



UNIVERSIDAD PERUANA
CAYETANO HEREDIA

Facultad de
MEDICINA

EFFECTO DE RESTRICCIONES DE HORAS ASISTENCIALES EN MÉDICOS
RESIDENTES SOBRE SEGURIDAD DEL PACIENTE, BIENESTAR DEL
RESIDENTE Y EDUCACIÓN DEL RESIDENTE: REVISIÓN SISTEMÁTICA
Y METAANÁLISIS

THE EFFECT OF RESIDENTS' WORKING-HOUR RESTRICTIONS ON
PATIENT SAFETY, RESIDENT WELL-BEING, AND RESIDENT
EDUCATION: A SYSTEMATIC REVIEW AND META-ANALYSIS

TRABAJO DE INVESTIGACIÓN PARA OPTAR POR EL TÍTULO
PROFESIONAL DE MÉDICO CIRUJANO

AUTORES:

MAITZA ROSARIO VIDAL MEZA
GABRIELA ZAVALA WONG

ASESORA:

MARÍA DE LOS ÁNGELES LAZO PORRAS

LIMA-PERÚ
2021

JURADO

Presidente: SILVANA VIRGINIA SARABIA ARCE

Vocal: RAY WILLY TICSE AGUIRRE

Secretario: CRISTIAN PAUL LEON RABANAL

Fecha de sustentación: 15 de diciembre de 2021

Calificación: Aprobado

ASESORA DEL TRABAJO DE INVESTIGACIÓN

Dra. María de los Ángeles Lazo Porras

CRONICAS Center of Excellence in Chronic Diseases

Universidad Peruana Cayetano Heredia

ORCID: 0000-0003-0062-5476

DEDICATORIA

A nuestras familias que nos apoyaron incondicionalmente en nuestra formación y en memoria de Bruno, por su eterna amistad.

AGRADECIMIENTOS

A la Dra. María Lazo Porras quien nos brindó innumerables consejos tanto para nuestro desarrollo profesional como personal.

FUENTES DE FINANCIAMIENTO

Autofinanciado

DECLARACIÓN DE CONFLICTO DE INTERÉS

Los autores declaran no tener conflictos de interés

TABLA DE CONTENIDOS

I.	INTRODUCTION	1
II.	METHODS	5
III.	RESULTS	11
IV.	DISCUSSION	19
V.	CONCLUSIONS.....	22
VI.	REFERENCES.....	23
VII.	FIGURES	28
VIII.	TABLES.....	37

RESUMEN

Antecedentes: Las restricciones de horas asistenciales de médicos residentes continúan siendo una controversia. Los turnos prolongados pueden tener un efecto perjudicial ligado a la fatiga, sobre la seguridad del paciente y el bienestar de los residentes. Por otro lado, reducir la exposición de los residentes podría tener un impacto negativo en su formación. **Métodos:** Se realizó una revisión sistemática en Embase, Cochrane, PubMed, Medline, Clinicaltrials.gov y Global Index Medicus. Se incluyeron estudios que evaluaban el efecto de las horas asistenciales restrictivas sobre la seguridad del paciente, la educación y el bienestar de los residentes. Se realizó una síntesis narrativa y los hallazgos se combinaron mediante un metaanálisis, si eran elegibles, estimando el cociente de riesgos (RR) o la diferencia de medias estandarizada (DME) según correspondía con intervalo de confianza al 95% (IC95%). **Resultados:** Se incluyeron 17 estudios de metodología y calidad variable. El metaanálisis fue factible para la mortalidad a 30 días (RR de 0.98; IC95% 0.95-1.00), readmisión a 30 días (RR de 0.98; IC95% 0.96-1.00), evaluaciones durante la residencia (DME 0.02, IC95% -0.04 -0.07), agotamiento (DME -1.21, IC95% -2.11- -0.31) y cantidad de sueño (DME 0.23, IC95% -0.06-0.51). No se observó una mejora en la mayoría de estos hallazgos como resultado de horarios reducidos. Se encontraron resultados variables al evaluar otros parámetros. **Conclusiones:** Las restricciones de turnos asistenciales de los residentes no siempre se traducen en una mejor seguridad del paciente, educación y bienestar de los residentes. La heterogeneidad entre los estudios es el principal obstáculo para sacar conclusiones.

Palabras clave: Internado y Residencia, Programación de Personal, Calidad de la Atención de Salud, Educación Médica (DeCS)

ABSTRACT

Background: Restrictions on the number of resident working hours continues to be a controversy. Extended duty hours might have a detrimental effect on patient safety and residents' well-being, which are fatigue related. On the other hand, limiting the exposure of residents to patient care might have a negative impact on their training. **Methods:** A systematic review was carried out on Embase, Cochrane Database, PubMed, Medline, Clinicaltrials.gov and Global Index Medicus. No language or time restrictions were applied. We included studies evaluating the effect of restrictive working hours on patient safety, resident education, and well-being. A narrative synthesis was performed, and findings were pooled via meta-analysis, if eligible, estimating the risk ratio (RR) or standardized mean difference (SMD) accordingly with 95% confidence intervals (95%CI). **Results:** We included 17 studies of variable methodology and quality. Meta-analysis was only feasible for 30-day mortality (RR of 0.98, 95%CI 0.95-1.00), 30-day readmission (RR of 0.98, 95%CI 0.96-1.00), in-training examinations (SMD 0.02, 95%CI -0.04 - 0.07), burnout (emotional exhaustion domain SMD -1.21, 95%CI -2.11 - -0.31), and sleep quantity (SMD 0.23, 95%CI -0.06 - 0.51). No overall improvement was observed in most of these outcomes as a result of restrictive duty hours. Mixed results were found when assessing other parameters. **Conclusions:** Residents' duty hour restrictions do not always translate into improved patient safety, resident education, and resident well-being. Heterogeneity among the studies is the main obstacle to draw conclusions. Further studies with more robust methodology and consistency are needed to guide future policies.

Keywords: Internship and Residency, Personnel Staffing and Scheduling, Patient Safety, Education Medical Graduate (MeSH Terms)

INTRODUCTION

Medical residents are physicians who are undergoing postgraduate medical training to prepare for independent practice as a specialist. Depending on their desired specialty and program, this training can vary in duration. Not only is medical residency a key aspect of medical education, but medical residents themselves comprise an integral part of the health care team. These trainees have the responsibility of providing patient coverage while working historically known extended shifts. Long duty hours have been previously associated with fatigue and sleep deprivation that can lead to deleterious effects (1). Prolonged shifts have been linked to reduced alertness and impaired performance of residents that translates into medical errors and adverse events (2-4). Trainees have been shown to demonstrate attentional failures (2), impaired cognitive functioning (5), reduced working memory capacity (6), and altered surgical dexterity (7,8) among others. However, the negative impact extends beyond patient safety into the resident's own personal safety. As such, prolonged working hours have been associated with motor vehicle collisions (9) and potential for workplace harm, including needle-stick injuries (10).

This historic issue was not adequately addressed until the unfortunate death of a patient in a New York teaching hospital in 1984 that led to statewide regulations establishing limits on shift duration in 1986 (11). Persistent concerns about the possibility of negative effects of duty hours on patient safety later translated into nationwide reforms introduced by the United States Accreditation Council for

Graduate Medical Education (ACGME) in 2003. It mandated that residents could not work for more than 80 hours per week, have shifts no longer than 24 hours, and have at least 10 hours of rest between shifts (12, 13). In 2011, the ACGME implemented further reforms restricting the duration of shifts (12, 14). Around the globe, steps were also taken towards reduction of working hours (15). Unfortunately, there are regions particularly involving low- and middle-income countries (LMICs) that are characterized by heterogeneous regulations across medical teaching centers where many times duty hours fail to be methodically logged (16).

Currently, there is a great controversy regarding the duration of hospital shifts and the balance between patient safety, preserving resident well-being and successful trainee education. On the one hand, prolonged working hours have a deleterious effect on patient safety by interfering with the residents' mental and physical capabilities (17). A high workload in a sleep deprived condition could impair the trainee's clinical reasoning, operative skills, and capability for complex decision-making (18). Similarly overworked residents may be more prone to suffer from anxiety, frustration, and ultimately burnout (19). If the residents were to be well-rested, they would be able to provide optimal care, ensure the safety of the patient, improve their well-being, and demonstrate adequate cognitive abilities, which would in turn allow them to consolidate clinical knowledge, skills, and expertise for their future independent practice. On the other hand, implementation of restricted working hours would implicate a larger number of transitions of care, which may represent an even bigger threat to patient safety given that

communication errors may arise during handoffs and continuity of care may be lost (20). Shortened shifts would also imply less time of direct patient contact, which is crucial for obtaining hands-on-experience in residents' education process. This is a particularly key aspect given that residents have the dual responsibility to act as care providers and obtain as many educational opportunities as possible. Various studies have attempted to explore the composite effects of resident work hour reforms on patient safety outcomes, resident well-being, and resident education. The Flexibility in Duty Hour Requirements for Surgical Trainees (FIRST) trial, which ran in the 2014-2015 academic year, demonstrated the non-inferiority of flexible, extended schedules with respect to patient mortality and morbidity outcomes (21). Based on the results of this trial, in 2017 ACGME updated the 2011 policy and allowed longer duration of shifts. Additional randomized controlled trials (RCTs) have been developed to explore similar outcomes yielding controversial results including the Individualized Comparative Effectiveness of Models Optimizing Patient Safety and Resident Education (iCOMPARE) trial and the Randomized Order Safety Trial Evaluating Resident-Physician Schedules (ROSTERS) (22, 23). An updated systematic review and meta-analysis are crucial to interpret this newly available, high-grade evidence from experimental studies that explore the impact of working hour restrictions on patient outcomes, resident well-being, and resident education. These parameters must be considered when developing policies regarding working hours in graduate medical education programs worldwide.

OBJECTIVES

The primary objective of this study:

1. To evaluate the effect of residents' working-hour restrictions on patient safety indicators.

Furthermore, the following secondary outcomes will be explored:

2. To assess the impact of the effect of residents' working-hour restrictions on residents' perceived well-being.
3. To assess the impact of the effect of residents' working-hour restrictions on performance on examinations and self-reported satisfaction with education received.

METHODS

This systematic review was conducted according to the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) statement (24). The protocol is available online (25).

1. Eligibility criteria

Types of studies

For the primary objective evaluating the effect of working hours on patient safety, only randomized controlled trials were included. For secondary objectives evaluating the effect of working hours on residents' education and well-being, we included randomized controlled trials as well as non-randomized intervention studies. No restrictions in language or year of publication were applied.

Participants

We only considered studies with subjects enrolled in a residency program aged 18 or older, and who work on reduced shifts (16-hour or less) or extended shifts (24-hour or more). No restriction was applied on their year of training or medical specialty. Other health professionals in a more senior level of training or from non-medical areas were excluded.

Intervention

Working hour restrictions were considered as shift length reduction (shifts of 16 hours or less) and/or implementation of a night float system. A specific minimum

or maximum duration of the intervention was not considered. Studies that implemented protected sleep periods, wellness or time management workshops were excluded.

Comparison

A traditional extended shift was considered as one which had no restrictions on the number of consecutive working hours (“flexible schedules”) and/or did not implement a night float system.

Outcomes

Primary outcomes were based on patient safety parameters including 30-day mortality, 30-day readmission, adverse events, medical errors, patient safety indicators, in-hospital complications, and length of stay.

Secondary outcomes were measured on residents’ self-reported job satisfaction, working hour satisfaction, sleepiness/fatigue, burnout, time for family/friends, time for hobbies, vacation time, overall health, mental health, somatic symptoms, risk for percutaneous/ attentional injuries and risk for vehicle crashes. With respect to education, evaluated outcomes included residents’ self-reported rounds/conference attendance, independent learning opportunities, missed operating room (OR) time, sense of autonomy, continuity of care, bedside teaching, teaching satisfaction, examination scores and overall satisfaction with educational experience.

Exclusion criteria

We excluded studies that were not randomized controlled trials or non-randomized intervention studies. We also excluded studies that evaluated working-hour restrictions in healthcare professionals other than medical residents, but which results were not disaggregated by occupation. Additionally, studies where co-interventions to working hours restrictions were implemented were also excluded. Lastly, studies where the full text was not accessible despite directly contacting the authors were also excluded.

2. Information sources and search strategy

We performed searches in PubMed, MEDLINE, EMBASE, Cochrane Library, Clinicaltrials.gov and Global Index Medicus using medical subject headings (MeSH terms) or equivalents in other medical databases and free text words related to resident working hours, resident well-being, resident education, and patient safety, randomized controlled trials and non-randomized intervention studies. The databases were searched on 30 July 2021 by the study authors (GZ, MV) with no restrictions in language or data of publication. Search strategies are detailed in the protocol of the present study.

3. Selection process

The results from the search in each database were imported into the Zotero citation management software where duplicate items were removed. Citation files organized by folders were then uploaded to the Rayyan QCRI reference management software. The initial screening process of titles and abstracts was

independently carried out by the two authors (GZ and MV). The reviewer specified the reason to exclude each study. Results were then compared, and disagreements were resolved. The second screening process of full text was then carried out in a similar manner. The two authors independently evaluated each study to identify whether they would be included. Discrepancies on eligibility in this phase were resolved by consulting with the third research collaborator (ML). A PRISMA flow chart was constructed to illustrate the number of articles included and/or excluded in each phase of the selection process.

4. Data collection process

Data extraction was carried out by two authors (GZ and MV) in an independent manner by filling out a Google Forms Spreadsheet. Any discrepancies in data collection were resolved with discussion between the two reviewers and with the third collaborator (ML), when needed. Data extracted included the following: (i) Study details: author, study title, journal, year of publication, number of sites, country, study type (RCT or non-randomized intervention study) (ii) Participant characteristics: sample size, medical specialty, rotation, rank (iii) Details of the intervention: type of intervention (shift length reduction or night float), schedule details, duration of intervention, control description (iv) Outcome measures for patient safety: 30-day mortality, 30-day readmission, adverse events, medical errors, patient safety indicators, in-hospital complications and length of stay for patient safety. Outcome measures recorded for secondary objectives were detailed above in the Eligibility criteria section.

