Supplementary Appendix

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

Supplement to: Basner M, Asch DA, Shea JA, et al. Sleep and alertness in a duty-hour flexibility trial in internal medicine. N Engl J Med 2019;380:915-23. DOI: 10.1056/NEJMoa1810641

Supplementary Appendix

Sleep and Alertness in a Duty-Hour Flexibility Trial in Internal Medicine

Mathias Basner, David A. Asch, Judy A. Shea, Lisa M. Bellini, Michele Carlin, Adrian J. Ecker, Susan K. Malone, Sanjay V. Desai, Alice L. Sternberg, James Tonascia, David M. Shade, Joel T. Katz, David W. Bates, Orit Even-Shoshan, Jeffrey H. Silber, Dylan S. Small, Kevin G. Volpp, Christopher G. Mott, Sara Coats, Daniel J. Mollicone, David F. Dinges, and the iCOMPARE Research Group

Corresponding author: Mathias Basner, MD, PhD, MSc, Department of Psychiatry, Perelman School of Medicine, University of Pennsylvania, 1019 Blockley Hall, 423 Guardian Drive, Philadelphia PA 19104, basner@pennmedicine.upenn.edu

Table of Contents

1. iCOMPARE Research Group

Margot Boigon, MD, FACP; Abington Memorial Hospital Program

Jill Patton, DO; Advocate Lutheran General Hospital Program

Donna Astiz, MD; Atlantic Health (Morristown) Program

Cheryl O'Malley, MD; University of Arizona College of Medicine-Phoenix Internal Medicine Program

Richard J. Hamill, MD; Baylor College of Medicine Program

Michael Rosenblum, MD, FACP; Baystate Medical Center/Tufts University School of Medicine Program

Charles Smith, MD; Beth Israel Deaconess Medical Center

Joel Katz, MD; Brigham and Women's Hospital Program

Maria Yialamas, MD; Brigham and Women's Hospital Program

Dominick Tammaro, MD; Brown University Program

Jennifer Bolyard, MD; Canton Medical Education Foundation/NEOMED Program

Claudia A. Kroker-Bode, MD, PhD; Carilion Clinic-Virginia Tech Carilion School of Medicine Program

Michael J. McFarlane, MD; Case Western Reserve University (MetroHealth) Program

Keith Armitage, MD; Case Western Reserve University/University Hospitals Case Medical Center Program

Amanda Ewing, MD, FACP; Cedars-Sinai Medical Center Program

Abby L. Spencer, MD, MS, FACP; Cleveland Clinic

Tammy Wichman, MD; Creighton University School of Medicine

Richard Paluzzi, MD; Drexel University College of Medicine/Hahnemann University Hospital Program

Aimee Zaas, MD, MHS; Duke University Hospital Program

Suzanne Kraemer, MD; East Carolina University

Benjamin Goodman, MD; Eastern Virginia Medical School Program

Lorenzo Di Francesco, MD; Emory University Program

Mary Harris, MD; Geisinger Health System Program

Jillian Catalanotti, MD, MPH, FACP; George Washington University Program

Sal Pindiprolu, MD; Georgetown University Hospital/Washington Hospital Center Program

Paul Foster, MD, FACP; Greater Baltimore Medical Center Program

Sean M. Drake, MD; Henry Ford Hospital/Wayne State University Program

Sanjay Desai, MD; Johns Hopkins University Program

Erica N. Johnson, MD; Johns Hopkins University/Bayview Medical Center Program

Elizabeth Nilson, MD; Lahey Clinic Program

William D. Surkis, MD, FACP; Lankenau Medical Center

Jonathon Doroshow, MD, FACP Lankenau Medical Center

Jatin M. Vyas, MD, PhD; Massachusetts General Hospital Program

Michael Frank, MD; Medical College of Wisconsin Affiliated Hospitals Program

Eric H. Green, MD, MSc; Mercy Catholic Medical Center Program

Cinnamon Bradley, MD; Morehouse School of Medicine Program

Soma Wali, MD; Olive View/UCLA Medical Center Program

Sapna Kuehl, MD; St. Agnes HealthCare Program

Harvey Friedman, MD, FCCP, FACP; St. Francis Hospital of Evanston Program

Ronald Witteles, MD; Stanford University Program

Marissa A. Blum, MD, MSHPR; Temple University Hospital Program

Curtis Mirkes, DO; Texas A&M College of Medicine-Scott and White Program

Michael Phy, DO; Texas Tech University (Lubbock) Program

Emily Stewart, MD, FACP; Thomas Jefferson University Program

Richard Kopelman, MD; Tufts Medical Center Program

Kari Roberts, MD; Tufts Medical Center Program

Jodi Friedman, MD; UCLA Medical Center Program

Brian Gable, MD; Robert Wood Johnson Medical School (Camden)/Cooper University Hospital Program

Eric Warm, MD; University Hospital/University of Cincinnati College of Medicine Program

Suzanne Brandenburg, MD; University of Colorado

Steven V. Angus, MD; University of Connecticut Program

Leigh M. Eck, MD; University of Kansas School of Medicine Program

Susan D. Wolfsthal, MD; University of Maryland Program

Richard M. Forster, MD, FACP; University of Massachusetts Program

James O'Dell, MD; University of Nebraska Medical Center College of Medicine Program

Debra L. Bynum, MD, MMEL; University of North Carolina Hospitals Program

Lisa Bellini, MD; University of Pennsylvania Program

Mark Pasanen, MD; University of Vermont/Fletcher Allen Health Care Program

Kenneth P. Steinberg, MD; University of Washington Program

Stephanie Ann Call, MD, MSPH; Virginia Commonwealth University Health System Program
Hal H. Atkinson, MD, MS; Wake Forest University School of Medicine Program
Melvin Blanchard, MD; Washington University/B-JH/SLCH Consortium Program
Thomas Ciesielski, MD; Washington University/B-JH/SLCH Consortium Program
Nathan Lerfald, MD; West Virginia University
Mark D. Siegel, MD; Yale - New Haven Medical Center Program

2. Description of Standard and Flexible Duty-Hour Rules

Table S1: Duty-Hour Policies for Inpatient Rotations in Flexible Programs and Standard Programs* (modified from Desai et al.[1])

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

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

PGY denotes postgraduate year.

3. Program Selection for the Sleep & Alertness Sub-study

Programs were selected for the Sleep & Alertness sub-study in a stepwise process until sample size requirements were met. Programs had to fulfill the following criteria to qualify for selection: (1) use the University of Pennsylvania as the IRB of record or have an IRB waiver; (2) train at least 50 interns; (3) implement flexible duty hours on general medicine, medical intensive care, cardiology, and/or coronary care rotations (flexible duty-hour programs only). The 6 programs from each arm of the study (flexible and standard rules) were selected to reflect diversity in program size and geographic region. All of the programs approached by the study team agreed to participate in the Sleep & Alertness sub-study and none dropped out. Selected programs are compared to those that were not selected in Table S2.

