ASERNIP-S

Australian Safety and Efficacy Register of New Interventional Procedures - Surgical

Systematic Review

The Effect of Fatigue on Surgeon Performance and Surgical Outcomes

ASERNIP-S REPORT NO. 68

August 2009

Australian Government
Department of Health and Ageing

Australian Safety & Efficacy Register of New Interventional Procedures – Surgical

The Royal Australasian College of Surgeons

The Effect of Fatigue on Surgeon Performance and Surgical Outcomes

ISBN 978-0-9806299-5-8

Published August 2009

This report should be cited in the following manner:

Sturm L, et al. The Effect of Fatigue on Surgeon Performance and Surgical Outcomes. ASERNIP-S Report No. 68. Adelaide, South Australia: ASERNIP-S, August 2009.

Copies of these reports can be obtained from:

ASERNIP-S

PO Box 553,

Stepney, SA 5069

AUSTRALIA

Ph: 61-8-8363 7513

Fax: 61-8-8362 2077

E-Mail: asernips@surgeons.org

http://www.surgeons.org/asernip-s

The Systematic Review of

The Effect of Fatigue on Surgeon Performance and Surgical Outcomes

Was ratified by the ASERNIP-S Advisory Committee on

June 1 2009

Approved by the Research Audit and Academic Surgery Board on

June 10 2009

Received by the Professional Development and Standards Board on June 23 2009

Noted by the Council of the Royal Australasian College of Surgeons on August 19 2009

Table of Contents

Executive Summary iv

The ASERNIP-S Classification System vii

The ASERNIP-S Review Group ix

1. Introduction 1

Objective 1

Fatigue 1

Causes of fatigue 1

Fatigue caused by sleep loss 1

Prevalence of fatigue in medical personnel 2

Sleep loss and performance 3

Workplace strategies for reducing fatigue in medical personnel 3

Summary 4

2. Methodology 5

Literature search protocol 5

Inclusion/exclusion criteria 5

Participants 5

Exposure 5

Intervention 5

Outcomes 5

Types of studies 6

Language restriction 6

Literature search strategies 7

Databases searched and search terms used 7

Search terms 7

Selection of studies 8

Data extraction and appraisal of study methodology 8

Data analysis 9

3. Studies included in the review 10

Literature search results 10

Critical appraisal 16

Clinical performance 16

Academic performance 17
Cognitive performance 17
Psychomotor skill performance 18
4. Results 20
Clinical performance 20
Retrospective patient complication rates 20
Self-assessed performance 23
Summary of clinical performance results 24
Academic performance 24
American Board of Surgery In-Training Examination (ABSITE) scores 24
Summary of academic performance results 25
Cognitive performance 25
Summary of cognitive performance results 27
Psychomotor skill performance 27
Overall performance scores 28
Performance time 28
Performance errors 31
Performance of movements 33
Manual dexterity 34
Summary of psychomotor skill performance results 35
Ongoing and unpublished trials 35
5. Discussion 36
Limitations of the evidence 36
Clinical performance 38
Academic performance 38
Cognitive performance 39
Psychomotor skill performance 39
Other considerations 40
6. Conclusions and Recommendations 43
Classification and Recommendations 43
Clinical and Research Recommendations 43
Acknowledgements 44
References 45
Appendix A – Excluded studies 53

Appendix B – Study profile tables 55

List of Tables

Table 1. NHMRC hierarchy of evidence 8

Table 2. Summary of included studies 11

Table 3. Time-of-day of assessment 12

Table 4. Duration of sleep received by study participants 13

Table 5. Definition of fatigue/sleep deprivation used in included studies 15

Table 6. Mortality rates for operations performed by sleep-deprived and non sleep-deprived surgeons 21

Table 7. Operative data and complications for operations performed by sleep-deprived and non sleep-deprived surgeons 22

Table 8. Complication* rates related to emergency/ non-emergency status of cases and residents' call status 23

Table 9. Cognitive performance 26

Table 10. Simulated performance - overall scores 28

Table 11. Simulated performance - pre-call and post-call performance time differences 30

Table 12. Simulated performance - pre-call and post-call performance errors 32

Table 13. Simulated performance - pre-call and post-call performance of movements 34

List of Figures

Figure 1. Process for selection of studies retrieved from the literature databases 10

Executive Summary

Objective

To investigate the effect of fatigue on surgeons and surgical outcomes, and to investigate the impact of fatigue on the cost of surgery and surgical training, through a systematic review of the literature.

Methods

Search strategy - Studies were identified by searching EMBASE, CINAHL, PubMed, The Cochrane Library and Current Contents from inception to June 2008. The Clinical Trials Database (US), NHS Centre for Research and Dissemination Databases (UK), National Research Register (UK), Meta Register of Controlled Trials, and the Australian Clinical Trials Registry were also searched in June 2008. Additional articles were identified through reference sections of the retrieved studies.

Study selection – Systematic reviews, randomised controlled trials (RCTs), non-randomised comparative studies and case series (pre-test/post-test outcomes) examining the effect of fatigue on clinical, academic, cognitive or psychomotor performance of surgeons or surgical trainees, and the effect of fatigue on the cost of surgery and surgical training were included.

Data collection and analysis - Data from the included studies were extracted by an ASERNIP-S researcher using standardised data extraction tables developed *a priori* and checked by a second researcher. Statistical pooling was not appropriate due to the study and result heterogeneity.

Results

A total of 20 studies were included for review: two RCTs, seven non-randomised comparative studies and 11 case series (pre-test/post-test outcomes). Studies were of variable quality and differed in study design. No economic evaluations were found.

Clinical performance (five studies): three non-randomised comparative studies failed to demonstrate any significant clinical differences between the sleep-deprived and non sleep-deprived groups. One non-randomised comparative study found that when residents operated on a not on-call day, complications were 45% more likely to occur when the resident had been on-call the day before ($p < 0.02$). The case series study found being on-call every other night was associated with significantly greater levels of fatigue ($p \leq 0.05$) and stress ($p \leq 0.05$) and less operating room participation (p-value not reported) and overall satisfaction ($p \leq 0.05$), when compared with the every fourth night with cross-cover schedule, but not with the frequency of reported errors.

Academic performance (two studies): two non-randomised comparative studies

demonstrated that being on-call the night before the American Board of Surgery In-Training Examination did not affect performance when compared with those not on-call the night before the examination.

Cognitive performance (five studies): RCT evidence (one study) indicated that sleep deprivation had no effects on factual recall and concentration. One non-randomised comparative study reported no differences within or between residents in relation to clear thinking, judgement, memory and learning when residents were acutely fatigued. Evidence from three case series studies suggested that were some variations in cognitive performance when participants were tired, but only for some variables in some studies, or only for certain individuals.

Psychomotor skill performance (11 studies): RCT evidence (two studies) reported no significant differences in psychomotor performance between rested and unrested groups. Non-randomised comparative studies (one study) and case series studies (eight studies) provided more mixed data: for performance time, hand movements and manual dexterity, approximately half of the studies found no significant differences or improvements between the rested and fatigued states post-call, while the other half reported decrements in performance when participants were fatigued. Errors were more likely to occur post-call. Surgical residents with less surgical training/experience appeared to be more affected by sleep deprivation than more senior residents.

Summary

There is a paucity of evidence investigating the effects of sleep loss and fatigue on the performance of surgeons and subsequent clinical outcomes. The overall weight of (poor) evidence shows that clinical, academic, and cognitive performance are not proven to be affected by sleep deprivation or fatigue and that psychomotor performance may or may not be. Variations in results were in some cases attributable to the level of training of participants, and between-subject differences. Many studies used surrogate markers to measure performance, although the relationship between these markers to actual clinical performance is unclear. It appears that fatigue can be compensated for in the acute operating room setting, but it is unclear what impact it has on *normal* functions. The search strategy did not identify any economic evaluations, resulting in an inability to comment on the financial effect of fatigue on surgery and surgical training.

We acknowledge that it would be beneficial to compare the results of this systematic review with data from professions other than the field of surgery. A systematic assessment of fatigue in other professions, such as aeronautics, transport, military and shift workers, was beyond the scope of this current assessment but, where available, have generally demonstrated similar findings to this review, although individual reports written within these industries do suggest detrimental effects of fatigue on performance.

Classification and Recommendations

On the basis of the evidence presented in this systematic review, the ASERNIP-S Review Group agreed on the following classification and recommendations concerning the effect of fatigue on the performance of surgeons and surgery:

Classifications

Evidence rating

The evidence-base in this review is rated as poor. The studies included were of variable quality, differed in study design, and many used surrogate markers to assess performance, resulting in an inability to draw solid conclusions.

Clinical and Research Recommendations

It is recommended that further research be done into:

- The identification of surrogate markers, if any, to actual clinical performance
    - The strength of the relationship between these surrogate markers (eg time of simulators, cognitive performance) and actual clinical performance
- The development of a clearer definition of fatigue and its relationship to sleep deprivation
- The development of common numerical values for acute and chronic sleep deprivation
- The effect of acute sleep deprivation on performance
- The effects of acute sleep deprivation on top of chronic partial sleep loss on performance
- Comparison of sleep-deprived surgeons with those who have had at least one week of normal sleep
- Comparison of performance at difference times of day to assess outcomes at different circadian points
- Comparison of performance of inexperienced surgeons with experienced surgeons with respect to fatigue and sleep loss
- Determine the impact of fatigue on the cost of surgery and surgical training.

