Supplementary Appendix

Supplement to: Fang Y, Lodi S, Hughes TM, et al. Work hours and depression in U.S. first-year physicians. N Engl J Med 2022;387:1522-4. DOI: 10.1056/NEJMc2210365

This appendix has been provided by the authors to give readers additional information about the work.

Table of Contents

Supplementary Methods

Design and Target Trial Specification

A target trial is a hypothesized trial designed to answer a specific causal question.[1] By emulating the target trial, an observational study can avoid many of the biases of standard methods. The key features of this approach are statistical methods that are robust to problems that can arise from modeling multiple correlated and time-varying factors,[17] eligibility criteria to optimize comparability between groups,[1,2] and alignment of eligibility, group assignment, and outcome assessment time periods.[3] It is worth noting that, the causal interpretation of results from a emulated trial still relies on the validity of certain assumptions, most critically the unobserved confounding assumption.

The ideal trial to elucidate the effect of work hours on well-being would randomly assign residents to specific work hours levels. Such a design is infeasible. We designed an observational study to emulate a sequentially randomized target trial in which, at each quarter, interns are randomly assigned to a work hour level and assessed changes in depressive symptoms from pre-internship levels. **Table S1** provides the components of the target trial and our emulation of that trial with the Intern Health Study data, following reporting recommendations.[2]

Data Source and Data Collection Procedures

The Intern Health Study is a cohort study, repeated annually, of interns working at health care institutions across the United States.[4] In total, 17,082 subjects from cohorts recruited during 2009 to 2020 had complete data on key variables necessary for the present analysis.

Interns received a study invitation by email 2-3 months prior to the start of internship and were assessed through online surveys in April-June (baseline), September (Quarter 1 [Q1]),

December (Q2), March (Q3) and June (Q4). Subjects provided informed consent and received between $50 and $125 in compensation, depending on the cohort year. The Institutional Review Board at the University of Michigan approved the study.

The present analysis was restricted to cohorts recruited from the spring of 2009 to the spring of 2020. Although the study began in 2007, data from 2007 and 2008 were not included because a covariate (an indicator of having had children) was not collected in those years. The annual number of interns included in the present analysis ranged from 452 (in 2009) to 2,721 (in 2016). See **Table S2** and Fang et al.[12] for recruitment details of the annual cohorts.

Measures

Depression symptom level was measured using the Patient Health Questionnaire (PHQ-9).[5] The PHQ-9 is a validated self-report measure of the nine symptoms of depression, as defined by the Diagnostic and Statistical Manual of Mental Disorders (DSM-5).[6] Interns indicated whether, during the previous two weeks, each of the 9 symptoms 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 item and a total score between 0 and 27.[20] A total score of 0-4, 5-9, 10-14, 15-19, and 20-27 correspond to minimal, mild, moderate, moderately severe, and severe depression, respectively.[5,7] The diagnostic validity of the PHQ-9 is comparable to clinician-administered assessments, with a cut-off of 10 achieving 88% sensitivity and specificity for a diagnosis of major depressive disorder.[5] The primary outcome of the study was change in depression symptom level from baseline to internship, calculated based on PHQ-9 scores at each of the quarters, minus baseline PHQ-9 score.

Work hours were self-reported by the interns at Q1-Q4 in response to the question "how many hours have you worked in the past week?" Prior analyses have found that self-reported daily work hours varied from electronic health record-derived measures by an average of 1.3 hours

among interns,[8] indicating relatively high accuracy of self-reported hours. To maximally emulate a randomized trial, in which investigators would assign subjects to categorical work hour groups, we categorized work hour levels as ≤20, 20+ to 40, 40+ to 45, 45+ to 50, 50+ to 55, 55+ to 60, 60+ to 65, 65+ to 70, 70+ to 75, 75+ to 80, 80+ to 85, 85+ to 90, and 90+, where "+" indicates more than the stated value. Assumptions of our measurement strategy are that the relationship between depressive symptoms and work hours, if it exists, is strongest for the work hours most proximate in time to the period when depressive symptoms are assessed, and that interns' past week work hours are representative of their work hours in the past quarter.

We included the following baseline covariates as possible confounders: (1) gender (woman or man); (2) specialty (surgical or non-surgical); (3) neuroticism from the NEO-Five Factor Inventory[9,10]; (4) self-reported history of depression; (5) early family environment from the Risky Families Questionnaire[11,12]; (6) age; (7) cohort calendar year; (8) marital status (married or unmarried); and (9) children (none, one or more). The internship variables, assessed at each quarter, were: (1) internship stressful life events (SLEs), a binary indicator of endorsing any one of 11 types of events (e.g., death of someone close, got married, assaulted)[13] and (2) self-reported medical errors from the question "are you concerned you have made any major medical errors in the last 3 months?" Although medical errors and SLEs were time-varying measures, they were not past-exposure-affected time-varying confounders[14,15] (which would require more complex modeling strategies), because both variables were only weakly associated with work hours at the prior time point (prior work hour and SLE: β=0.31; prior work hour and medical error: β=0.59). Internship quarter and its interaction with work hour level were also included as covariates to account for the potential difference in the associations of work hour with depression during different periods of the year. We excluded observations with any missing data on PHQ-9, work hour or covariates (see **Table S2** for the number of subjects by year). The bias introduced by this exclusion was accounted for via sampling weights.

