366(15)00505-2
45. **Fu WW, Gauger PG, Newman EA.** Mental illness and stigma in surgical residencies—an unspoken truth. JAMA Surg. 2021;156:117-118. [PMID: 33052384] doi:10.1001/jamasurg.2020.2965
46. **Magaard JL, Seeralan T, Schulz H, et al.** Factors associated with help-seeking behaviour among individuals with major depression: a systematic review. PLoS One. 2017;12:e0176730. [PMID: 28493904] doi:10.1371/journal.pone.0176730
47. **Seidler ZE, Dawes AJ, Rice SM, et al.** The role of masculinity in men's help-seeking for depression: a systematic review. Clin Psychol Rev. 2016;49:106-118. [PMID: 27664823] doi:10.1016/j.cpr.2016.09.002
48. **Galdas PM, Cheater F, Marshall P.** Men and health help-seeking behaviour: literature review. J Adv Nurs. 2005;49:616-23. [PMID: 15737222]
49. **Barzansky B, Etzel SI.** Medical schools in the United States, 2017-2018. JAMA. 2018;320:1042-1050. [PMID: 30208435] doi:10.1001/jama.2018.11679

Author Contributions: Conception and design: Y. Fang, A.S.B. Bohnert, E. Frank, S. Sen.
Analysis and interpretation of the data: Y. Fang, K. Pereira-Lima, E. Frank, Z. Zhao, W. Dempsey, S. Sen.
Drafting of the article: Y. Fang.
Critical revision of the article for important intellectual content: Y. Fang, A.S.B. Bohnert, K. Pereira-Lima, J. Cleary, E. Frank, S. Sen.
Final approval of the article: Y. Fang, A.S.B. Bohnert, K. Pereira-Lima, J. Cleary, E. Frank, Z. Zhao, W. Dempsey, S. Sen.
Provision of study materials or patients: S. Sen.
Statistical expertise: Y. Fang, A.S.B. Bohnert, W. Dempsey.
Obtaining of funding: S. Sen.
Administrative, technical, or logistic support: S. Sen.
Collection and assembly of data: E. Frank, Z. Zhao, S. Sen.

Appendix: Appendix Methods and Results

Recruitment

Each year after the residency match (mid-March), a convenience sample of U.S. health care institutions provided names and e-mail addresses of graduating medical school students or incoming interns. Additional lists of names and e-mail addresses of eligible incoming interns were gathered from publicly available medical school match lists.

Variable Definitions

Race

The self-reported races are categorized into White, Asian, and underrepresented minorities in the analysis. Underrepresented minorities include African, Latino/a, Native American, Pacific Islander, Arab/Middle Eastern, other, and mixed.

Surgical and Nonsurgical Specialties

Surgical specialties were assigned on the basis of the American College of Surgeons classification (50). Specifically, for this study, interns who were in the following specialties were classified as "surgical": general surgery, gynecology and obstetrics, neurological surgery, ophthalmic surgery, orthopedic surgery, otolaryngology, plastic surgery-integrated, urology, vascular surgery-integrated, and surgery preliminary.

Interns from the following specialties were classified as nonsurgical: internal medicine, pediatrics, psychiatry, neurology, emergency medicine, combined internal medicine and pediatrics, family practice, anesthesiology, transitional, and other nonsurgical.

Neuroticism

The neuroticism subscale of the NEO Five-Factor Inventory (25) is a 14-item scale assessing negative emotional constructs, such as anxiety, moodiness, worry, envy, and jealousy. Each evaluation yields a score from 0 (strongly disagree) to 4 (strongly agree), resulting in a total score of 0 to 56. Higher scores on the neuroticism subscale indicate an increased likelihood of experiencing negative emotions, especially in response to environmental stress.

Early Family Environment

The Risky Families Questionnaire (26) evaluates the family environment during childhood and early adolescence (age 5 to 15 years). A total of 13 questions assess the frequency of 10 adverse events, such as insults, drinking, quarreling, neglecting, and chaotic household, and 3 positive events, such as loving and hugging, in the early family environment, on a scale of 1 (not at all) to 6 (very often). The total score equals [10 * Mean(adverse events) − 10] + [18 – 3 * Mean(positive events)], resulting in a scale of 0 to 65. Higher scores indicate a more chaotic, harsh, and abusive early family environment, whereas lower scores indicate a more organized, supportive, and nurturing early family environment.

Non–Internship-Related Stressful Life Events

The stressful life events question in the follow-up quarterly survey asks: "During the past 3 months, have you experienced any of the non–internship-related stressful life events: death of family or friends, being ill or injured, ending of relationship, being in a violent relationship, financial loss or debt, being assaulted or attacked, getting married, the pregnancy or birth of a child?" The results were classified as a binary outcome of yes or no.

Timely and Proper Faculty Feedback

The faculty feedback question in the resident questionnaire (28) asks the interns to indicate whether they agree with the statement of the instrument, "I get timely and appropriate feedback from faculty" using a 5-point Likert scale ranging from 1 ("strongly disagree") to 5 ("strongly agree").