Important note

The information contained in this report is a distillation of the best available evidence located at the time the searches were completed as stated in the protocol.

The ASERNIP-S Classification System

Evidence Rating

The evidence for ASERNIP-S systematic reviews is classified as Good, Average or Poor, based on the quality and availability of this evidence. High quality evidence is defined here as having a low risk of bias and no other significant flaws. While high quality RCTs are regarded as the best kind of evidence for comparing interventions, it may not be practical or ethical to undertake them for some surgical procedures, or the relevant RCTs may not yet have been carried out. This means that it may not be possible for the evidence on some procedures to be classified as good.

Good

Most of the evidence is from a high quality systematic review of all relevant randomised trials or from at least one high quality RCT of sufficient power. The component studies should show consistent results, the differences between the interventions being compared should be large enough to be important, and the results should be precise with minimal uncertainty.

Average

Most of the evidence is from high quality pseudo-RCTs, or from non-randomised comparative studies without significant flaws, such as large losses to follow-up and obvious baseline differences between the comparison groups. There is a greater risk of bias, confounding and chance relationships compared to high-quality RCTs, but there is still a moderate probability that the relationships are causal.

An inconclusive systematic review based on small RCTs that lack the power to detect a difference between interventions and RCTs of moderate or uncertain quality may attract a rating of average.

Poor

Most of the evidence is from case series, or studies of the above designs with significant flaws or a high risk of bias. A poor rating may also be given if there is insufficient evidence.

Research Recommendations

It may be recommended that an audit or a controlled (ideally randomised) clinical trial be undertaken in order to strengthen the evidence base.

Clinical Recommendations

Additional recommendations for use of the new intervention in clinical practice may be provided to ensure appropriate use of the procedure by sufficiently qualified/

experienced centres and on specific patient types (where appropriate).

* A comparator may be the current 'gold standard' procedure, an alternative procedure, a non-surgical procedure or no treatment (natural history)

The ASERNIP-S Review Group

ASERNIP-S Director

Professor Guy Maddern
Royal Australasian College of Surgeons
ASERNIP-S
Stepney, South Australia

Advisory Panel Members

Professor Drew Dawson
Director of the Centre for Sleep Research
University of South Australia
Adelaide, South Australia

Professor Richard Vaughan
Principal Medical Adviser
Health Workforce – Medical
Department of Health
Western Australia

Mr Peter Hewett
Colorectal Surgical Unit
Royal Adelaide Hospital
Adelaide, South Australia

Associate Professor Andrew Hill
Department of Surgery
University of Auckland
Middlemore Hospital
Auckland, New Zealand

Mr John Graham
St Vincents Medical Centre
Lismore, New South Wales, Australia

ASERNIP-S Researcher

Ms Lana Sturm
Royal Australasian College of Surgeons
ASERNIP-S
Stepney, South Australia

Conflict of Interest

No conflicts of interest were declared.

1. Introduction

Objective

The aim of this review is to investigate the effect of fatigue on surgeons, the impact that a fatigued surgeon has on patient outcomes, and the impact of fatigue on the cost of surgery and surgical training, through a systematic review of the literature.

Fatigue

Fatigue has been shown to adversely affect the performance of individuals in various situations. Fatigue has been widely studied in relation to poor performance outcomes in drivers (Lal & Craig 2001), pilots (Holley et al 2003) and industrial workers (Frank 2000). There is growing concern that fatigue and extended surgical working hours may contribute to poor performance in surgery.

Fatigue is a complex, multicausal, multidimensional, non-specific, and subjective phenomenon for which no one definition is widely accepted (Tiesinga et al 1996). Lal & Craig (2001) describe fatigue as a transitory period between wake and sleep, which, if uninterrupted can lead to sleep. Jensen & Given (1991) define fatigue as a subjective feeling which exists at one point in time on a continuum from weariness to complete exhaustion, resulting from physical, mental, or emotional activity. Lal & Craig (2001) state that when a person is fatigued, there is no desire for physical or mental effort and there is an associated heavy, drowsy feeling. Piper (1993) differentiates fatigue from normal feelings of tiredness, and defines fatigue as unusual, abnormal or excessive whole body tiredness, disproportionate to, or unrelated to, activity or exertion.

Physical fatigue, which is considered synonymous with muscle fatigue and is characterised by reduced muscular power and movement (Lal & Craig 2001), will not be addressed in this review. "Time-on-task" fatigue describes fatigue that is accumulated during the working period, from such activities as long surgical cases, and will also not be addressed in this review.

Causes of fatigue

There are many, often interrelated, causes of fatigue. Much of the literature regarding fatigue focuses on fatigue caused by sleep loss, sleep deprivation, or long working hours.

This review will focus on sleep related causes of fatigue.

Fatigue caused by sleep loss

Research suggests that sleep requirements are idiosyncratic, with wide variation across individuals (Aeschbach et al 2003). Most adults require between six to 10

hours of sleep per 24 hour period, with most people acquiring approximately seven and a half hours of sleep a day (Kripke et al 2002). Infants generally require more sleep than adults, with approximately 14 hours of sleep a day, while teenagers require approximately nine hours of sleep a day (Jenni et al 2004). Women in the first three months of pregnancy often need several more hours of sleep than usual (Hedman et al 2002).

Sleep deprivation results when a person fails to get the required amount of sleep. A person who is sleep-deprived will have difficulty staying awake at the point of the circadian cycle where sleep is normally induced (Weigner & Ancoli-Israel 2002). The amount of sleep a person needs increases if he or she has been deprived of sleep in previous days. Getting too little sleep creates a sleep debt and eventually the body will require that the lost sleep, or at least part thereof, is replaced. Total sleep deprivation occurs when an individual gets no sleep, and is more likely to occur in acute situations (eg studying or working all night) (Weigner & Ancoli-Israel 2002). Partial sleep deprivation refers to a night of reduced or interrupted sleep, which may be due to work schedules, lifestyle, sleep disorders, medical conditions, or medications (Weigner & Ancoli-Israel 2002). Prolonged periods without adequate amounts of sleep can result in chronic sleep deprivation.

Individuals with constantly changing schedules, such as shift workers, suffer desynchronosis, where their circadian rhythm becomes out of phase with the environment (Kuhn 2001). Adaption to a new sleep/wake pattern occurs at a rate of approximately one hour per day (Akerstedt, 2003). The amount of sleep, or the amount of sleep disruption following a work shift varies with the type of shift (morning, afternoon, or night), direction of rotation, and time of changeover (Akerstedt 2003). For rotating schedules, adaptation to one shift may not be complete before a further shift change occurs, and sleep disturbances and fatigue may continue into rest days.

Prevalence of fatigue in medical personnel

Long hours and overnight call are common in many professions, and are a prominent feature in many medical professions, particularly during medical residencies (Philibert 2005). A survey completed by 1412 junior doctors in New Zealand in 2003 revealed that 13% of respondents documented working weeks exceeding 70 hours per week, with the remaining 87% working between 50 and 70 hours per week (Gander et al 2007). Lockley et al (2006) conducted a survey of 2737 first year post graduate residents in the US between 2002 and 2003 and found that working after more than 40 hours without sleep was reported more than 1000 times (Lockley et al 2006). The survey also found that residents reported being in the hospital for 70.2 ± 26.0 hours per week, and were asleep for 3.2 ± 4.2 hours of this time.

In 2001, the Australian Medical Association published the results of an audit of junior doctors' working hours (AMA 2001). Results of this audit indicated that 78%

of junior doctors had rostering and work practice variables that placed them into the significant level of risk or the higher level of risk categories of fatigue and performance impairment. The AMA later conducted a similar audit of all doctors, and found that the proportion of doctors in the significant level of risk or the higher level of risk was 62% (AMA 2006).

Long work hours are almost considered a tradition during residency training and are a feature of common life once training is complete (Jha et al 2001). The discipline of surgery appears to require a greater number of hours of work than other medical disciplines. A survey of otolaryngology residents found that residents worked on average 79 to 90 hours per week (Strunk et al 1991). Longer work hours were reported by Whang et al (2003) who surveyed general surgical residents in Canada in 2002, and found that residents worked an average of 105 hours per week.

Sleep loss and performance

When healthy adults receive an average of less than five hours sleep per night, the homeostatic drive to sleep rises sharply, and cognitive performance begins to decline (van Dongen et al 2003). Sustained wakefulness of 24 hours has been found to result in a decline in cognitive psychomotor performance equivalent to that found at a blood alcohol concentration of 0.10% (Dawson & Reid 1997). In the non-medical workforce, night work has shown to be associated with a greater relative risk of accidents and injuries from impaired alertness and performance caused by lack of sleep than either the morning or afternoon shift (Folkard et al 2005).

Fatigue caused by sleep deprivation has been shown to negatively impact performance in both non-medical workers (Philibert 2005) and medical residents (Leonard et al 1998). Prolonged duty hours and sleep deprivation have been associated with increased attention failures in intensive care interns (Lockley et al 2004), increased risk of percutaneous injuries in medical trainees and residents in their first year of clinical training (Fisman et al 2007; Ayas et al 2006), and increased risk of interns having motor vehicle accidents during a commute from work (Barger et al 2005). A randomised controlled trial (RCT) comparing rates of serious medical errors made by internal medicine interns working 24 hour or more shifts compared with every third night found that substantially more serious medical errors were made when the interns worked frequent long shifts (Landrigan et al 2004).