Table S2: Comparison of Selected Flexible and Standard Programs to Remaining Programs

Characteristic	Programs						
	Flexible (n = 32)			Standard (n = 31)			
	Not selected (N=26)	Selected (N=6)	P (6 vs. 26)	Not selected (N=25)	Selected (N=6)	P (6 vs. 25)	P (6 Flexible vs. 6 Standard)
N of Programs	26	6		25	6		
Program Type – N (%)			0.11*			0.82*	1.00*
Community	1 (4)	0 (0)		3 (12)	0 (0)		
University	14 (54)	6 (100)		15 (60)	5 (83)		
Both Community and University	11 (42)	0 (0)		7 (28)	1 (17)		
Geographic Region N (%)			0.45*			0.17*	0.07*
Northeast	6 (23)	2 (33)		10 (40)	4 (67)		
Midwest	8 (31)	0 (0)		3 (12)	2 (33)		
South	10 (38)	3 (50)		8 (32)	0 (0)		
Mountain or Pacific	2 (8)	1 (17)		4 (16)	0 (0)		
Mean N of Residents per Program (±SD)	86.6 ± 37.7	128.8 ± 25.1	0.01†	93.4 ± 46.0	121.7 ± 41.0	0.18†	0.72†
Resident-to-bed ratio	0.57	0.77	0.07†	0.60	0.56	0.74	0.18†

*Fisher's exact test; †Student's t-test (equal variances); SD: Standard Deviation: N: Number

4. Smartphone App Screenshots

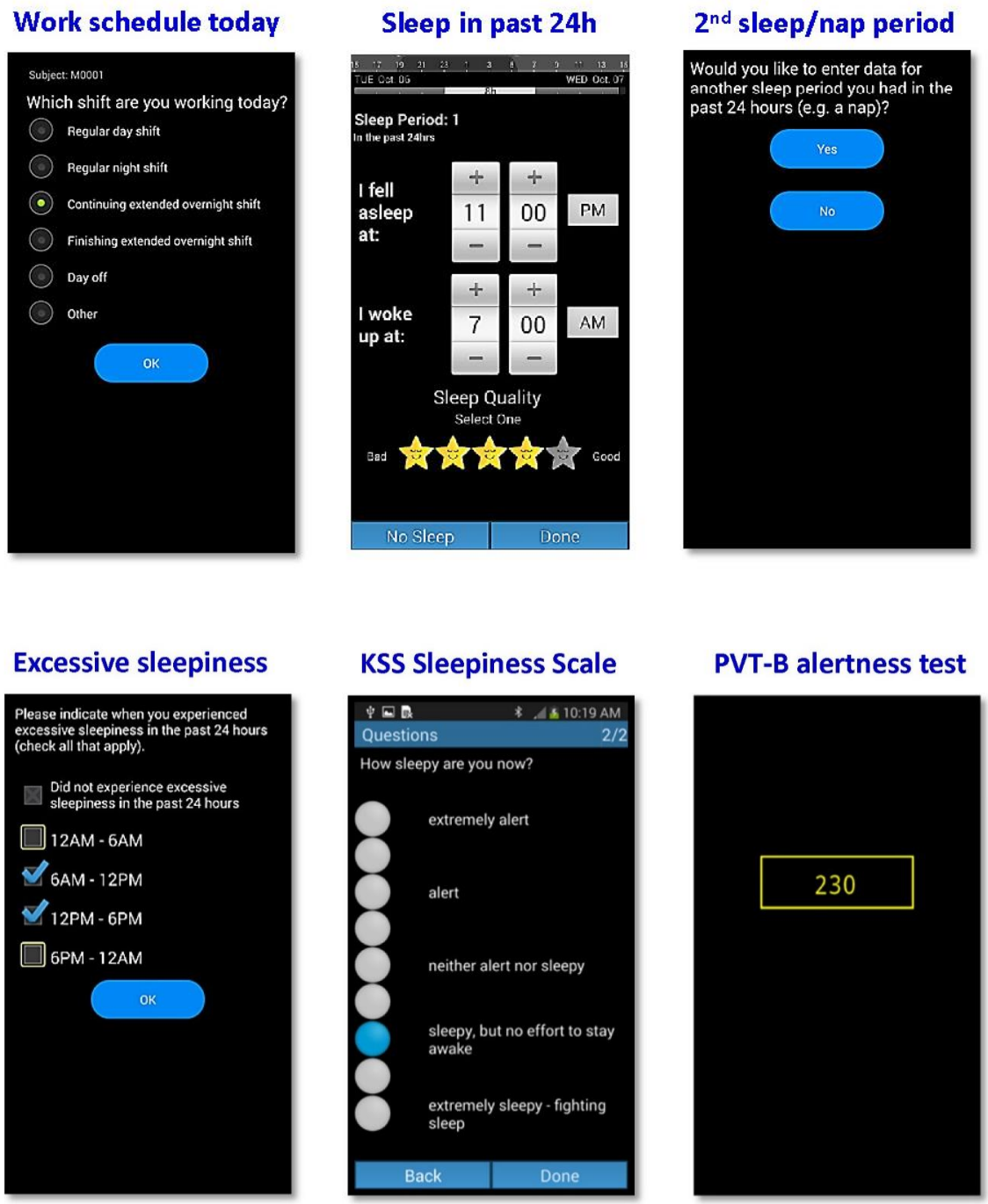

Figure S1: Smartphone App Screenshots.

KSS: Karolinska Sleepiness Scale[2]; PVT-B: Brief Psychomotor Vigilance Test[3].