Statistical analysis

Prior to analysis, we generated two sets of weights: post-stratification weights, based on the demographic characteristics of all entering U.S. interns in these years, to reduce potential bias due to non-representative sampling; and attrition weights, to account for differences observations included in analysis and those excluded due to missing data. The two sets of weights were multiplied and included in the model estimation. Details of weights generation are described in our prior publication.[4]

We used standardization[16] to estimate the association between self-reported work hours at each quarter and the change in PHQ-9 score between that quarter and baseline. Specifically, we first fitted a linear regression model using a generalized estimating equation (GEE; with exchangeable working correlation matrix; 'geeglm' function in R package 'geepack') with quarterly repeated measures (clustered by individual) and weights. The independent variables in the model were the work hour categorical variable and all baseline and internship covariates listed under "Measures." We then used the model to compute the standardized change in PHQ-9 score for each level of work hours by calculating the mean predicted PHQ-9 change assuming all subjects at all quarters were assigned that level of work hour. We used non-parametric bootstrap with 500 replications to obtain 95% confidence intervals of the estimates of PHQ-9 change.

Because surgery training has been reported to have longer work hours and a different set of stressors,[4,17] we conducted a secondary analysis within specialty groups. First, we repeated the analyses separately for surgical and non-surgical interns. Second, we estimated an interaction between specialty type and work hours, with work hours defined as a continuous, rather than categorical, variable in the model. To ensure that the estimates were not overly influenced by the small number of surgical interns at the lower work hour levels (see **Figure S1**), we repeated this analysis restricting to interns and quarters in which 50 hours or greater were reported.

There is also the potential for correlation among subjects from the same residency institution. We conducted a sensitivity analysis adding residency institution (collected in the baseline survey) as a higher-level clustering factor. We first updated the GEE model by change the cluster factor from individual to residency institutions. Because individuals are fully nested within institutions, clustering only at the higher level is sufficient.[18] We then repeated the standardization and bootstrapping with the updated model to estimate the mean predicted PHQ-9 change assuming all subjects at all quarters were assigned that level of work hour.

Supplementary Results

The 17,082 interns (median age, 27 years [IQR 26-28]; 52.3% women; 19.6% surgical interns) contributed 53,862 follow-up observations (average 3.2 follow-up observations per intern). The unweighted mean and median work hours reported at quarterly surveys was 63.0 hours and 67.0 hours. Median work hours were higher among surgical interns (75.0, IQR: 66.0-80.0) than non-surgical interns (65.0, IQR: 50.0-75.0). At baseline, interns had a median PHQ-9 score of 2.0 and mean of 2.7 (SD=3.1), at the low end of the score range of 0 to 27. The median and mean of the highest PHQ depressive symptom score across quarterly surveys reported by each intern was 7.0 and 8.0 (SD=4.9). In the unadjusted data, interns were more likely to fall into higher severity depression categories (i.e., moderate depression or greater) if they were working greater numbers of hours (**Figure S2**). See detailed characteristics of all interns, surgical interns, and non-surgical interns in **Table S3**.

Figure 1 (in the main text) shows the estimated adjusted change in PHQ-9 scores from baseline to internship at each work hour level after standardization. The change in PHQ-9 score from baseline between two proximate 5-hour interval work hour categories ranged from 0.12 (95%CI: -0.11-0.35; comparison of 45+ to 50 hours/week and 50+ to 55 hours/week), to 0.55 (95%CI: 0.31-0.80; comparison of 75+ to 80 hours/week and 80+ to 85 hours/week). On average, 5

additional hours of work were associated with a 0.25 (95%CI: 0.24-0.26) point higher depressive symptom increase.

The association between greater work hours and greater average increases in PHQ-9 scores was present for both surgical and non-surgical interns (**Figure S3**). However, the increase in PHQ-9 scores associated with higher work hours was larger for non-surgical interns than surgical interns (beta for interaction= -0.012). Specifically, 5 additional hours of work was associated with 0.26 (95% CI: 0.25-0.27) points higher depressive symptom increase in non-surgical interns and 0.19 (95% CI: 0.16-0.22) points higher depressive symptom increase in surgical interns. This finding was consistent in sensitivity analyses including only 50 hours or greater (beta for interaction= -0.016; **Figure S4**). Additionally, the effect of work hours was slightly larger for later quarters compared to earlier quarters (beta for interaction between continuous hours and quarter= 0.0035).

In our sensitivity analyses account for clustering among subjects from the same residency institution, the estimated adjusted change in PHQ-9 scores from baseline to internship at each work hour level after standardization were almost identical to the original results (Figure S5). At 40+ to 45 hours/week, the estimated increase in depressive symptoms was 1.8 (95% CI: 1.6-2.0), while at 90+ hours/week the increase was 5.2 (95% CI: 4.9-5.6).