Workplace strategies for reducing fatigue in medical personnel

Working hour obligations and on-call duties vary between specialties, hospitals, and with the number of support staff available (RACS 2007). Internationally, efforts to reduce working hours for medical professionals are wide-ranging. Doctors working in European Union Member States come under the provisions of the European Working Time Directive (UK Department of Health Website 2007). This Directive was enacted into UK law in 1998 as the Working Time Regulations, and limits the

number of hours that doctors are allowed to work over an average week. The Government negotiated an extension for compliance with the Regulations of up to twelve years to prepare for full implementation for doctors in training. As of August 2003 all junior doctors are limited by contract to 56 hours of active work (UK Department of Health Website 2007).

The Accreditation Council for Graduate Medical Education (ACGME) in the US implemented requirements regarding working hours for resident training in July 2003 (ACGME 2002). Included in these requirements is that duty hours must be limited to 80 hours per week; in-house call must not occur more often than every third night; continuous on-site duty must not exceed 24 consecutive hours, and a minimum 10 hour rest period between duty periods (ACGME 2007).

In 1999, as part of the Safe Hours Campaign, the Australian Medical Association adopted the National Code of Practice – Hours of Work, Shiftwork and Rostering for Hospital Doctors (AMA 2005) to assist doctors assess the risk associated with their working hours. In 2007, The Royal Australasian College of Surgeons (RACS) developed standards to address safe working hours, specifically for Fellows, surgical trainees and international medical graduates (RACS 2007). These standards include that normal working hours should be less than 70 hours per week; on-calls should be no more than a one in four on-call rotation and; day and night shifts should be a maximum of 14 and 12 hours long respectively.

Since the introduction of mandated working hours, concerns have been raised regarding the potential negative impact on the economics of health care, discontinuity of patient care (Fletcher et al 2004), and the overall quality of patient care (Laine et al 1993; Nuckols & Escarce 2005). In addition to this, there are concerns that restricted working hours may lead to restricted access to health care practitioners through a reduction in the labour supply, increased sleep restriction in senior physicians, and increases in error rates due to work intensification (Dawson & Zee 2005). There are also concerns that efforts to minimise the impact of fatigue on individuals can compromise the quality of medical education and training (Weatherby et al 2007) as well as increase the costs associated with training (Schenarts et al 2006).

Summary

Data from non-medical fields suggest that sleep deprivation and disturbances of circadian rhythms lead to poor performance. This link has not yet been clearly established in surgery. The aim of this review is to assess the impact of fatigue on surgeon performance and surgical outcomes through a systematic review of the literature. The review will also aim to assess the financial impact that fatigue has on surgery and surgical training.

2. Methodology

Literature search protocol

Inclusion/exclusion criteria

Articles were selected for inclusion in this systematic review on the basis of the following criteria.

Participants

Participants conducting surgical procedures, or enrolled in a surgical training program.

Exposure

Fatigue as a result of sleep loss or sleep deprivation, prolonged working hours or training (ie on-the-job surgical training) hours, or any strategy designed to induce fatigue in the study participants.

Intervention

Studies examining the effect of work hour restrictions were not included.

Outcomes

Studies that measure performance through patient outcomes, simulator performance, academic test scores or cognitive functioning. At least one of the following outcomes were included:

- **Psychomotor skills** of participant which included, but were not limited to:
  - time to complete skill/test/technique/procedure
  - accuracy of skill/technique/test/procedure
  - error rates
- **Academic performance** of participant which included, but were not limited to:
  - examination scores and results
- **Cognitive performance** of participant which included, but were not limited to:
  - attention
  - memory
  - reasoning
- **Outcomes of patients** under the care of fatigued/non-fatigued individuals, which included, but were not limited to:
  - morbidity
  - mortality
  - continuity of care
- Cost/resource use

- Any studies which examined the impact of fatigue on the cost of surgery and/or surgical training

Types of studies

Systematic reviews, RCTs, pseudo-RCTs, non-randomised comparative studies, and case series (with pre- and post- test outcomes) were included for review.

Study design

Systematic reviews were defined as those studies that meet all the following criteria as defined by Cook *et al* (1997):

1. Focused clinical question
2. Explicit search strategy
3. Use of explicit, reproducible and uniformly applied criteria for article selection
4. Critical appraisal of the included studies
5. Qualitative or quantitative data synthesis.

RCTs published after the search dates of the most recent systematic review were also included.

RCTs and pseudo-RCTs were considered eligible for inclusion and critical appraisal. A study was deemed to be an RCT if the author(s) stated explicitly (usually by some variant of the term 'random' to describe the allocation procedure used) that the groups compared in the trial were established by random allocation (Higgins & Green 2008). Studies in which the method of allocation was known but was not considered strictly random (for example, alternation, date of birth and medical record number) were classified as pseudo-RCTs (Higgins & Green 2008).

When overlapping patient groups were reported in studies, only the paper quoting the most complete data were used.

Background information

Where appropriate; additional relevant published material in the form of letters, conference material, commentary, editorials and abstracts were included as background information.

Language restriction

Searches were conducted without language restriction.

Literature search strategies

Databases searched and search terms used

- The Cochrane Library
- Entrez-PubMed from 1953
- Ovid EMBASE from 1980
- Webspirs CINAHL from 1982
- ISI Current Contents Connect from 1993
- Clinical Trials Databases (US)
- National Research Register (UK)
- NHS CRD databases

Search terms

In the Cochrane Library the search terms used are:

sleep and (performance or work* hours)

For MEDLINE, EMBASE, CINAHL and Current Contents Connect the following search terms were used:

1. performance (MeSH)

   job performance/ or motor performance/ or task performance/ or psychomotor performance / or mental performance / or performance / or physical performance /

2. error (MeSH)

   medical error/ or surgical error / or error /

3. patient safety (MeSH)

   patient care/ or health Care Quality/ or patient safety/

4. duty hours (MeSH)

   medical practice/ or medical education/ or working time/ or residency education/ or workload/ or work schedule/

5. rest (MeSH)
6. 1 OR 2 OR 3 OR 4 OR 5
7. 6 AND [fatigue (MeSH) OR sleep (MeSH) sleep/ or sleep deprivation/]
8. 7 AND surg* (keyword)

The NHS CRD databases were searched using the above terms. The National Research Register, Clinicaltrials.gov, Meta-Register and the Australian Clinical Trials Registry were

also searched using the above search terms for RCTs in progress.

Note: * is a truncation character that retrieves all possible suffix variations of the root word e.g. surg* retrieves surgery, surgical, surgeon, etc. In Cochrane the truncation character is *; in Current Contents, EMBASE, CINAHL and MEDLINE (Ovid) it is $.

Selection of studies

One reviewer applied the inclusion criteria to identify those studies potentially eligible for selection and appraisal based on their abstracts; these studies were retrieved as full text. The selection criteria were then applied fully to the retrieved studies to identify those to be appraised and included in the review. Full publications subsequently found not to meet the inclusion criteria were excluded and reasons for exclusion were documented.

The bibliographies of all publications retrieved were manually searched for relevant references that may have been missed in the database search (pearling).

Data extraction and appraisal of study methodology

Data from all included studies were extracted by one reviewer and checked by a second reviewer using standardised data extraction tables that were developed *a priori*. The studies included in the review were classified according to the National Health and Medical Research Council (NHMRC) hierarchy of evidence (NHMRC 2000) (Table 1). Any differences were resolved through discussion.

Table 1. NHMRC hierarchy of evidence

Level of Evidence	Study Design
I	Evidence obtained from a systematic review of all relevant randomised controlled trials.
II	Evidence obtained from at least one properly designed randomised controlled trial.
III-1	Evidence obtained from well-designed pseudo-randomised controlled trials (alternate allocation or some other method).
III-2	Evidence obtained from comparative studies (including systematic reviews of such studies) with concurrent controls and allocation not randomised, cohort studies, case-control studies, or interrupted time-series with a control group.
III-3	Evidence obtained from comparative studies with historical control, two or more single arm studies, or interrupted time series without a parallel control group.
IV	Evidence obtained from case-series, either post-test or pre-test/post-test.

(NHMRC 2000)

If systematic reviews were eligible for inclusion in the review, the methodology of these secondary studies was evaluated with respect to the following factors:

- Did the review ask a focused research question that incorporated the elements of the patient population, intervention, comparator intervention and outcomes (PICO)?
- Were the inclusion and exclusion criteria of included studies clearly stated?
- Did the review use a clear and comprehensive search strategy?
- Did the review assess the validity of included studies, and if so which validity criteria were used?
- Was the analysis or synthesis of the results appropriate?

- Did the review include a summary of its main results, including a discussion of its strengths and limitations?