5. Actigraphy Sleep Scoring

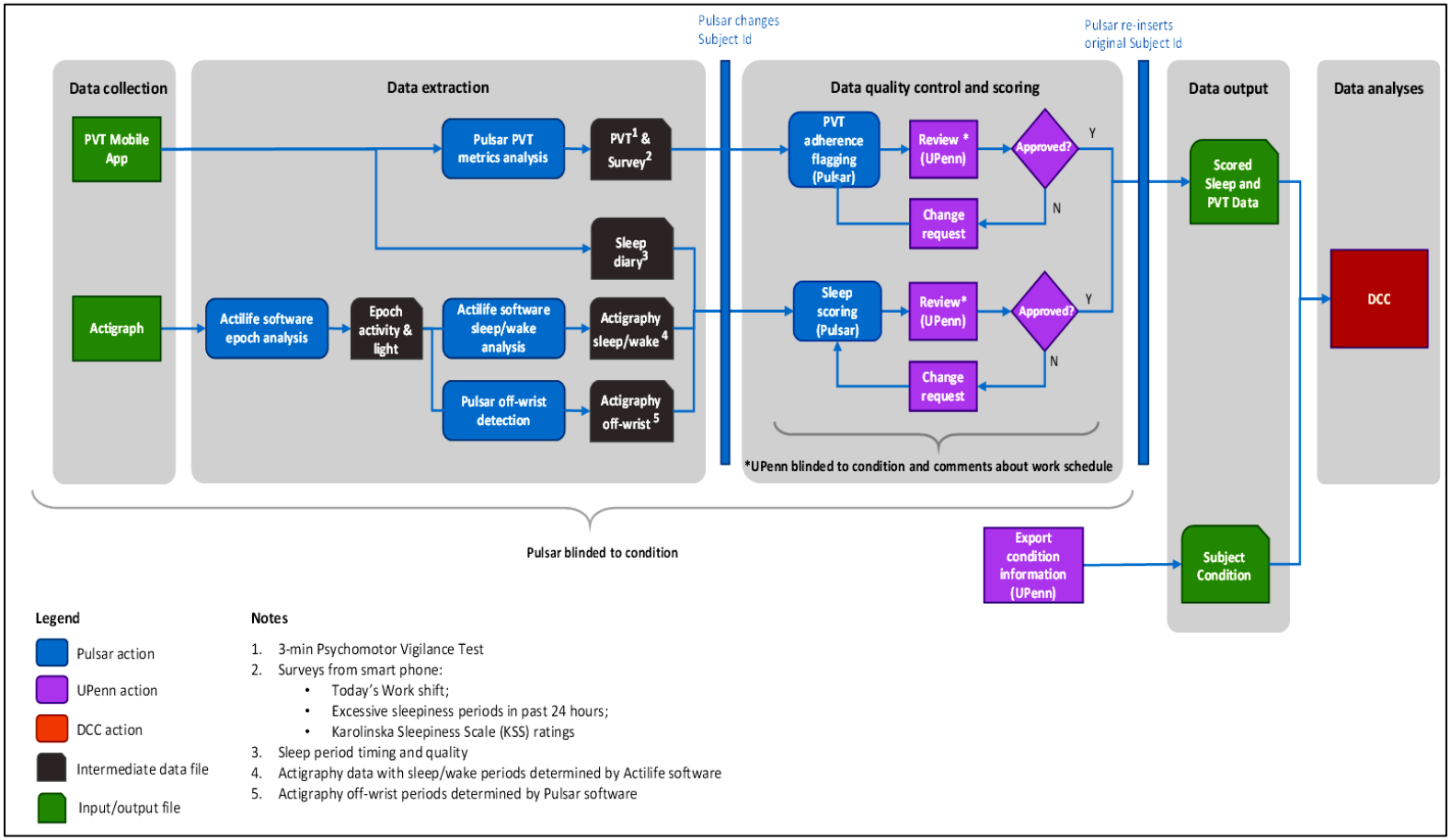

Figure S2: Scoring and Review Process of Actigraphy Data.

Pulsar: Pulsar Informatics Inc.; UPenn: Clinical Coordinating Center at the University of Pennsylvania; DCC: Data Coordinating Center at the Johns Hopkins University.

The scoring procedures for actigraphy data are shown in Figure S2. We intended to monitor wrist movements and ambient light intensity data for each intern for 13 consecutive 24-hour periods with the gt3x actiwatch from the "Actigraph" corporation. Watches were handed out to interns on day 1 and collected from them again on day 15. We expected thirteen 24 hour actigraphy measurement periods between 9 pm on day 1 and 9 pm on day 14. Wrist movement and ambient light data were recorded at a sample rate of 30 Hz and 1 Hz, respectively. After the Clinical Coordinating Center (CCC) received the actiwatches back from the interns, the data were downloaded and forwarded to Pulsar Informatics Inc. (Pulsar) for data processing. Pulsar used the software of the "Actigraph" corporation (Actilife software version 6.13.3 standard settings) and the Sadeh algorithm[4] to transform 30 Hz data into 1-minute activity epochs and to perform an automatic sleep/wake/off-wrist analysis with the Actilife software algorithm. The actiwatches have built-in off-wrist detection, which, however, was found to be unreliable. For this reason, Pulsar developed its own off-wrist detection algorithm. The automatic sleep/wake scoring of the Actilife software was combined with information derived from Pulsar's off-wrist detection to generate a revised sleep/wake/off-wrist scoring. Finally, information from sleep logs relevant for sleep/wake scoring (i.e., sleep periods entered by interns and time of day when the sleep log was filled out) was extracted from data collected with the Smartphone. Based on actigraphy sleep/wake/off-wrist scoring and sleep log information, an algorithm automatically classified each 1-minute epoch into sleep, wake, or unknown (missing) according to the matrix shown in Figure S3.

	Sleep Diary had sleep time entered (S)	Sleep Diary did not have sleep time entered (W)
Actigraphy Reviewed indicated sleep (S)	Scored as Sleep (S) based on agreement of actigraphy and diary	Scored as Sleep (S) based on actigraphy
Actigraphy Reviewed indicated wake (W)	Scored as Wake (W) based on actigraphy	Scored as Wake (W) based on actigraphy
Actigraphy Reviewed indicated missing or off-wrist (O)	Scored as Sleep (S) based on diary	Scored as Unknown (U)

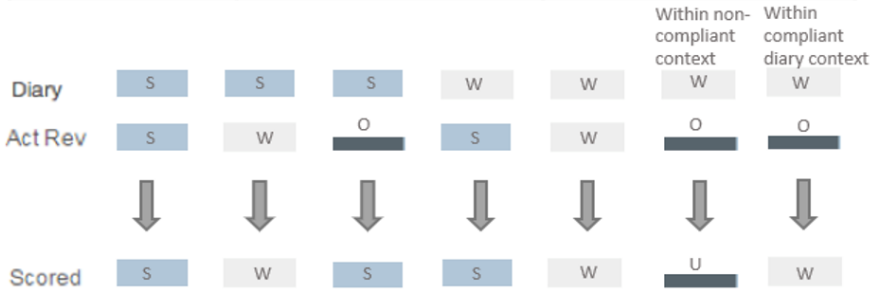

Figure S3: Sleep Scoring Matrix.

When there was agreement between the sleep diary (Diary) and the actiwatch (Act Rev) scoring of state was straightforward (see data columns 1 and 5 at the bottom of Figure S3). When they disagreed, objective information gathered with the actiwatch (Act Rev) was used instead of subjective sleep diary information (see data columns 2 and 4 at the bottom of Figure S3). If no actigraphy information was available, sleep diary information was used to classify wake/sleep times (see data columns 3 and 7 at the bottom of Figure S3). If no sleep period was recorded in the diary for 24 hours or longer (i.e., no smartphone indication from the subject of sleep times) sleep time per the diary was classified as non-compliant (see data column 6 at the bottom of Figure S3). Only if both the diary sleep time was classified as non-compliant and the actigraph was off-wrist (or not collecting data due to a technical failure), was the sleep/wake state scored as unknown (see data column 6 at the bottom of Figure S3). In a small number of cases (0.17%), exceptions were made to the rules represented by the columns at the bottom of Figure S3 and the

Scored row was adjusted instead of the Act Rev row. An exception would be made, for example, if a device was recording but being inconsistently worn by a subject. In this case the Actilife software might generate a sleep and wake pattern from the fragmented data without factoring in non-compliance. A correction would then be made in the Scored row. All available information was plotted for visual data inspection by Pulsar Informatics (see Figure S4 for an example).