5. Data synthesis

A qualitative narrative synthesis was employed to summarize the key findings, population, and methodology of studies using text and tables. Each outcome was individually evaluated to see if results from at least three different studies and/or three different intervention schedules could be meta-analyzed. This included outcome measurement (e.g.: scale used) and value reported (mean, median, proportion).

6. Statistical analysis

Analysis was carried out using the RevMan version 5.4 following a random effects analysis model. The summary outcome measures reported varied according to the type of data reported. For continuous data, we calculated the standardized mean difference (SMD) and applied fixed-effect (Mantel-Haenszel method). For dichotomous data, the risk ratio (RR) and 95% confidence interval (95%CI) was estimated, and an inverse variance method was used. Heterogeneity was assessed using Higgins and Thompson I^2 . Studies with high heterogeneity (>75%) were excluded for meta-analysis. P values <0.05 were considered significant.

7. Risk of bias assessment

The assessment of studies' quality followed the Cochrane Risk-of-Bias tool for randomized controlled trials (RoB 2) and Risk of Bias in Non-Randomized Studies-of Intervention (ROBINS-I) (26, 27). For randomized controlled trials, risk of bias was assessed using the tool's algorithm for suggested judgment. For non-randomized studies, it was assessed using the tool's interpretation tables. Two

authors (GZ and MV) independently evaluated the studies and determined risk of bias for each domain and overall. Disagreements were resolved by discussion.

RESULTS

Study selection

A total of 1220 articles were identified from medical databases, 844 were selected for title and abstract screening after eliminating duplicates. During this initial screening a total of 789 studies were excluded, with 55 studies remaining. Out of these 55, seven registered clinical trials were identified; these had a total of 24 associated publications which were added to the remaining pool totaling 79 studies for full text screening. Finally, 62 studies were excluded, and the remaining 17 were eligible for inclusion. The number of articles included and/or excluded in each phase of the selection process can be visualized in the flowchart (Figure 1).

Study characteristics

Twelve of 17 studies were conducted in the United States (US), with the remaining taking place in Canada (n=3), Saudi Arabia (n=1) and Singapore (n=1). Included studies were published in the years 2004 through 2021 and, despite the absence of language restrictions, all of them were published in English. Regarding participants, Internal Medicine was the most common studied specialty (n=6), followed by surgery (n=5) and pediatrics (n=3). The remaining three studies evaluated residents from diverse specialties. The most common study design was RCT (n=11); the rest of studies were non-randomized interventional designs including pre-and-post studies (n=3), non-randomized trials (n=1) and prospective cohort (n=1). Of note, from the 11 RCTs, six studies followed a noninferiority design and three followed a crossover design (Table 1).

The most common intervention was shift length reduction (n=6), out of which one study (30) featured more than one intervention schedule (12-hour shifts and another for 16-hour shifts). Three studies implemented only night float as the intervention, while two studies (33, 36) featured two intervention schedules one of which was night float. The remaining six studies implemented a “flexible” extended schedule as the intervention and a control group with “standard” reduced shifts. In these studies, as specified in Table 1, the intervention and control groups were switched for homogeneity when meta-analyzed. The outcomes evaluated were patient safety (n=7), resident education (n=11), and resident well-being (n=11). In total, eight out of the 17 included studies contained data eligible for meta-analysis and the outcomes that were possible to be meta-analyzed were 30-day mortality, 30-day readmission, in-training examination scores, burnout, and sleep quantity.

Effects of Residents Working Hours Restrictions on Patient Safety

We reviewed seven studies that met the criterion of inclusion to evaluate patient safety. Meta-analysis was only feasible for two patient safety outcomes: 30-day mortality and 30-day readmission.

Parshuram *et al.*, Bilimoria *et al.* and Silber *et al.* evaluated 30-day mortality and were eligible for meta-analysis. Of note, Parshuram *et al.* had two interventions (a 16-hour and a 12-hour group compared to a control schedule of 24-hour shifts), each of which was considered as an independent entry for analysis purposes. The overall effect of restricted resident working hours was estimated at RR of 0.98, 95%CI 0.95-1.00 (Figure 2). This means that there is no significant effect of resident working hour restrictions on 30-day mortality.

Desai *et al.* 2013 and Silber *et al.* evaluated 30-day readmission and were eligible for meta-analysis. In a similar manner, Desai *et al.* 2013 had two interventions (a Q5 schedule and a night float rotation compared to an extended control schedule), once again considered as independent entries for analysis. The effect on 30-day readmission was estimated at a RR of 0.98 (95%CI 0.96-1.00), which is not statistically significant (Figure 3). Hence, there is no significant effect of resident working hour restrictions on 30-day readmissions.

The other patient safety outcomes could not be meta-analyzed (7-day readmission, near misses, preventable adverse events, serious medical errors, in-hospital complications, length of stay and patient safety indicators.). A summary of the intervention effect on these different patient safety parameters can be found in Table 2. Of note, studies did not find a significant change in 7-day readmission, in-hospital complications, or Agency for Healthcare Research and Quality (AHRQ) patient safety indicators after duty hour reforms. We found mixed results on whether a shorter shift length had an effect on the number of preventable adverse events, serious medical errors, and length of stay. Only one study evaluated near misses and found that it apparently increased after restrictions on shift length were implemented; however, when this was adjusted for residents' workloads as a possible confounding factor, there was no longer an association.

Effects of Residents Working Hours Restrictions on Educational Outcomes

Eleven studies evaluated the educational impact of reducing working hours in residents. Meta-analysis was only feasible for one resident education outcome: in-training examination scores.

Rajaram *et al.*, Blay *et al.*, and Desai *et al.* 2018 objectively assessed the impact of shift reduction on trainees' education through in-training examinations. When meta-analysis was performed for these three studies, the SMD was estimated at 0.02 (95%CI -0.04 - 0.07), which is not significant among shifts (Figure 4). Hence, there was no significant difference in in-training examination scores between groups assigned to different duty hour schedules.

Board examinations were reported by two studies. Blay *et al.* concluded that there was no difference in trainees' performance regarding the type of schedule they had. Similarly, Rajaram *et al.* found no significant change in oral boards passing rate (80.9% in the intervention group vs. 81.7% in the control group, $p=0.21$); however, controversial results were encountered when evaluating the written boards passing rate (Table 3).

The other educational outcomes evaluated were self-reported by participants in the form of surveys. A summary of the intervention effect on the most representative resident education outcomes can be found in Table 3. Studies revealed that independent learning, particularly protected research time, improved in residents whose schedules were reduced. Only one study assessed trainees' missed OR time and found that 16-hour or less shifts led to a higher frequency of missed operations. Educational conferences attendance, resident's autonomy, continuity of care, opportunities for bedside teaching, teaching satisfaction and overall educational experience had mixed results.

Effects of Residents Working Hours Restrictions on Residents' Well-being

Eleven studies assessed the impact of shift reduction on residents' well-being. Meta-analysis was only feasible for two resident well-being outcomes: burnout and sleep quantity.

Burnout was evaluated in five studies (Table 4). Two studies found that residents in intervention groups (reduced shifts and/or night float) rated burnout as worse than those in the control group. One study found no significant differences between groups with respect to this parameter. The remaining two studies (Desai *et al.* 2018 and Parshuram *et al.* with two interventions each of which was considered as an independent entry) were eligible for meta-analysis and evaluated this outcome objectively with the Maslach Burnout Inventory (MBI) that evaluated three domains: depersonalization, emotional exhaustion, and personal achievement. While higher scores in depersonalization and emotional exhaustion correlate with worse burnout, lower scores in personal achievement correspond to worse burnout. When the three parameters of the MBI were meta-analyzed, only the SMD in the emotional exhaustion score was found to be significant -1.21 (95%CI $-2.11 - -0.31$) (Figures 5.1-5.3) and lower in residents with reduced working hours. The latter suggests that, in this particular domain, residents with shorter shifts presented lower degrees of burnout.

Sleep quantity was evaluated in five studies (Table 4). Two studies found no significant difference in residents' amount of sleep between the intervention groups. One study found that residents in the intervention group reported significantly more sleep than those in the control group. The remaining two studies (Basner *et al.* and Desai *et al.* 2013 with two interventions each considered as an independent entry)

were eligible for meta-analysis. The overall effect of shift duration on sleep quantity was not significant, SMD of 0.23 (95%CI -0.06 - 0.51) (Figure 6). In a similar manner, seven studies subjectively evaluated the degree of sleepiness and fatigue in residents. Five studies found no significant difference in this parameter between intervention groups. Basner *et al.* reported that “flexible” extended schedule policies were noninferior to “standard” reduced hours policies with respect to daytime sleepiness. On the other hand, Alshime *et al.* reported that residents considered that extended schedules contributed significantly more to overall fatigue.

A summary on the intervention effect on the other resident well-being outcomes can be found in Table 4. Of note, in three out of four studies, residents reported that shorter shifts improved their time for family, friends, and hobbies. Only one study reported about driving safety after shifts, which was reported as being more impaired in trainees with extended schedules. Overall health was also reported, in four out of five studies, to be better in reduced shifts. Results for job satisfaction were controversial. Lastly, two of the three studies that evaluated working hours satisfaction showed an improvement in the group of residents with shorter shifts.

Risk of bias assessment

Results for the eleven RCTs included are shown in Figure 7. Seven of the RCTs (63.6%) were considered as having an overall low risk of bias, three (27.2%) were considered as having some concerns overall and one (9.1%) was considered as having an overall high risk of bias. Of note, Landrigan *et al.* 2020 was considered as having some concerns for bias arising from the randomization process due to a

baseline difference (number of patients per resident) between intervention groups. This observation led to a classification of ‘some concern’ for overall risk of bias. Additionally, Desai *et al.* 2013 was labeled as high risk for bias due to deviations from intended intervention given that there were significant deviations that arose (one of the intervention schedules was terminated early) and there was no information regarding whether an appropriate analysis used to estimate the effect of assignment to intervention.

Results for the six non-randomized intervention studies are shown in Figure 8. Out of the six, four studies (66.7%) were considered as having an overall serious risk of bias while two studies (33.3%) were considered as having an overall moderate risk of bias. It is important to note that for all these studies the protocols were not available, and the evaluation was based in the published article. In Rajaram *et al.*, there was no information provided on deviations from intended intervention because no details are given on programs and residents’ actual adherence to the schedule throughout the study period or how they applied the duty hour reforms in their particular hospitals. Additionally, the risk of bias in selection of the reported result was deemed to be serious because the reported effect estimate seems to have likely been selected from multiple analyses (they disregard unadjusted results in their conclusions). Also of note, Auger *et al.* was considered to have a serious risk of bias due to missing data given that they do not report absolute values of survey responses and no appendix was available; the extracted data were estimates from graphs. Concerns for Ming Low *et al.* included significant difference in baseline burnout and sleepiness measurements between intervention groups and a high dropout rate. Lastly, Zahrai *et al.* had significant deviation from the intended

intervention given that even though the study period was six months, the residents on the intervention team only completed three weeks of the intervention schedule.

DISCUSSION

Extended shifts are often reported to have adverse consequences on patient safety, residents' education, and residents' well-being (42). Nonetheless, this present systematic review and meta-analysis reveals that restrictions in working schedules do not always translate into improved patient safety parameters nor residents' education and overall health.

Overall, most of the studies reported that working on a shorter schedule improved residents' self-directed learning and attendance to educational conferences. But when this reported perception is compared to objective ways of measuring residents' performance, no changes are found. The two studies that assessed both oral and written boards when comparing residents in a short or extended schedule revealed that scores did not vary. Furthermore, we meta-analyzed in-training examinations and results supported those same conclusions. It is important to highlight the findings in one of the RCTs that evaluate the impact of reducing working hours on missing an operation. Bilimoria *et al.* found that when interns (first year residents) work in 16-hour schedules they were more likely to miss an operation. This is particularly important to surgical specialties when the mastery of the procedures requires time and practice. Even some studies suggest that 10,000 hours are required to achieve complex technical tasks (43). This might generate an adverse consequence on postgraduate first year residents whose role in the operating room is further limited by these types of duty restrictions, therefore, making them feel less prepared for next year responsibilities (44). If this happens in five-year surgical programs, we can expect additional limitations in three-year programs, as is the case of many LMICs.

This review of literature indicates that residents reported an improvement in their overall health in most of the studies. However according to Bilimoria *et al.*, no significant change was found when they were asked about job and working hours satisfaction. Surprisingly, shorter shifts did not necessarily translate into better sleep quantity or a decrease in sleepiness (21,30,34,35). Four of five studies (30,35,37,40), two of them RCTs (30,35), found that restricting working hours did not have a major impact on residents' burnout. However, our meta-analysis reveals that having a reduced shift length improves the mean score on MBI emotional exhaustion. Even though most of the well-being outcomes are similar in both kinds of schedules, shift length is concerning due to the detrimental effects it has on residents' quality of life.

The present review has limitations. They are related to the heterogeneity (Figures 9-15) among the included studies, duty hours restrictions were applied over diverse time intervals and various clinical contexts, adherence to duty hours was rarely considered or evaluated, and that for assessing resident wellness and resident education sometimes randomization was not possible given the different regulations among medical programs. The latter represented a problem as these interventional studies were categorized as low quality, which ultimately originates unfavorable conditions to generate a meaningful meta-analysis of the literature related to the impact of residents working hours restrictions. Because of this, we were not able to draw definitive conclusions. In addition, many of the included studies did not correlate the working hours with working load (patients per resident), which is an

important factor, especially when analyzing patient safety indicators. Also, they did not report, in most of the cases, if the total hours per week (80 hours) were respected. Unfortunately, the impact of other parameters such as attentional failures, alertness, time spent on direct patient care, professionalism and costs were out of the scope of this review.