Where primary studies were eligible for inclusion in the review, the following criteria were used to appraise their methodology, where applicable:

- Were the objectives of the study clearly defined?
- Were the inclusion and exclusion criteria clearly described?
- Was there a clear description of the interventions used?
- Were the characteristics of patients included in the study clearly described?
- Were patients randomly assigned to intervention groups, and if so was the method of randomisation described?
- Was the randomised assignment of patients to intervention groups concealed from both patients and staff administering the study until recruitment was complete?
- Was an attempt made to blind both patients and staff responsible for measuring outcomes of the intervention to the interventions patients received?
- Were the number of patients who withdrew or dropped out of the study reported, and the characteristics of these patients described?
- Were the main outcomes of interest adequately reported?
- Were point estimates and measures of variability presented for the primary outcome measures?

Non-randomised studies were assessed for other features of study design or execution that may have introduced bias, such as comparability of patient groups at baseline, method of patient selection and comparability of timing of outcome assessment.

Data analysis

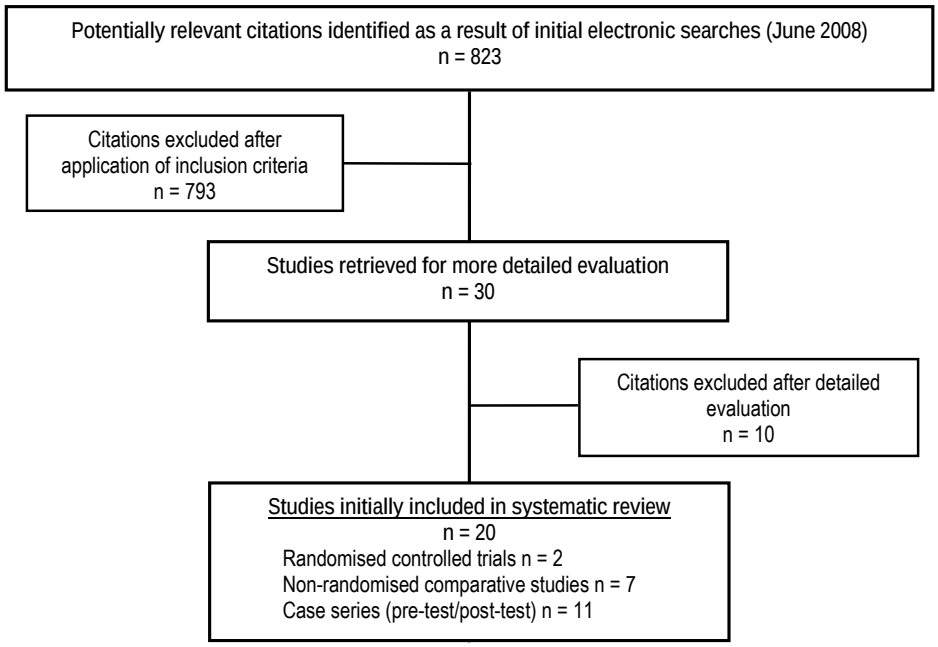
If the data were suitable for statistical pooling, meta-analyses of the main outcomes were performed. If possible, the data were stratified into clinically relevant groups. Otherwise, data for the main outcomes were reported narratively.

3. Studies included in the review

Literature search results

Details of the searching and retrieval process are shown in Figure 1.

Figure 1. Process for selection of studies retrieved from the literature databases



Description of studies

A total of 20 studies were included in this review (Table 2): two RCTs, seven non-randomised comparative studies and 11 case series.

Excluded studies are listed in Appendix A. The study profiles for the included studies are given in Appendix B.

No economic evaluations were found. The included studies examined a total of 706 academic scores, 20795 operations (number of surgeons unknown), and 332 operating surgeons (many surgeons may have been represented more than once). The time-of-day at which participants were assessed is given in Table 3, the duration of sleep the participants had is given in Table 4, and the definition of the rested and unrested states is given in Table 5.

Table 2. Summary of included studies

Study	L	Assessment technique	N	Rested state	Not rested state
				n	n
CLINICAL PERFORMANCE					
Ellman et al 2005 (USA)	III-2	Retrospective complication rates and mortality	7323	Not sleep-deprived	Sleep-deprived
				7094 operations	229 operations
Ellman et al 2004 (USA)	III-2	Retrospective complication rates and mortality	6751	Not sleep-deprived	Sleep-deprived
				6412 operations	339 operations
Haynes et al 1995 (USA)	III-2	Retrospective complication rates	6451	Not sleep-deprived	Sleep-deprived
				6541 operations	359 operations
Schieman et al 2008 (Canada)	III-2	Retrospective complication rates and mortality	270	Rested	Fatigued
				248 operations	22 operations
Sawyer et al 1999 (USA)	IV	Participant and faculty surveys	19	On-call every 2, 3 or 4 nights	
ACADEMIC PERFORMANCE					
Minion et al 2007 (USA)	III-2	ABSITE	282	Not on-call night before exam	On-call night before exam
				213	69
Stone et al 2000 (USA)	III-2	ABSITE	424	Not on-call night before exam	On-call night before exam
				354	70
COGNITIVE PERFORMANCE					
Reznick & Folse 1987 (USA)	II	Factual recall, concentration, manual dexterity*	21	Not sleep-deprived	Sleep-deprived
				21	21
Light et al 1989 (USA)	III-2	Neuropsychological tests, manual dexterity,* mood†	42	Rested	Fatigued
				21	21
Deaconson et al 1988 (USA)	IV	Psychometric tests, manual dexterity*	26	Not sleep-deprived	Sleep-deprived (on-call)
				26	26
Deary & Tait 1987 (Scotland)	IV	Cognitive and mood† tests	12	Off-duty / On-call / Waiting for activity	
				12 / 12 / 12	
Wesnes et al 1997 (UK)	IV	Cognitive and mood† function	10	Weekend not on-call	Weekend on-call
				10	10
PSYCHOMOTOR SKILL PERFORMANCE					
Uchal et al 2005 (Norway)	II	Suturing stomach in pelvic trainer	64	Post-work	Post-call
				32	32
DeMaria et al 2005 (USA)	IV	MIST-VR	17	Pre-call	Post-call
				17	17
Eastridge et al 2003 (USA)	IV	SCMIS, MIST-VR	35	Not on-call (rested)	Post-call (acute sleep deprivation)
				35	35
Grantcharov et al 2001 (Denmark)	IV	MIST-VR	14	Day shift	Post-call
				14	14
Jakubowicz et al 2005 (USA)	IV	ESS	8	Pre-call	Post-call
				8	8
Kahol et al 2008 (USA)	IV	ProMIS, FLS	37	Pre-call	Post-call
				37	37
Leff et al 2008 (UK)	IV	MIST-VR	21	7 consecutive night shifts	
				21	
Taffinder et al 1998 (UK)	IV	ICSAD, MIST-VR	6	Not sleep-deprived / Sham night on-call / Night no sleep	
				12 / 12 / 12	

ABSITE, American Board of Surgery In-Training Examination; ESS, Endoscopic Sinus Surgery Simulator; FLS, Fundamentals of Laparoscopic Surgery; ICSAD, Imperial College Surgical Assessment Device; L, Level of Evidence; MIST-VR, Minimally Invasive Surgery Trainer – Virtual Reality; SCMIS, Southwestern Centre for Minimally Invasive Surgery

*Mood outcomes were not included in this review.

† Results for manual dexterity for Reznick & Folse (1987), Light et al (1989) and Deaconson et al (1988) appear in the psychomotor skill performance results category.

Table 3. Time-of-day of assessment

Study	L	Assessment technique	N	Rested state	Not rested state	
				Time assessed	Time assessed	
CLINICAL PERFORMANCE						
Ellman et al 2005	III-2	Complication rates and mortality	7323	Not sleep-deprived	Sleep-deprived	
				NA	NA	
Ellman et al 2004	III-2	Complication rates and mortality	6751	Not sleep-deprived	Sleep-deprived	
				NA	NA	
Haynes et al 1995	III-2	Complication rates	6541	Not sleep-deprived	Sleep-deprived	
				NA	NA	
Schieman et al 2008	III-2	Complication rates and mortality	270	Rested	Fatigued	
				NA	NA	
Sawyer et al 1999	IV	Sleep logs, surveys, faculty surveys	19	On-call every 2, 3 or 4 nights		
				NA		
ACADEMIC PERFORMANCE						
Minion et al 2007	III-2	ABSITE	282	Not on-call night before exam	On-call night before exam	
				NR	NR	
Stone et al 2000	III-2	ABSITE	424	Not on-call	On-call	
				NR	NR	
				Weekly sleep logs and monthly questionnaires		
COGNITIVE PERFORMANCE						
Reznick & Folse 1987	II	Factual recall, concentration, manual dexterity	21	Not sleep-deprived	Sleep-deprived	
				0800 h – 1800 h during working day	0800 h – 1800 h during working day	
Light et al 1989	III-2	Neuropsychological tests	42	Rested	Fatigued	
				NR	NR	
Deaconson et al 1988	IV	Psychometric tests	26	Not sleep-deprived	Sleep-deprived (on-call)	
				0600 – 0800 h	0600 – 0800 h	
Deary & Tait 1987	IV	Cognitive tests	12	Off-duty	On-call	Waiting for activity
				1400 – 1700 h	1400 – 1700 h	1400 – 1700 h
Wesnes et al 1997	IV	Cognitive function	10	Weekend not on-call	Weekend on-call	
				Monday morning	Monday morning	
PSYCHOMOTOR SKILL PERFORMANCE						
Uchal et al 2005	II	Suturing stomach in pelvic trainer	64	Post-work	Post-call	
				1600 – 1700 h	0800 – 0900 h	
DeMaria et al 2005	IV	MIST-VR	17	Pre-call	Post-call	
				Morning of day on-call	Late the following day	
Eastridge et al 2003	IV	SCMIS, MIST-VR	35	Not on-call (rested)	Post-call (acute sleep deprivation)	
				0800 h morning before on-call & morning of day on-call (combined data)	0800 h	
Grantcharov et al 2001	IV	MIST-VR	14	Day shift	Post-call	
				Daytime hours	0930 h	
Jakubowicz et al 2005	IV	ESS	8	Pre-call	Post-call	
				NR	NR	
Kahol et al 2008	IV	ProMIS, FLS	37	Pre-call	Post-call	
				NR	NR	
Leff et al 2008	IV	MIST-VR	21	7 consecutive night shifts		
				0800 – 1030 h after each night shift		
Taffinder et al 1998	IV	ICSAD, MIST-VR	6	Not sleep-deprived	Sham night on-call	Night no sleep
				1700 – 1800 h & 0800 – 0900 h	1700 – 1800 h & 0800 – 0900 h	1700 – 1800 h & 0800 – 0900 h

