Supplementary file 4. Detailed study outcomes

Physician health and wellness outcomes and associations with fatigue

Study Risk of Bias (RoB)	Study design	Exposures or interventions Assessment measure and time points	 Baseline	Outcomes Assessment measure and time points	Associations between exposure and outcome
Surgeons					
Jackson, 2017 RoB: high	CS	Not feeling well rested: self-reported as 'unhealthy' Time points NR	71% healthy, 28% unhealthy in terms of being well rested	Job satisfaction: Abridged Job in General Scale; grouped into more or less satisfied using the median Time points NR	Job satisfaction in those more vs. less satisfied: Healthy (well rested): 85% vs. 58%, p<0001; Unhealthy (not well rested): 15% vs. 42%, p<0.001.
Nishimura, 2014 RoB: unclear	CS	Sleep hours/night: self-reported (continuous) Time points NR	Mean±SD sleep: 5.94±1.08h	Burnout: Japanese MBI (severe: EE >4.0 and either DP >2.6 or PE <4.17) Time points NR	1) Mean±SD sleep for not burned out vs. mild to moderate vs. severe: 6.07±1.15 vs. 5.88±0.94 vs. 5.63±0.94, p<0.05; 2) Association between sleep and burnout (OR (95% CI)): bivariate 0.67 (0.61-0.73), p<0.001; multivariate including work characteristics and mental health: 0.84 (0.75-0.94), p=0.002.
Sargent, 2009 RoB: high	CS	Sleep deprivation: self-reported on a 4-point scale (none, a little, quite a bit, a lot) Time points NR	21% none, 48% a little, 23% quite a bit, 8% a lot	Burnout: MBI (norms NR); Marital satisfaction: RDAS; Psychological morbidity: GHQ-12 score ≥4 Time points NR	1) Positive correlation between sleep deprivation and EE, DP, psychological distress, lower marital satisfaction, all p<0.001. No relationship with PA.
Anesthesiologists[a]					
Lederer, 2006 RoB: high	BA	24-h shift with on-call duty; Sleep hours and interruptions: self-reported; Tiredness: VAS from 0 (low) to 100 (high) Assessed pre- and post-duty	Mean±SD sleep: 4.1±1.7h; Number of interruptions: 0.8±1.1; Tiredness pre- vs. post-duty: 30.9±27.5 vs. 59.5±18.9, p=0.01.	Stress during duty: 4-point scale from 'calm' to 'very demanding' Assessed post-duty	1) Mean stress score during duty: 2.1.

Study Risk of Bias (RoB)	Study design	Exposures or interventions Assessment measure and time points	 Baseline	Outcomes Assessment measure and time points	Associations between exposure and outcome
Leitchtfried, 2011 RoB: high	TS	24-h shift; Sleepiness: ESS (range: 0-24); Sleep hours: self-reported (continuous) Sleepiness assessed pre-shit, sleep hours pre, during and post-shift	ESS (mean (range)): 7.4 (4-12); Mean±SD sleep hours: 1) pre-study: 7.74±1.35h; 2) Pre-24-h shift (11h00 on day 1: 0.13±0.35h, 19:00 on day 1: 6.99±0.68h); 3) During the 24-h shift (07h00 on day 2: 0.0±0.0h, 19h00 on day 2, 5.49±1.95h); 4) Post-24-h shift (11h00 on day 3: 0.5±0.71h, 19h00 on day 3: 7.06±1.18h).	aMT6-s: urinalysis Assessed at 4-h intervals from 07:00 to 11:00	1) aMT6-s over shift, mean (95% CI): higher at 11:00AM pre- (12.2 (6.3-8.1)) and post-shift (9.3 (3.7-14.9)) vs. during, p=0.016; 2) Correlations between sleep and aMT6-s (data NR): mild for sleep duration the night prior with aMT6-s at 3PM the following day; sleep on night 2 with aMT6-s at 3PM the next day; total sleep with aMT6-s at 11AM on third day; moderate for sleep on first night with aMT6-s at 7AM and 11AM pre-shift, 11PM during 24-h shift and 11AM post-shift; total sleep pre-shift and nocturnal sleep during 24-h shift with aMT6-s at 11PM during shift; total sleep with aMT6-s at 3PM on first and second day, 11PM on second day; 3) Correlations between ESS and aMT6-s: moderate for aMT6-s at 7AM during shift, 11AM on day off.
Beaujouan, 2005 RoB: high	CS	Sleep deprivation: 4-point scale (always, frequently, rarely, never) Time points NR	48.8% always or frequently feel sleep deprived	Substance abuse: 93-item addiction and substance abuse questionnaire Time points NR	1) 60.6% with drug dependence vs. 46.0% of those without reported sleep difficulties, p<0.001. 2) OR (95% CI) of addiction for frequently/always vs. rarely/never sleep deprived: tobacco 1.42 (1.04-1.94); tranquilizer/hypnotics 3.26 (2.12-5.02).
Doppia, 2011 RoB: low	CS	Insufficient sleep: 4-point scale (no, not really, sort of, yes) Time points NR	28.9% reported insufficient sleep during work time	Burnout: CBI (mild: 1-2.4, moderate: 2.5-3.5, severe: 3.6-5) Time points NR	1) Frequency of burnout by response for sleep sufficiency: 47.6% for no/not really, 16.3% for sort of/yes, p<0.001.
Lindfors, 2006 RoB: low	CS	Sleep hours/day: self-reported to the nearest 0.5h; Adequacy of sleep and rest: self-reported (yes/no)	Sleep hours (mean (range)): 7 (5-9)	Stress: MOSQ on a 3-point scale (no, to some extent, clearly); Thoughts of suicide: 4-point scale ('never' to 'have tried')	1) Sleep sufficiency predicted stress symptoms: bivariate β=-0.362, p<0.001; multivariate including gender, sick leave, suicide β=-0.269, p<0.001; 2) Sleep disturbance associated with thoughts of suicide, p=0.009.