Supplemental Discussion

The magnitude of the association between work hours and change in depressive symptoms suggests that work hour reduction should be a primary intervention to reduce clinician depression. The 2011 Accreditation Council for Graduate Medical Education (ACGME) national policy implementing a cap on the maximum shift length[19] did not appear to have substantially improved resident wellbeing but also did not reduce mean work hours.[4] Our finding that estimated depression levels decreased along the entire continuum of work hour levels suggests

that this policies that decrease the work hours overall may result in greater improvements to mental health than policies only eliminating extreme work hour levels, consistent with the approach to preventing a disease by "shift[ing] the whole distribution of exposure in a favorable direction."[20]

We found a smaller magnitude in the association of work hours with depression symptoms among surgical interns compared to non-surgical interns. We also found that surgical interns had higher depression symptom scores during internship than non-surgical interns, indicating that the modestly smaller effect of work hours is not due to lower depression overall. Differences between surgical and non-surgical interns in predisposing factors, the content of work hours, and stressful internship experiences, such as harassment, discrimination, and abuse, should be assessed as potential factors important in the different magnitude of the effect of work hours between specialties.[21]

Strengths of this study are the large, national sample size and the measurement of work hours, rather than work schedule policies. We used design and analytic approaches that emulate a target trial with observational data, but there is still potential for unmeasured confounding. Several other limitations should be noted. We used self-reported work hours, which has the potential for systematic mismeasurement, although a prior comparison to medical records-derived work hours found accurate recall.[8] We only had work hours information for the last week before each quarterly survey, and were not able to measure other aspects of work schedule, such as the timing and setting of the work. Night shifts may be particularly disruptive and could be associated with further increases in depressive symptoms, given the effects of sleep and circadian disruption on depression.[22]

Table S1. Specification of the target trial and its emulation.

	Target Trial	**Emulation**
Eligibility Criteria	Intern physicians at U.S. institutions who began internship during 2009-2020	Intern physicians at U.S. institutions who agreed to participate in the Intern Health Study during spring recruitment periods in 2009-2020
Work Hour Strategies	Work hours levels of: ≤20, 20+ to 40, 40+ to 45, 45+ to 50, 50+ to 55, 55+ to 60, 60+ to 65, 65+ to 70, 70+ to 75, 75+ to 80, 80+ to 85, 85+ to 90, 90+ hours, where "+" indicates more than the stated value	Same categories as in the target trial. We used past week work hours as an approximation of typical work hours in the past quarter
Assignment	Random assignment at each quarter to one of the 13 work hour levels	There is no randomization in this observational study. Validity of results rests on the assumption of exchangeability (no confounding) after adjustment
Outcome and Follow-up	Repeated measures of change in PHQ-9 total score from baseline to 3-, 6-, 9-, and 12-months into internship	Same as for the target trial
Causal Contrast	There are two types of contrasts that could be tested: 1) Intent-to-treat effect, i.e., compare work hour levels to which the interns were randomized, regardless of adherence; 2) Per-protocol effect, i.e., compare work hour levels to which the interns were randomized and adhered to their assigned level	Observational analog of the per-protocol effect. (Like the per-protocol effect in a trial, this estimate is potentially biased by factors that differ across treatment groups)
Statistical analysis	Average treatment effect of work hours levels on changes in PHQ-9 score, pooled over cohort years and quarters. Sub-group analysis by surgical and non-surgical specialty	Average change in PHQ-9 scores under each work hour level, pooled over cohort years and quarters. Estimated through standardization by generating the expected mean PHQ-9 change for the entire sample, given their covariate values, conditional on each specific work hours level. Sub-group analysis by surgical and non-surgical specialty

Table S2. Intern population size and unweighted sample size of interns participating in the Intern Health Study, from annual cohorts of 2009 to 2020.

Cohort Year	Total number of first-year residents in the US*	Total invited	Enrolled at baseline	Enrollment rate	Complete at least one follow-up survey	Follow-up rate	No missing data in baseline and follow-up surveys	Number of sponsor residency institutions
2009	25,198	1,156	748	64.71%	607	81.15%	452	75
2010	25,201	1,448	739	51.04%	629	85.12%	606	95
2011	25,745	2,071	810	39.11%	673	83.09%	622	109
2012	25,686	2,336	1,342	57.45%	1,191	88.75%	1,125	187
2013	26,212	2,518	1,457	57.86%	1,268	87.03%	942	176
2014	26,825	1,758	1,092	62.12%	960	87.91%	761	132
2015	27,936	4,855	3,122	64.30%	2,681	85.87%	2,511	292
2016	28,830	5,375	3,288	61.17%	2,802	85.22%	2,721	295
2017	29,943	4,996	2,846	56.97%	2,473	86.89%	2,439	271
2018	30,692	4,347	2,127	48.93%	1,843	86.65%	1,831	245
2019	30,246	2,725	1,685	61.83%	1,200	71.22%	1,187	225
2020	31,857	4,771	2,472	51.81%	1,896	76.70%	1,885	234
Total	**334,371**	**38,356**	**21,728**	**56.65%**	**18,223**	**83.87%**	**17,082**	**463****

*Source: Association of American Medical Colleges (AAMC)

**Number of unique sponsor institutions in 2009-2020 cohorts

Table S3. Unweighted characteristics of surgical and non-surgical interns participating in the Intern Health Study during 2009-2020.