We found that current evidence is not sufficient to be the base of new policy and reforms on residents' duty hours. Even though there has been an increase in RCTs publications regarding this matter, methodology varies across programs, and they do not always use validated measures or even evaluate the same outcomes. It is essential to future research to be more methodologically robust, to measure in a constant manner the same outcomes and to include process evaluation outcomes to conduct a more effective synthesis of evidence and to understand the barriers and facilitators for different stakeholders to implement the studied interventions. Involving other stakeholders (i.e., program directors and attendings), which is often ignored in most studies, helps to have a more integrated view of the impact on residents' duty hours facilitating policymaking. Finally, we encourage other countries to conduct these types of studies, as evidence included in this systematic review came exclusively from high-income countries, which makes it harder to generalize to other settings, particularly LMICs.

CONCLUSIONS

The present systematic review and meta-analysis included 17 studies that evaluated the effect of trainees' reduced working hours on patient safety parameters, resident education, and resident well-being. The synthesized studies suggest that restrictions on residents' working hours do not often result in an improvement of patient safety parameters. Reported evidence varies across the studies, mainly due to the heterogeneity of their methodology. When it comes to residents' education, the review of literature reveals that boards and in-training examinations were not affected by the length of working shifts. However, residents' subjective perception of educational experience is diverse. The evidence suggests that restrictive duty hours generally improve residents' sense of well-being. This topic is of great importance to medical education and residency programs should individually evaluate these parameters in their particular settings when implementing duty hour reforms.

REFERENCES

1. Philibert I. Sleep loss and performance in residents and nonphysicians: a meta-analytic examination. *Sleep* 2005;28:1392-402.
2. Barger LK, Ayas NT, Cade BE, Cronin JW, Rosner B, Speizer FE, et al. Impact of Extended-Duration Shifts on Medical Errors, Adverse Events, and Attentional Failures. *PLOS Medicine*. 2006 dic;3(12):e487.
3. Gaba DM, Howard SK. Fatigue among Clinicians and the Safety of Patients. *New England Journal of Medicine*. 2002 Oct 17;347(16):1249–55.
4. Weinger MB, Ancoli-Israel S. Sleep deprivation and clinical performance. *JAMA* 2002; 287: 955-7.
5. Arnedt JT, Owens J, Crouch M, Stahl J, Carskadon MA. Neurobehavioral performance of residents after heavy night call vs after alcohol ingestion. *JAMA*. 2005 Sep 7;294(9):1025–33.
6. Gohar A, Adams A, Gertner E, Sackett-Lundeen L, Heitz R, Engle R, et al. Working Memory Capacity is Decreased in Sleep-Deprived Internal Medicine Residents. *J Clin Sleep Med*. 2009 Jun 15;5(3):191–7.
7. Taffinder NJ, McManus IC, Gul Y, Russell RC, Darzi A. Effect of sleep deprivation on surgeons' dexterity on laparoscopy simulator. *Lancet*. 1998 Oct 10;352(9135):1191.
8. Eastridge BJ, Hamilton EC, O'Keefe GE, Rege RV, Valentine RJ, Jones DJ, et al. Effect of sleep deprivation on the performance of simulated laparoscopic surgical skill. *Am J Surg*. 2003 Aug;186(2):169–74.
9. Barger LK, Cade BE, Ayas NT, Cronin JW, Rosner B, Speizer FE, et al. Extended Work Shifts and the Risk of Motor Vehicle Crashes among Interns [Internet]. <http://dx.doi.org/10.1056/NEJMoa041401>. Massachusetts Medical Society; 2009 [cited 2021 Nov 12]. Available from: <https://www.nejm.org/doi/10.1056/NEJMoa041401>
10. Ayas NT, Barger LK, Cade BE, Hashimoto DM, Rosner B, Cronin JW, et al. Extended Work Duration and the Risk of Self-reported Percutaneous Injuries in Interns. *JAMA*. 2006 Sep 6;296(9):1055–62.

11. Spritz N. Oversight of physicians' conduct by state licensing agencies. Lessons from New York's Libby Zion case. *Ann Intern Med.* 1991 Aug 1;115(3):219–22.
12. Imrie KR, Frank JR, Parshuram CS. Resident duty hours: past, present, and future. *BMC Medical Education.* 2014 Dec 11;14(1):S1.
13. Philibert I, Friedmann P, Williams WT, for the members of the ACGME Work Group on Resident Duty Hours. New Requirements for Resident Duty Hours. *JAMA.* 2002 Sep 4;288(9):1112–4.
14. Nasca TJ, Day SH, Amis ES. The New Recommendations on Duty Hours from the ACGME Task Force. *New England Journal of Medicine.* 2010 Jul 8;363(2):e3.
15. Temple J. Resident duty hours around the globe: where are we now? *BMC Medical Education.* 2014 Dec 11;14(1):S8.
16. Inga-Berrospi F, Toro-Huamanchumo CJ, Arestegui Sanchez L, Torres-Vigo V, Taype-Rondán A. Características de la residencia médica en sedes docentes de Lima, Perú. *Educ Med Super [Internet].* 2016 Jun;30(2). Available from: http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S0864-21412016000200014&lang=es
17. Institute of Medicine (US) Committee on Optimizing Graduate Medical Trainee (Resident) Hours and Work Schedule to Improve Patient Safety. Resident Duty Hours: Enhancing Sleep, Supervision, and Safety [Internet]. Ulmer C, Miller Wolman D, Johns MME, editors. Washington (DC): National Academies Press (US); 2009 [cited 2021 Oct 12]. Available from: <http://www.ncbi.nlm.nih.gov/books/NBK214948/>
18. Lockley SW, Cronin JW, Evans EE, Cade BE, Lee CJ, Landrigan CP, et al. Effect of Reducing Interns' Weekly Work Hours on Sleep and Attentional Failures. *New England Journal of Medicine.* 2004 Oct 28;351(18):1829–37.
19. Thomas NK. Resident Burnout. *JAMA.* 2004 Dec 15;292(23):2880–9
20. Vidyarthi AR, Vidyarthi AR, Arora V, Schnipper JL, Wall SD, Wachter RM. Managing discontinuity in academic medical centers: Strategies for a safe and effective resident sign-out. *Journal of Hospital Medicine*

[Internet]. 2006 Jul 1 [cited 2021 Nov 12];1(4). Available from:
<https://www.journalofhospitalmedicine.com/jhospmed/article/128167/strategies-safe-and-effective-resident-sign-out>

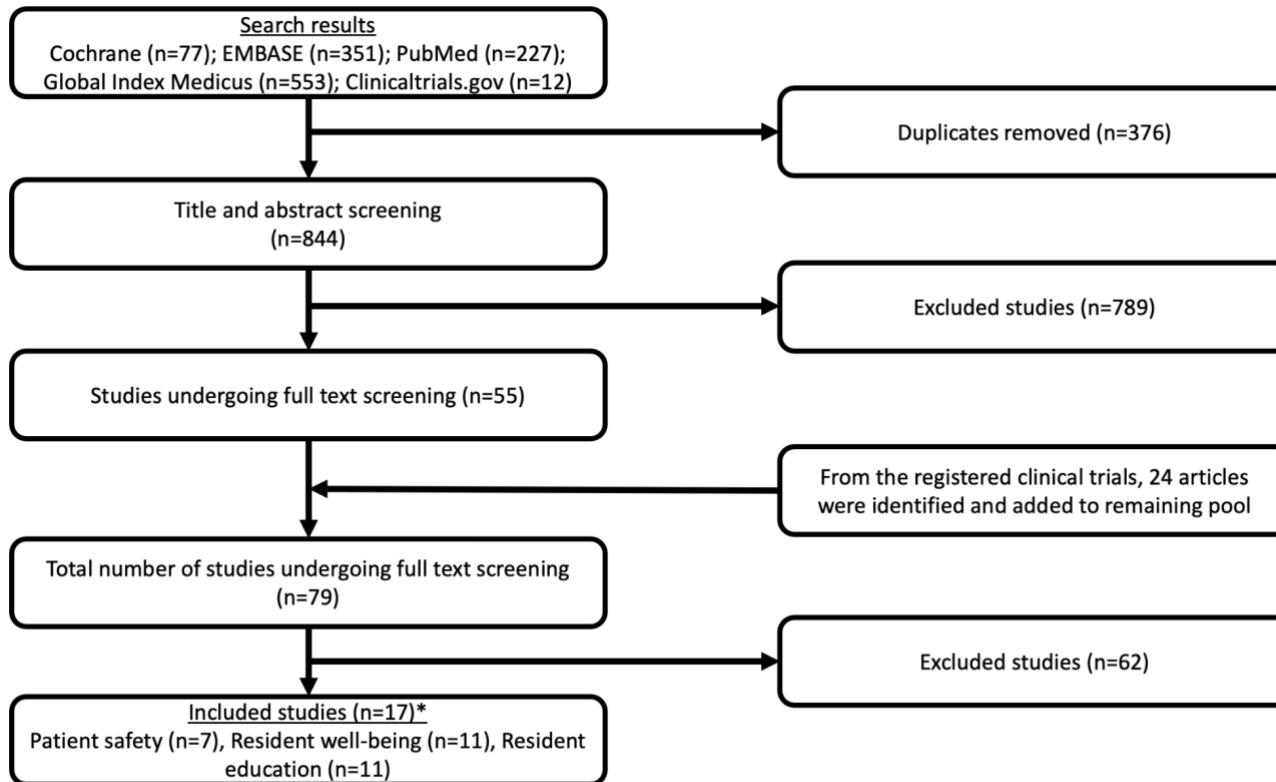
21. Bilimoria KY, Chung JW, Hedges LV, Dahlke AR, Love R, Cohen ME, et al. National Cluster-Randomized Trial of Duty-Hour Flexibility in Surgical Training. *New England Journal of Medicine*. 2016 Feb 25;374(8):713–27.
22. Silber JH, Bellini LM, Shea JA, Desai SV, Dinges DF, Basner M, et al. Patient Safety Outcomes under Flexible and Standard Resident Duty-Hour Rules. *New England Journal of Medicine*. 2019 Mar 7;380(10):905–14.
23. Landrigan CP, Rahman SA, Sullivan JP, Vittinghoff E, Barger LK, Sanderson AL, et al. Effect on Patient Safety of a Resident Physician Schedule without 24-Hour Shifts. *New England Journal of Medicine*. 2020 Jun 25;382(26):2514–23.
24. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *PLOS Medicine*. 2021 Mar 29;18(3):e1003583.
25. Zavala-Wong G, Vidal-Meza M, Lazo-Porras M. The Effect of Residents' Working-Hour Restrictions on Patient Safety, Resident Well-Being, and Resident Education: Protocol for a Systematic Review and Meta-Analysis [Internet]. 2021 Aug [cited 2021 Nov 12] p. 2021.08.26.21262689. Available from:
<https://www.medrxiv.org/content/10.1101/2021.08.26.21262689v1>
26. Sterne JAC, Savović J, Page MJ, Elbers RG, Blencowe NS, Boutron I, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. *BMJ*. 2019 Aug 28;366:l4898.
27. Sterne JA, Hernán MA, Reeves BC, Savović J, Berkman ND, Viswanathan M, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. *BMJ*. 2016 Oct 12;355:i4919.
28. Barger LK, Sullivan JP, Blackwell T, O'Brien CS, St Hilaire MA, Rahman SA, et al. Effects on resident work hours, sleep duration, and work experience in a randomized order safety trial evaluating resident-physician schedules (ROSTERS). *Sleep*. 2019 Aug 1;42(8):zsz110.

29. Landrigan CP, Rothschild JM, Cronin JW, Kaushal R, Burdick E, Katz JT, et al. Effect of reducing interns' work hours on serious medical errors in intensive care units. *N Engl J Med*. 2004 Oct 28;351(18):1838–48.
30. Parshuram CS, Amaral ACKB, Ferguson ND, Baker GR, Etchells EE, Flintoft V, et al. Patient safety, resident well-being and continuity of care with different resident duty schedules in the intensive care unit: a randomized trial. *CMAJ*. 2015 Mar 17;187(5):321–9.
31. Rajaram R, Chung JW, Jones AT, Cohen ME, Dahlke AR, Ko CY, et al. Association of the 2011 ACGME resident duty hour reform with general surgery patient outcomes and with resident examination performance. *JAMA*. 2014 Dec 10;312(22):2374–84.
32. Stulberg JJ, Pavey ES, Cohen ME, Ko CY, Hoyt DB, Bilimoria KY. Effect of Flexible Duty Hour Policies on Length of Stay for Complex Intra-Abdominal Operations: A Flexibility in Duty Hour Requirements for Surgical Trainees (FIRST) Trial Analysis. *J Am Coll Surg*. 2017 Feb;224(2):143-148.e1.
33. Desai SV, Feldman L, Brown L, Dezube R, Yeh H-C, Punjabi N, et al. Effect of the 2011 vs 2003 duty hour regulation-compliant models on sleep duration, trainee education, and continuity of patient care among internal medicine house staff: a randomized trial. *JAMA Intern Med*. 2013 Apr 22;173(8):649–55.
34. Basner M, Asch DA, Shea JA, Bellini LM, Carlin M, Malone SK, et al. A randomized trial on the effects of standard and flexible duty-hour rules on intern sleep and alertness. *Sleep Sci*. 2019;12((Basner M.; Asch D.A.; Shea J.A.; Bellini L.M.; Carlin M.; Malone S.K.; Volpp K.G.; Dinges D.F.) University of Pennsylvania, Philadelphia, PA, United States):6.
35. Desai SV, Asch DA, Bellini LM, Chaiyachati KH, Liu M, Sternberg AL, et al. Education Outcomes in a Duty-Hour Flexibility Trial in Internal Medicine. *N Engl J Med*. 2018 Apr 19;378(16):1494–508.
36. Moeller A, Webber J, Epstein I. Resident duty hour modification affects perceptions in medical education, general wellness, and ability to provide patient care. *BMC Med Educ*. 2016 Jul 13;16:175.