ABSITE, American Board of Surgery In-Training Examination; ESS, Endoscopic Sinus Surgery Simulator; FLS, Fundamentals of Laparoscopic Surgery; ICSAD, Imperial College Surgical Assessment Device; MIST-VR, Minimally Invasive Surgery Trainer – Virtual Reality; NA, not applicable; NR, not reported; SCMIS, Southwestern Centre for Minimally Invasive Surgery.

Table 4. Duration of sleep received by study participants

Study	L	N	Method of sleep determination	Rested state Sleep duration	Not rested state Sleep duration
CLINICAL PERFORMANCE					
Ellman et al 2005	III-2	7323		Rested	Fatigued
			NR	NR	NR
Ellman et al 2004	III-2	6751		Rested	Fatigued
			NR	NR	NR
Haynes et al 1995	III-2	126*		Most recent on-call (n = 15)	'Usual' sleep when on-call (n = 15)
			Self-reported, mean hours	1.8†	1.3†
Schieman et al 2008	III-2	270		Rested	Fatigued
			NR	NR	NR
Sawyer et al 1999	IV	19		Frequency of on-call	
			Weekly sleep log, hours sleep/ week	Every 2 nights 38 ± 1 ‡ §	Every 3 nights 37 ± 1 ‡	Every 4 nights 41 ± 1
			P-value	NR	NR	NR
ACADEMIC PERFORMANCE					
Minion et al 2007	III-2	282		Not on-call night before exam	On-call night before exam
			NR	NR	NR
Stone et al 2000	III-2	424		Not on-call night before exam	On-call night before exam
			NR	NR	NR
COGNITIVE PERFORMANCE					
Reznick & Folse 1987	II	21		Not sleep-deprived	Sleep-deprived
			NR. mean ± SD	6.95 ± 0.97	1.75 ± 0.89
Light et al 1989	III-2	42		Rested	Fatigued
			NR	NR	NR
Deaconson et al 1989	IV	26	Total sleep (sleep diary) ‖ , mean min ± SD	Not sleep-deprived 419 ± 106	Sleep-deprived 184 ± 113
			P-value		< 0.05
Deary & Tait 1987	IV	12	Self-reported, median (range)	Off-duty 7 (6 – 10)	On-call 5 (0 – 7)	Waiting 1.5 (0 – 7)
			P-value (compared with off-duty)¶	-	< 0.005**	< 0.0005
Wesnes et al 1997	IV	10		Weekend off-call	Weekend on-call
			Sleep diary, mean hours and minutes	Saturday: 8 h 36 min Sunday: 6 h 27 min	Saturday: 4 h 3 min Sunday: 4 h 9 min
PSYCHOMOTOR SKILL PERFORMANCE					
Uchal et al 2005	II	64	Self reported, median (range)	Post-work (n = 32) 6.5 (5 – 9)	Post-call (n = 32) 1.5 (0 – 3)
			P-value		0.05
DeMaria et al 2005	IV	17		Pre-call	Post-call
			Self-reported, mean (range)	6.2 (5 – 8)††	3.6 (2 – 6)‡‡
Eastridge et al 2003	IV	35		Pre-call	Post-call
			Self-reported, mean ± SEM	6.5 ± 0.3§§	1.5 ± 0.3‖ ‖
			P-value		< 0.001
Grantcharov et al 2001	IV	14		Pre-call (daytime)	Post-call
			NR, median hours (range)	NR	1.5 (0 – 3)
Jakubowicz et al 2005	IV	8	Sleep log, mean hours/preceding 24 hr	Pre-call 5.8	Post-call 3.6 ¶¶
			P-value		< 0.005
Kahol et al 2008	IV	37		Pre-call	Post-call
			NR	NR	NR
Leff et al 2008	IV	21	Self-reported (questionnaire) mean hours ± SD	Week before night shifts 44.5 ± 7.1	Week of night shifts*** 49.0 ± 5.8†††
			P-value		0.035
Taffinder et al 1998	IV	6		Control (n = 12)	Sham on-call(n=12)	No sleep (n = 12)
			NR	NR	NR	NR
			P-value		NR	NR

NR, not reported; SD, standard deviation; SEM, standard error of the mean.

Table continued over page

* Only a subset of resident cases were included in the retrospective analysis.
† Residents assessment of the severity of sleep deprivation on day-to-day performance (0, least severe; 4, most severe) revealed a mean adverse effect rating of 1.5.
‡ $p \leq 0.05$ versus on-call every 4 nights.
§ Fatigue on a scale of 0 (lowest) to 5 (highest) on-call was 1.8 ± 0.2 and off-call 2.4 ± 0.2. Both significantly greater than when on-call every 4th night ($p \leq 0.05$).
‖ Longest uninterrupted sleep interval (minutes): Not sleep-deprived 132 ± 71 vs sleep-deprived 379 ± 85 ($p < 0.05$). There were no significant differences in the number of interruptions, fatigue score or motivation score.
¶ On-call nights allowed more sleep than waiting nights ($p < 0.005$).
** Sleep was disturbed a median of once (range 0 – 3 times). Approx half of calls required participant to leave his/her bed, and calls lasted a median of 1 hour (range 0.5 – 7 hours).
†† Pre-call, 14 (82%) residents reported that they were well rested or rested.
‡‡ Post-call 13 (77%) residents reported that they were not well rested. The longest period of uninterrupted sleep during a night on-call was mean 2.4 hours (range 0.5 – 5 hours).
§§ < 90% participants had 1 interruption during the night. Participants had a mean 6 hours continuous sleep. Subjective fatigue levels measured on a scale of 1, none to 10, exhausted and were mean 2.3 ± 0.3.
‖‖ 43% participants reported ≤ 1 interruption (range, 0 – 10 interruptions). Subjective fatigue levels were mean 6.8 ± 0.3 ($p < 0.001$ compared with pre-call).
¶¶ Participants woken on average twice per night (for 3.5 pages). Participants awake almost another 5 hours before assessment. Sleep and alertness were rated on a scale of 1, bad to 5, best. Average sleep was rated 2.8. Average rating of alertness was 3.3.
*** Subjective sleepiness scores generally increased across the week of night shifts reaching abnormal levels (according to Epworth Sleepiness Scale) following nights 3, 4, 6 and 7. The greatest drive to sleep was observed on the 7th night (greatest sleep debt). No statistical analyses reported in this data.
††† It was reported that participants had significantly less sleep during their week of night shifts compared with the week preceding night duty. The values reported in the text do not correlate with this statement. The raw data supplied indicates that sleep may have been higher in the week of night duty (approx 63 minutes/night). Raw sleep values for the week before night shifts were not reported.