Study Risk of Bias (RoB)	Study design	Exposures or interventions Assessment measure and time points	 Baseline	Outcomes Assessment measure and time points	Associations between exposure and outcome
		Time points NR		Time points NR	
Saadat, 2015 RoB: low	CS	Sleep deprivation (<7h/24-h) due to 17-h overnight shift; Sleepiness and alertness: VAS from 0 (not at all) to 100 (extremely) All assessed on a regular day and a post-call day	Mean±SD sleepiness on a regular day vs. post-call day: 2.99±2.18 vs. 6.79±2.30, p<0.001	Simple cognitive tests: VAS from 0 (not at all) to 100 (extremely); Mood disturbance: PMS (scoring NR) All assessed on a regular day and a post-call day	Regular day v. post-call day, mean±SD scores: 1) Simple cognitive tests: energetic 6.04±2.27 vs. 2.53±1.87, confident 7.03±1.83 vs. 4.98±2.29, irritable 2.03±1.94 vs. 4.86±2.16, sleepy 2.99±2.18 vs. 6.79±2.30, talkative 4.46±1.74 vs. 2.41±1.97, all p<0.001; jittery 1.44±1.74 vs. 3.12±2.34, p=0.003; anxiousness ns; 2) PMS: tension 13.48±2.71 vs. 15.43±4.46, p=0.049; anger 15.24±4.41 vs. 18.14±5.92, p=0.005; fatigue 10.14±2.63 vs. 20.05±6.87, p<0.001; confusion 10.57±1.69 vs. 12.57±4.24, p=0.025; vigor 24.05±6.75 vs.16.67±5.70, p<0.001; depression: ns; total mood disturbance: 42.57±15.26 vs. 70.90±6.91, p<0.001.
ER or ICU physicians					
Dutheil, 2013 RoB: high	RCT	14-h or 24-h shift; Sleep hours: self-reported sleep and wake time; Sleep quality: VAS from 1 (low) to 100 (high); Mental and physical fatigue: VAS from 1 (low) to 100 (high) Assessed on day prior to shift; during shift; each day of protocol (work, off, clerical, control)	1) Sleep duration and quality lower during shifts (14h and 24h) than any other day, and lower during the 24-h vs. 14-h shift (p<0.05); 2) Mental and physical fatigue higher after 14-h and 24-h shift vs. control day (data NR).	Stress: VAS from 0 (low) to 100 (high); IL-8: urinalysis Assessed at 08:30 and 18:30 on each day of protocol	1) Stress: higher following 14-h and 24-h shifts vs. the control day, p<0.05 (data NR); 2) IL-8: higher following 24-h shift vs. control (p=0.007) and 14-h shift (p=0.015); ns difference between 14-h shift and control day; 3) Correlations with IL-8: sleep hours pre-24-h shift, r=-0.627, p=0.007; poor sleep quality during 14-h and 24-h shifts, r=0.452, p=0.031; 4) Multivariable regression: 24-h shift increased IL-8 by 1.9ng vs. control day, p=0.007; ns association with 14-h shift, mental or physical fatigue, sleep deprivation, 14-h shift.
Sende, 2012 RoB: high	CS	Fatigue and sleep deprivation as sources of stress	NR	Most important sources of stress among 4 categories (work-related, patient-	1) 78% indicated that sleep loss and fatigue were sources of stress.

Study Risk of Bias (RoB)	Study design	Exposures or interventions Assessment measure and time points	 Baseline	Outcomes Assessment measure and time points	Associations between exposure and outcome
		Time points NR		related, organizational, individual) Time points NR	
Generalists[b]					
Harbeck, 2015 RoB: unclear	CS	24-hours on-call shift with sleep disturbance: self-reported number of sleep disturbances and hours of sleep per night Assessed before a normal day shift, and after a 24-h on call shift	1) Sleep hours on a normal day vs. following a 24-h shift: <2 hours: 0 vs. 5.9%; 2-4 hours: 5.9% vs. 47.1%; 4-6 hours: 11.8% vs. 35.3%; >6 hours: 82.4% vs. 11.8% 2) Number of sleep disturbances a normal day vs. following a 24-h shift: 0: 82.4% vs. 11.8%; 1: 11.8% vs. 35.3%; 2: 5.9% vs. 47.1%; 3: 0% vs. 5.9%; 4: 0% vs. 0%; >4: 0% vs. 0%	Biochemical (laboratory values) and physiological (heart rate variability, skin resistance, blood pressure) stress parameters Assessed before a normal day shift, and after a 24-h on call shift	Before a normal shift vs. after overnight call shift: 1) Biochemical parameters: no changes in any parameter except for thyroid stimulating hormone which was higher after the on-call shift (p = 0.049, data NR); 2) Physiological parameters: no significant changes in any parameter
Pit, 2014 RoB: unclear	CS	Work-related sleep disturbance: 7-point scale from 'never' to 'every day' Time points NR	Work-related sleep disturbance: 41% never, 59% a few times a year to every day	Early retirement (<65 years) intentions (yes/no) Time points NR	For sleep disturbance a few times a year to every day vs. never: 1) Intention to retire early: 74% vs. 26%, p<0.01; 2) Association with intention to retire early (OR (95% CI)): univariate 3.6 (1.47-8.80), p<0.01; multivariate including work, occupational, individual factors 2.91 (1.11-7.6), p<0.05; 4) RR (95% CI) for intention to retire early: 2.0 (1.18-3.49); attributable fraction: 50.0%; population attributable fraction: 37.1%.
Pit, 2016 RoB: unclear	CS	Work-related sleep disturbance: 7-point scale from 'never' to 'every day'	Work-related sleep disturbance: 41% never, 59% a few times a year to every day	Sickness presenteeism: 'yes' response indicated 1 or more days	For sleep disturbance a few times a year to every day vs. never: 1) Sickness presenteeism: 32% vs. 68%, p=0.018;