Learning Experience in Inpatient Rotations

The inpatient rotation question in the resident questionnaire (28) asks the interns to indicate whether they agree with the statement of the instrument, "The inpatient ward rotations are generally a good learning experience" using a 5-point Likert scale ranging from 1 ("strongly disagree") to 5 ("strongly agree").

Missing Data Imputation

Missing variables were imputed by multiple imputation (R "mice" package [51]). Before imputation, to create the predictor matrix, the variables used to impute each incomplete variable were determined with LASSO (least absolute shrinkage and selection operator) regression (R "glmnet" package [52, 53]). The imputation method was predictive mean matching for continuous variables, logistic regression for binary variables, and polynomial regression for categorical variables. The number of multiple imputations was 5.

The variables include the following:

1) Continuous: age, baseline sleep time, baseline work hours, baseline PHQ-9 score, baseline neuroticism score, difficult early family score, internship PHQ-9 score, internship sleep time, internship work hours, rating of timely and proper faculty feedback, and rating of learning experience in inpatient rotations.

2) Binary: gender, specialty (surgical or nonsurgical), baseline stressful life events (yes or no), depression history at baseline (yes or no), internship stressful life events (yes or no), medical errors (yes or no), and mental health treatment during internship (yes or no).

3) Categorial: race (White, Asian, or underrepresented minority).

Survey Weights

Poststratification

We obtained the population data of first-year residents in the United States from the Association of American Medical Colleges between 2007 and 2019. The data included the total number of residents and the number of surgical and nonsurgical residents from each cohort year. Within each specialty group from each cohort, we also obtained the numbers of women and men and numbers of White, Asian, and underrepresented minority residents.

Using the data from the Association of American Medical Colleges as the reference population, we did a 3-step poststratification raking on Intern Health Study data to develop weights such that applying those weights results in a sample with a distribution of cohort year, specialty, gender, and race that matches that of the Association of American Medical Colleges data. The first step was to generate between-cohort weights (w1b) with the raking variable to be cohort year, the second step was to generate weights (w1wa) with the raking variables to be specialty within each cohort, and the last step was to generate weights (w1wb) with the raking variables to be gender and race within each specialty group (surgical and nonsurgical) in each cohort. The poststratification rakings were performed with the R package "anesrake" (30).

Attrition Weights

Among 19 993 enrolled participants, 16 965 completed at least 1 of the 4 follow-up surveys. Through LASSO regression, we identified 5 baseline variables that were significantly associated with the completion of follow-up surveys: cohort year (2019 vs. other), race, specialty, baseline PHQ-9 score, and baseline neuroticism score. We estimated the propensity score of follow-up survey completion with R package "twang" (31) using gradient-boosted models and then extracted the attrition weights (w2) from the propensity score with the "get.weights" function in R. The weight for each participant who completed at least 1 follow-up survey—that is, the participants we included in our analysis—is 1 / p, where p is the propensity score.

Total Weights

Total weights = w1b * w1wa * w1wb * w2.

Statistical Analysis

We used the R "survey" package (36, 54) to implement the key steps of the statistical analysis. Specifically, the "svydesign" function was used to incorporate the survey weights and cluster identifications into the data set (at levels of clustering—by residency institution within year), the "svyglm" function was used to fit the models, and the "marginpred" function was used for the marginal prediction and standardization. In addition, the basis matrix for the natural splines regression model was generated by the R function "ns" with the degrees of freedom equal to 3.

Supplementary Assessments

Variation Across Residency Institutions

We assessed whether the depressive symptom change with internship varies across the residency institutions with the nonparametric Kruskal–Wallis 1-way analysis of variance (55). We found that although there was significant variation between institutions ($P < 0.001$), the institutions explained only 2% of the variance in depressive symptom change after adjustment for individual characteristics, including gender, race, and specialty.

Effects of Number and Season of Follow-up Surveys

We assessed whether the number of completed follow-up surveys or the season of the follow-up surveys completed was correlated with the depressive symptom change of the participants.

We found that the number of completed follow-up surveys was negatively correlated with the average PHQ-9 change (mean score change [± SE]: one follow-up, 3.6 ± 0.10; two, 3.5 ± 0.08; three, 3.4 ± 0.06; and four, 3.1 ± 0.03). However, the mean numbers (ranged from 3.1 to 3.2) of completed follow-up surveys were similar across 13 cohorts, and the proportions of participants who completed 1, 2, 3, and 4 follow-up surveys did not differ significantly across the cohorts (Appendix Table 7). Further, there was no clinically significant difference in PHQ-9 score across quarterly surveys (mean score [± SE]: first quarter, 5.7 ± 0.04; second quarter, 5.8 ± 0.04; third quarter, 5.8 ± 0.04; and fourth quarter, 5.6 ± 0.04). In addition, all of the institutions with more than 20 enrolled participants (>80% of all included institutions) had a follow-up rate above 70%.