37. Auger KA, Landrigan CP, Gonzalez del Rey JA, Sieplinga KR, Sucharew HJ, Simmons JM. Better rested, but more stressed? Evidence of the effects of resident work hour restrictions. *Acad Pediatr*. 2012 Aug;12(4):335–43.
38. Blay EJ, Hewitt DB, Chung JW, Biester T, Fiore JF, Dahlke AR, et al. Association Between Flexible Duty Hour Policies and General Surgery Resident Examination Performance: A Flexibility in Duty Hour Requirements for Surgical Trainees (FIRST) Trial Analysis. *J Am Coll Surg*. 2017 Feb;224(2):137–42.
39. Alsohime F, Alkhalaf H, Almuzini H, Alyahya M, Allhidan R, Assiry G, et al. Pediatric resident's perception of night float system compared to 24 hours system, a prospective study. *BMC Med Educ*. 2021 Jan 6;21(1):23.
40. Jia-Ming LOW, Mae-Yue TAN, Kay-Choong SEE, Marion-M AW. Sleep, activity and fatigue reported by Postgraduate Year 1 residents: a prospective cohort study comparing the effects of night float versus the traditional overnight on-call system. *Singapore medical journal*. 2018;652–5.
41. Zahrai A, Chahal J, Stojimirovic D, Schemitsch EH, Yee A, Kraemer W. Quality of life and educational benefit among orthopedic surgery residents: A prospective, multicentre comparison of the night float and the standard call systems. *Can J Surg*. 2011;54(1):25–32.
42. Caruso CCH, Hitchcock EM, Dick RB, Russo JM, Schimit JM. Overtime and Extended Work Shifts: Recent Findings on Illnesses, Injuries, and Health Behaviors. Cincinnati: National Institute for Occupational Safety and Health; 2004.
43. Anders Ericsson K, KrampeR, Tesch-Romer C. The role of deliberate practice in the acquisition of expert performance. *Psychol Rev*. 1993;100:363–406.
44. Drolet BC, Christopher DA, Fischer SA. Residents' response to duty-hour regulations: a follow-up national survey. *N Engl J Med*. 2012;366:e35.

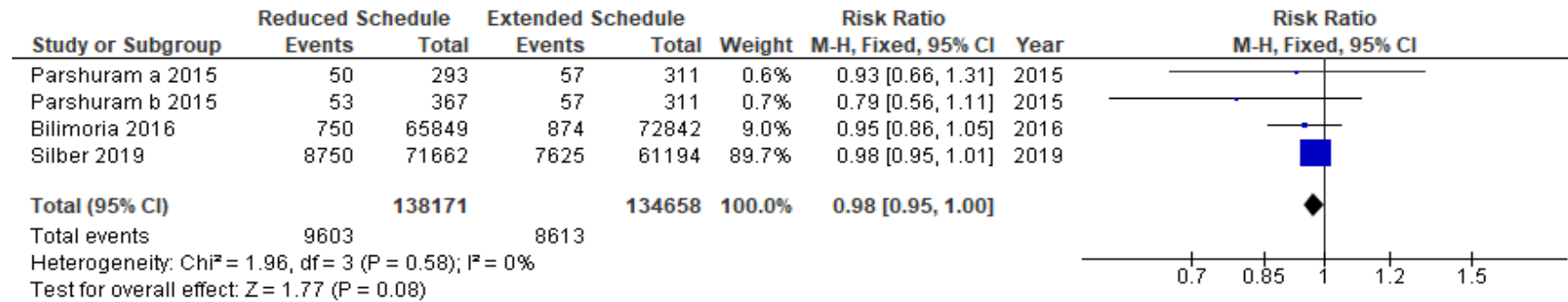
FIGURES

Figure 1. Study selection PRISMA flowchart



**Some studies reported more than one outcome of interest*

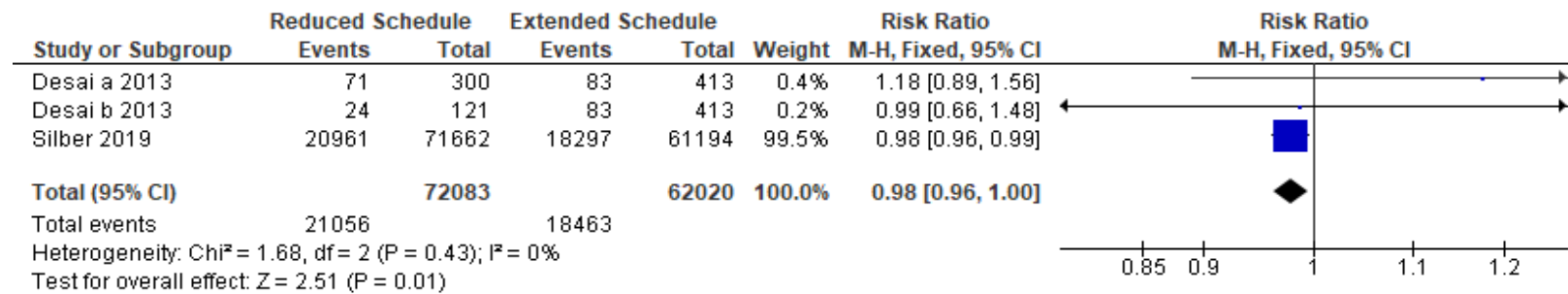
Figure 2. Meta-analysis of three studies for 30-day mortality



*Parshuram *et al.* had two intervention groups:

- Parshuram *et al.* a: intervention was 16hr shifts. Control was the 24hr shift schedule.
- Parshuram *et al.* b: intervention was 12hr shifts. Control was the 24hr shift schedule.

Figure 3. Meta-analysis of studies for 30-day readmission



*Desai *et al.* had two intervention groups:

- Desai *et al.* 2013 a: intervention was the Q5 schedule (every fifth night overnight call).
- Desai *et al.* 2013 b: intervention was the night float schedule.

Figure 4. Meta-analysis of studies for in-training examinations

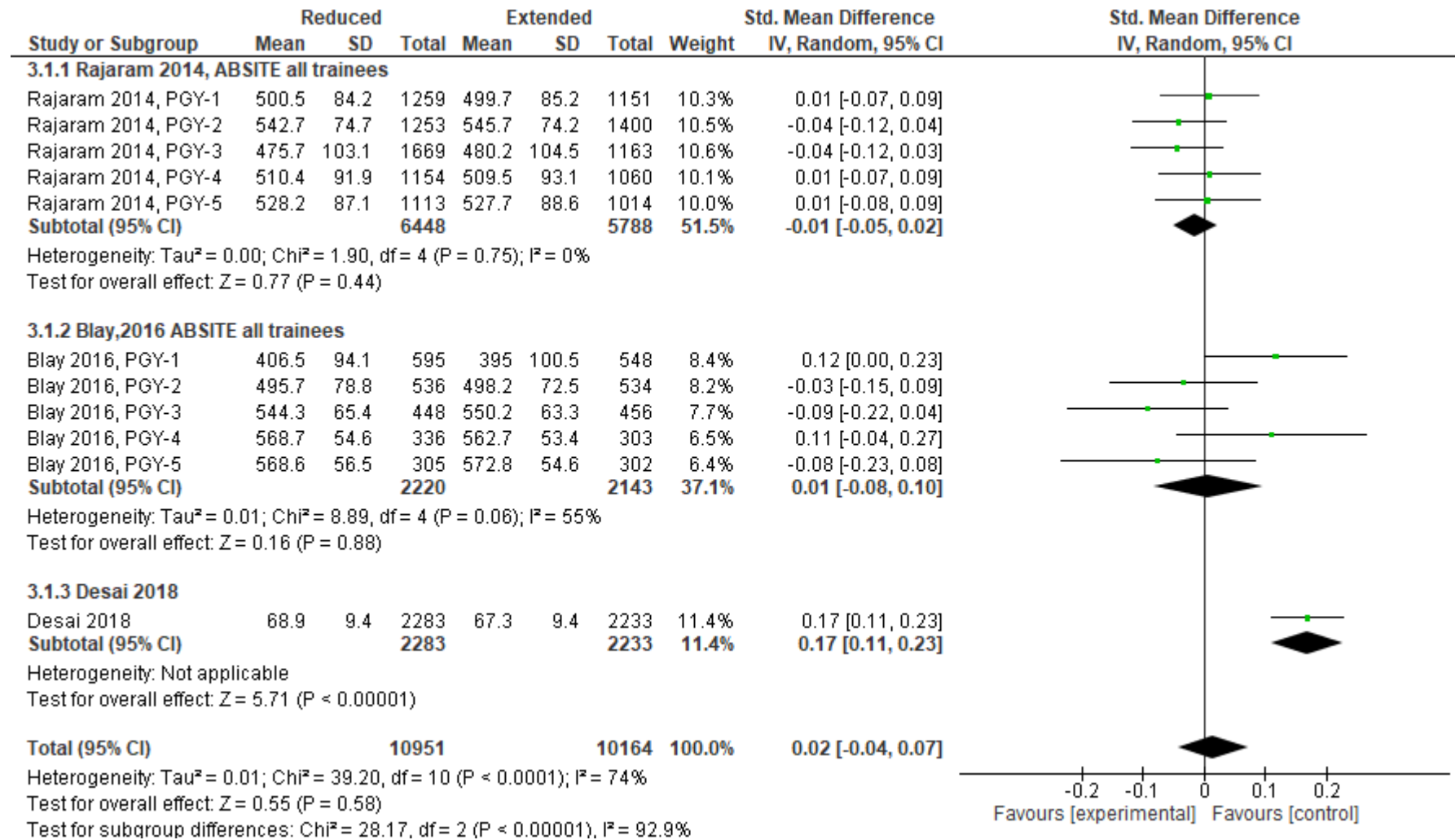


Figure 5.1. Meta-analysis for Mean Score on Maslach Burnout Inventory Depersonalization

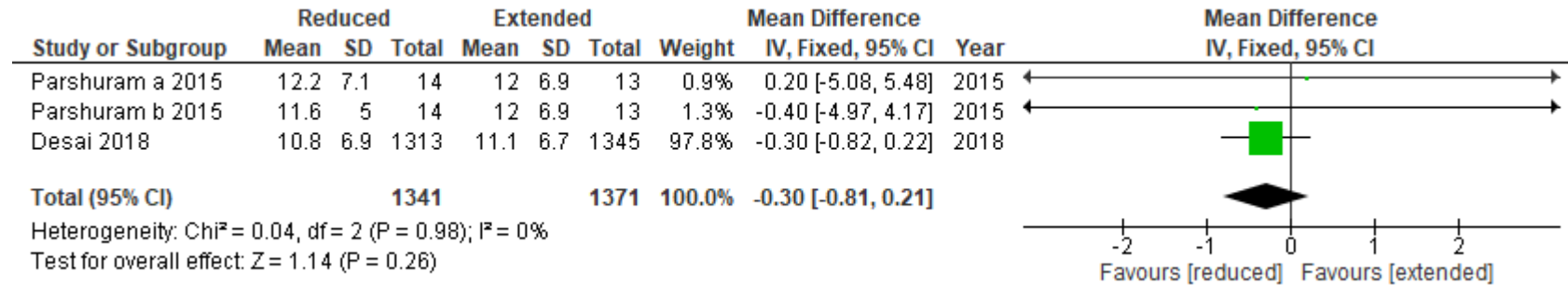


Figure 5.2. Meta-analysis for Mean Score on Maslach Burnout Inventory Emotional Exhaustion

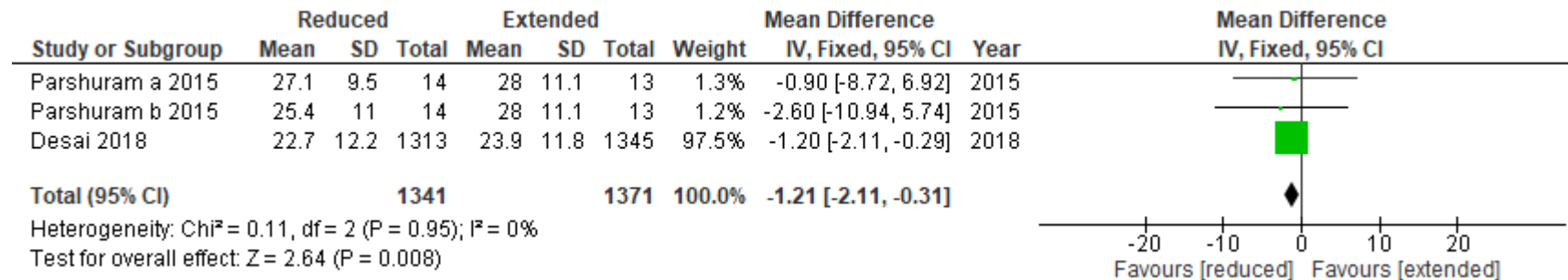


Figure 5.3. Meta-analysis for Mean Score on Maslach Burnout Inventory Personal Achievement