Characteristic	Overall	Surgical Interns	Non-Surgical Interns
	N = 17,082	N = 3,351	N = 13,731
Baseline			
Age: median years (IQR)	27 (26-28)	27 (26-28)	27 (26-28)
Female Gender: n (%)	8,928 (52.3%)	1,773 (52.9%)	7,155 (52.1%)
Race: n (%)			
White	10,325 (60.4%)	2,115 (63.1%)	8,210 (59.8%)
Asian	3,678 (21.5%)	590 (17.6%)	3,088 (22.5%)
Underrepresented Minority	3,079 (18.0%)	646 (19.3%)	2,433 (17.7%)
Married: n (%)	7,076 (41.4%)	1,297 (38.7%)	5,779 (42.1%)
Parent: n (%)	1,238 (7.3%)	207 (6.2%)	1,031 (7.5%)
Depressive Symptoms at Baseline: median PHQ-9 score[a] (IQR)	2.0 (0.0-4.0)	2.0 (0-4)	2.0 (0-4)
Neuroticism: median NEO Five-Factor[b] (IQR)	22 (16-28)	21 (15-27)	22 (16-29)
Difficult Early Family Life: median Risky Families Questionnaire[c] (IQR)	10 (6-17)	10 (6-17)	10 (6-17)
History of Depression: n (%)	7,932 (46.4%)	1,510 (45.1%)	6,422 (46.8%)
Internship			
Experienced any Stressful Event During the Year: n (%)	7,934 (46.4%)	1,507 (45.0%)	6,427 (46.8%)
Any Self-Reported Medical Errors During the Year: n (%)	6,309 (36.9%)	1,177 (35.1%)	5,132 (37.4%)

IQR = interquartile range; PHQ-9 = 9-item Patient Health Questionnaire.
[a] PHQ-9 score 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.
[b] NEO Five-Factor Inventory Neuroticism sub-scale range, 0–56.
[c] Risky Families Questionnaire range, 0–65.

Figure S1. Distribution of work hours for 17,082 interns in the Intern Health Study cohorts of 2009-2020, cumulative for all quarterly follow-up surveys (n=53,862 total observations), overall and by specialty.