Web References

50. **American College of Surgeons.** Surgical specialties. Accessed atwww.facs.org/member-services/join/specialties on 10 June 2021.

51. **van Buuren S, Groothuis-Oudshoorn K.** Mice: multivariate imputation by chained equations in R. J Stat Softw. 2011;45:1-67. doi:10.18637/jss.v045.i03

52. **Tibshirani R.** Regression shrinkage and selection via the lasso. J R Stat Soc Series B Stat Methodol. 1996;58:267-88. doi:10.1111/j.2517-6161.1996.tb02080.x

53. **Friedman J, Hastie T, Tibshirani R.** Regularization paths for generalized linear models via coordinate descent. J Stat Softw. 2010;33:1-22. [PMID: 20808728]

54. **Lumley T.** Complex Surveys: A Guide to Analysis Using R. J Wiley; 2010.

55. **Hollander M, Wolfe DA.** The one-way layout. In: Nonparametric Statistical Methods. J Wiley; 1973:115-20.

Appendix Table 1. Weighted Summary Statistics of Baseline Depressive Symptoms, Average Internship Depressive Symptoms, and Change in Depressive Symptoms From Baseline to Internship, Among Groups Stratified by Baseline and Internship Risk Factors

Characteristic	Baseline Depressive Symptoms (Group Mean)	Average Internship Depressive Symptoms (Group Mean)	Mean Change in Depressive Symptoms From Baseline to Internship	
			Group Mean (95% CI)	Mean Difference Between Groups (95% CI)
Baseline				
Age	–	–	–	0.2 (0.1 to 0.3)
≤Median	2.5	5.7	3.2 (3.1 to 3.4)	–
>Median	2.8	6.3	3.5 (3.3 to 3.7)	–
Gender	–	–	–	0.7 (0.5 to 0.9)
Men	2.5	5.5	3.0 (2.8 to 3.2)	–
Women	2.7	6.4	3.7 (3.4 to 3.9)	–
Specialty	–	–	–	0.2 (−0.0004 to 0.4)
Nonsurgical	2.6	5.9	3.3 (3.1 to 3.5)	–
Surgical	2.5	6.0	3.5 (3.3 to 3.7)	–
Race				
White	2.6	5.9	3.3 (3.1 to 3.5)	Reference
Asian	2.5	5.9	3.4 (3.1 to 3.6)	0.05 (−0.2 to 0.3)
URM	2.7	6.1	3.4 (3.2 to 3.5)	0.04 (−0.2 to 0.2)
Depression history	–	–	–	0.9 (0.7 to 1.2)
No	1.9	4.8	2.9 (2.8 to 3.0)	–
Yes	3.5	7.3	3.8 (3.6 to 4.1)	–
Neuroticism: NEO Five-Factor Inventory score	–	–	–	0.9 (0.7 to 1.0)
≤Median	1.5	4.4	2.9 (2.8 to 3.1)	–
>Median	3.8	7.6	3.8 (3.6 to 4.1)	–
Difficult early family environment: Risky Families Questionnaire score	–	–	–	0.4 (0.3 to 0.6)
≤Median	2.0	5.1	3.1 (3.0 to 3.3)	–
>Median	3.3	6.7	3.6 (3.3 to 3.8)	–
Internship				
Weekly work hours, self-reported	–	–	–	1.3 (1.2 to 1.5)
≤Median	2.6	5.2	2.7 (2.5 to 2.8)	–
>Median	2.6	6.6	4.0 (3.8 to 4.2)	–
Daily sleep hours, self-reported	–	–	–	−1.2 (−1.4 to −1.1)
≤Median	2.8	6.7	3.9 (3.7 to 4.1)	–
>Median	2.3	5.0	2.6 (2.5 to 2.8)	–
Medical error, self-reported	–	–	–	1.2 (1.0 to 1.4)
No	2.4	5.2	2.9 (2.7 to 3.0)	–
Yes	3.0	7.0	4.1 (3.8 to 4.3)	–
Stressful life events	–	–	–	0.3 (0.2 to 0.5)
No	2.3	5.5	3.2 (3.0 to 3.4)	–
Yes	2.9	6.4	3.5 (3.3 to 3.7)	–
Sought mental health treatment when depressed	–	–	–	−0.3 (−0.5 to −0.2)
No	3.8	10.0	6.2 (5.9 to 6.4)	–
Yes	5.2	11.0	5.8 (5.5 to 6.2)	–
Timely and proper faculty feedback	–	–	–	−0.6 (−0.8 to −0.4)
≤Median	2.6	6.0	3.4 (3.2 to 3.6)	–
>Median	2.1	4.9	2.8 (2.6 to 3.0)	–
Inpatient rotation learning experience	–	–	–	−0.9 (−1.1 to −0.8)
≤Median	2.7	6.2	3.5 (3.3 to 3.7)	–
>Median	2.1	4.7	2.6 (2.4 to 2.8)	–

URM = underrepresented minority.