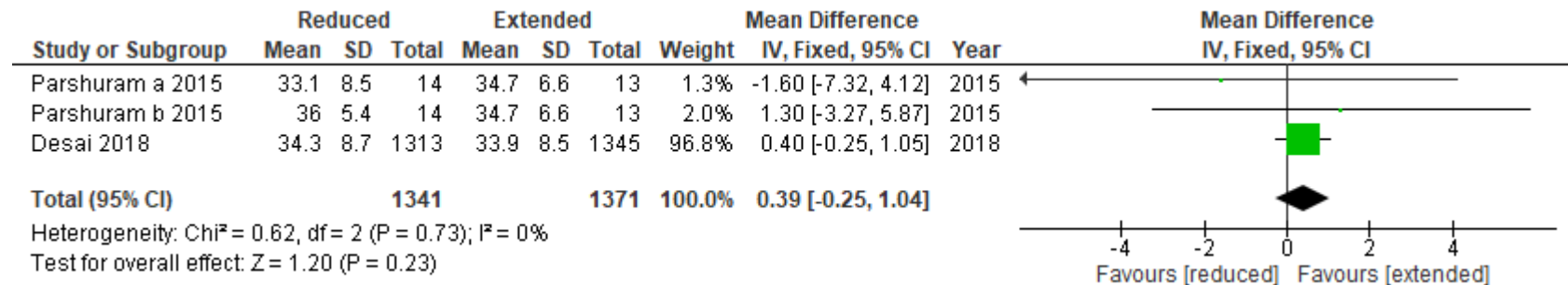


Figure 6. Meta-analysis of studies for sleep quantity

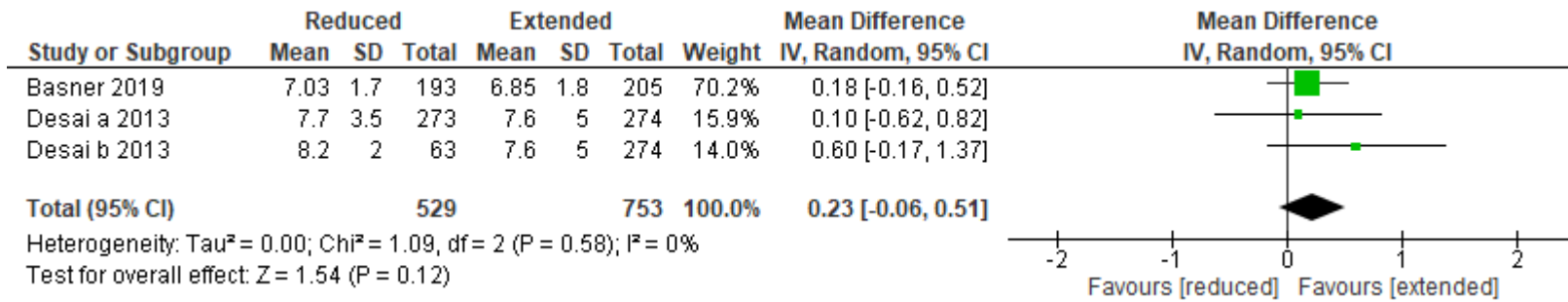


Figure 7: Risk of Bias for Randomized Controlled Trials (RCTs)

Study	Risk of bias domains						Overall
	D1	D1b	D2	D3	D4	D5	
Landrigan 2020	⊖	⊕	⊕	⊕	⊕	⊕	⊖
Silber 2019	⊕	⊕	⊕	⊕	⊕	⊕	⊕
Barger 2019	⊕	⊕	⊕	⊕	⊕	⊕	⊕
Landrigan 2004	⊕	○	⊕	⊕	⊕	⊕	⊕
Parshuram 2015	⊕	○	⊕	⊕	⊕	⊕	⊕
Bilimoria 2016	⊕	⊕	⊕	⊕	⊕	⊕	⊕
Stulberg 2016	⊕	⊕	⊕	⊕	⊕	⊖	⊖
Desai 2013	⊕	○	⊗	⊕	⊕	⊕	⊗
Basner 2019	⊕	⊕	⊕	⊕	⊕	⊕	⊕
Desai 2018	⊕	⊕	⊕	⊕	⊕	⊕	⊕
Blay 2016	⊕	⊕	⊕	⊕	⊕	⊖	⊖

Domains:
D1 : Bias arising from the randomization process.
D1b: Bias arising from the timing of identification and recruitment of individual participants in relation to timing of randomization.
D2 : Bias due to deviations from intended intervention.
D3 : Bias due to missing outcome data.
D4 : Bias in measurement of the outcome.
D5 : Bias in selection of the reported result.

Judgement
⊗ High
⊖ Some concerns
⊕ Low
○ Not applicable

Domain 2 (bias due to deviations from intended intervention): we considered that 10 studies (90.1%) had a low risk for bias. However, it is important to note that it was not feasible to blind the intervention in any of the studies, given that all trainees were aware of their rotation schedule, as well as program directors and other health personnel.

Domain 4 (bias in measurement of outcome): we considered that all 11 studies (100%) had a low risk of bias. Nonetheless, it is key to consider that except for three studies (Silber *et al.*, Stulberg *et al.* and Blay *et al.*) that involved only secondary data collection, the rest of the eight studies (72.7%) did not involve blinded data collectors. In these mentioned studies, outcome assessors were aware of the intervention received by participants (either in the form of resident self-reported outcomes or direct observation by physicians) which could represent a potential source of bias but was deemed low risk given that blinding was not feasible in this scenario and assessment of the outcome is not likely to have been influenced by knowledge of intervention received.





Domain 5 (bias in selection of the reported result): Stulberg *et al.* was labeled as having some concerns given that it was not specified whether the primary data analysis to examine the association between length of stay and flexibility in resident duty hours was outlined prior to the FIRST trial's conclusion and availability of unblinded data. In a similar manner, Blay *et al.* was considered as having some concerns given that it was uncertain if a pre-specified analysis plan for resident examination scores (the outcome evaluated in the study) was finalized before unblinded data was available for analysis.

* Out of the eleven RCTs, eight were cluster-randomized trials, for which the specific version of the tool was employed that included an additional domain of bias arising from the timing of identification and recruitment of individual participants in relation to timing of randomization (Domain 1b).

Figure 8: Risk of Bias for Non-Randomized Intervention Studies

Study	Risk of bias domains							Overall
	D1	D2	D3	D4	D5	D6	D7	
Rajaram 2014	-	+	+	?	+	+	×	×
Moeller 2016	-	+	+	+	-	-	+	-
Auger 2012	-	+	+	×	×	-	+	×
Alsohime 2021	-	+	+	+	-	-	+	-
Ming Low 2018	×	+	×	+	×	-	+	×
Zahrai 2011	-	+	+	×	+	-	-	×

Domains:
D1: Bias due to confounding.
D2: Bias due to selection of participants.
D3: Bias in classification of interventions.
D4: Bias due to deviations from intended interventions.
D5: Bias due to missing data.
D6: Bias in measurement of outcomes.
D7: Bias in selection of the reported result.

Judgement
 Serious
 Moderate
 Low
 No information

Domain 1 (bias due to confounding): we considered that all studies had at least a moderate risk of bias due to confounding because no information was provided on potential confounding factors, co-interventions and/or baseline characteristics for the participants.

Domain 5 (bias due to missing data): we considered that four studies (66.7%) had at least a moderate risk of bias since outcome data was not available for all participants and/or participants were excluded due to missing data needed for the analysis.

Domain 6 (bias in measurement of outcomes): we considered that five studies (83.3%) had a moderate risk because data was self-reported in the form of surveys, so outcome assessors were the residents themselves. The former implies that outcome assessors were aware of the intervention received by study participants so the possibility that results were influenced by this knowledge cannot be excluded. In a similar manner, given that these studies implemented surveys, there is a possibility of recall bias.

Figure 9. Funnel Plot of Comparison: 30-day mortality

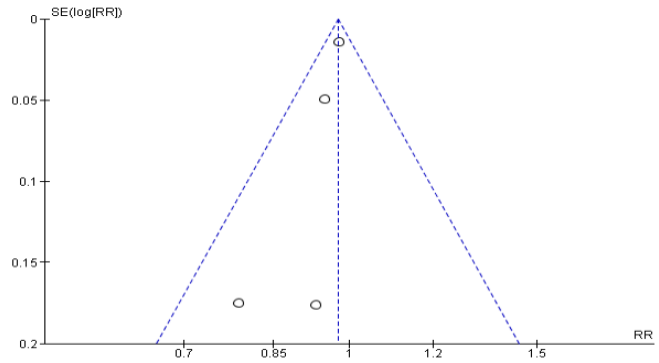


Figure 11. Funnel Plot of Comparison: In-training examinations

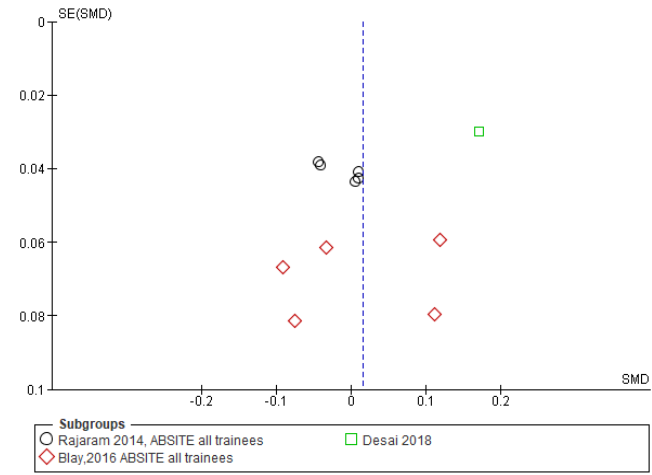


Figure 10. Funnel Plot of Comparison: 30-day readmission

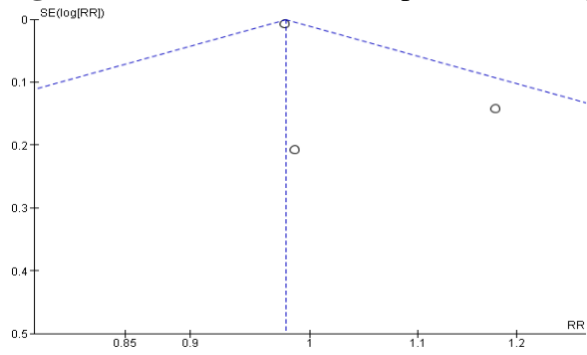


Figure 12. Funnel Plot of Comparison: Burnout, depersonalization

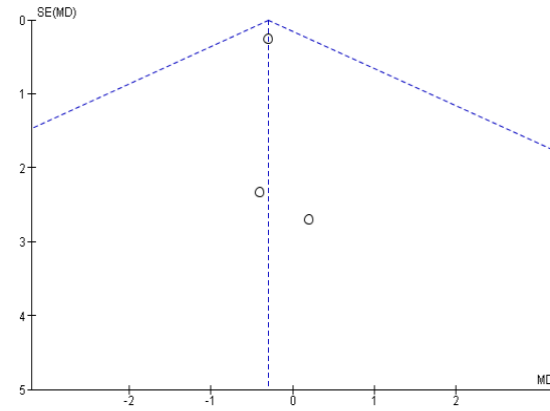


Figure 13. Funnel Plot of Comparison: Burnout, emotional exhaustion

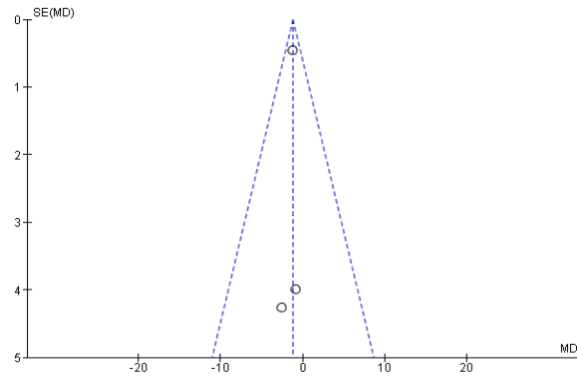


Figure 14. Funnel Plot of Comparison: Burnout, personal achievement

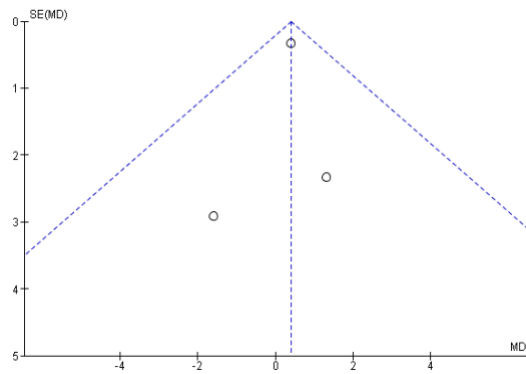
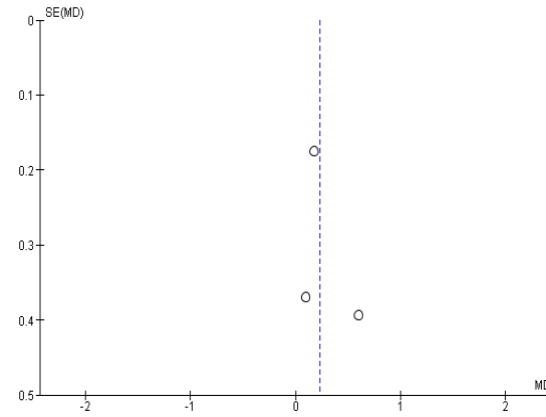