Study Risk of Bias (RoB)	Study design	Exposures or interventions Assessment measure and time points	 Baseline	Outcomes Assessment measure and time points	Associations between exposure and outcome
		Time points NR		Assessed for the past 12 months	2) Association with sickness presenteeism (OR (95% CI)): 2.92 (1.19-7.16), p=0.02.
Roberts, 2014 RoB: unclear	CS	Fatigue: LAS from 0 (low) to 10 (high) Assessed for the past week	Mean (SD) score: 5.8 (2.4) for hospitalists; 5.9 (2.4) for general internists	Impact of fatigue on daily activities (falling asleep while driving) (yes/no) Time points NR	1) 8.7% of hospitalists and 4.3% of outpatient general internists had fallen asleep while driving due to fatigue.
Vela-Bueno, 2008 RoB: low	CS	Sleep Quality: PSQI (Spanish): score ≥5 indicates low quality (range; 0 to 21); Insomnia: DSM-IV criteria Time points NR; insomnia symptoms in past month	Prevalence (% (95% CI)): 1) Sleep-onset latency >30 minutes: 8.4 (4.8-11.9); 2) Wake time after sleep onset >30 minutes: 15.4 (10.8-19.9); 3) Early morning awakening: 22.5 (19.5-30.4); 4) Nonrestorative sleep: 22.5 (17.2-27.7); 5) Daytime impairment for ≥5 days in past month: 14.2 (9.7-18.6); 6) Insomnia: 18.8 (13.8-23.7).	Burnout: PBM with a 7-point scale from 1 (never) to 7 (always) Time points NR	Low vs. high burnout, mean±SD: 1) Global PSQI: 2.72±2.22 vs. 7.24±4.17, p<0.001; 2) PSQI subscores: sleep quality: 0.54±0.57 vs. 1.40±0.83, p<0.001; sleep latency: 0.51±0.80 vs. 1.38±1.03, p=0.002; sleep duration: 0.45±0.64 vs. 1.16±0.92, p=0.003; sleep efficiency: 0.21±0.57 vs. 0.77±0.98, p=0.018; sleep disturbance: ns; use of medication: 0.14±0.49 vs. 0.57±0.83, p=0.032; daytime dysfunction: 0.52±0.73 vs. 1.57±0.88, p=0.002. 3) Prevalence (95% CI) of insomnia symptoms: sleep latency: 5.5% (2.5-11.5%) vs. 21.1% (10.5-31.6%), p=0.015; wake time >30 min after sleep onset: 9.4% (1.6-17.1%) vs. 25.5% (14.2-37.7%), p=0.029; early awakening: 14.5% (5.1-23.8%) vs. 45.6 (32.7-58.4%), p<0.001; somewhat/very dissatisfied with sleep: 5.5% (2.5-11.5%) vs. 50% (37.1-62.8%), p<0.001; day impairment: 5.5% (2.5-11.5%) vs. 38.2% (25.6-50.7%), p<0.001; insomnia: 7.3% (0.4-14%) vs. 39.7% (27.1-52.2%), p<0.001.
Oncologists					
Shanafelt, 2005 RoB: unclear	CS	Fatigue: LASA QOL ≤7; Sleep deprivation: 10-point Likert scale from 0 (not at all) to 10 (stressful as can be)	75% had a high level of fatigue; Mean±SD sleep score: 4.5±2.65.	Wellbeing: 10-item LASA QOL, high ≥8 vs. low ≤7 Time points NR	1) Sleep deprivation for high vs. low overall well-being (mean±SD): 3.9±2.57 vs. 5.1±2.60, p=0.0004; 2) Lower fatigue predicted overall wellbeing in a multivariate model including personal and professional characteristics, p=0.002.

Study Risk of Bias (RoB)	Study design	Exposures or interventions Assessment measure and time points	 Baseline	Outcomes Assessment measure and time points	Associations between exposure and outcome
		Time points NR			
Shanafelt, 2014 RoB: unclear	CS	Fatigue: 10-point LAS (lower scores indicate greater fatigue) Time points NR	Mean±SD fatigue score: 5.7±2.4	Satisfaction with WLB: 5-point Likert scale from 'strongly agree' to 'strongly disagree' Time points NR	1) OR (95%CI) of lower satisfaction predicted by high fatigue (vs. not) in multivariate model including personal and work-related factors, and burnout: 0.489 (0.337-0.710), p<0.001.
Mixed groups of physicians					
Aziz, 2004 RoB: high	CS	Working while fatigued: 5-point scale from 'extreme' to 'a little' Time points NR	NR	Stress: 47-item questionnaire with a 5-point scale from 'extreme' to 'a little' Time points NR	1) Sources of stress: working while fatigued had a mean±SD score of 2.44±1.20, factor loading: 0.653, in factor analysis; 2) Inverse correlation between stress and working while fatigued: r=-0.270 (significance level NR).
Chen, 2008 RoB: high	CS	Sleepiness: ESS score ≥11 Time points NR	Mean±SD ESS score: 7.8±4.0, range: 0-20, 23% had scores ≥11.	Impact on work and personal life: Impact Questionnaire with a 5-point Likert scale from 1 (strongly agree) to 5 (strongly disagree) Time points NR	1) Impact score correlated with ESS, r=0.31, p<0.05; 2) ESS score was higher among physicians who agree/strongly agree vs. other response: worried about having a car accident while driving home post-call: 5.4 vs. 7.0, p<0.001; sleep loss has a major impact on personal life: 8.4 vs. 7.0, p=0.01; 3) Higher ESS scores predicted by impact score in multivariate regression including personal and work-related factors: β=0.11, p=0.005.
Elovaino, 2015 RoB: low	CS	Sleeping problems: Jenkins Scale with a 6-point scale from 1 (never) to 6 (every night) Assessed in 2006 and 2010	Mean±SD score: 2006: 2.30 (1.00); 2010: 2.35 (1.05).	Jobs demands: 5 items scored on a 5-point scale from 1 (strongly disagree) to 5 (strongly agree); Job control: 3 items derived from the Karasek Job Questionnaire	There was no association between sleeping problems in 2006 and job demands or control in 2010.