Appendix Table 2. Population and Unweighted Sample Size of Interns From 2007 to 2019

Cohort Year	Total First-Year Residents in the United States, *n**	Total Invited, *n*	Enrolled at Baseline, *n*	Enrollment Rate, %	Completed ≥1 Follow-up Survey, *n*	Follow-up Rate, %	Residency Institutions, *n*
2007	25 040	401	242	60.35	217	89.67	7
2008	25 295	853	495	58.03	421	85.05	73
2009	25 198	1156	748	64.71	607	81.15	96
2010	25 201	1448	739	51.04	629	85.12	109
2011	25 745	2071	810	39.11	673	83.09	127
2012	25 686	2336	1342	57.45	1191	88.75	209
2013	26 212	2518	1457	57.86	1268	87.03	236
2014	26 825	1758	1092	62.12	960	87.91	164
2015	27 936	4855	3122	64.30	2681	85.87	346
2016	28 830	5375	3288	61.17	2802	85.22	347
2017	29 943	4996	2846	56.97	2473	86.89	318
2018	30 692	4347	2127	48.93	1843	86.65	298
2019	30 246	2725	1685	61.83	1200	71.22	264
Total	352 849	34 839	19 993	57.39	16 965	84.85	645†

* Source: Association of American Medical Colleges.
† Number of unique residency programs in 13 cohorts.

Appendix Table 3. Comparisons Between Participants, Stratified by Follow-up Survey Completion

Summary Statistic	No Follow-up	Follow-up	*P* Value From *t* Test or χ^2 Test
	***n* = 3028**	***n* = 16 965**	
Raw, unweighted			
Mean age (±SE), *y*	27.6 ± 0.05	27.5 ± 0.02	0.21
Gender: women, %	48.4	51.7	<0.001
Non-White race, %	47.5	39.1	<0.001
Surgical specialty, %	23.2	19.6	<0.001
Mean baseline depressive symptom score (±SE)	3.0 ± 0.06	2.6 ± 0.02	<0.001
Mean neuroticism score (±SE)	23.2 ± 0.2	22.0 ± 0.07	<0.001
Mean difficult early family environment score (±SE)	13.6 ± 0.2	12.9 ± 0.07	<0.001
Depression history, %	48.2	46.0	0.028
	Weighted *n* = 18 313	**Weighted *n* = 19 867**	
Weighted*			
Mean age (±SE), *y*	27.7 ± 0.10	27.6 ± 0.04	0.24
Female gender, %	45.9	48.8	0.080
Non-White race, %	42.7	41.7	0.33
Surgical specialty, %	19.2	19.0	0.75
Mean baseline depressive symptom score (±SE)	2.6 ± 0.08	2.6 ± 0.05	0.85
Mean neuroticism score (±SE)	22.1 ± 0.3	22.0 ± 0.3	0.85
Mean difficult early family environment score (±SE)	13.2 ± 0.2	13.1 ± 0.1	0.73
Depression history, %	46.2	45.8	0.76

* Both sample weights and attrition weights applied.

Appendix Table 4. Trends of Annual Average Depressive Symptom Change With Internship From 2007 to 2011 and From 2011 to 2019

Population	Expected Mean Change in Depressive Symptom Score (95% CI)			Difference 2011 vs. 2007 (95% CI)	Difference 2019 vs. 2011 (95% CI)
	2007	2011	2019		
All interns	4.1 (3.8 to 4.4)	3.1 (3.0 to 3.3)	3.0 (2.7 to 3.4)	−1.0 (−1.3 to −0.6)	−0.1 (−0.5 to 0.3)
Men	3.5 (3.3 to 3.7)	2.7 (2.5 to 2.9)	2.9 (2.3 to 3.4)	−0.8 (−1.1 to −0.5)	0.2 (−0.4 to 0.7)
Women	4.7 (4.2 to 5.1)	3.5 (3.3 to 3.6)	3.3 (3.1 to 3.4)	−1.2 (−1.7 to −0.8)	−0.2 (−0.5 to 0.1)
Surgical interns	4.0 (3.7 to 4.4)	3.1 (2.6 to 3.5)	3.2 (3.0 to 3.5)	−1.0 (−1.5 to −0.4)	0.2 (−0.3 to 0.7)
Nonsurgical interns	4.1 (3.7 to 4.4)	3.1 (3 to 3.2)	3.0 (2.6 to 3.4)	−1.0 (−1.3 to −0.6)	−0.2 (−0.5 to 0.3)

Appendix Table 5. Comparisons of Residency Institution–Related Internship Factors Between Full Sample and Common Institution Subsample

Factor	Full Sample	Common Institution Subsample
Median weekly work hours (IQR)	65 (56–74)	67 (57–75)
Median rating of timely and proper faculty feedback (IQR)	4 (3–4)	4 (3–4)
Median rating of learning experience in inpatient rotations (IQR)	4 (4–4)	4 (4–4)

IQR = interquartile range.