Figure 15. Funnel Plot of Comparison: Resident well-being, sleep quantity



TABLES

Table 1: Characteristics of included studies

	Author	Year	Country	Study Design	N (residents)	Medical Specialties	Intervention	Control	Patient Safety	Resident Well-being	Resident Education
1	C.P. Landrigan, <i>et al.</i> (ROSTERS) (23)	2020	USA	RCT, cluster-randomized, crossover	333	Pediatrics	Shift length reduction (\leq 16h shifts)	\geq 24h shifts	x		
2	J.H. Silber, <i>et al.</i> * (iCOMPARE) (22)	2019	USA	RCT, cluster-randomized, noninferiority	5040	Internal Medicine	“Flexible” extended shifts (>16h shifts)	“Standard” reduced shifts (\leq 16h shifts)	x		
3	L.K. Barger, <i>et al.</i> (ROSTERS) (28)]	2019	USA	RCT, cluster-randomized, crossover	302	Pediatrics	Shift length reduction (\leq 16h shifts)	\geq 24h shifts		x	x
4	C.P. Landrigan, <i>et al.</i> (29)	2004	USA	RCT	7	Internal Medicine	Shift length reduction (\leq 16h shifts)	\geq 24h shifts	x		
5	C.S. Parshuram, <i>et al.</i> (30)	2015	Canada	RCT	47	Internal Medicine, Emergency Medicine, Surgery and Anesthesia	Shift length reduction; two intervention groups (12h or 16h shifts)**	\geq 24h shifts	x	x	x
6	K.Y. Bilimoria, <i>et al.</i> * (FIRST)	2016	USA	RCT, cluster-randomized, noninferiority	4330	Surgery	“Flexible” extended shifts (>16h shifts)	“Standard” reduced shifts	x	x	x

(21)											(≤16h shifts)
7	R. Rajaram, <i>et al.</i> (31)	2014	USA	Retrospective observational study (pre and post)	12236	Surgery	Shift length reduction/ post 2011 reform (≤16h shifts)	Pre 2011 reform (≥ 24h shifts)			x
8	J.J. Stulberg, <i>et al.</i> * (FIRST) (32)	2016	USA	RCT, cluster-randomized, noninferiority	4330	Surgery	“Flexible” extended shifts (>16h shifts)	“Standard” reduced shifts (≤16h shifts)	x		
9	S.V. Desai, <i>et al.</i> (33)	2013	USA	RCT, crossover	43	Internal Medicine	Shift length reduction (≤16h shifts) and night float; two intervention groups (Q5 and NF)**	≥ 24h shifts	x	x	x
10	M. Basner, <i>et al.</i> * (iCOMPARE) (34)	2019	USA	RCT, cluster-randomized, noninferiority	398	Internal Medicine	“Flexible” extended shifts (>16h shifts)	“Standard” reduced shifts (≤16h shifts)		x	
11	S.V. Desai, <i>et al.</i> * (iCOMPARE) (35)	2018	USA	RCT, cluster-randomized, noninferiority	6313	Internal Medicine	“Flexible” extended shifts (>16h shifts)	“Standard” reduced shifts (≤16h shifts)		x	x
12	A. Moeller, <i>et al.</i> (36)	2016	Canada	Prospective interventional study (pre and post)	27	Internal Medicine	Shift length reduction and night float; two intervention groups (11h	24h shifts		x	x

							day shift and 15h night shift)			
13	K.A. Auger, <i>et al.</i> [37)	2012	USA	Prospective, interventional study	11	Pediatrics, Family medicine	Shift length reduction (<12h shifts)	30h shifts every 4 th night	x	x
14	E. Blay, <i>et al.</i> * (FIRST) (38)	2016	USA	RCT, cluster-randomized, noninferiority	4363	Surgery	“Flexible” extended shifts (>16h shifts)	“Standard” reduced shifts (≤16h shifts)		x
15	F. Alsohime, <i>et al.</i> (39)	2021	Saudi Arabia	Prospective interventional study (pre and post)	42	Pediatrics	Night float (16h shifts)	24h shifts (24h on-call system)	x	x
16	J.M.Low, <i>et al.</i> (40)	2018	Singapore	Non-randomized controlled trial	49	Internal medicine, Pediatrics, Orthopedics, Obstetrics and Gynecology, and Surgery	Night float (12h shifts)	24h shifts (24h on-call system)	x	
17	A. Zahrai, <i>et al.</i> (41)	2011	Canada	Non-randomized controlled trial	16	Orthopedics	Night float (14h shifts)	≥ 24h shifts	x	x

* In these studies, the intervention was considered as a “flexible” extended schedule whereas the control group had “standard” reduced shifts. For homogeneity, the intervention and control groups were switched in the following tables and for meta-analysis purposes, such that all intervention groups represent residents with reduced duty hours.

**These studies had two intervention groups and for analysis purposes they were labeled as:

- Parshuram *et al.* a: intervention was 16hr shifts. Control was the 24hr shift schedule.
- Parshuram *et al.* b: intervention was 12hr shifts. Control was the 24hr shift schedule.
- Desai *et al.* 2013 a: intervention was the Q5 schedule (every fifth night overnight call)
- Desai *et al.* 2013 b: intervention was the NF (night float schedule)

Table 2: Patient safety outcomes reported in included studies

<u>Study</u>	<u>Measurement scale</u>	<u>Intervention group</u> (reduced duty hours and/or night float)	<u>Control group</u> (extended duty hours)	<u>Conclusion</u>
7-day readmission or death				
Silber <i>et al.</i> (22)	Proportion of patients who presented 7-day readmission or death, obtained from Medicare claim records. Difference-in-difference analysis.	N= 70,972 Pretrial year 16.7% CI 95% [13.00%, 20.34%] N= 71,662 Trial year 16.6% CI 95% [11.78%, 21.42%] Difference = -0.1%	N= 60,757 Pretrial year 16.6% CI 95% [12.10%, 21.16%] N= 61,194 Trial year 16.9% CI 95% [10.82%, 22.95%] Difference = 0.3%	Difference in change of 0.3% 95% CI (-∞ - 1.0) [p>0.05]. Non-inferiority criterion met. “Flexible” extended schedule policies were noninferior to “standard “reduced hours policies with respect to 7-day readmission or death.
Near misses¹				
Landrigan <i>et al.</i> 2020 (23)	Only the relative risk is reported. Near misses are observed by direct observation and chart review.	N= 3,310	N= 3,267	RR of 1.42, 95% CI (1.26-1.59)*
Preventable adverse events^{2,3}				
Landrigan <i>et al.</i> 2020 (23)	Only the relative risk is reported. Preventable adverse events are evaluated by direct observation and chart review.	N= 3,310	N= 3,267	RR of 4.03 95% CI (2.94-5.53)*

Parshuram <i>et al.</i> a (30)	Number of events per 1000 patient-days was reported. Preventable adverse events were identified by daily prospective screening with a multimodal approach (direct observation and chart review).	N= 293 0 per 1000 patient-days	N= 311 0.5 per 1000 patient-days	There were no significant differences between schedules for rates of adverse events.
Parshuram <i>et al.</i> b (30)		N= 367 3.2 per 1000 patient-days	N= 311 0.5 per 1000 patient-days	There were no significant differences between schedules for rates of adverse events.
Serious medical errors⁴				
Landrigan <i>et al.</i> 2020 (23)	Number of events per 1000 patient-days and relative risk were reported. Resident-related serious medical errors were evaluated by direct observation and chart review.	N= 3,310 97.1 per 1000 patient-days at risk	N= 3,267 79.0 per 1000 patient-days at risk	RR of 1.53 95%CI (1.37-1.72)*
Landrigan <i>et al.</i> 2004 (29)	Number of events per 1000 patient-days was reported. Intern-related serious medical errors were evaluated by direct observation and chart review.	N= 227 100.1 per 1000 patient-days	N= 354 136.0 per 1000 patient-days	Interns made 35.9 percent more serious medical errors during the traditional extended schedule than during the intervention reduced schedule (p<0.001)
In-hospital complications				

Bilimoria <i>et al.</i> 2016 (21)	Proportion of patients with in-hospital complications (“any morbidity mean rate”). This data was obtained for general surgery cases from American College of Surgeons National Surgical Quality Improvement Program (ACS NSQIP).	N= 65,849 8.48% CI 96% [7.69-9.27]	N= 72,842 8.21% CI 96% [7.51-8.91]	Odds ratio (OR) for control group of 0.94 92% CI (0.84 -1.06) “Flexible” extended schedule policies were noninferior to “standard “reduced hours policies with respect to any complication.
Length of stay				
Silber <i>et al.</i> (22)	Absolute value (mean number of days) obtained from Medicare claim records. Difference-in-difference analysis.	N= 70,972 Pretrial year mean LOS (days): 5.61 CI 95% [4.55, 6.67]	N= 60,757 Pretrial year mean LOS (days): 6.07 CI 95% [4.27, 7.86]	Difference in change of 0.80% 95% CI (-∞ - 3.16) Non-inferiority criterion was not met.
		N= 71,662 Trial year mean LOS (days) 5.64 CI 95% [4.33, 6.95] Difference: 0.47%	N= 61,194 Trial year mean LOS (days) 6.14 CI 95% [4.35, 7.92] Difference: 1.27%	
Stulberg <i>et al.</i> (32)	Absolute value (mean length of stay). Patient-level data were obtained through the ACS NSQIP.	N= 12,202 Mean LOS 6.21 days (SD=5.82) median=5	N= 14,421 Mean LOS 6.03 days (SD= 5.78) median=5	Incidence rate ratio for Flexible vs Standard of 0.982; 95% CI (0.939-1.026); p=0.41. There was no statistically significant difference in overall mean LOS between study arms (p=0.74).
Desai <i>et al.</i> 2013 a (33)	Absolute value (median length of stay).	N= 300 Median LOS: 3 days IQR: 2-4	N=413 Median LOS: 3 days IQR: 2-5	Only values are reported. No commentary on this particular outcome is provided by authors.

Desai <i>et al.</i> 2013 b (33)		N=121 Median LOS: 3 days IQR: 2-6	N=413 Median LOS: 3 days IQR: 2-5	
Patient safety indicators⁵				
Silber <i>et al.</i> (22)	Proportion of patients presenting at least one patient safety indicator, obtained from Medicare claim records. Difference-in-difference analysis.	N= 70,972 Pretrial year 0.74% CI 95% [0.01%, 1.47%] N= 71,662 Trial year 0.67% CI 95% [-0.09%, 1.43%] Difference = -0.1%	N= 60,757 Pretrial year 0.96% CI 95% [0.01%, 1.90%] N= 61,194 Trial year 0.90% CI 95% [-0.01%, 1.82%] Difference = -0.1%	Difference in change of <0.1% 95%CI (-∞ - 0.2) Non-inferiority criterion was met. “Flexible” extended schedule policies were noninferior to “standard “reduced hours policies with respect to patient safety indicators.

*Intervention schedules that eliminated extended shifts had a higher risk of near misses, preventable adverse events, and medical errors. However, secondary analysis showed that when adjusted for the number of patients per resident as a possible confounding factor, intervention schedules were no longer associated with an increase in these events.

¹Near miss is defined as "an error in care that has substantial potential to cause harm but does not, either because it is intercepted or because it unexpectedly causes no apparent harm despite reaching the patient"

²In Landrigan *et al.* 2020, preventable adverse events were defined as an "injury caused by an error in medical management".

³In Parshuram *et al.*, preventable adverse events were defined as “adverse events that could have been avoided given current knowledge and standards of care”.

⁴In Landrigan *et al.* 2020 and 2004, resident-related serious medical errors were defined as ones "that cause harm or have substantial potential to cause harm (i.e., the sum of preventable adverse events plus near misses)".

⁵In Silber *et al.* patient safety indicators were defined according to AHRQ criteria. These include rates of pressure ulcers, iatrogenic pneumothorax, bloodstream infection from a central venous catheter, hip fracture, hemorrhage, or hematoma, physiologic or metabolic derangement, respiratory failure, pulmonary embolism or deep-vein thrombosis, sepsis, and accidental puncture or laceration.

- N represents the total number of patients evaluated in the study

Table 3: Resident education outcomes reported in included studies

<u>Study</u>	<u>Measurement scale</u>	<u>Intervention group</u> (reduced duty hours and/or night float)	<u>Control group</u> (extended duty hours)	<u>Study's conclusion</u>
Board examination scores				
Rajaram <i>et al.</i> (31)	Proportion of residents who passed written boards examination	Post-reform group 2012 cohort N=1082 88.6% CI 95%: 86.7-90.5%	Pre-reform group 2010 cohort N=1063 83.1% CI 95%: 80.8-85.3%	Although the difference was significant (p<0.001) when comparing the intervention group (2012 and 2013 examinees) with the control group (2010 and 2011), the authors suggest that this was due to abnormally poor results in the 2010 cohort and, when this year was excluded from the analysis, the difference in written board passing rates between intervention groups was no longer significant (p=0.41).
		2013 cohort N=1097 88.1% CI 95%: 86.1-90.0%	2011 cohort N=1038 87.5% CI 95%: 85.5-89.5%	
	Proportion of residents who passed oral boards examination	Post-reform group 2012 cohort N=1046 82.5% CI 95%: 80.2-84.8%	Pre-reform group 2010 cohort N=906 81.7% CI 95%: 79.2-84.2%	No significant change was detected in oral boards passing rate (p=0.21).
		2013 cohort N=1154 80.9% CI 95%: 78.7-83.2%	2011 cohort N=1086 74.7% CI 95%: 72.1-77.3%	
Blay <i>et al.</i> (38)	Proportion of residents who passed written boards examination	N= 293 90.4%	N= 283 90.5%	No significant change was detected in written boards or oral boards passing rate (p=0.99 and p=0.24, respectively)
	Proportion of residents who passed oral boards examination	N= 272 86.3%	N= 261 88.6%	

Rounds/teaching conference attendance				
Bilimoria <i>et al.</i> 2016 (21)	Proportion of residents who reported negative effect of hours on attendance at educational conferences on survey	N=1886 22.9%	N=1780 12.2%	OR for flexible group (control) of 0.47 95%CI (0.36-0.62) Residents in extended “flexible” schedules were significantly less likely to perceive a negative effect of duty-hour policies on conference attendance (p<0.001).
Desai <i>et al.</i> 2018 (35)	Proportion of all trainees reporting a negative effect of duty hours on Ability to attend required educational conferences on survey	N=1411 12.7%	N=1435 20.1%	OR for flexible group (control) of 1.72 95%CI (1.24-2.38) Residents in extended “flexible” schedules were more likely to perceive a negative effect of duty-hour policies on conference attendance.
Auger <i>et al.</i> (37)	Proportion of residents who rated Ability to attend didactic conferences (unchanged/improved)	N=5 20%	N=6 66.7%	Residents in the intervention group rated ability to attend didactic conferences worse compared to the control group (p>0.05).
Zahrai <i>et al.</i> (41)	Proportion of residents who attended ≥4 conferences	N=7 71.4%	N=5 60.0%	Only values are reported. No data analysis or commentary on this particular outcome is provided by authors.
Independent/self-directed learning				
Bilimoria <i>et al.</i> 2016 (21)	Proportion of residents who reported negative effect of hours on participation in research on survey	N= 1888 9.1%	N= 1780 21.0%	OR for flexible group (control) of 2.81 95%CI (2.12-3.73) Residents in extended “flexible” schedules were more likely to perceive a negative effect of duty-hour policies on research participation (p<0.001).
Desai <i>et al.</i> 2018 (35)	Proportion of all trainees reporting negative effect of hours on ability to participate in research on survey	N=1411 11.1%	N=1435 21.4%	OR for flexible group (control) of 2.14 95%CI (1.60-2.87) Residents in extended “flexible” schedules were more likely to perceive a negative effect of duty-hour policies on research participation.
Auger <i>et al.</i> (37)	Proportion of residents who rated Ability to reflect on clinical concepts (unchanged/improved)	N=5 40%	N=6 66.7%	Residents in the intervention group rated ability to reflect on clinical concepts worse compared to the control group (p>0.05).