Table 5. Definition of fatigue/sleep deprivation used in included studies

Study	Type	Description of rested state	Description of not rested state
CLINICAL PERFORMANCE			
Ellman et al 2005	III-2	All cases (besides sleep deprived cases) were considered not sleep-deprived cases.	Sleep deprivation: if thoracic surgical resident preformed a case that started between 11 pm and 5 am, or ended a case between 11 pm and 7:30 am, or if resident performed a subsequent case within the next 24 hours.
Ellman et al 2004	III-2	All cases (besides sleep deprived cases) were considered not sleep-deprived cases.	Sleep deprivation: If attending cardiac surgeon preformed a case that started between 10 pm and 5 am, or ended a case between 11 pm and 7:30 am, or if the resident performed a subsequent case within the next 24 hours.
Haynes et al 1995	III-2	NR	Sleep deprivation: When required to remain in-house for duration of a 24-hour day.
Schieman et al 2008	III-2	All cases except when surgeon was considered fatigued.	Fatigued: If surgeon billed for clinical work after 10 pm the night before.
Sawyer et al 1999	IV	NA (did not compare with a rested state)	Long call: overnight in-hospital duties
ACADEMIC PERFORMANCE			
Minion et al 2007	III-2	Not on-call night before exam	On-call night before exam
Stone et al 2000	III-2	Not on-call night before exam	On-call night before exam: if resident had to be in hospital overnight without scheduled sleep period before exam
COGNITIVE PERFORMANCE			
Reznick & Folse 1987	II	> 5hours sleep in 24 hour period	Sleep deprivation: < 3 hours sleep in 24 hour period
Light et al 1989	III-2	NR	Acute sleep deprivation: < 4 hours sleep
Deaconson et al 1988	IV	> 4 hours continuous sleep in preceding 24 hours	Sleep deprivation: ≤ 4 hours continuous sleep during preceding 24 hours
Deary & Tait 1987	IV	Off-duty	On-call or waiting for activity
Wesnes et al 1997	IV	Weekend off-duty	Weekend on-call
PSYCHOMOTOR SKILL PERFORMANCE			
Uchal et al 2005	II	Post-work: 0800 – 1600 h working day with a previous undisturbed night spent at home	Post-call: 0800 – 0800 h on-call duty in hospital
DeMaria et al 2005	IV	Pre-call: a regular working day before a night on-call	Post-call: the day after a night on-call
Eastridge et al 2003	IV	Pre-call: morning before scheduled 24 h in-house call & on-call: the morning of 24 h in-house call (data combined)	Post-call: morning after 24 h in-house call
Grantcharov et al 2001	IV	NR	< 3 hours sleep
Jakubowicz et al 2005	IV	Pre-call	After 24 hour on-call period
Kahol et al 2008	IV	Pre-call: before performing night duties	Post-call: after performing night duty
Leff et al 2008	IV	Days prior to start of night shifts	7 consecutive night shifts
Taffinder et al 1998	IV	Undisturbed night	Sham night on-call: disturbed at 0000 h, 0300 h and 0600 h. Night no sleep

NR, not reported

Critical appraisal

Clinical performance

Non-randomised comparative studies

Four retrospective non-randomised comparative studies compared patient complication rates from surgery when participants were rested or not rested (Ellman et al 2005; Ellman et al 2004; Haynes et al 1995; Schieman et al 2008). All of these studies compared participants who were not rested with concurrent rested controls.

Three studies were conducted in single centres (Ellman et al 2005; Ellman et al 2004; Haynes et al 1995), and one study was conducted in two centres (Schieman et al 2008).

Ellman et al (2005) recruited thoracic surgical residents; Ellman et al (2004) attending cardiac surgeons, Haynes et al (1995) surgical residents, and Schieman et al (2008) fellowship trained colorectal surgeons. None of these studies reported how many surgeons were included in the studies, and instead reported only the number of operations.

Three studies examined all surgeries of a specific type and then correlated the outcomes of these surgeries with the on-call or rested state of the surgeon (Ellman et al 2005; Ellman et al 2004; Schieman et al 2008). Another study (Haynes et al 1995) examined all emergency and elective procedures performed by surgery residents and identified which cases had complications. Complications had previously been entered into a database by a senior resident and attending surgeon, which may have resulted in selection bias. The operations with the complications were correlated with the on-call status of the resident.

There were large differences in the number of rested and unrested groups in all four of these studies which may have skewed the results. No sample size calculations were reported to determine the optimal sample size required to show a difference between the two groups.

No aspects of blinding were reported in any of the studies. Three studies reported that the patients were well matched (Ellman et al 2005; Ellman et al 2004; Schieman et al 2008), but no study reported whether there were any baseline differences between the participants conducting the operations.

Losses to follow up were not applicable in the four studies as they were case reviews. The study period was reported by all four studies. One study reported using outcome measures that had not been validated (Haynes et al 1995). The other studies did not report whether the outcome measures had been validated or not.

Inclusion criteria included the cases or procedures that participants had performed and were of interest (Ellman et al 2005; Ellman et al 2004; Haynes et al 1995; Schieman et al 2008). Exclusion criteria were procedures not of interest to the study (Ellman et al 2005; Schieman et al 2008; Haynes et al 1995) or were not reported (Ellman et al 2004).

Case series (pre-test/post-test outcomes)

One single study case series study prospectively compared errors, job satisfaction, stress

and fatigue when first year surgical residents took call every second, third and fourth night (Sawyer et al 1999). Monthly surveys were given to residents and asked questions related to hours of sleep, frequency of fatigue, number of sleepless nights, errors, inability to complete work, operating room performance and stress levels. It was not reported how participants were selected for inclusion into the study. No participant demographics were reported. Inclusion criteria were not reported. Exclusion criteria were interns on holidays, non surgery interns and stressful months of the year. It was not reported whether the assessment tools had been previously validated.

Academic performance

Non-randomised comparative studies

Two studies retrospectively evaluated academic examination scores and the effect of on-call status (Minion et al 2007; Stone et al 2000). One study evaluated the scores of general surgery residents from a single centre (Minion et al 2007), and the other study evaluated scores of general surgery residents from 21 general surgery programs (Stone et al 2000). Scores from residents who were on-call the night before the examination were compared with concurrent controls (residents who were not on-call the night before the examination).

One study did not report whether any data were lost to assessment (Minion et al 2007), while the multicentre study reported that six centres did not provide data, and that one program director of a centre did not respond to the survey (Stone et al 2000). It was not reported whether the outcome methods used in the studies were validated. The study periods were reported. No demographic details of participants were reported for either study. Besides stating who was included in the studies, no other inclusion or exclusion criteria were reported.

Cognitive performance

Randomised controlled trials

One RCT examined factual recall, concentration and manual dexterity of resident surgeons when sleep-deprived or not. Reznick & Folze (1987) used a random number table to prospectively assign surgical residents from a single centre to one of two groups according to whether they would be tested in the sleep-deprived or non sleep-deprived condition first. Allocation concealment, power calculations and intention-to-treat analysis were not reported. Participant eligibility, selection and participation rates were not reported. Twelve participants were lost after randomisation, but the reasons for these losses were not reported. No blinding of assessors was reported. The factual recall test was reported to be stable and reliable; the concentration ability and manual dexterity tasks were found to be stable but not reliable; and the Purdue pegboard task was found to be reliable. Inclusion criteria included residents in training in the department of surgery. No exclusion criteria or participant demographics were reported. It was not reported whether the assessment tasks had been previously validated. It was reported that there was no learning effect for the manual dexterity task.

Non-randomised comparative studies

One prospective non-randomised comparative study investigated cognitive performance of 21 paired surgical house staff when rested or not rested (Light et al 1989). It was not reported whether the study was a single- or multi-centre, whether there were any differences between the groups, or how participants were recruited. The study period and participant demographics were also not reported. It was not reported whether the Purdue Pegboard had been validated.

Case series (pre-test/post-test outcomes)

Three case series (pre-test/post-test outcomes) studies reported performance outcomes when participants were rested or not rested (Deaconson et al 1998; Deary & Tait 1987; Wesnes et al 1997). All of these studies were conducted in single centres (Deaconson et al 1998; Deary & Tait 1987; Wesnes et al 1997). Two studies reported recruiting volunteers from surgical training programs (Deary & Tait 1987; Wesnes et al 1997). One study did not report how participants were recruited (Deaconson et al 1998). Not one study reported losses to assessment or the reasons for the losses. Three studies did not report using validated assessment tools (Deaconson et al 1988; Deary & Tait 1987; Wesnes et al 1997).

Deaconson et al (1998) reported using a study design to detect a difference of 10% or less for each of the five psychometric tests. When comparing sleep-deprived and non sleep-deprived participants, there was a 90% chance (power equals 0.90) of detecting a difference for each of the five variables when there was a 10% difference between the two sleep states. No study reported elements of blinding (when appropriate).

Besides reporting who was included in the study, no further inclusion or exclusion criteria were reported for any of the studies. Gender was the only reported participant demographic in one study (Deary & Tait 1987). Another study reported age and gender, with 25/26 (96%) being males who were predominately right handed (Deaconson et al 1988). One study did not report any participant demographics (Wesnes et al 1997). The study period was reported in one study (Deaconson et al 1998), and not in another study (Wesnes et al 1997). The year was not reported in one study (Deary & Tait 1987).

Psychomotor skill performance

Randomised controlled trials

One multi-centre study examined performance in a pelvic trainer after participants had conducted a day of work compared with a night on-call (Uchal et al 2005). A computer generated random sample was used to assign subjects to either post-call or post-work study arms. Allocation concealment was ensured by giving identity numbers to the participants, and the generator of assignment was separated from its executor. A power calculation from a previous pilot study was used to determine the sample size. An 80% power parallel block randomisation design at alpha = 0.05 indicated that 60 subjects would be needed to detect significant differences in operating time. Outcome measures were evaluated by two assessors blinded to participants' identity and arm assignment. Intention-to-treat analysis

was not reported. Interrater reliability was ≥ 0.75 (Kendall's tau concordance coefficient). Accuracy error, tissue damage, leak rates, goal-directed actions, non goal-directed actions, and operating time were previously tested for validity and reliability. Two patients withdrew after the pre-test and were excluded from the results. The study period was not reported. No exclusion criteria were reported. Construct validity was determined by the ability of the simulator to distinguish between trained (surgeons) and untrained (nurses) subjects. The nurses in the post-call arm and in the post-work arm were well matched for age, gender, practice duration, and ESS and MIST-VR scores, but not for hours slept in the previous 24 hours (post-call 1.3 (0 – 3 hours) vs post-work 7.2 (5 – 10 hours), p = 0.04).