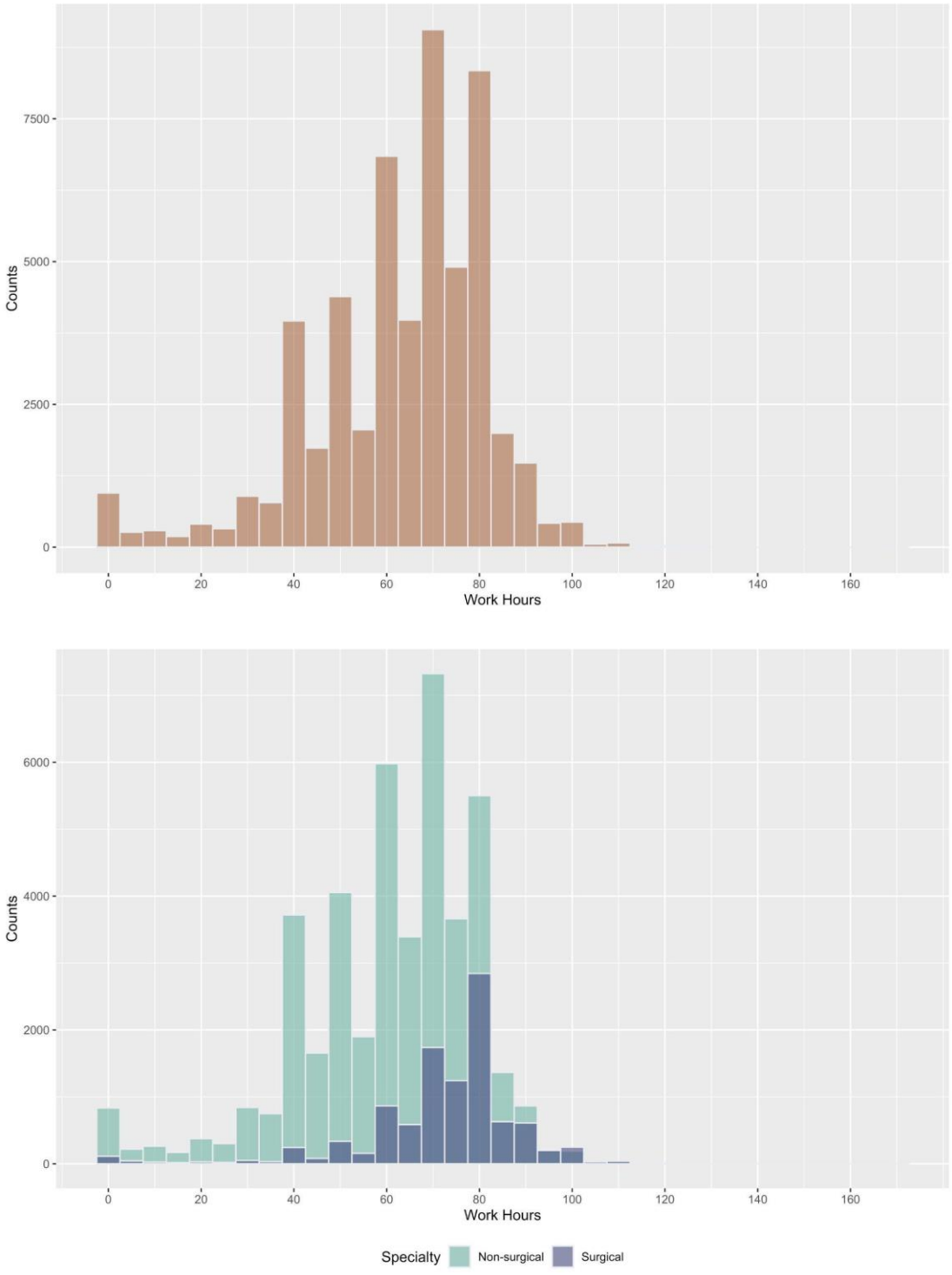
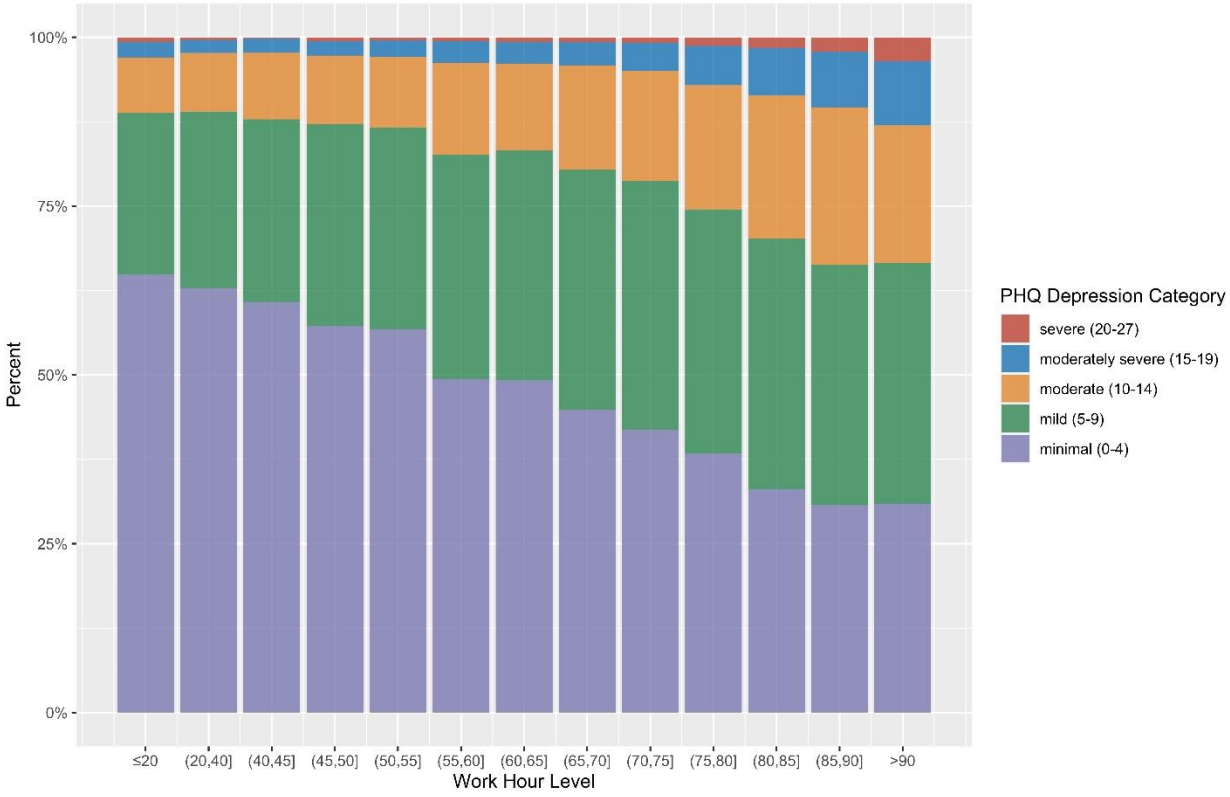
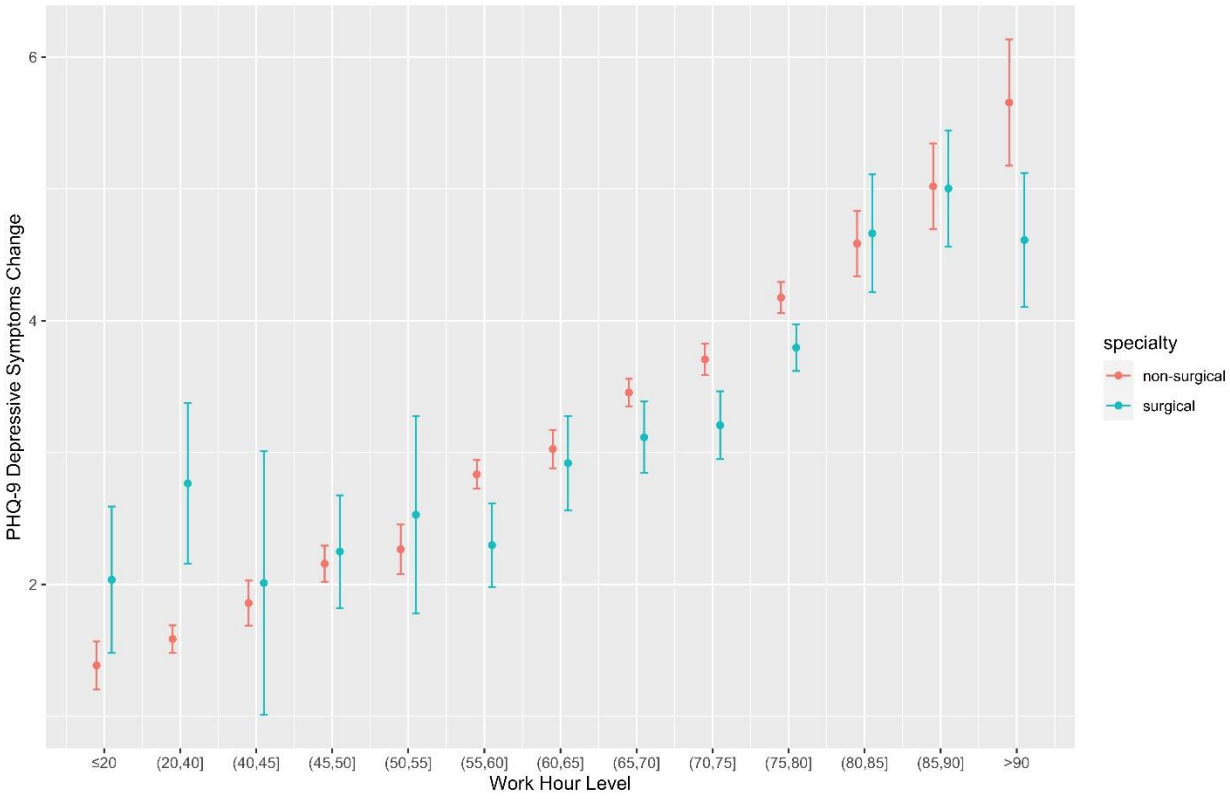