Appendix Table 6. Weighted Baseline and Internship Characteristics of Interns, Stratified by Number of Completed Follow-up Surveys

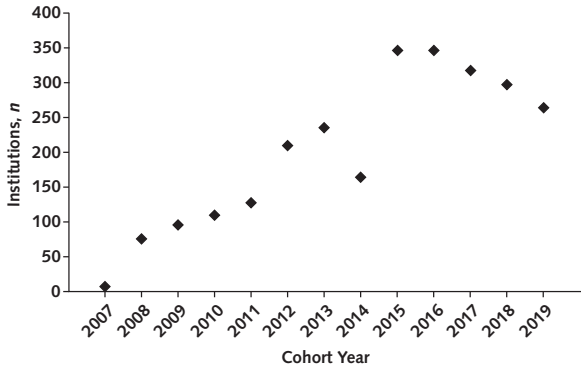
Sample Characteristic	Number of Completed Follow-up Surveys				
	0	1	2	3	4
Sample size					
Unweighted, *n*	3028	2278	2278	3168	9241
Weighted, *n*	18 313	2597	2668	3848	10 574
Baseline					
Median age (IQR), *y*	27 (26–29)	27 (26–29)	27 (26–29)	27 (26–29)	27 (26–29)
Gender, *n (%)*					
Men	9905 (54.1)	1367 (52.6)	1370 (51.4)	2011 (52.2)	5417 (50.4)
Women	8408 (45.9)	1230 (47.4)	1298 (48.6)	1838 (47.8)	5337 (49.6)
Specialty, *n (%)*					
Nonsurgical	14 789 (80.8)	2073 (79.8)	2123 (79.6)	3076 (79.9)	8814 (82.0)
Surgical	3523 (19.2)	524 (20.2)	545 (20.4)	772 (20.1)	1940 (18.0)
Race, *n (%)*					
White	10 502 (57.3)	1316 (50.7)	1384 (50.6)	2100 (54.6)	6790 (63.1)
Asian	4100 (22.4)	679 (26.2)	693 (29.8)	963 (25.0)	2136 (19.9)
URM	3711 (20.3)	601 (23.1)	591 (15.7)	784 (20.4)	1828 (17.0)
Depressive symptoms: median PHQ-9 score (IQR)*	2 (0–4)	2 (0–4)	2 (0–4)	2 (0–4)	2 (0–3)
Depression history, *n (%)*					
No	9858 (53.8)	1360 (52.4)	1455 (54.5)	1991 (51.7)	5967 (55.5)
Yes	8454 (46.2)	1237 (47.6)	1213 (45.5)	1858 (48.3)	4797 (44.5)
Neuroticism: median NEO Five-Factor Inventory score (IQR)†	22 (16–28)	23 (17–28)	22 (16–28)	22 (16–28)	21 (15–28)
Difficult early family environment: median Risky Families Questionnaire score (IQR)‡	10 (6–18)	11 (6–19)	11 (6–19)	11 (6–18)	10 (6–17)
Internship					
Average internship depressive symptoms: median PHQ-9 score (IQR)*	NA	5 (3–9)	5.5 (3–9)	5.3 (3–8.7)	5 (2.6–8)
Median self-reported weekly work hours (IQR)	NA	70 (57.5–80)	67.5 (56.5–76)	65.7 (56.7–74)	64.3 (56.1–72.3)
Median self-reported daily sleep hours (IQR)	NA	6 (6–7)	6.5 (6–7)	6.3 (6–7)	6.5 (6–7)
Responses endorsing medical error, *n (%)*					
No	NA	2014 (77.5)	1822 (68.3)	2381 (61.9)	6096 (56.7)
Yes	NA	583 (22.5)	845 (31.7)	1468 (38.1)	4658 (43.3)
Responses endorsing stressful life event, *n (%)*					
No	NA	1823 (70.2)	1585 (59.4)	1969 (51.2)	5218 (48.5)
Yes	NA	773 (29.8)	1083 (40.6)	1879 (48.8)	5536 (51.5)
Sought mental health treatment when depressed, *n (%)*					
No	NA	512 (80.6)	772 (83.8)	1134 (77.1)	3035 (76.7)
Yes	NA	123 (19.4)	149 (16.2)	337 (22.9)	922 (23.3)
Timely and proper faculty feedback: median rating (IQR)§	NA	4 (3–4)	4 (3–4)	4 (3–4)	4 (3–4)
Learning experience in inpatient rotations: median rating (IQR)§	NA	4 (4–4)	4 (4–4)	4 (4–4)	4 (4–4)

IQR = interquartile range; NA = not applicable; URM = underrepresented minority.
* Range, 0–27.
† Range, 0–56.
‡ Range, 0–65.
§ Range, 1–5.