Case series (pre-test/post-test outcomes)

Seven prospective case series studies measured differences during rested and unrested states. Five of the seven studies reported being single centre studies (DeMaria et al 2005; Eastridge et al 2003; Grantcharov et al 2001; Jakubowicz et al 2008; Taffinder et al 1998). The other two studies did not report whether they were single- or multi-centre studies (Kahol et al 2008; Leff et al 2008). All of the studies reported using volunteers or recruited people on a particular surgical rotation. Losses to follow up were not reported in all but one study, but the reasons for the losses were not reported (DeMaria et al 2005). The study period was reported by two of the studies (Eastridge et al 2003; Jakubowicz et al 2005), but not by the other studies.

Assessment of psychomotor skill was done by the use of simulators which calculated performance scores for each parameter, limiting assessor bias. Outcome measures were previously validated in six (DeMaria et al 2005; Eastridge et al 2003; Grantcharov et al 2001; Leff et al 2008; Taffinder et al 1998) of the seven studies (although of these, DeMaria et al (2005) and Grantcharov et al (2001) did not explicitly state that the methods had been previously validated). One study reported using outcome measures of which many, but not all, had been previously validated (Kahol et al 2008). One study did not report whether the outcomes measures had been validated or not (Jakubowicz et al 2005). Inclusion criteria were limited to who had been recruited into the study. No exclusion criteria were reported for any study. It was reported by four of the seven studies that participants had varying degrees of experience in laparoscopic surgery (DeMaria et al 2005; Grantcharov et al 2001; Leff et al 2008) and simulator use (Eastridge et al 2003). Demographic characteristics were limited to reporting participants' levels of training/years of training, age and gender, or a combination of these characteristics.

4. Results

The results are presented in four main sections according to the type of performance that was assessed: clinical, academic, cognitive, and psychomotor skill. The results are then categorised by the level of the evidence.

Clinical performance

Three studies reported clinical outcomes for when a surgeon was rested or not (Ellman et al 2004 & 2005; Haynes et al 1995; Schieman et al 2008), and one study reported self-assessed clinical performance (Sawyer et al 1999).

Retrospective patient complication rates

Four retrospective studies compared patient complication rates with either the sleep deprivation of the operating surgeon (Ellman et al 2004 & 2005), the on-call status of the operating surgeon (Haynes et al 1995), or whether the surgeon was fatigued or not (Schieman et al 2008).

Non-randomised comparative studies

Ellman et al (2004) conducted a review of all cases performed by attending cardiac surgeons in one hospital between January 1994 and April 2003. Ellman et al (2005) conducted a review of all cases performed by thoracic surgical residents between January 1994 to March 2004 at the same hospital. A surgeon was considered sleep-deprived if he or she performed a case: that started between 2200 h and 0500 h (Ellman et al 2004), or 2300 h and 0500 h (Ellman et al 2005); ended a case between 2300 h and 0730 h (both studies); or performed a subsequent case within the next 24 hours (both studies). Schieman et al (2008) conducted a review of all patients undergoing anterior resection for rectal cancer between 1994 and 2005 at two institutions and compared complication rates with the time of the operation. A surgeon was considered fatigued if he or she billed for clinical work after 2200 h the night before the operation.

In these three studies, complication rates of cases performed by sleep-deprived surgeons and non sleep-deprived surgeons were compared. There were no significant differences in patient mortality rates when the surgeon was sleep-deprived or not (Table 6).

Table 6. Mortality rates for operations performed by sleep-deprived and non sleep-deprived surgeons

Study	Operation	Mortality, n/N (%)		P-value
		Cases performed by SD surgeons	Cases performed by non-SD surgeons	
Ellman et al 2004 Level III-2		n = 339	n = 6412	
	CABG	4/223 (2%)	133/4206 (3%)	NS
	Valve	2/45 (4%)	32/443 (4%)	NS
	CABG-valve	3/22 (13%)	34/443 (8%)	NS
	Other	8/49 (16%)	97/939 (10%)	NS
Ellman et al 2005 Level III-2		n = 229	n = 7094	
	CABG	3/141 (2%)	143/4452 (3%)	NS
	Valve	0/36 (0%)	44/890 (5%)	NS
	CABG-valve	2/12 (14%)	38/483 (7%)	NS
	Other	7/28 (20%)	98/946 (9%)	NS
Schieman et al 2008 Level III-2		n = 22	n = 248	
	Anterior rectal resection	0/22 (0%)	3/248 (1%)	NS

CABG, coronary artery by-pass; SD, sleep-deprived; NS, not significant; NSD, not sleep-deprived

There were no significant differences in operative variables and complications for two studies (Table 7).

In both Ellman et al 2005 & 2005, univariate analyses of all neurologic, renal, pulmonary, and infectious complications were performed. These analyses failed to demonstrate any significant differences between the patients treated by sleep-deprived and non sleep-deprived groups. In Ellman et al (2005), forwards and backwards logistic regression analysis was performed to determine whether surgeon sleep deprivation alone was a risk factor for mortality.

This analysis demonstrated that the type of operation performed, presence of an operative complication, male sex, age, and length of hospital stay were all significantly associated with mortality ($p < 0.001$), but that sleep deprivation was not ($p = 0.62$).

Haynes et al (1995) examined complications of surgical procedures from one institution from January 1985 to April 1988 and the on-call status of the operating surgeon (Table 8). Complication rates were consistently higher for emergency cases than for non-emergency cases. It was found that the lowest incidence of complications occurred when residents were operating on days when they were not on-call and had not been on-call the night before (ie not sleep-deprived), and that the highest complication incidence occurred when residents operated the day after being on-call but were not on-call again that day (ie sleep-deprived). No statistical analyses were conducted on these data. When residents operated on a not on-call day, complications were 45% more likely to occur when the resident had been on-call the day before ($p < 0.02$) compared to when they were not.

Table 7. Operative data and complications for operations performed by sleep-deprived and non sleep-deprived surgeons

Study	Perioperative data and complications	Cases performed by SD surgeons	Cases performed by non-SD surgeons	P-value
Ellman et al 2004 Level III-2		n = 339	n = 6412	
	Pulmonary by-pass time, mean ± SE	107.7 ± 3.0	107.4 ± 0.7	0.91
	Aortic cross-clamp time, mean ± SE	74.9 ± 2.4	73.6 ± 0.5	0.56
	Blood products (yes) (%)	49%	49%	0.82
	Operation to discharge, days mean ± SE	7.1 ± 0.4	6.9 ± 0.1	0.58
	Operative complications, n (%)	29 (9%)	480 (7%)	0.47
	Neurologic complications, n (%)	53 (16%)	809 (13%)	0.11
	Renal complications, n (%)	25 (7%)	480 (7%)	0.94
	Pulmonary complications, n (%)	69 (20%)	1232 (19%)	0.6
	Infectious complications, n (%)	23 (7%)	421 (7%)	0.87
Ellman et al 2005 Level III-2		n = 229	n = 7094	
	Pulmonary by-pass time, mean ± SE	106.5 ± 3.6	107.4 ± 0.7	0.92
	Aortic cross-clamp time, mean ± SE	74.4 ± 2.4	74.1 ± 0.6	0.83
	Blood products (yes) (%)	49%	49%	0.91
	Operation to discharge, mean days ± SE	7.5 ± 0.6	7.7 ± 0.1	0.83
	Operative complications, n (%)	15 (7%)	541 (8%)	0.76
	Neurologic complications, n (%)	35 (15%)	840 (12%)	0.16
	Renal complications, n (%)	14 (6%)	502 (7%)	0.71
	Pulmonary complications, n (%)	39 (17%)	1342 (19%)	0.66
	Infectious complications, n (%)	15 (7%)	467 (7%)	0.81
Schieman et al 2008 Level III-2		n = 22	n = 248	
	Intraoperative complications, n (%)	3 (14%)	44 (18%)	0.6
	Estimated blood loss > 1 L, n	2	16	NR
	Injury to other organs, n	0	13	NR
	Spillage of stool, n	0	11	NR
	Stapler misfire/ anastomotic problem, n	0	4	NR
	Tumour transection, n	1	0	NR
	Mean estimated blood loss, mean ml	448	430	0.8
	Operative time, mean min	114	148	0.5
	Postoperative complication rate, n (%)	12 (55%)	164 (66%)	0.27
	Major complication rate (≥ 3), n (%)	2 (9%)	37 (15%)	0.68
	Mean length of stay, mean days	9.3	11.4	0.29
	Leak/abscess rate, n (%)	2 (9%)	24 (9.6%)	0.9
	Long term complications, n (%)	7 (31%)	77 (31%)	0.9

SD, sleep-deprived; NSD, not sleep-deprived

When Haynes et al (1995) examined complication rates by resident call status, the complication rate for emergency procedures was lowest when done on-call (10%), and the highest rate occurred in procedures where the resident had not been on-call the day of, or the day before the operation (ie rested) (12%). Complication rates in the non-emergency group were lowest in the not on-call either day group (3%) and highest in the post-call group (4%). Although the authors stated that neither difference was statistically significant, no statistical analyses were reported on these data.