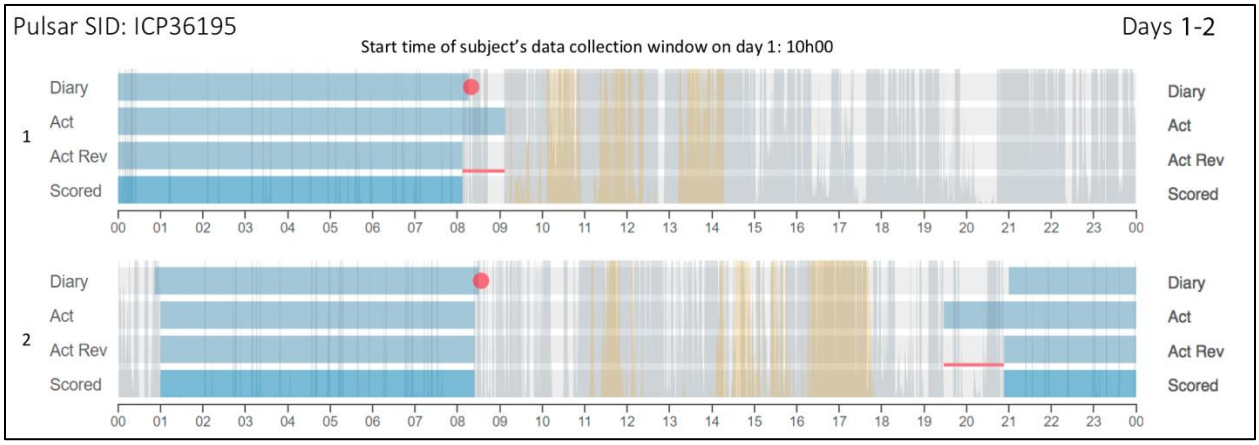

Figure S4: Example Data Review Plot for Two Data Collection Days in One Intern.

Actigraph activity counts are shown as vertical grey spikes and light intensity is shown as vertical yellow spikes for each 1-min epoch. Red dots represent the time of day when the sleep diary information was entered by interns. Sleep periods are shown in blue horizontal bars for self-reported diary entries (Diary), for automatic Actilife software sleep/wake/off-wrist scoring (Act), and for the revised sleep/wake/off-wrist scoring based on Pulsar's off-wrist algorithm (Act Rev). Manual corrections to the Act Rev scoring are shown as horizontal thin red lines. The final sleep/wake scoring appears in the row named "Scored" (i.e., sleep time = blue bar at the bottom of the figure).

Obvious Actilife software classification errors were manually corrected. These corrections were individually documented in an Excel spreadsheet and are shown as thin red lines in the Act Rev scoring in Figure S4. The manually corrected review plots and Excel spreadsheets were then provided to CCC sleep experts, Dr. Dinges and Dr. Basner, for review while blinded to intervention arm. They could request changes to Pulsar's scoring and request additional changes to the scoring based on their expert judgment from years of using actiwatches and sleep diaries.

These were documented in the Excel spreadsheet. Pulsar then addressed the changes requested by Dr. Dinges or Dr. Basner and circulated a new version of the revised data review plot and Excel spreadsheet. This process was repeated until all changes were approved by Dr. Dinges and Dr. Basner. In the example in Figure S4 above, the automatic actigraphy algorithm indicated that sleep time ended around 9 am on day 1. However, the subject completed the diary shortly after 8 am (and thus must have been awake). The subject also indicated a wake-up time of around 8 am in the diary, and activity counts clearly indicate waking activity shortly after 8 am. The scoring was thus revised from sleep to wake.

6. Measures Taken to Prevent Bias in Scoring of Actigraphy and PVT-B Data

As the data review process described in detail above includes assessments by trained human experts, several measures were taken to prevent systematic bias.

The CCC was provided with a list of subject IDs by the Data Coordinating Center. These subject IDs were randomly assigned to interns from Standard and Flexible programs by the CCC. The only information about individual interns available to Pulsar Informatics was this subject ID. At no time during the data acquisition or analysis process did Pulsar Informatics have knowledge about study site, study arm, or other characteristics (like age or gender) that could have potentially identified an intern. Therefore, Pulsar was blinded to study condition.

Before sharing sleep/wake scoring sheets with sleep experts at the CCC, Pulsar Informatics assigned a new unique subject ID to each intern. Furthermore, date information was removed from the sheets, and days were instead counted from 1 to 15. That way, sleep experts at the CCC were also blinded relative to condition when they reviewed Pulsar's initial sleep/wake scoring. For PVT-B data, the same process was adopted (i.e., a new subject ID was assigned and date information was removed). In addition, subject comments that could have revealed a subject's condition (Standard or Flexible) were obscured by Pulsar before PVT-B data review with the CCC.

Actigraphy, PVT-B and KSS data files were sent to the DCC by Pulsar Informatics using the original subject ID assigned by the CCC. The subject information itself (i.e., age, gender, and ethnicity of the intern) were sent to the DCC by the CCC in a separate data file. Importantly, letters A and B were randomly assigned to study arms (i.e., Standard or Flexible). Therefore, the

DCC was also blinded to condition during data analysis. The DCC was unblinded by the CCC after data analysis was completed.

7. PVT-B Scoring and Review Process

For each intern, the following summary scores over all PVT-Bs performed by the intern were plotted (see Figure S5 for an example):

a) Number of errors of commission (reaction times <130 ms; false starts).

b) Number of errors of omission (reaction times ≥355 ms; lapses).

c) Range from average fastest 10% reaction times to average slowest 10% reaction times.

d) Average reaction time.

e) Comments entered in the smartphone by an intern after completing the PVT-B.