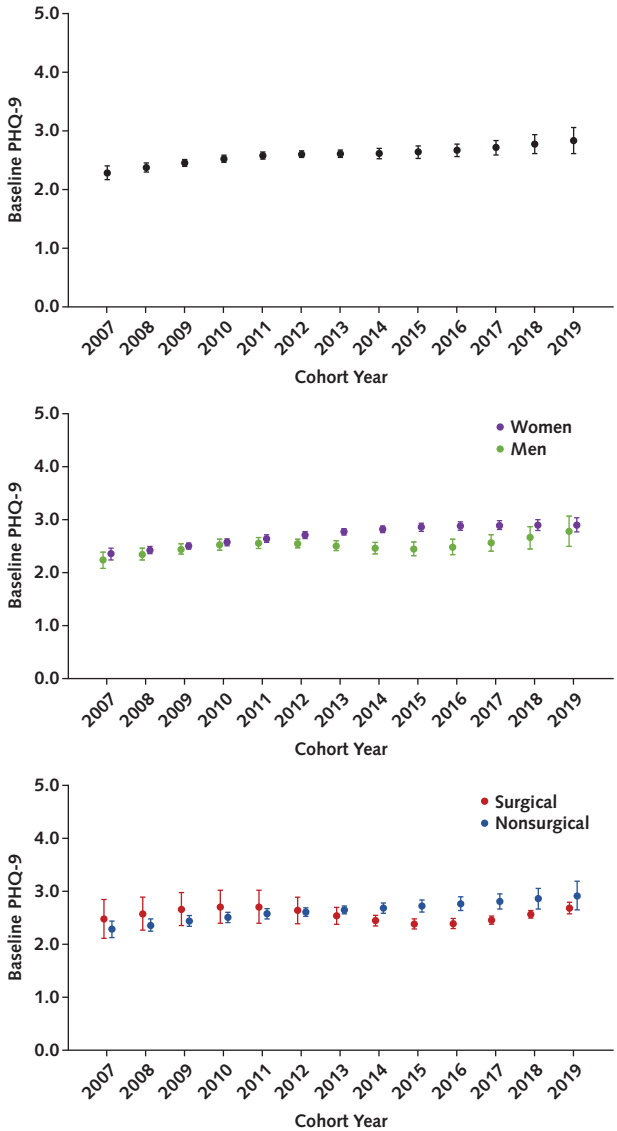
Appendix Table 7. Proportion of Participants Who Completed 1 to 4 Follow-up Surveys, in Full Sample and in Each Cohort Year

Cohort Year	Percentage				Mean Follow-up Surveys Completed, *n*	χ^2 Test *P* Value (Reference 2007)
	Completed 1 Follow-up Survey	Completed 2 Follow-up Surveys	Completed 3 Follow-up Surveys	Completed 4 Follow-up Surveys		
All	13	13	19	54	3.1	–
2007	12	17	19	53	3.1	–
2008	14	12	21	53	3.1	0.3
2009	13	12	19	56	3.2	0.46
2010	9	12	25	53	3.2	0.11
2011	11	12	18	58	3.2	0.32
2012	11	12	21	56	3.2	0.31
2013	15	14	19	52	3.1	0.56
2014	13	16	20	51	3.1	0.93
2015	14	12	19	56	3.2	0.2
2016	15	15	18	51	3.1	0.52
2017	11	13	18	58	3.2	0.4
2018	16	14	16	53	3.1	0.27
2019	15	12	16	57	3.2	0.12

Appendix Figure 1. Number of enrolled residency institutions in the Intern Health Study from 2007 to 2019.