Figure S2. Unadjusted distribution of PHQ-9 total scores in raw data (n=53,862 total observations) categorized by depression severity level* and within each work hour level.



*Interpretation of PHQ-9 score levels: minimal (0-4) - no depressive disorder; mild (5-9) - subthreshold depressive disorder; moderate (10-14) - probable major depressive disorder, treatment should be considered; moderately severe (15-19) or severe (20-27) - major depressive disorder highly likely, treatment indicated.[7]

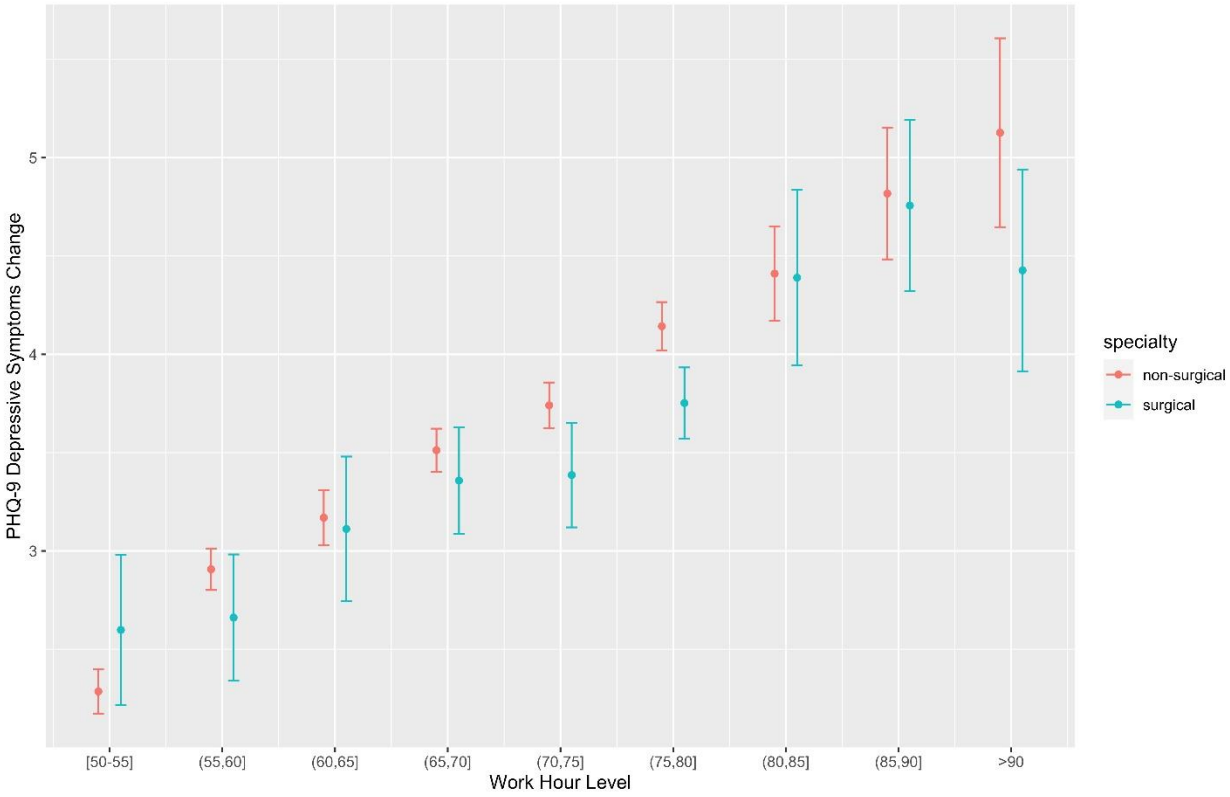
Figure S3. Estimated mean depression score change during internship and 95% CI[a], using standardization[b] with weights, under varying work hour levels among all 17,082 interns, grouped by specialty type, in the Intern Health Study, 2009-2020.