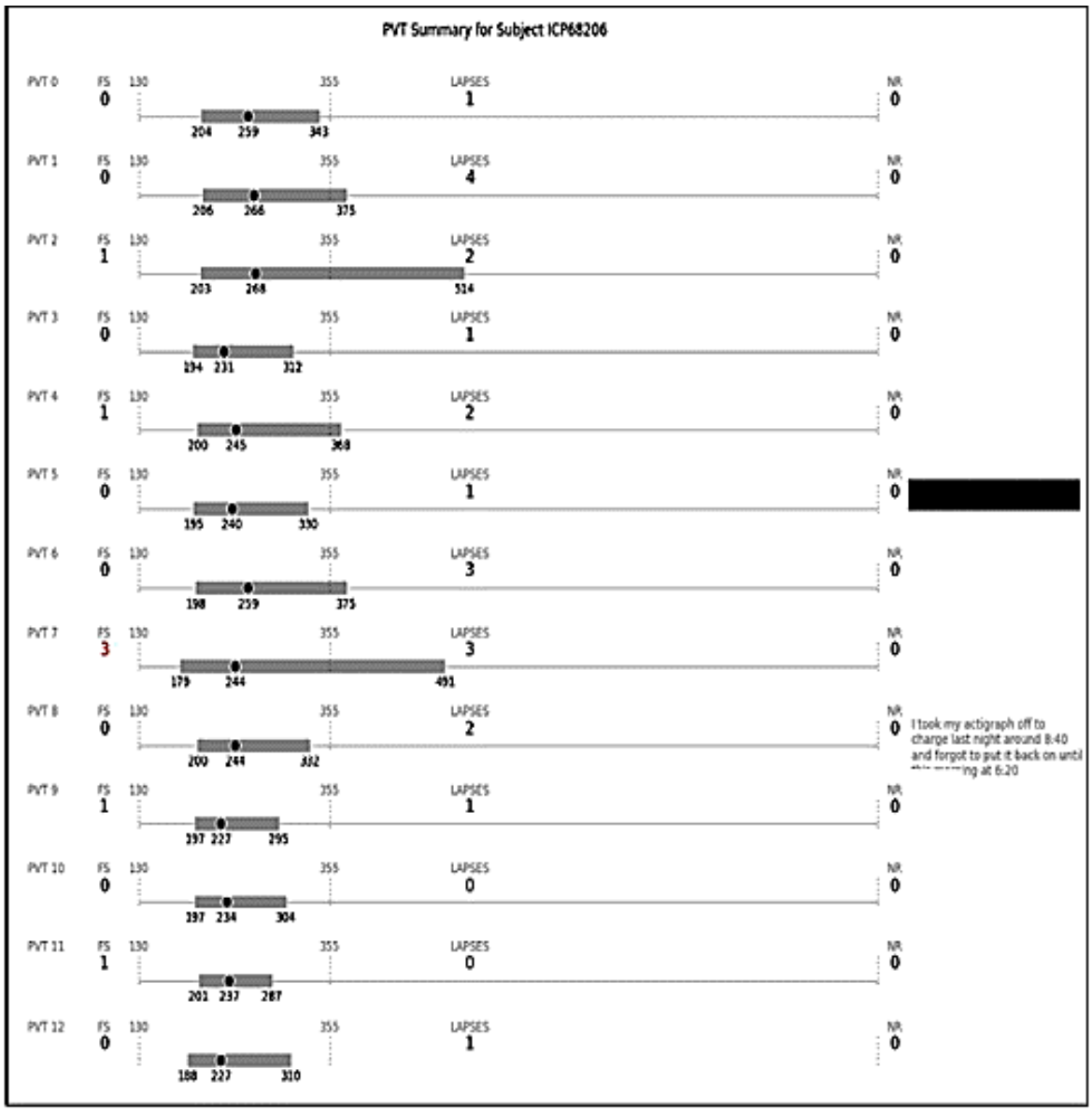

Figure S5: Consecutive PVT-B Test Bout Results for a Sample Single Intern.

Each row shows the extracted results of a single 3-minute PVT-B bout. The number of false start (FS) premature responses per test bout is shown on the far left (FS column). The average 10% fastest reaction times (RT), average RTs and slowest 10% RTs are shown from left to right along each gray horizontal bar (block dot is the average RT in milliseconds). The number of lapses of attention (RTs>355ms) per test bout are shown in the middle of each plot. The number of non-responses (NR) before the stimulus timed out to 30,000 milliseconds are shown on the far right. Any comments about the test the subject entered into the smartphone after the test are also shown in the far right column. In the case of this subject, two test bouts had comments. Test bout PVT 5 had a comment from the subject about work schedule, so this was blacked out by Pulsar before Drs. Dinges and Basner reviewed the data. Test bout PVT 8 had a comment from the subject about forgetting to put the actiwatch on.

In several review meetings, Pulsar Informatics and CCC investigators, Drs. Dinges and Basner, reviewed these PVT-B summary plots, blind to subject condition. Based on the data shown in the plots, individual subjects were classified (using all PVT-Bs the individual subject performed) according to the extent to which their PVT-B data indicated they were adherent to the task instructions to respond as quickly as possible to the light stimulus (i.e., millisecond counter), but not to respond prematurely (i.e., before the light stimulus turned on). Definitions of the three PVT-B adherence categories follow:

a) Adherent: PVT-B data reflect an effort to do the task correctly, and comments left by the subject do not suggest non-adherence.
b) Likely non-adherent: PVT-B data reflect a consistently poor effort to do the task correctly, but comments left by the subject do not suggest non-adherence (e.g., performing the task while brushing teeth).
c) Non-adherent: PVT-B data reflect a consistently poor effort to do the task correctly, and comments left by the subject do suggest non-adherence (e.g., performing the task while brushing teeth).

Furthermore, if an intern left a comment after performing the PVT-B, each comment was classified in one of the following categories (comments that could have revealed the study arm were hidden by Pulsar Informatics Inc., to avoid a biased classification by the study team):

a) No comment
b) Subject reported distraction or engaged in secondary activity at time of test (e.g., noisy environment, putting on coat, boarding a bus)
c) Subject reported non-fatigue related impairment (e.g., physical injury to hand, pain, intoxication, illness)

d) Subject reported other comment (e.g., thinking about something else)

These classification variables were used for adjusting in statistical models for sensitivity analyses.

8. Single Imputation of Actigraphy Data

Missing actigraphy data were first imputed with sleep log data (for those instances where interns entered a sleep period in their sleep log and actigraphy information was available, agreement between actigraphy sleep-wake scoring and the sleep log was 94.1%). After imputation with sleep log data, the sleep-wake state was known in 95.1% and 94.2% of standard and flexible programs, respectively. For the remaining 1-minute epochs with unknown sleep-wake state, we used single imputation stratifying by program (standard/flexible), shift type (day, night, off, etc.), and time of day. For example, if for a given standard program intern, sleep-wake state at 10:53 pm was unknown on a shift classified as a day shift by the intern, we imputed 0.545 minutes sleep for this minute, which reflects the percent of interns with known sleep-wake state in the same program, on the same shift, and at the same time of day sleeping. The averages used for imputation are shown in Figure S6 and Figure S7 for Standard and Flexible programs, respectively.