Study Risk of Bias (RoB)	Study design	Exposures or interventions Assessment measure and time points	 Baseline	Outcomes Assessment measure and time points	Associations between exposure and outcome
Heponiemi, 2014 RoB: low	CS	Sleeping problems: Jenkins Scale[81] with a 6-point scale from 1 (never) to 6 (every night) Assessed in 2006	Mean±SD (range) score: 2.30±1.00 (1-6)	Psychological distress: GHQ-12 with a 4-point scale (low to high); Job satisfaction: JDS with a Likert scale from 1 (strongly disagree) to 5 (strongly agree) Assessed in 2010	1) Sleeping problems associated with job satisfaction, β=-0.12, p<0.001, psychological distress, β=0.18, p<0.001; 2) Total indirect effect of on-call duty through two mediators (sleeping problems, work interference with family) (R^2 (95% CI)): job satisfaction 0.06 (-0.059, -0.016), p<0.001; psychological distress 0.16 (0.023, 0.081), p<0.001.
Mahmood, 2016 RoB: high	CS	Sleep deprivation: self-reported mean hours of sleep when on call Assessed at 4 years, 10 years, and 15 years post-graduation	Mean±SD hours: 4 years: 4.52 (2.79); 10 years: 5.38 (6.36); 15 years: 6.41 (7.14).	Alcohol use disorders: Modified 9-item version of the Alcohol Use Disorder Identification Test (AUDIT) ≥6 for men and ≥5 for women. Assessed at 4 years, 10 years, and 15 years post-graduation	There was no association between hours of sleep when on call and hazardous drinking behaviours (p=0.732)
Shirom, 2010 RoB: low	CS	Tiredness and exhaustion: SMBM Physician Fatigue Subscale on a 7-point scale from 1 (almost never) to 7 (always) Time points NR	NR	Burnout: SMBM on a 7-point scale from 1 (almost never) to 7 (always)	1) Correlation between physical fatigue subscale and overall burnout: 0.88, p<0.05; 2) In a predictive structural model for burnout, physical fatigue accounted for unique variance in the burnout items, not accounted for by total burnout (R^2=0.24).
Smith, 2017 RoB: unclear	CS	Sleep deprivation: self-reported via open-ended comments Time points NR	NR	Mental and physical illness: self-reported via open-ended comments Time points NR	Some physicians reported developing mental illness (e.g., bipolar disorder, alcohol misuse) due to tiredness and stress at work; others developed physical health problems due to sleep deprivation, poor eating habits and lack of exercise.

Study Risk of Bias (RoB)	Study design	Exposures or interventions Assessment measure and time points	 Baseline	Outcomes Assessment measure and time points	Associations between exposure and outcome
Starmer, 2016 RoB: low	CS	Sleep deprivation: <7 hours sleep in a typical 24-h period (self-reported) Time points NR	27.7% sleep deprived	Burnout, satisfaction with career and life, balanced personal and professional commitments: Each on a 5-point Likert scale (strongly agree to strongly disagree) Time points NR	≥7-h vs. <7-h sleep: 1) Burnout (% strongly agree/agree): 26.4% vs. 39.6%, p<0.05; career satisfaction (% strongly agree/agree): ns; life satisfaction (% completely/very satisfied): 76.4% vs. 55.9%, p<0.05; balanced personal and professional commitments (% completely/very satisfied): 49.7% vs. 26.1%. 2) <7-h sleep (vs. ≥7-h) (OR, 95% CI) associated with life satisfaction 0.44 (0.29-0.67), p<0.05; balanced personal/professional commitments 0.46 (0.31-0.71), p≤0.05, in a model including work and personal factors.
Tokuda, 2009 RoB: low	CS	Sleep hours/day: self-reported (continuous) Time points NR (included weekday and weekends)	Mean±SD (range) sleep hours/day: 6±0.9 (3-8)	Burnout: MBI (Japanese) with a 7-point Likert scale: 0 (none) to 6 (every day); Job satisfaction: JHPSS with a 5-point Likert scale: 1 (strongly disagree) to 5 (strongly agree) Time points NR	Maximum likelihood estimates±SE: 1) Sleeping time to job satisfaction: group 0.990±0.458, p=0.031; ns for men; women 1.711±0.805, p=0.034; 2) Sleeping time to EE: group -0.219 ±0.070, p=0.002; men -0.215±0.082, p=0.009; ns for women.
Wada, 2010 RoB: unclear	CS	Sleep hours/day: Self-reported (continuous) Assessed for past month when not completing overnight work	<5 hours: 8.7% men, 9.9% women; 5 to <6 hours: 32.3% men, 34.6% women; 6 to <7 hours: 46.0% men, 43.7% women; ≥7 hours: 13.0% men, 11.8% women.	Depression: QIDS-SR; Japanese score <5 (no symptoms) to >20 (very severe symptoms) Assessed for past 7 days	1) Sleep hours for those with vs. without depressive symptoms: <5: 18.7% vs. 7.7% men, 20.5% vs. 8.7% women; 5 to <6: 33.7% vs. 32.2% men, 38.6% vs. 34.2% women; 6 to <7: 35.1% vs. 46.9% men; 31.8% vs. 45.1% women; 2) Association between <5h sleep (vs. 6-7h) and depressive symptoms (OR (95% CI)): univariate 2.79 (1.96-3.95) for men, 2.65 (1.47-4.78) for women; multivariate (including age and workload

Study Risk of Bias (RoB)	Study design	Exposures or interventions Assessment measure and time points	Baseline	Outcomes Assessment measure and time points	Associations between exposure and outcome
					factors) 2.70 (1.82-4.03) for men, 2.38 (1.11-5.10) for women.

[a]Includes studies of anesthetists, where these were physicians.

[b]Includes primary care physicians, internal medicine physicians, and general practitioners.