Alsohime <i>et al.</i> (39)	Residents' mean score on Likert-scale of perception of hours restricting time available for research	N=42 Mean: 2.52 SD: 1.27	N=42 Mean: 4.10 SD: 0.98	Residents considered that extended schedules were more restrictive in terms of time allotment for research than night float system (p<0.001)
Zahrai <i>et al.</i> (41)	Proportion of residents who spent average ≥3 hours studying/reading	N=7 57.1%	N=5 0%	Only values are reported. No data analysis or commentary on this particular outcome is provided by authors.
Missed OR				
Bilimoria <i>et al.</i> 2016 (21)	Proportion of residents who reported missing an operation during past month owing to duty-hour regulations on survey	N= 1944 42.0%	N= 1821 29.9%	OR for flexible group (control) of 0.56 95%CI (0.45-0.69) Residents in extended "flexible" schedules were significantly less likely to miss an operation owing to duty-hour policies (p<0.001).
Resident autonomy				
Bilimoria <i>et al.</i> 2016 (21)	Proportion of residents who reported negative effect of duty hours on resident autonomy on survey	N= 1888 35.1%	N=1782 13.0%	OR for flexible group (control) of 0.26 95%CI (0.20-0.34) Residents in extended "flexible" schedules were significantly less likely to report a negative effect of duty hours on autonomy (p<0.001).
Desai <i>et al.</i> 2018 (35)	Proportion of all trainees reporting negative effect of duty hours on autonomy on survey	N= 1411 8.8%	N= 1435 6.0%	OR for flexible group (control) of 0.69 95%CI (0.46-1.04) Residents in extended "flexible" schedules were significantly less likely to report a negative effect of duty hours on autonomy (p<0.001).
Moeller <i>et al.</i> (36)	Resident's mean score on Likert-scale of perceiving hours allowing staff physician supervision	N=23 Mean: 3.38 SD: 0.59	N=23 Mean: 3.23 SD: 0.75	No significant change was detected in perception that resident duty hour reform allowed attending supervision (p=0.37).
Alsohime <i>et al.</i> (39)	Residents' mean score on Likert-scale of perception of clinical skills being observed by an attending	N=42 Mean: 3.10 SD: 1.16	N=42 Mean: 2.33 SD: 1.00	Residents considered that night float promoted more observation by attendings than extended schedules (P<0.001)

Continuity of care¹				
<i>Bilimoria et al.</i> 2016 (21)	Proportion of residents who perceived a negative effect on continuity of care	N= 1892 55.7%	N= 1786 19.0%	OR for flexible group (control) of 0.16 95%CI (0.12–0.21) Residents in extended “flexible” schedules were significantly less likely to report a negative effect of duty hours on continuity of care (p<0.001).
	Proportion of residents who reported dissatisfaction with quality and ease of handoffs and transitions in care	N= 1873 10.1%	N= 1766 7.0%	OR for flexible group (control) of 0.69 95%CI (0.52–0.92) Residents in extended “flexible” schedules were significantly less likely to report dissatisfaction with transitions in care (p=0.01).
<i>Desai et al.</i> 2018 (35)	Proportion of residents who perceived a negative effect on continuity of care	N= 1411 33.6%	N= 1435 14.9%	OR for flexible group (control) of 0.37 95%CI (0.25–0.55) Residents in extended “flexible” schedules were less likely to report a negative effect of duty hours on continuity of care.
	Proportion of residents who reported dissatisfaction with quality and ease of handoffs and transitions in care	N= 1411 8.5%	N= 1435 5.8%	OR for flexible group (control) of 0.69 95%CI (0.48–1.01) No significant change was detected when comparing the intervention group with the control group.
<i>Moeller et al.</i> (36)	Mean values represent location on 5-point Likert scale. Surveyed residents believed that the change in duty hours allows continuity of patient care.	N=23 Mean: 4.22 SD: 0.6	N=23 Mean: 4.13 SD: 0.87	No significant change was detected in perception that resident duty hour reform allowed for continuity of patient care (p=0.60).
<i>Alsohime et al.</i> (39)	Mean values represent location on 5-point Likert scale. Residents reported that duty-hours change promotes continuity of patient care.	N=42 Mean: 4.19 SD: 0.77	N=42 Mean: 2.93 SD: 1.24	Residents considered that night float promoted more continuity of patient care than the extended schedules (P<0.001)
Bedside teaching				

Auger <i>et al.</i> (37)	Proportion of residents who rated Bedside teaching (unchanged/improved)	N=5 60%	N=6 66.7%	Residents in the intervention group rated bedside teaching worse compared to the control group. (p>0.05).
Zahrai <i>et al.</i> (41)	Proportion of residents who reported ≥6 attending teaching interactions lasting >5 mins	N=7 57.1%	N=5 40.0%	Only values are reported. No data analysis or commentary on this particular outcome is provided by authors.
Teaching satisfaction				
Barger <i>et al.</i> (28)	Resident's mean score on scale (range 0-60 where higher scores represent a more positive experience) of perceiving opportunity to obtain skills and knowledge	N=183 Mean: 44.8 SD: 8.3	N=167 Mean: 45.2 SD: 8.2	No significant change was detected between groups (p=0.50).
Bilimoria <i>et al.</i> 2016 (21)	Proportion of residents who reported perception of negative effect of hours on clinical skills acquisition on survey	N= 1888 36.4%	N= 1777 13.1%	OR for flexible group (control) of 0.24 95% CI (0.19–0.31) Residents in extended “flexible” schedules were significantly less likely to report a negative effect of duty hours on clinical skills acquisition (p<0.001).
Desai <i>et al.</i> 2018 (35)	Proportion of all trainees reporting negative effect of duty hours on Ability to acquire clinical skills on survey	N= 1411 13.9%	N= 1435 9.9%	OR for flexible group (control) of 0.81 95% CI (0.55–1.19) No significant change was detected between groups.
Moeller <i>et al.</i> (36)	Resident's mean score on Likert-scale of perceiving hours allowing clinical skills expertise	N=23 Mean: 3.91 SD: 0.52	N=23 Mean: 3.39 SD: 0.72	Residents perceived improvement in clinical skills expertise with reduced duty hours (p=0.0004).
Alsohime <i>et al.</i> (39)	Residents’ mean score on Likert-scale of perception of being able to manage complex medical patients appropriately	N=42 Mean: 3.98 SD: 0.87	N=42 Mean: 3.0 SD: 1.08	Residents in the night float group felt more confident to manage complex patient than those assigned to an extended schedule (p<0.001).

Zahrai <i>et al.</i> (41)	Proportion of residents who perceive that current rotation provides better opportunity to improve clinical decision-making and diagnostic skills than previous rotations (agree/strongly agree)	N=7 57.1%	N=5 40%	Only values are reported. No data analysis or commentary on this particular outcome is provided by authors.
Overall educational experience				
Barger <i>et al.</i> (28)	Proportion of residents who rated the educational experience as poor/fair	N=183 37.7%	N=167 17.4%	Residents in the intervention group were more likely to rate the education experience as poor/fair (p= 0.0001).
Bilimoria <i>et al.</i> 2016 (21)	Proportion of residents who reported dissatisfaction with overall quality of resident education on survey	N= 1874 10.7%	N= 1768 11.0%	OR for flexible group (control) of 1.08 95%CI (0.77–1.52) No significant change was detected between groups (p=0.64).
Desai <i>et al.</i> 2013 a (33)	Resident's mean score on Likert-scale of satisfaction with education	N=40 Mean: 4.1 CI 95%: 3.9-4.3	N=44 Mean: 4.3 CI 95%: 4.1-4.4	No significant change was detected between Q5 intervention and control group (p=0.27).
Desai <i>et al.</i> 2013 b (33)		N=12 Mean: 3.9 CI 95%: 3.5-4.3	N=44 Mean: 4.3 CI 95%: 4.1-4.4	Residents' satisfaction with education was higher in the control group than in the night float intervention group (p=0.04)
Desai <i>et al.</i> 2018 (35)	Proportion of all trainees reporting dissatisfaction with overall quality of resident education on survey	N= 1411 7.7%	N= 1435 13.4%	OR for flexible group (control) of 1.75 95%CI (1.11–2.75) Residents in extended “flexible” schedules were more likely to report dissatisfaction with education.
Moeller <i>et al.</i> (36)	Resident's mean score on Likert-scale of perceiving hours allowing successful learning	N=23 Mean: 4.00 SD: 0.56	N=23 Mean: 3.52 SD: 0.57	Residents perceived more successful learning with reduced duty hours (p=0.003)
Auger <i>et al.</i> (37)	Proportion of residents who rated Quality of education (very good/excellent)	N=5 20%	N=6 66.7%	Residents in the intervention group rated quality of education worse compared to the control group. (p>0.05).

Alsohime <i>et al.</i> (39)	Residents' mean score on Likert-scale of perception of educational experience on call being satisfying	N=42 Mean: 3.95 SD: 0.91	N=42 Mean: 2.83 SD: 0.99	Residents considered that night float (intervention) allowed for a more satisfying educational experience than extended schedules (p<0.001)
Zahrai <i>et al.</i> (41)	Proportion of residents who perceive that current rotation provides better overall educational experience than previous rotations (agree/strongly agree)	N=7 57.1%	N=5 40.0%	Only values are reported. No data analysis or commentary on this particular outcome is provided by authors.

¹ For the continuity of care outcome, two additional studies were included that are not listed in this table, including Parshuram *et al.* and Desai *et al.* 2013. Parshuram *et al.* reported the number of mean days that a resident looked after a patient and found no significant difference between intervention groups (p=0.1). Desai *et al.* 2013 reported continuity of care as measured by the minimal number of hands-off and by the number of interns per patient; it found that both parameters increased in the intervention schedule (Q5 and night float schedules). This data was omitted from the table because this outcome domain focused on the residents' subjective perceptions on education rather than objective measurements.

Table 4: Resident well-being outcomes reported in included studies

<u>Study</u>	<u>Measurement scale</u>	<u>Intervention group</u> (reduced duty hours and/or night float)	<u>Control group</u> (extended duty hours)	<u>Study's conclusion</u>
Burnout (studies not eligible for meta-analysis)				
Moeller <i>et al.</i> (36)	Resident's mean score on Likert-scale of perceiving expenditure of emotional labor	N=23 Mean: 1.75 SD: 0.56	N=23 Mean: 1.84 SD: 0.74	No significant difference was detected in perception that resident duty hour reform caused expenditure of emotional labor (p=0.58).
Auger <i>et al.</i> (37)	Proportion of residents who rated Likelihood of burnout (never/not very likely/somewhat likely)	N=5 20%	N=6 66.7%	Residents in the intervention group rated burnout as worse compared to the control group(p>0.05).
Ming Low <i>et al.</i> (40)	Proportion of residents with High burnout score on ProQOL scale ¹	N=16 31.1%	N=10 10%	More night float residents (intervention group) reported a higher burnout score on the ProQOL scale (p=0.211).
Sleep quantity (studies not eligible for meta-analysis)				
Barger <i>et al.</i> (28)	Average amount of sleep in hours per week per resident Measured through actigraphy	N= 162 residents Mean: 52.9 SD: 6.0	N= 134 residents Mean: 49.1 SD: 5.8	Residents in the intervention group reported significantly more sleep than those in the control group (p<0.0001)
Auger <i>et al.</i> (37)	Average amount in hours of intern sleep reported per 24 hours	N=5 Mean: 7.5	N=6 Mean: 7.3	No significant difference was detected in the total sleep time (p=0.63).
Ming Low <i>et al.</i> (40)	Amount of sleep logged in minutes	N= 21 logs Median: 361 Range: 149-630	N= 37 logs Median: 380 Range: 175-484	No significant difference was detected in amount of sleep (p= 0.369).
Job satisfaction				
Barger <i>et al.</i> (28)	Proportion of residents who rated quality of work experience as poor/fair on survey	N= 183 29.8%	N= 167 11.4%	Residents in the intervention group were significantly more likely to report a negative work experience (p=0.0001).
Bilimoria <i>et al.</i> 2016 (21)	Proportion of residents who had perception of negative	N= 1888 13.9%	N= 1782 12.7%	OR for flexible group (control) of 0.94 95% CI (0.73-1.23). No significant difference was detected in