[a] confidence interval widths have not been adjusted for multiplicity and may not be used in place of hypothesis testing

[b] Adjusted for the baseline factors of gender, neuroticism, pre-internship history of depression, early family environment, age, cohort calendar year, marital status, and children, and time-varying factors of stressful life events and medical errors.

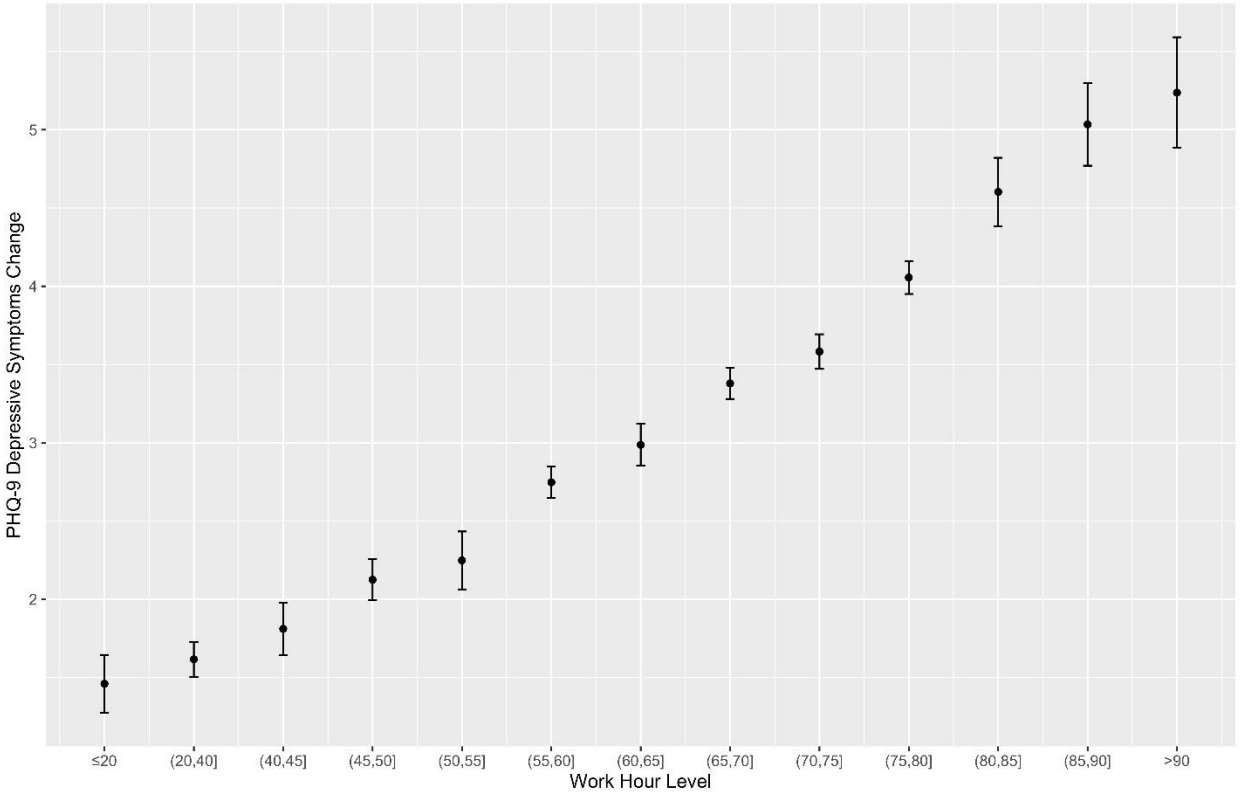
Figure S4. Estimated mean depression score change during internship and 95% CI[a], using standardization[b] with weights, under varying work hour levels among all 17,082 interns, grouped by specialty type, in the Intern Health Study, 2009-2020, based on models excluding observations with <50 hours/week.