AM: morning; aMT6-s: melatonin metabolite; BA: before-after; CI: confidence interval; CBI: Copenhagen Burnout Inventory; CS: cross-sectional; DP: depersonalization; DSM: Diagnostic and Statistical Manual of Mental Disorders; EE: emotional exhaustion; ER: emergency; ESS: Epworth Sleepiness Scale; GHQ: General Health Questionnaire; h: hour(s); ICU: intensive care unit; IL-8: interleukin-8; JDS: Job Diagnostic Survey; JHPSS: Japanese Hospital Physicians Satisfaction Scale; LAS: linear analog scale; LASA: linear analog assessment scales; MBI: Maslach Burnout Inventory; MOSQ: Modified Occupational Stress Questionnaire; min: minute(s); NA: not applicable; NR: not reported; ns: not statistically significant; OR: odds ratio; PA: personal achievement; PBM: Pines Burnout Measure; PE: professional efficacy; PM: afternoon; PMS: Profile of Mood States; PSQI: Pittsburgh Sleep Quality Index; QIDS-SR: Quick Inventory Depressive Scale – Self-Reported; QOL: Quality of Life; RCT: randomized controlled trial; RDAS: Revised Dyadic Adjustment Scale; RoB: Risk of Bias; SD: standard deviation; SE: standard error; SMBM: Shirom-Melamed Burnout Measure; TS: time series; US: United States of America; VAS: visual analog scale; vs.: versus; WLB: work-life balance

Performance and safety outcomes related to fatigue or sleep loss among physicians in independent practice

Study Risk of Bias (RoB)	Study design	Exposures or intervention Assessment measure and time points	Baseline	Outcomes Assessment measure and time points	Associations between exposure and outcome
Surgeons					
Uchal, 2005 RoB: unclear	RCT	Sleep deprivation from a 24-h call shift vs. 8-h work; Sleep hours: self-reported (continuous); Sleepiness: ESS (moderate: 10-15, severe: ≥16) Assessed post-call and post-work	Median (range) sleep hours: 1.5 (0-3) post-call vs. 6.5 (5-9) post-work, p<0.05; Median ESS score: 7.0 post-call vs. 5.5 post-work, ns.	Surgical performance: laparoscopic surgical simulator(Minimally Invasivs Surgical Trainer-Virtual Reality) for product quality, procedure effectiveness Assessed post-call and post-work	Post call vs. post-work: 1) Product quality: no difference in accuracy error, tissue damage, leak rate; 2) Procedure effectiveness: no difference in goal-directed actions, non-goal directed actions, operating time.
Chu, 2011 RoB: low	CO	Sleep deprivation: self-reported hours, moderate (3-6h) or severe (<3h) Assessed the night before surgery	Of 4,047 procedures, 83 (2.1%) performed by severely sleep-deprived and 1,595 (39.4%) moderately sleep-deprived surgeons	Surgical performance: CABG, ACC Assessed during surgery	For 0-3 vs. 3-6 vs. >6 hours of sleep: no difference in CABG or ACC.
Ellman, 2004 RoB: low	CO	Sleep deprivation: performed a case starting 22:00 to 05:00, or ending 22:00 to 07:30 and another case in the next 24-h	Of 6,751 procedures, 339 (5%) performed by sleep-deprived surgeons	Surgical performance: CABG, ACC Assessed during surgery	Sleep deprived vs. non-sleep deprived: no difference in CABG or ACC.
Govindarajan, 2015 RoB: low	CO	Sleep deprivation: treated patients from midnight to 07:00 and performed a subsequent case on the same day	NR	Surgical performance: duration of surgery	Sleep deprived vs. non-sleep deprived: no difference in duration of surgery, even after stratification by type of procedure.
Amirian, 2014 RoB: high	BA	17-h night call shift; Sleep hours during the shift: Wrist-mounted Micro-Mini-Motionlogger; Sleepiness: KSS	Naps pre-call: 11 (37%) napped for median (IQR) 90 (58-128) min; Median (IQR) sleep: 91 (62-123) min on the pre-call night vs. 430 (329-449) on	Surgical performance: LapSimGyn laparoscopic simulation for time, blood loss, instrument path; D2 test of attention and concentration	Pre- vs. post-call: 1) LapSimGyn: no difference in total time, blood loss, instrument path length, instrument angular path; napping did not affect performance;

Study Risk of Bias (RoB)	Study design	Exposures or intervention Assessment measure and time points	 Baseline	Outcomes Assessment measure and time points	Associations between exposure and outcome
		Assessed on pre-call and on-call day; sleepiness assessed during shift	the on-call night, p<0.001; Sleep on-call: 12 (40%) slept for median (IQR) 98 (39-135) min; Significant development of sleepiness during shift (p<0.001), plateau score of 7 at 04:00 to 08:00.	Assessed on pre-call and on-call day	2) D2 test: improvement in concentration, p<0.05. No changes in any other parameters; 3) ns difference in laparoscopic simulation time in those who slept during the shift vs. not.
Gerdes, 2008 RoB: high	BA	On-call shift; Fatigue: questionnaire designed by Behrenz & Monga, 1999; Sleep hours: self-reported (continuous) Assessed in 3 sessions pre- and post-call	Fatigue differential from pre- to post-call (range): 1-7 (units unclear); Sleep during call (range): 1-5h	Psychomotor performance: virtual ring transfer task for gesture-level proficiency, hand movement smoothness, tool movement smoothness, elapsed time Assessed in 3 sessions pre- and post-call	1) Pre- to post-call: decrease in all measures of psychomotor proficiency (p<0.05, data NR) except elapsed time; no change in number of psychomotor errors; increase cognitive errors (p<0.05, data NR); 2) Cognitive errors increased exponentially as fatigue ratings increased (R^2=0.9219) and as hours of sleep declined (R^2=0.933).
Shanafelt, 2010 RoB: unclear	CS	Degree of fatigue as a contributor to errors (self-reported) Assessed for the past 3 months	NR	Perceived recent major medical errors (self-reported) Assessed for the past 3 months	1) Prevalence of perceived recent major medical error: 8.9%; 2) Of those reporting an error, 6.9% listed degree of fatigue as the greatest contributing factor.
Anesthesiologists[a]					
Lederer, 2006 RoB: high	BA	24-h shift, on-call duty; Sleep hours and interruptions: self-reported; Tiredness: VAS from 0 (low) to 100 (high) Assessed pre- and post-duty	Mean±SD sleep: 4.1±1.7h; Number of interruptions: 0.8±1.1; Tiredness pre- vs. post-duty: 30.9±27.5 vs. 59.5±18.9, p=0.01.	Psychomotor performance: reaction time, critical flicker fusion, response measure, peripheral awareness; Concentration ability: scale of 0 (low tiredness) to 100 (maximum tiredness) Assessed pre- and post-duty	Pre- vs. post-duty, mean±SD: 1) Psychometric testing: recognition reaction time (ms): 439.6±50.8 vs. 480.3±58.9; motor reaction time (ms): 252.8±39.3 vs. 465.4±65.0; total reaction time (ms): 690.8±73.4 vs. 746.5±113.7; critical flicker fusion (Hz): 29.0±2.3 vs. 28.7±3.7; response measure (pixels): 647.8±126.7 vs. 598.3±138.1,