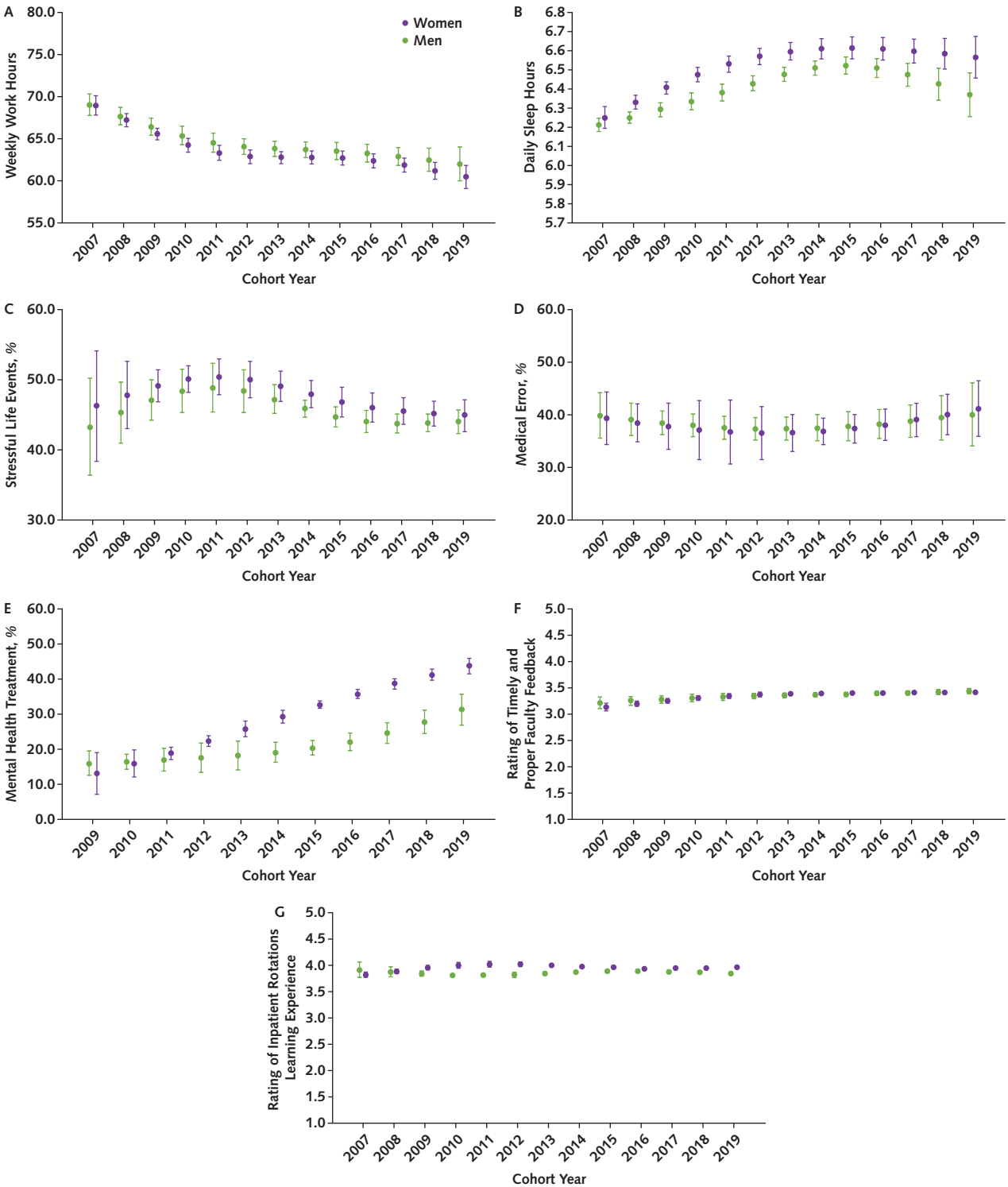


Appendix Figure 2. The weighted, model-based, expected baseline depressive symptoms across cohorts: full sample.



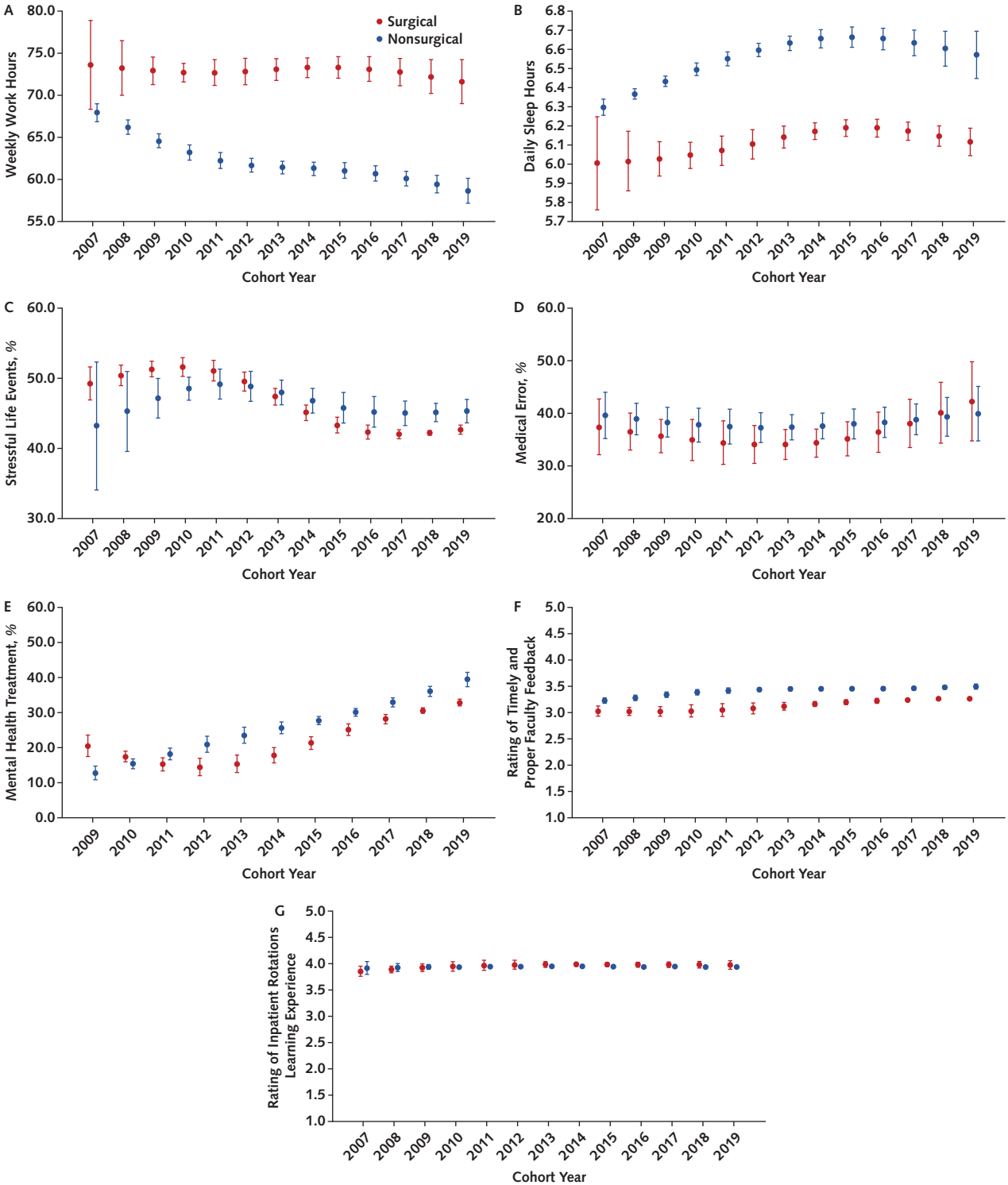
The natural splines fitting of baseline PHQ-9 depressive symptoms between 2007 and 2019 cohorts among all participants (*top*), stratified by gender (*middle*), and stratified by specialty (*bottom*). Error bars represent 95% CIs. PHQ-9 = 9-item Patient Health Questionnaire.