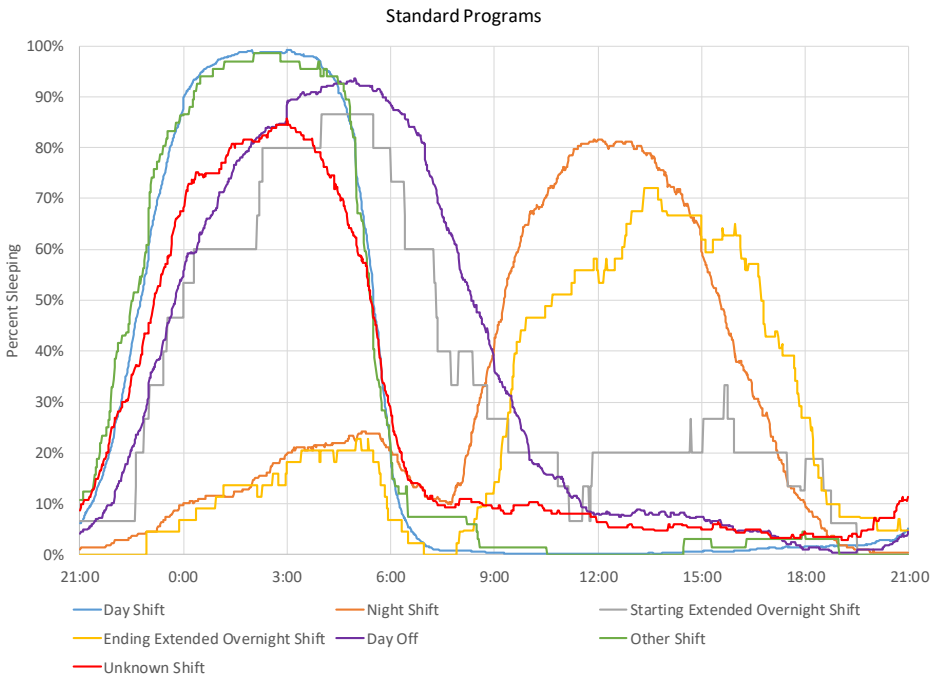

Figure S6: Percent Sleeping by Shift Type and Time of Day (Standard Programs).

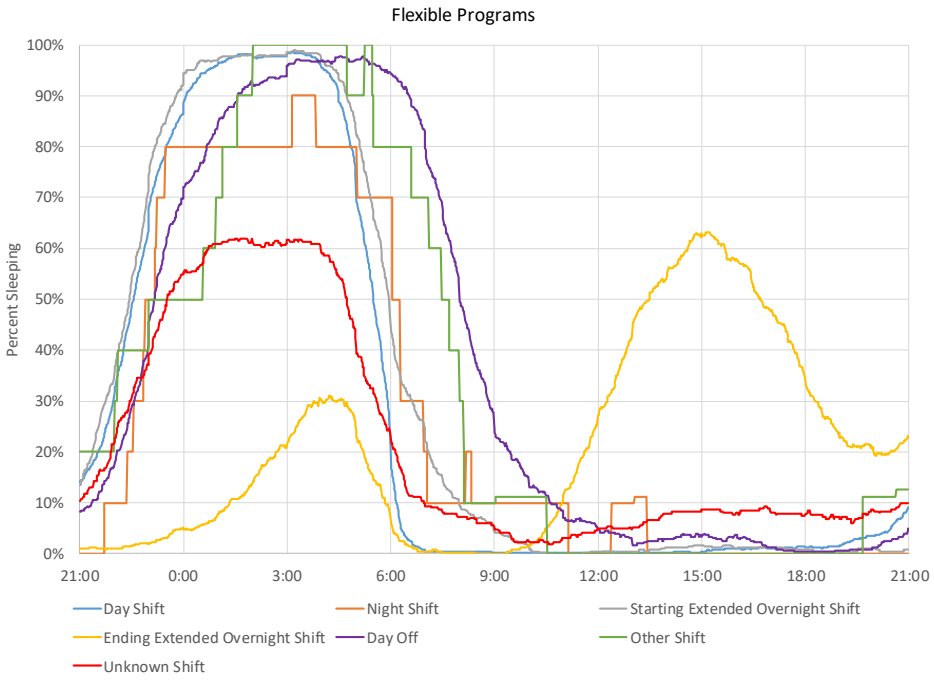

Figure S7: Percent Sleeping by Shift Type and Time of Day (Flexible Programs).

9. Classification of Shifts

Table S3 shows the classification of shifts by interns in standard and flexible programs. A minority of interns in standard programs indicated "starting or finishing an extended overnight shift". It is likely that these were misclassifications by the intern rather than a breach of the standard duty-hour rules. These ratings, together with shifts classified as "other" or "missing", were re-classified as "other". Likewise, a minority of interns in flexible programs indicated a "regular night shift" in the Smartphone App. These ratings, together with shifts classified as "other" or "missing", were re-classified as "other" before analyses by shift type were performed.

Table S3: Shift Classifications by Interns in Flexible Programs (N=2664) and Standard Programs (N=2509).

	Flexible Programs	**Standard Programs**
Regular Day Shift	38.8%	63.3%
Regular Night Shift	0.4%	8.8%
Starting Extended Overnight Shift	16.5%	0.6%
Finishing Extended Overnight Shift	16.5%	1.8%
Day off	13.6%	11.8%
Other	0.4%	2.7%
Missing	13.9%	10.9%

10. Participant Flow Chart

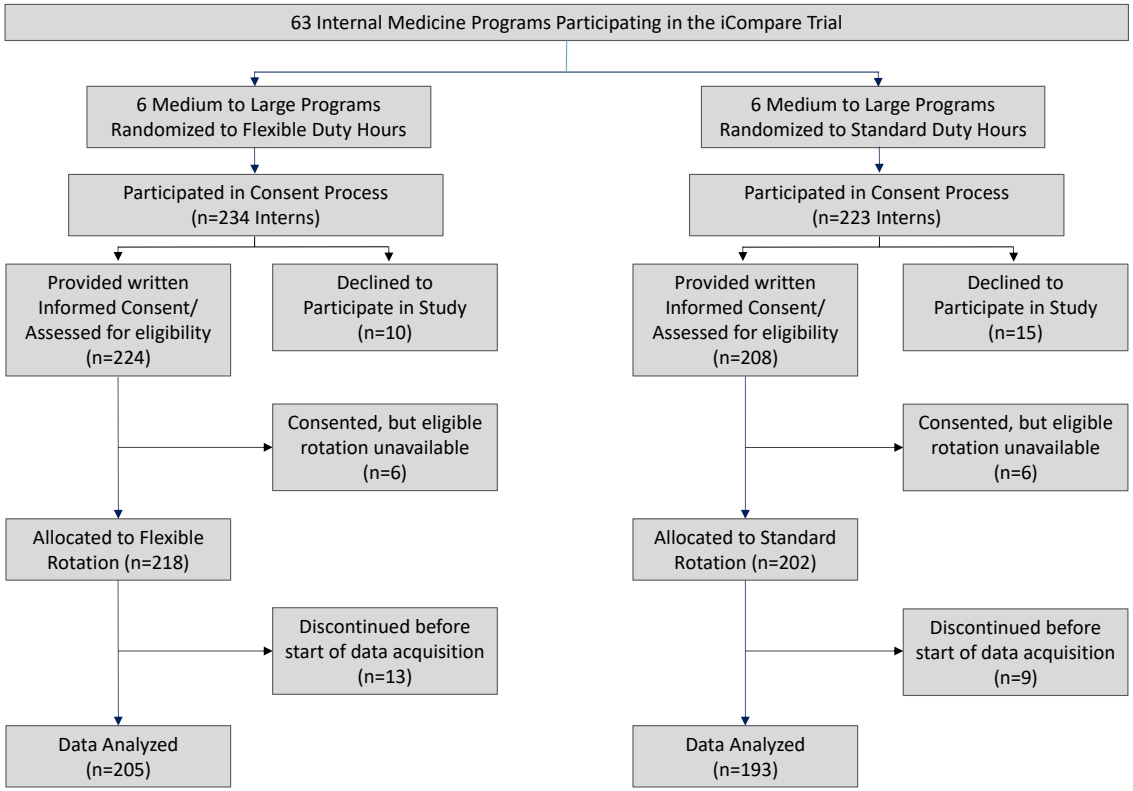

Figure S8: Participant flow chart

11. Characteristics of Interns and Completeness of Data

Table S4: Characteristics of Interns and Completeness of Data for Programs by Policy Group