[a] confidence interval widths have not been adjusted for multiplicity and may not be used in place of hypothesis testing

[b] Adjusted for the baseline factors of gender, neuroticism, pre-internship history of depression, early family environment, age, cohort calendar year, marital status, and children, and time-varying factors of stressful life events and medical errors.

Figure S5. Estimated mean depression score change during internship and 95% CI[a], using standardization[b] with weights, under varying work hour levels among all 17,082 interns, clustered by residency institutions, in the Intern Health Study, 2009-2020.



[a] confidence interval widths have not been adjusted for multiplicity and may not be used in place of hypothesis testing

[b] Adjusted for the baseline factors of gender, neuroticism, pre-internship history of depression, early family environment, age, cohort calendar year, marital status, and children, and time-varying factors of stressful life events and medical errors.

References to the Supplement

1. Hernán MA. Methods of Public Health Research — Strengthening Causal Inference from Observational Data. N Engl J Med 2021;385(15):1345–8.

2. Hernán MA, Robins JM. Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available: Table 1. Am J Epidemiol 2016;183(8):758–64.

3. Hernán MA, Sauer BC, Hernández-Díaz S, Platt R, Shrier I. Specifying a target trial prevents immortal time bias and other self-inflicted injuries in observational analyses. Journal of Clinical Epidemiology 2016;79:70–5.

4. Fang Y, Bohnert ASB, Pereira-Lima K, et al. Trends in Depressive Symptoms and Associated Factors During Residency, 2007 to 2019: A Repeated Annual Cohort Study. Ann Intern Med 2021;M21-1594.

5. Kroenke K, Spitzer RL, Williams JBW. The PHQ-9: Validity of a brief depression severity measure. J Gen Intern Med 2001;16(9):606–13.

6. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders [Internet]. Fifth Edition. American Psychiatric Association; 2013 [cited 2021 Dec 5]. Available from: https://psychiatryonline.org/doi/book/10.1176/appi.books.9780890425596

7. Kroenke K. Enhancing the clinical utility of depression screening. CMAJ 2012;184(3):281–2.

8. Soleimani H, Adler-Milstein J, Cucina RJ, Murray SG. Automating Measurement of Trainee Work Hours. J Hosp Med [Internet] 2021 [cited 2021 Dec 2];16(7). Available from: https://www.journalofhospitalmedicine.com/jhospmed/article/238685/hospital-medicine/automating-measurement-trainee-work-hours

9. McCrae RR, Costa PT. A contemplated revision of the NEO Five-Factor Inventory. Personality and Individual Differences 2004;36(3):587–96.

10. Costa Jr. PT, McCrae RR. Stability and Change in Personality Assessment: The Revised NEO Personality Inventory in the Year 2000. Journal of Personality Assessment 1997;68(1):86–94.

11. Taylor SE, Way BM, Welch WT, Hilmert CJ, Lehman BJ, Eisenberger NI. Early Family Environment, Current Adversity, the Serotonin Transporter Promoter Polymorphism, and Depressive Symptomatology. Biological Psychiatry 2006;60(7):671–6.

12. Taylor SE, Eisenberger NI, Saxbe D, Lehman BJ, Lieberman MD. Neural Responses to Emotional Stimuli Are Associated with Childhood Family Stress. Biological Psychiatry 2006;60(3):296–301.

13. 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(6):557.

14. Clare PJ, Dobbins TA, Mattick RP. Causal models adjusting for time-varying confounding—a systematic review of the literature. International Journal of Epidemiology 2019;48(1):254–65.

15. Mansournia MA, Etminan M, Danaei G, Kaufman JS, Collins G. Handling time varying confounding in observational research. BMJ 2017;j4587.

16. Hernán MA, Robins JM. Causal Inference: What If. Boca Raton: Chapman & Hall/CRC;

17. Guglielmetti LC, Gingert C, Holtz A, Westkämper R, Lange J, Adamina M. Nationwide study on stress perception among surgical residents. British Journal of Surgery 2021;108(Supplement_4):znab202.091.

18. Miglioretti DL, Heagerty PJ. Marginal Modeling of Nonnested Multilevel Data using Standard Software. American Journal of Epidemiology 2006;165(4):453–63.

19. Sen S, Kranzler HR, Didwania AK, et al. Effects of the 2011 Duty Hour Reforms on Interns and Their Patients: A Prospective Longitudinal Cohort Study. JAMA Intern Med 2013;173(8):657.

20. Rose G. Sick individuals and sick populations. International Journal of Epidemiology 2001;30(3):427–32.

21. Hu Y-Y, Ellis RJ, Hewitt DB, et al. Discrimination, Abuse, Harassment, and Burnout in Surgical Residency Training. N Engl J Med 2019;381(18):1741–52.

22. Fang Y, Forger DB, Frank E, Sen S, Goldstein C. Day-to-day variability in sleep parameters and depression risk: a prospective cohort study of training physicians. npj Digit Med 2021;4(1):28.