Appendix Figure 3. Comparisons of the weighted, model-based, expected internship factors across cohorts: men vs. women.



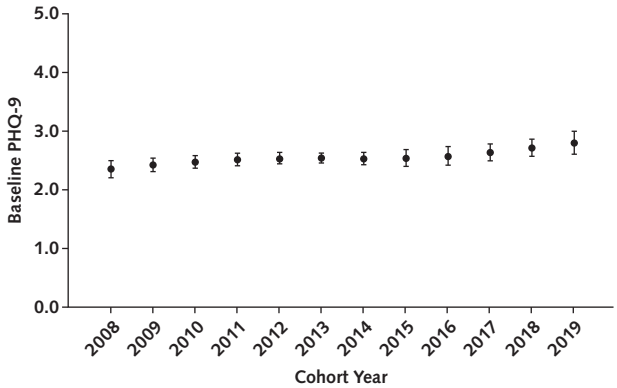
The natural splines fitting of internship risk factors: weekly work hours (*A*), daily sleep hours (*B*), stressful life events (*C*), medical errors (*D*), mental health treatment (*E*), faculty feedback (*F*), and inpatient rotation learning experience (*G*), stratified by gender across cohorts. Error bars represent 95% CIs.

Appendix Figure 4. Comparisons of the weighted, model-based, expected internship factors across cohorts: surgical vs. nonsurgical interns.



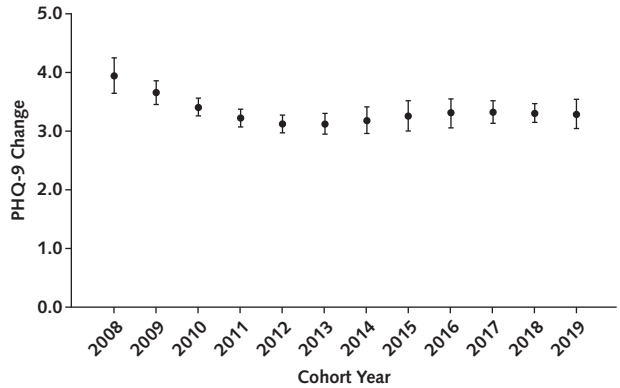
The natural splines fitting of internship risk factors: weekly work hours (*A*), daily sleep hours (*B*), stressful life events (*C*), medical errors (*D*), mental health treatment (*E*), faculty feedback (*F*), and inpatient rotation learning experience (*G*), stratified by specialty across cohorts. Error bars represent 95% CIs.

Appendix Figure 5. The weighted, model-based, expected baseline depressive symptoms across cohorts: common institution subsample.



The natural splines fitting of baseline PHQ-9 depressive symptoms between 2008 and 2019 cohorts among participants in the common institution subsample (institutions that were included in all cohorts from 2008 to 2019). PHQ-9 = 9-item Patient Health Questionnaire.

Appendix Figure 6. The weighted, model-based, expected changes in depressive symptoms with internship across cohorts: common institution subsample.



The natural splines fitting of annual average changes in PHQ-9 depressive symptoms with internship between 2008 and 2019 cohorts among participants in the common institution subsample (institutions that were included in all cohorts from 2008 to 2019). PHQ-9 = 9-item Patient Health Questionnaire.