	Flexible Programs	Standard Programs
Interns		
Number of Participating Interns	205	193
Age*	27.9 (0.2) years	27.8 (0.2) years
% female	46.3%	51.8 %
Actigraphy		
Technical Failure [% of expected 13 days]†	0.14 (0.07) Days [1.1%]	0.16 (0.08) Days [1.2%]
Off-wrist [% of expected 13 days]†	1.52 (0.21) Days [11.7%]	1.37 (0.21) Days [10.5%]
Imputed from Sleep Log [% of expected 13 days]†	0.88 (0.09) Days [6.8%]	0.90 (0.09) Days [6.9%]
Known Sleep-Wake State [% of expected 13 days]†	12.24 (0.21) Days [94.2%]	12.36 (0.21) Days [95.1%]
PVT-B‡		
Number of collected PVT-Bs per Intern [% expected]§	11.9 (0.5) PVT-Bs [85.0%]	12.4 (0.5) PVT-Bs [88.6%]
Interns Classified as at Least Likely Non-Adherent, N [%]‖	6 [2.9%]	12 [6.2%]
PVT-Bs with intern comments indicating distraction or non-fatigue related impairment [%]	5.5% (1.1%)	4.3% (1.1%)
PVT-Bs collected between 6 am and 9 am	57.7% (5.7%)	59.0% (5.6%)

Values represent means (standard errors).

All tests comparing characteristics in flexible and standard programs were statistically not significant (i.e., all P>0.05).

*The age of 4 interns (all in flexible programs) was imputed with the average age of 28 years.

†relative to 13 twenty-four hour periods from 9:00 pm on study day 1 until 8:59 pm on study day 14

‡The PVT-B was performed after the survey. Survey data (KSS) were collected without corresponding PVT-B data in two instances.

§relative to 14 expected tests on mornings of study days 2-15

‖Only one intern from the standard policy group was classified as non-adherent, all other interns were classified as likely non-adherent.

12. Sensitivity Analyses for Noninferiority Tests

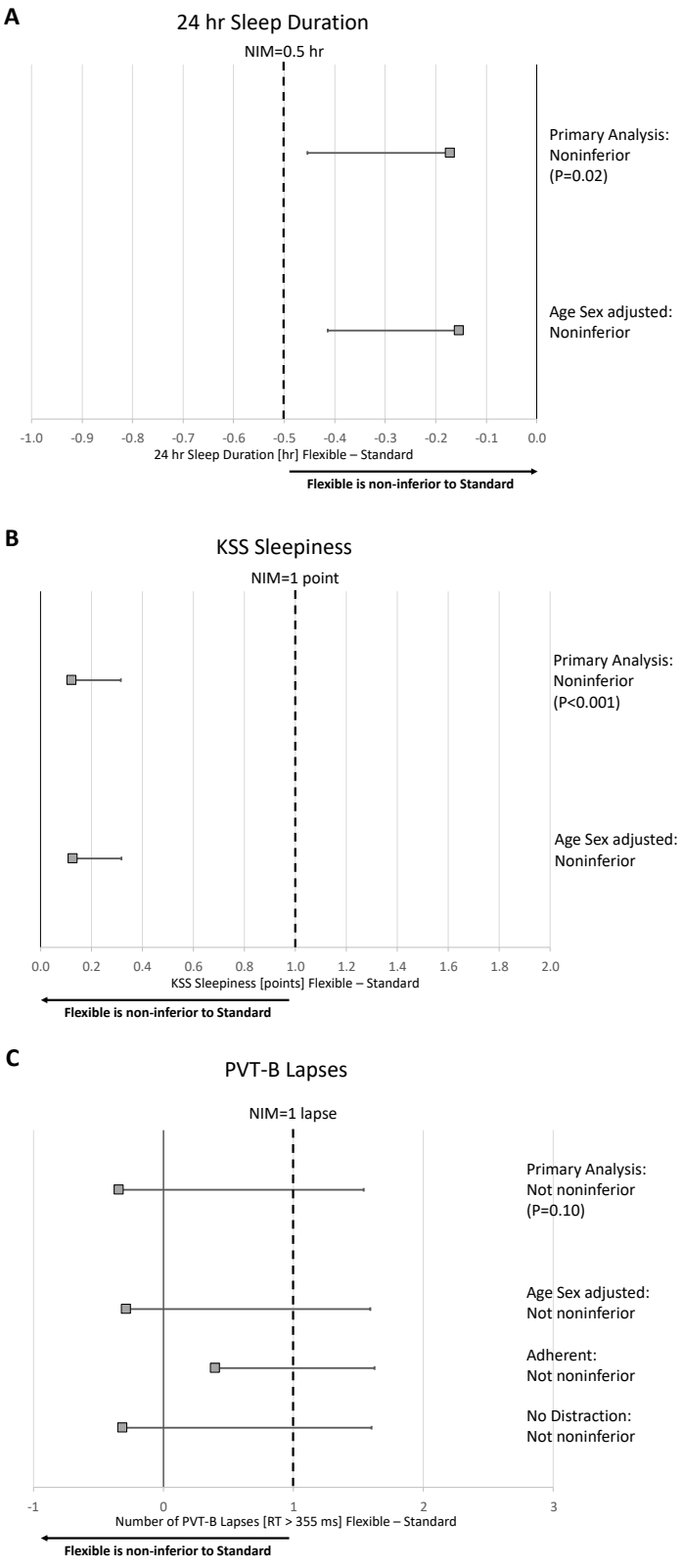

Figure S9: Noninferiority analysis results are shown for average 24h sleep duration (Panel A), average Karolinska Sleepiness Scale score (Panel B; higher values reflect higher levels of sleepiness), and average Psychomotor Vigilance Test (PVT-B) lapses (Panel C; higher values reflect lower levels of alertness). Noninferiority tests were one-sided with noninferiority margins (indicated by NIM in the figures) of 0.5 h, 1 point on the 9-point KSS scale, and 1 additional PVT-B lapse, respectively. Primary and sensitivity analyses indicate that 24 h sleep duration and subjective ratings of sleepiness were noninferior in flexible compared to standard programs, whereas findings for objectively assessed alertness via PVT-B lapses were inconclusive. Unadjusted one-sided 95% confidence intervals and P-values (reflecting noninferiority tests) are shown for the 3 primary outcomes; sleep duration and KSS sleepiness in flexible programs remained noninferior to standard programs at α=0.05 after Benjamini-Hochberg adjustments[5] for multiple testing (N=3 comparisons). The figure also shows unadjusted 95% confidence intervals for age and sex adjusted sensitivity analyses of each of the 3 primary outcomes and analysis of average PVT-B for those classified as adherent and those classified as not distracted; these confidence intervals have not been adjusted for multiple comparisons and inferences drawn from these intervals may not be reproducible.