Study Risk of Bias (RoB)	Study design	Exposures or intervention Assessment measure and time points	 Baseline	Outcomes Assessment measure and time points	Associations between exposure and outcome
					peripheral awareness task recognition time: 58.9±59.2 vs. 51.6±47.5; 2) Concentration ability: 26.4±23.5 vs. 56.3±23.0, p=0.007.
Chang, 2013 RoB: unclear	CS	15-h in-house overnight call; Sleepiness pre-call: ESS ≥9; Sleep hours: self-reported (continuous) Sleepiness assessed pre-call, sleep hours during call	Median (IQR) ESS: 9 (9), 64% scored ≥9; Median (IQR) hours slept during shift: 1 (0-3).	Psychomotor performance: reaction time; CCPT II; N-back; HVLT (3 trials of 12 words) Assessed at baseline and pre- and post-call	1) Afternoon baseline vs. pre-call: no difference in reaction time, CCPT, N-back, of HVLT; Morning baseline vs. post-call: 1) No change in auditory or visual reaction time; 2) CCPT (t-scores): No change in detectability, response style, hit reaction time, omissions/commissions; 3) N-back % accuracy: no change for auditory, visual, or mean N-value; 4) HVLT (t-score): mean for trials 1-3: 48.6±7.6 vs. 41.5±9.9 (p=0.04); delayed recall: ns; 5) No correlation between ESS scores pre-call or sleep during shift and any measure of psychomotor performance.
Gander, 2000 RoB: low	CS	Nights of work-related sleep disturbance: self-reported (continuous) Assessed for the past 6 months	NR	Risk of fatigue-related errors: questionnaire modelled after Gravenstein et al., 1990 Assessed for the past 6 months	1) Risk of fatigue-related errors increased with increasing nights of work-related sleep disturbance: RR: 1.25, 95% CI: 1.06-1.49.
Saadat, 2017 RoB: low	CS	Sleep deprivation due to an overnight call shift	NR	Reaction time: PVT Assessed after an overnight call shift and the morning of a regular (non-call) day	Mean (SD) reaction time was slower post-call (297.76 (83.75)) vs. on a regular day (266.58 (38.35)), p=0.047.

Study Risk of Bias (RoB)	Study design	Exposures or intervention Assessment measure and time points	 Baseline	Outcomes Assessment measure and time points	Associations between exposure and outcome
Gander, 2008 RoB: unclear	NC	Sleep loss across consecutive working days or on-call work: Wrist-mounted Actiwatch (Mini Mitter, Bend, Oregon, US), sleep and duty diary Assessed over a 2-week period including a weekend of rostered shifts or on-call	≥2 hours sleep <baseline: 8% of 24-h periods that included day work vs. 14% that included day + call; Sleep hours: mean 0.6h less sleep when working day shifts (p=0.014) and 0.8h less sleep when working day shifts + call (p=0.013) vs. off.	Psychomotor performance: PVT Assessed within 2 hours pre- and post-call	1) In fixed model analysis for reaction time including sleep, time since waking, work hours: acute sleep loss associated with slower median reaction time, $F_{(1,184)}$=5.70, p<0.05; longer time since waking associated with poorer performance on the slowest 10%, $F_{(1,185)}$=5.13, p<0.05; 2) Reaction time across 12 consecutive work days: no change in pre-duty reaction times but post-duty reaction times slowed linearly, median -3.38, p<0.001; decline in performance across 10 minutes became progressively steeper both pre- and post-duty, p=0.020.
ER or ICU physicians					
Sanches, 2015 RoB: high	CS	Acute sleep deprivation (<5h of night sleep after a night shift of 12h) Sleep hours: 7-day Actigraphy via SenseWear® Pro2 Armband; Sleepiness: ESS; Sleep quality: PSQI Assessed the week and night before the psychomotor tests	Non-sleep deprived vs. sleep deprived: PSQI >5: 0% vs. 33%, ns; ESS≥10: 11% vs. 67% Sleep time (mean±SD) in week before tests: duration and number of naps higher in sleep deprived group, but diurnal sleep hours lower, 428.6±30.1 vs. 375.8±55.9, p=0.038; Sleep quality (mean±SD): week before tests: 3.3±0.7 vs. 2.6±0.3, p=0.013; night before tests: 3.1±0.8 vs. 1.9±1.0, p=0.020.	Psychomotor performance via Battery Test Reaction 5 (v1): StimulTest, InstrucTest, MovemTest; TP test of visual attention Assessed on morning after night shift 8	Sleep deprived group vs. non-sleep deprived, mean±SD: 1) InstrucTest: correct answers: 169.4 (16.0) vs. 148.3 (28.3), p=0.070; wrong answers: ns; perfection index (%): 99.6 (0.3) vs. 98.9 (1.3), p=0.021; response latency (sec/click): ns; 2) StimulTest: correct answers: 170.7 (21.9) vs. 145.1 (17.9), p=0.022; wrong answers: ns; perfection index (%): ns; response latency (sec/click): 1.06 (0.1) vs. 1.24 (0.1), p=0.022; 3) MovemTest: ns for any parameter; 4) TP: omitted symbols: 34.2±18.4 vs. 62.7±44.0, p=0.034; concentration index (%): 14.1±8.9 vs. 30.0±25.9, p=0.019; quality index (%): 13.8±8.6 vs. 29.2±26.4, p=0.031; correct/wrong symbols: ns; Correlations between sleep and tests: 1) TP for sleep hours nights 1-6: omitted symbols: r=-0.686, p=0.011 for non-sleep-