	effect of duty hours on job satisfaction on survey			perception of negative effects of duty hours on job satisfaction (p=0.43).
Desai <i>et al.</i> 2018 (35)	Proportion of all trainees who had perception of negative effect of duty hours on job satisfaction on survey	N= 1411 7.5%	N= 1435 20.4%	OR for flexible group (control) of 3.17 95%CI (2.30-4.37). Residents in extended “flexible” schedules were more likely to perceive a negative effect of duty-hour on job satisfaction.
Auger <i>et al.</i> (37)	Proportion of residents who rated job satisfaction as very good/excellent	N= 5 20%	N= 6 66.6%	Residents in the intervention group rated job satisfaction as worse compared to the control group(p>0.05).
Working hours satisfaction				
Bilimoria <i>et al.</i> 2016 (21)	Proportion of residents who reported dissatisfaction with work hours and scheduling on survey	N= 1874 12.6%	N= 1767 12.1%	OR for flexible group (control) of 0.95 95%CI (0.71-1.27). No significant difference was detected in dissatisfaction with work hours and scheduling. (p= 0.76).
Desai <i>et al.</i> 2018 (35)	Proportion of all trainees reporting dissatisfaction with work hours and scheduling on survey	N=1411 11.3%	N=1435 20.0%	OR for flexible group (control) of 1.95 95%CI (1.41-2.70). Residents in extended “flexible” schedules were more likely to report dissatisfaction with work hours and scheduling.
Moeller <i>et al.</i> (36)	Resident's mean score on Likert-scale of perceiving hours allow work efficiency	N=23 Mean: 4.14 SD: 0.33	N=23 Mean: 3.72 SD: 0.75	Residents perceived improvement in work efficiency with reduced duty hours. (p=0.001)
Percutaneous/attentional injuries				
Alshime <i>et al.</i> (39)	Resident's mean score on Likert-scale of perception of increased potential for workplace harm, such as needle-stick injuries	N=42 Mean: 2.21 SD: 1.22	N=42 Mean: 4.10 SD: 1.10	Residents perceived that extended schedules had more potential for workplace harm compared with the night float schedule (p<0.001)
Sleepiness/Fatigue				
Parshuram <i>et al.</i> a (30)	Residents’ mean score on Stanford sleepiness scale ² in the daytime	N= 350 measurements (15 residents) Mean: 2.61 SD: 1.17	N= 468 measurements (15 residents) Mean: 2.33 SD: 1.20	No significant difference was detected in daytime sleepiness between the three intervention groups (p=0.3).

Parshuram <i>et al.</i> b (30)		N= 468 measurements (17 residents) Mean: 2.30 SD: 0.99	N= 468 measurements (15 residents) Mean: 2.33 SD: 1.20	
Bilimoria <i>et al.</i> 2016 (21)	Proportion of residents who report that fatigue always or often affects personal safety	N= 1878 9.3%	N= 1774 10.6%	OR for flexible group (control) of 1.15 95%CI (0.91-1.47). No significant difference was detected in reporting that fatigue affected personal safety (p= 0.26).
Basner <i>et al.</i> (34)	Mean score on the Karolinska Sleepiness scale ³ across shifts	N=193 Mean: 4.7 CI 95%: 4.6-4.9	N=205 Mean: 4.8 CI 95%: 4.7-5.0	Between-group difference = 0.12 points One-sided upper limit 95%CI 0.31 points Non-inferiority criterion met (p<0.001) “Flexible” extended schedule policies were noninferior to “standard “reduced hours policies with respect to daytime sleepiness.
Desai <i>et al.</i> 2018 (35)	Proportion of all trainees perceiving that their fatigue Almost always or often affected their personal safety	N=1411 10.6%	N=1435 14.0%	OR for flexible group (control) of 1.40 95%CI (0.99-2.00) No significant difference was detected in perception that fatigue affected personal safety.
Auger <i>et al.</i> (37)	Proportion of residents who rated Making mistakes due to fatigue (somewhat likely/ very likely/ always)	N=5 20%	N=6 66.7%	Residents in the intervention group rated that they were less likely to make mistakes due to fatigue compared to the control group(p>0.05).
Alsohime <i>et al.</i> (39)	Residents’ mean score on Likert-scale of perception of hours contributing to overall fatigue level	N=42 Mean: 2.33 SD: 1.26	N=42 Mean: 4.74 SD: 0.59	Residents considered that extended schedules contributed significantly more to overall fatigue than night float (p<0.001)
Ming Low <i>et al.</i> (40)	Proportion of residents with Epworth Sleepiness Scale (ESS) ⁴ score >10	N=13 53.8%	N=13 15.4%	No significant difference was detected in daytime sleepiness between the intervention and control groups (p=0.103).

Safety while driving

Alshime <i>et al.</i> (39)	Residents' mean score on Likert-scale of perception of hours impairing safety while driving home after an on-call	N=42 Mean: 2.57 SD: 1.19	N=42 Mean: 4.45 SD: 0.86	Residents perceived that extended schedules led to significantly greater impairment of safety while driving compared to night float ($p<0.001$).
Time for family/friends				
Bilimoria <i>et al.</i> 2016 (21)	Proportion of residents who reported negative effect of duty hours on time with family and friends on survey	N= 1888 8.9%	N=1779 24.8%	OR for flexible group (control) of 3.66 95%CI (2.70-4.97). Residents in extended "flexible" schedules were more likely to report a negative effect of duty hours on time with family and friends ($p<0.001$).
Desai <i>et al.</i> 2018 (35)	Proportion of all trainees reporting a negative effect of duty hours on time with family and friends on survey	N= 1411 7.1%	N= 1435 28.9%	OR for flexible group (control) of 5.07 95%CI (3.57-7.20). Residents in extended "flexible" schedules were more likely to report a negative effect of duty hours on time with family and friends.
Auger <i>et al.</i> (37)	Proportion of residents who rated Work-life balance (good/very good/excellent)	N=5 20%	N=6 66.7%	Residents in the intervention group rated work-life balance as worse compared to the control group ($p>0.05$).
Alshime <i>et al.</i> (39)	Residents' mean score on Likert-scale of perception of hours providing opportunities for spending time with family	N=42 Mean: 4.26 SD: 1.08	N=42 Mean: 2.00 SD: 1.15	Residents considered that night float (intervention group) was significantly more family friendly than extended schedules ($p<0.001$)
Time for hobbies				
Bilimoria <i>et al.</i> 2016 (21)	Proportion of residents who reported dissatisfaction with time for rest on survey	N= 1875 14.9%	N= 1768 18.6%	OR for flexible group (control) of 1.41 95%CI (1.06-1.89). Residents in extended "flexible" schedules were significantly more likely to be dissatisfied with time for rest ($p=0.02$)
	Proportion of residents who reported negative effect of duty hours on time for extracurricular activities on survey	N= 1886 9.1%	N= 1779 25.7%	OR for flexible group (control) of 3.81 95%CI (2.84-5.11). Residents in extended "flexible" schedules were more likely to report a negative effect of duty hours on time for extracurricular activities ($p<0.001$).
Desai <i>et al.</i> 2018 (35)	Proportion of all trainees reporting dissatisfaction with time for rest on survey	N= 1411 17.4%	N= 1435 29.8%	OR for flexible group (control) of 1.96 CI 95% (1.44-2.66). Residents in extended "flexible" schedules were

				significantly more likely to be dissatisfied with time for rest.
	Proportion of all trainees reporting a negative effect of duty hours on time for hobbies and outside interests on survey	N= 1411 7.5%	N= 1435 27.7%	OR for flexible group (control) of 4.32 95%CI (2.98–6.27). Residents in extended “flexible” schedules were more likely to report a negative effect of duty hours on time for extracurricular activities.
Auger <i>et al.</i> (37)	Proportion of residents who rated Ability to take breaks (rarely/never)	N=5 100%	N=6 66.7%	Residents in the intervention group had a decreased ability to take breaks compared to the control group (p>0.05).
Alsohime <i>et al.</i> (39)	Residents’ mean score on Likert-scale of perception of hours allowing free time to accomplish non-work-related errands	N=42 Mean: 4.07 SD: 1.20	N=42 Mean: 2.57 SD: 1.33	Residents considered that night float (intervention group) allowed significantly more free time for errands than extended schedules (p<0.001).
General health				
Bilimoria <i>et al.</i> 2016 (21)	Proportion of residents who reported negative effect of duty hours on health on survey	N= 1883 6.8%	N= 1778 18.3%	OR for flexible group (control) of 3.22 95%CI (2.37–4.36). Residents in extended “flexible” schedules were more likely to report a negative effect of duty hours on health (p<0.001).
Desai <i>et al.</i> 2018 (35)	Proportion of all trainees reporting a negative effect of duty hours on health on survey	N=1411 6.7%	N=1435 26.1%	OR for flexible group (control) of 4.60 95%CI (3.16–6.69). Residents in extended “flexible” schedules were more likely to report a negative effect of duty hours on health.
Moeller <i>et al.</i> (36)	Resident's mean score on Likert-scale of perceiving hours allowing general wellness	N=23 Mean: 3.06 SD: 0.71	N=23 Mean: 2.70 SD: 0.64	Residents perceived improvement in general wellness with reduced duty hours. (p=0.04).
Alsohime <i>et al.</i> (39)	Residents’ mean score on Likert-scale of perception of hours adversely affecting health	N=42 Mean: 2.45 SD: 1.27	N=42 Mean: 4.48 SD: 0.92	Residents perceived that extended schedules had a greater adverse effect on health than night float (p<0.001).
Zahrai <i>et al.</i> (41)	Residents' mean score on General health on SF-36	N=9 Mean: 56.43 SD: 24.89	N=7 Mean: 84.20 SD: 16.50	No significant difference was detected in the general health domain after controlling for between-group differences at baseline (p=0.41).

general quality-of-life questionnaire				
Bodily pain				
Parshuram <i>et al.</i> a (30)	Frequency of somatic symptoms of at least moderate severity	N= 114 assessments Mean: 0.38 SD: 0.75	N= 138 assessments Mean: 1.15 SD: 1.71	Residents in the extended schedules (control) reported a higher frequency of symptoms of at least moderate severity (p = 0.04).
Parshuram <i>et al.</i> b (30)		N= 135 assessments Mean: 0.28 SD: 0.73	N= 138 assessments Mean: 1.15 SD: 1.71	
Zahrai <i>et al.</i> (41)	Residents' mean score on Bodily pain on SF-36 general quality-of-life questionnaire	N=9 Mean: 61.71 SD: 20.69	N=7 Mean: 87.20 SD: 20.86	Residents in the night float (intervention group) had significantly worse bodily pain after controlling for between-group differences at baseline (p=0.032)
Overall well-being and mental health				
Barger <i>et al.</i> (28)	Resident's mean score on scale ⁶ of perceiving hours negatively affected day-to-day activities	N= 183 assessments Mean: 8.8 SD: 3.8	N= 167 assessments Mean: 8.4 SD: 4.1	No significant difference was detected in perception of negative effect of duty hours on day-to-day activities (p =0.27).
Bilimoria <i>et al.</i> 2016 (21)	Proportion of residents who reported dissatisfaction with overall well-being on survey	N= 1876 12.0%	N= 1769 14.9%	OR for flexible group (control) of 1.31 95%CI (0.99–1.74). No significant difference was detected in dissatisfaction with overall well-being (p=0.10).
Desai <i>et al.</i> 2018 (35)	Proportion of all trainees reporting dissatisfaction with overall well-being on survey	N= 1411 14.6%	N= 1435 26.3%	OR for flexible group (control) of 2.04 95%CI (1.48–2.81). Residents in extended “flexible” schedules were more likely to report dissatisfaction with overall well-being.
Moeller <i>et al.</i> (36)	Resident's mean score on Likert-scale of perceiving hours allowing healthy relationships	N=23 Mean: 3.13 SD: 0.81	N=23 Mean: 2.70 SD: 0.88	No significant difference was detected in perception that resident duty hour reform allowed for healthy relationships (p=0.09).
Alsohime <i>et al.</i> (39)	Residents' mean score on Likert-scale of perception of hours allowing healthy interpersonal relationships	N=42 Mean: 4.17 SD: 0.88	N=42 Mean: 2.55 SD: 1.06	Residents considered that night float (intervention group) allowed for more healthy relationships than extended schedules (p<0.001).

Zahrai <i>et al.</i> (41)	Residents' mean score on Mental health on SF-36 general quality-of-life questionnaire	N=9 Mean: 52.00 SD: 15.49	N=7 Mean: 60.80 SD: 11.45	No significant difference was detected in the mental health domain after controlling for between-group differences at baseline (p=0.72).
------------------------------	--	---------------------------------	---------------------------------	--

¹ In the Professional Quality of Life (ProQOL) scale, burnout scores ≥ 57 (inefficiency and feeling overwhelmed) were defined as high.

² Stanford Sleepiness Scale (SSS) is a 7-point scale that ranges from 1=wide awake to 7=no longer fighting sleep

³ Karolinska Sleepiness scale is a 9-point scale that ranges from 1=extremely alert to 9=extremely sleepy. For data analysis, a non-inferiority margin of 1 point on the 9-point scale was determined by the study.

⁴ Epworth Sleepiness Scale (ESS) is a scale where a score of > 10 depicts increased fatigue.

⁵ Short Form-36 (SF-36) general quality-of-life questionnaire measures 8 health concepts: physical function, physical role limitations, bodily pain, social functioning, mental health, emotional role limitations, vitality/energy, and general health. Lower scores mean poorer health, worse bodily pain, worse mental health, etc.

⁶ Scale ranging from 0-25 where higher scores represent a more negative experience.