13. Analyses of Additional Outcomes Stratified by Shift Type

Table S5: Sleep Duration, Sleepiness, and Sleep Quality Among Interns by Shift Type and Duty-Hour Policy Group

	Shift Type	**Sleep Quality†** Estimate (95% CI) §	**Excessive Sleepiness‡** % Days (95% CI)§	**High KSS Score (8 or 9)** % Days (95% CI)§	**Sleep Duration <7 h** % Days (95% CI)§	**Sleep Duration <6 h** % Days (95% CI)§
Flexible Programs	Day	3.7 (3.5; 3.8)	59.4 (51.8; 67.1)	5.4 (2.2; 8.6)	50.4 (45.9; 54.9)	20.5 (15.5; 25.5)
	Day 1 Overnight	3.7 (3.6; 3.8)	47.7 (40.0; 55.4)	5.2 (1.9; 8.5)	31.8 (27.1; 36.5)	8.6 (3.5; 13.8)
	Day 2 Overnight	2.4 (2.3; 2.6)	87.7 (80.0; 95.4)	38.6 (35.3; 41.9)	74.2 (69.5; 78.9)	54.9 (49.8; 60.0)
	Off	4.1 (4.0; 4.3)	59.7 (52.1; 67.4)	7.0 (3.7; 10.2)	12.8 (8.2; 17.3)	4.6 (0.0; 9.6)
	Other*	4.0 (3.7; 4.3)	41.9 (27.7; 56.1)	10.0 (0.7; 19.3)	50.8 (45.1; 56.5)	39.4 (33.6; 45.2)
	Across Shifts	3.6 (3.5; 3.8)	61.5 (53.1; 69.9)	12.1 (10.3; 14.0)	49.1 (43.8; 54.5)	28.4 (22.1; 34.7)
Standard Programs	Day	3.5 (3.4; 3.7)	53.2 (45.6; 60.8)	7.8 (4.6; 11.1)	56.4 (51.7; 61.1)	22.6 (17.5; 27.7)
	Night	3.5 (3.3; 3.7)	61.7 (51.4; 71.9)	13.5 (7.6; 19.5)	46.0 (36.8; 55.3)	26.8 (18.4; 35.3)
	Off	3.9 (3.8; 4.1)	53.3 (45.6; 61.1)	2.3 (0.0; 5.8)	16.7 (11.8; 21.7)	6.7 (1.4; 11.9)
	Other*	3.4 (3.2; 3.6)	65.7 (55.9; 75.5)	11.1 (5.6; 16.7)	52.7 (46.7; 58.6)	25.8 (19.9; 31.7)
	Across Shifts	3.6 (3.4; 3.7)	55.2 (46.9; 63.5)	8.3 (6.4; 10.1)	53.5 (48.3; 58.8)	22.0 (15.7; 28.2)

*In Flexible programs, days with missing shift information or classified by the interns as a regular night shift were re-classified as "other"; in Standard programs, days with missing shift information or classified by the interns as starting or finishing an extended overnight shift were re-classified as "other" (see Appendix Section 9 above).

†Sleep quality was measured on a 5-point scale with anchors bad (1) and good (5).

‡ at least one period reported

§ Confidence intervals have not been adjusted for multiple testing and inferences drawn from the intervals may not be reproducible.

14. Comparison of Day and Night Shift in Standard Programs

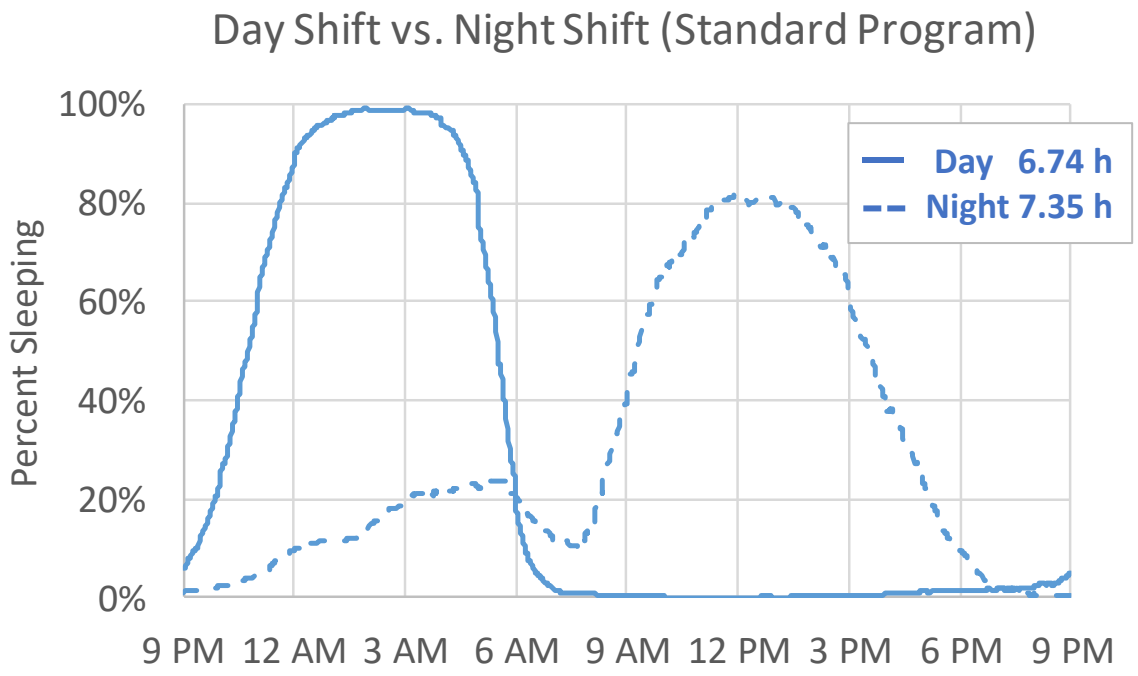

Figure S10: Percent of interns sleeping by time of day. Standard program interns received on average 0.61 hours more sleep per 24 hours on night shift rotations compared to day shift rotations.

15. References

1. Desai SV, Asch DA, Bellini LM, et al. Education Outcomes in a Duty-Hour Flexibility Trial in Internal Medicine. N Engl J Med 2018;378:1494-508.
2. Akerstedt T, Gillberg M. Subjective and objective sleepiness in the active individual. Int J Neurosci 1990;52:29-37.
3. Basner M, Mollicone DJ, Dinges DF. Validity and sensitivity of a brief Psychomotor Vigilance Test (PVT-B) to total and partial sleep deprivation. Acta Astronaut 2011;69:949-59.
4. Sadeh A, Sharkey KM, Carskadon MA. Activity-based sleep-wake identification: an empirical test of methodological issues. Sleep 1994;17:201-7.
5. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B (Methodological) 1995;57:289-300.