Study Risk of Bias (RoB)	Study design	Exposures or intervention Assessment measure and time points	Baseline	Outcomes Assessment measure and time points	Associations between exposure and outcome
					deprived, ns for sleep-deprived; concentration index (%): r=-0.359, p=0.037 for sleep-deprived, ns for non-sleep deprived; r=-0.359, p=0.037 for the group; no other significant correlations; 2) No correlation between PSQI, ESS and any of the psychomotor tests.
Generalists[b]					
Harbeck, 2015	CS	24-hours on-call shift with sleep disturbance: self-reported number of sleep disturbances and hours of sleep per night Assessed before a normal day shift, and after a 24-h on call shift	1) Sleep hours on a normal day vs. following a 24-h shift: <2 hours: 0 vs. 5.9%; 2-4 hours: 5.9% vs. 47.1%; 4-6 hours: 11.8% vs. 35.3%; >6 hours: 82.4% vs. 11.8% 2) Number of sleep disturbances a normal day vs. following a 24-h shift: 0: 82.4% vs. 11.8%; 1: 11.8% vs. 35.3%; 2: 5.9% vs. 47.1%; 3: 0% vs. 5.9%; 4: 0% vs. 0%; >4: 0% vs. 0%	Neurocognitive parameters: computerized attentional test (vigilance, alertness); D2 letter cancellation test (divided attention); Trail Making Test (visual attention, task switching); Digit Span, Digit Symbol Substitution Test, Weschler Memory Scale (memory functions) Assessed before a normal day shift, and after a 24-h on call shift	Intrinsic alertness, focused attention and vigilance were similar on both occasions; Phasic alertness improved following the on-call shift: mean (SD) 24.8 (15.6) vs. 38.3 (21.5), p = 0.022.
Mixed specialties or undefined populations					
Chen, 2008 RoB: high	CS	Sleepiness: ESS score ≥11 Time points NR	Mean±SD ESS score: 7.8±4.0, range: 0-20, 23% had scores ≥11.	Impact on work and personal life: Impact Questionnaire with a 5-point Likert scale from 1 (strongly agree) to 5 (strongly disagree) Time points NR	1) Impact score correlated with ESS, r=0.31, p<0.05; 2) ESS score was higher among physicians who agree/strongly agree vs. other response: written an incorrect order: 8.8 vs. 7.3, p=0.02; might fall asleep while examining a patient: 13.2 vs. 7.7, p=0.001; look forward to sleeping at grand rounds: 10.4 vs. 7.4, p=0.002;

Study Risk of Bias (RoB)	Study design	Exposures or intervention Assessment measure and time points	 Baseline	Outcomes Assessment measure and time points	Associations between exposure and outcome
					3) No difference in ESS score for those who agree/strongly agree vs. other response: work is unaffected by sleep loss and fatigue, thinking is unaffected by sleep loss, sleep loss and fatigue affect my medical decisions, have heard of others making medical errors due to sleep loss and fatigue, never make errors in prescriptions on post-call days, have made medical errors because of sleep loss and fatigue; 4) Higher ESS scores predicted by impact score in multivariate regression including personal and work-related factors: β=0.11, p=0.005.
Heponiemi, 2014 RoB: low	CS	Sleeping problems: 4-item Jenkins Scale on 6-point scale from 1 (never) to 6 (every night) Assessed in 2006	Mean±SD (range) score: 2.30±1.00 (1-6)	Work ability: Work Ability Index on scale from 1 (could not work at all) to 10 (best work ability) Assessed in 2010	1) On-call duty had an indirect effect on work ability (R^2=0.11, 95% CI: -0.122, -0.031, p<0.001) through two mediators (work interference with family, sleeping problems); 2) Sleeping problems inversely associated with work ability, β=-0.29, p<0.001.
Kanieta, 2011 RoB: unclear	CS	Sleep hours: self-reported (continuous) Sleepiness and sleep difficulties: 5-point scale from 1 (never) to 5 (always); Insomnia: ≥3 sleep difficulties Assessed for the past month	Insufficient rest: 32.5%; Daytime sleepiness: 3.5%; Insomnia: 20.0%; Sleep time (mean±SD min): 279.8±60.9	Self-reported medical incidents: 4-point scale from 1 (never) to 4 (often) Assessed for the past month	1) Prevalence of medical incidents (% (95% CI)): sleep deprived (26.8% (24.2, 29.4)) vs. not (15.2% (13.7, 16.7)), p<0.01; insomnia (24.8% (21.6, 28.0)) insomnia vs. not (17.6% (16.2, 19.0)), p<0.01; ≥6h sleep (18.3% (16.8, 19.8)) vs. <6h (21.7% (18.8, 24.6)), p=0.03; 2) Predictors of medical incidents in multivariate model including personal and work-related factors (OR (95% CI)): lacking rest due to sleep deprivation vs. not (1.65 (1.33-2.04)), p<0.01); insomnia vs. not (1.45 (1.16-1.82), p<0.01); ns for sleep hours.

Study Risk of Bias (RoB)	Study design	Exposures or intervention Assessment measure and time points	Baseline	Outcomes Assessment measure and time points	Associations between exposure and outcome
Sexton, 2001 RoB: high	CS	Fatigue as a factor impacting performance Time points NR	NR	Performance effectiveness measured by 1 question: agree, neutral, disagree Time points NR	1) "When fatigued, I perform effectively during critical phases of operations/patient care": Anesthetic: 47% agree; 15% neutral; 38% disagree; Surgical: 70% agree; 12% neutral; 18% disagree.
Shirom, 2006 RoB: low	CS	Tiredness and exhaustion: SMBM Physician Fatigue Subscale on a 7-point scale from 1 (almost never) to 7 (always) Time points NR	NR	Quality of care: Adapted 15-item SERVQUAL with a 5-point Likert scale from 1 (very small extent) to 5 (very large extent) Time points NR	1) Quality of care positively predicted by fatigue in a model incorporating several other components of burnout, $\beta=0.17$, $p<0.05$.
Smith, 2017 RoB: moderate	CS	Sleep deprivation: self-reported via open-ended comments Time points NR	NR	Perceived competence: self-reported via open-ended comments Time points NR	Some physicians indicated that continual tiredness and exhaustion led to concerns that it would affect their competence; some felt that professional performance was compromised at times of physical and mental fatigue.
Tanti, 2017 RoB: high	CS	Fatique: questionnaire on contributors to prescribing errors, with a 5-point Likert scale (very high to very low association) Time points NR	NR	Prescribing errors: questionnaire on contributors to prescribing errors, with a 5-point Likert scale (very high to very low association) Time points NR	Perception of the contribution of fatigue to prescribing errors differed by physician type ($p<0.05$): 34% of community doctors, 96% hospital doctors, 8% of office-working doctors perceived a very high or high association between fatigue and prescribing errors.

[a]Includes studies of anesthetists, where these were physicians.
[b]Includes primary care physicians, internal medicine physicians, and general practitioners.
ACC: aortic cross-clamp time; BA: before-after; CABG: cardiopulmonary bypass time; CCPT II: Connor's Continuous Performance Test II; CI: confidence interval; CO: cohort; CS: cross-sectional; ER: emergency; ESS: Epworth Sleepiness Scale; h: hour(s); HVLT: Hopkin's Verbal Learning Task; Hz: Hertz; ICU: intensive care unit; IQR: interquartile range; KSS: Karolinska Sleep Scale; min: minutes; ms: millisecond(s); N-back: Dual N-back test; NA: not applicable; NR: not reported; ns: not statistically significant; OR: odds ratio; PSQI: Pittsburgh Sleep Quality Index; PVT: Psychomotor vigilance Performance Task; RR: risk ratio; RCT: randomized controlled trial; RoB: Risk of Bias; SD: standard deviation; SE: standard error; SERVQUAL: Service Quality Measure; SMBM: Shirom-Melamed Burnout Measure; TP: Toulouse-Piéron test; TS: time series; US: United States of America; vs.: versus

Patient outcomes related to fatigue or sleep restriction among physicians in independent practice

Study Risk of Bias (RoB)	Study design	Exposures Intervention or assessment scale and time points	 Baseline	Outcome Measures Assessment scale and time points	Associations between exposure and outcome
Surgeons					
Chu, 2011 RoB: low	CO	Sleep deprivation: moderate (3-6 h) or severe (<3-h) sleep deprivation the night before surgery (self-reported hours)	Of 4,047 procedures, 83 (2.1%) performed by severely sleep-deprived, 1,595 (39.4%) by moderately sleep-deprived surgeons	Chart review: mortality, surgical complications, length of stay Assessed during and post-surgery	1) 0-3 vs. 3-6 vs. >6 hours of sleep: No difference in incidence of mortality, incidence of 10 major complications (except septicemia, 3.6% vs. 0.9% vs. 0.8%, p=0.03), ICU length of stay; in-hospital length of stay (days): 7.0 vs. 6.0 vs. 7.0, p<0.001.
Ellman, 2004 RoB: low	CO	Sleep deprivation: performed a case starting 22:00 to 05:00, or ending 22:00 to 07:30 and performed a subsequent case in the next 24-h	Of 6,751 procedures, 339 (5%) were performed by sleep deprived surgeons	Chart review: mortality, surgical complications, length of stay Assessed during and post-surgery	1) Sleep deprived vs. non-sleep deprived: no difference in mortality, need for blood products, complications (operative, neurologic, renal, infectious, pulmonary), in-hospital length of stay.
Govindarajan, 2015 RoB: low	CO	Sleep deprivation: treated patients from midnight to 07:00 and performed a subsequent case on the same day	NR	Chart review: mortality, surgical complications, readmission, length of stay Assessed during and post-surgery	1) Sleep deprived vs. non-sleep deprived: no difference in mortality, surgical complications, readmissions within 30 days, or length of stay.
Rothschild, 2009 RoB: low	CO	Sleep deprivation: daytime procedures following an overnight procedure; Sleep opportunity: 0-6h, <6h	NR	Chart review: frequency of adverse surgical complications Assessed during and post-surgery	1) Post-nighttime vs. control: no difference in number of procedures with complications, total number of complications, preventable complications, type of complications; 2) Operating room procedures with complications, OR (95% CI): 8.5% for 0-6h sleep vs. 3.1% for >6h sleep, 2.70 (1.13-6.48), p=0.03; 3) All procedures with complications, OR (95% CI): 6.2% for 0-6h sleep vs. 3.4% for >6h sleep, 1.72 (1.02-2.89), p=0.04.

Study Risk of Bias (RoB)	Study design	Exposures		Outcome Measures	Associations between exposure and outcome
		Intervention or assessment scale and time points	Baseline	Assessment scale and time points	
Schieman, 2007 RoB: low	CO	Fatigue: surgeon billed for clinical work after 22:00 the night before surgery	Of 270 procedures, 22 (8%) were performed by fatigued surgeons	Chart review: surgical complications, length of stay, mortality, cancer recurrence Assessed during and post-surgery	1) Fatigued vs. non-fatigued surgeons: no difference in intra- or post-operative complication rate, length of stay, in-hospital length of stay, cancer recurrence.
Vinden, 2014 RoB: low	CO	Sleep deprivation (at risk): surgeon worked 00:00 to 07:00 and performed surgery 07:00 to 18:00	Of 94,183 surgeries, 2,078 (2.2%) were performed by surgeons who were 'at risk'	Chart review: conversion to open procedure (from laparoscopic), iatrogenic injuries, mortality Assessed during and post-surgery	1) At risk vs. not at risk surgeon: no difference in incidence of conversion to open procedure, iatrogenic injuries, mortality, in either univariate or multivariate analyses.
Obstetricians					
Rothschild, 2009 RoB: low	CO	Sleep deprivation: daytime procedures following an overnight procedure; Sleep opportunity: 0-6h, <6h	NR	Chart review: frequency of adverse obstetric complications Assessed during and post-delivery	1) Post-nighttime vs. control: no difference in number of procedures with complications, total complications, preventable complications, type of complications; 2) No association between sleep deprivation and proportion of procedures with complications, nor difference for 0-6h vs. >6h of sleep opportunity.

CI: confidence interval; CO: cohort; h: hours; NR: not reported; OR: odds ratio; RoB: Risk of Bias; SD: standard deviation; US: United States of America; vs.: versus