Table 8. Complication* rates related to emergency/ non-emergency status of cases and residents' call status

Haynes et al 1995 Level III-2		Sleep-deprived			Rested	Total
		On-call day before case, and on-call day of case	On-call day before case, not on-call day of case	On-call day of case, not on-call day before case	Not on-call either day	
	Emergency cases					
	Number of procedures, n	NR	677	742	428	NR
	Complications, n (%)	NR	75 (11.1)	71 (9.6)	50 (11.7)	NR
	Non-emergency cases					
	Number of procedures, n	NR	1022	1222	2280	NR
	Complications, n (%)	NR	39 (3.8)	41 (3.4)	75 (3.3)	NR
	Total					
	Number of procedures, n	141	1728	1969	2703	6541
	Complications, n (%)	8 (5.7)	114 (6.6)	112 (5.7)	125 (4.6)	359 (5.5)

*A complication was defined as any adverse intraoperative of postoperative effect or outcome submitted as a complication to the weekly death and complications conference at or shortly after the time of surgery. All complications were recorded by senior resident and attending surgeon. Surgical mortality, either as a direct result of the operation of because of the occurrence of a complication was not considered.

Haynes et al (1995) further reported that the mean complication incidence was lower in operations done by residents during their internship year (2%) and rose to 14% for PGY 5. Complication rates for PGY 1, 2 and 3 were similar, as were those for PGY 4 and 5. Complication rates were higher for upper-level residents than junior residents and interns (11% vs 2%, $p < 0.001$) (likely to be attributed to senior residents dealing with more complex procedures prone to a greater likelihood of complications).

Self-assessed performance

Case series

One study reported results of weekly sleep/operative logs and monthly surveys for participants that were on-call every second, third or fourth night (Sawyer et al 1999). Being on-call every other night was associated with significantly greater levels of fatigue ($p \leq 0.05$) and stress ($p \leq 0.05$) and less operating room participation (p-value not reported) and overall satisfaction ($p \leq 0.05$), when compared with the every fourth night with a cross-cover schedule. The frequency of reported errors made while on-call was similar between groups. Multiple regression analysis revealed that the number of errors made while on-call was most strongly related to on-call fatigue ($p \leq 0.001$). Analysis of other variables, suggested that on-call fatigue was related to off-call fatigue ($p \leq 0.001$) and the number of sleepless nights per month ($p \leq 0.01$) but not to call schedule. Operating room participation was inversely related to the number of call nights taken per week ($p \leq 0.001$) and the degree of fatigue experienced while not on-call ($p \leq 0.01$), whereas stress was related to fatigue off-call ($p \leq 0.001$) and patient census ($p \leq 0.05$). Overall satisfaction was most strongly associated with infrequent call ($p \leq 0.01$) and operating room participation ($p \leq 0.05$). The number of hours of sleep obtained while on-call was related only to service census by multiple regression analysis ($p = 0.002$). Attending physicians noted little difference in intern fatigue on the basis of the on-call schedule.

Summary of clinical performance results

1. Three non-randomised comparative studies failed to demonstrate any significant clinical differences between the sleep-deprived and non sleep-deprived groups. One study found that when residents operated on a not on-call day, complications were significantly more likely to occur when the resident had been on-call the day before.
2. One case series study found being on-call every other night was associated with subjectively reported significantly greater levels of fatigue and stress, less operating room participation and less overall satisfaction, when compared with the every fourth night with cross-cover schedule, but not with the frequency of reported errors.

Academic performance

Two retrospective studies investigated surgical resident examination performance with on-call status (Minion et al 2007; Stone et al 2000).

American Board of Surgery In-Training Examination (ABSITE) scores

The ABSITE is a written, multiple-choice examination designed to measure surgical residents' knowledge of basic science and the management of clinical problems related to general surgery. Clinical directors use ABSITE scores as an evaluation instrument to assess residents' progress.

Non-randomised comparative studies

Minion et al (2007) analysed ABSITE scores from 1999 to 2006 from a single institution. Of the 282 ABSITE scores, 69 (24%) residents were on-call the night before the examination. Although standard scores improved significantly with each subsequent year of training (ie from PGY 1 to PGY 5) ($p < 0.001$), no statistically significant differences were found in standard scores or percentile scores at any PGY level attributable to call status.

Mean percentile rankings of the categorical and preliminary residents were compared using the Mann-Whitney test, and showed that there was a significant difference in ABSITE scores (mean rank; categorical residents 123/208 and preliminary residents (PGY 1-3) 84/208 ($p < 0.001$). The mean rank of the categorical residents on-call was 51/109 and 56/109 for those residents not on-call, which was not statistically different. A similar analysis for the preliminary residents also found that call status had no effect on ABSITE performance.

Scores for general surgery residents were matched for years and with on-call status, and there was no statistically significant difference. The best predictor of ABSITE score was found to be PGY level, followed by the United States Medical Licensing Examination (USMLE) step 2 and categorical versus preliminary resident.

Stone et al (2000) obtained ABSITE scores from 1994 from multiple institutions and also compared them with the call status of residents. Of the 424 ABSITE scores, 70 (17%) residents were on-call the night before the examination. For all residents examined in the aggregate, there were statistically significant differences in total test and clinical

management scores between residents off-call and on-call ($p < 0.02$), but not for the basic science component. Categorical residents performed significantly better than preliminary residents in all measures of all three test components ($p < 0.01$). When analysed alone, off-call residents performed significantly better in terms of standard score on both the total test and the clinical management component ($p < 0.01$). There was no statistical difference between off-call and on-call residents on the basic science component. There was no significant difference between on-call and off-call residents at each PGY level, except for PGY 2 ($p < 0.03$). The association between PGY level and score appeared linear and was modelled as a linear effect. ANCOVA revealed that differences in training track (categorical verses preliminary, with categorical residents scoring higher in all test areas) and PGY level contributed significantly to the differences in scores. Call status was no longer a significant factor in clinical management, basic science, or total test after adjusting for PGY level and track. Call status was a significant factor in the variation of clinical management scores only in the PGY 2 cohort after adjusting for track ($p = 0.02$). Individual training program was not a significant contributing factor. Similar ANCOVA analyses for all residents and for categoricals only in each PGY level and within junior (PGY 1 – 3) and senior (PGY 4 - 5) revealed the importance of training track and PGY level, but no significant effect of being on-call. Four program directors (for 118 residents) reported that they changed their program's call schedule for the ABSITE, giving all of their residents the night off before the examination. Performance of residents in PGY 1 to PGY 5 that adjusted either their call or exam schedule before the ABSITE examination were compared with all other residents and with all others who were off-call and there was no significant difference.

Summary of academic performance results

- Two non-randomised comparative studies found that being on-call the night before the ABSITE did not affect performance when compared with those not on-call the night before the examination.

Cognitive performance

Four studies reported cognitive performance outcomes (Reznick & Folse 1987; Deary & Tait 1987; Deaconson et al 1989; Light et al 1989; Wesnes et al 1997).

Randomised controlled trials

One RCT reported cognitive outcomes. Reznick & Folse (1987) reported no significant differences in factual recall and concentration when participants were sleep-deprived or not (Table 9). There was no significant effect in the experimental condition ($p = 0.90$); group membership ($p = 0.73$); or learning effect ($p = 0.11$), for the 12 subjects that performed under both conditions. For combined speed and accuracy scores for the concentration ability task, split plot ANOVA revealed no significant condition effect, group membership effect or learning effect for either accuracy score or combined speed and accuracy score. When the results of performance of all 21 subjects were analysed, there were no significant differences between the mean scores in the two conditions ($p = 0.81$).

Non-randomised comparative studies

One non-randomised comparative study reported that when the in an acute sleep-deprived state there were no differences within or between resident levels on the functional testing of clear thinking, judgement, memory and learning, although no actual data were reported (Light et al 1989).

Case series (pre-test/post-test)

Three case series reported cognitive outcomes (Deaconson et al 1989; Deary & Tait 1987; Wesnes et al 1997) (Table 9).

Table 9. Cognitive performance

Reznick & Folse 1987 Level II N = 21	Measure	Not sleep-deprived (n = 21), mean score ± SD		Sleep-deprived (n = 12), mean score ± SD		P-value	
	Factual recall	23 ± 5		22 ± 5		NS*	
	Concentration (combined speed and accuracy)	54 ± 4		55 ± 4		NS*	
Deaconson et al 1989 Level IV N = 26	Measure	Sleep-deprived (mean ± SD) n = 26		Not sleep-deprived (mean ± SD) n = 26		P-value	
	Trail-Making	41.4 ± 17.2		40.2 ± 16.4		NS	
	Grammatical Reasoning	20.1 ± 5.8		21.2 ± 5.4		< 0.05	
	Minnesota Paper Form Board	12.4 ± 2.4		12.6 ± 2.2		NS	
	Paced Auditory Serial Addition	78.2 ± 15.9		81.5 ± 10.2		< 0.05	
Deary & Tait 1987 Level IV N = 12	Measure	Test score, mean ± SD					
		Off-duty	P-value (vs waiting)	On-call	P-value (vs off-duty)	Waiting	P-value (vs on-call
	Digit span (forward)	6.7 ± 0.9	NS	6.8 ± 1.1	NS	6.4 ± 0.9	NS
	Digit span (backward)	5.0 ± 1.1	NS	5.2 ± 0.9	NS	5.1 ± 1.4	NS
	Serial 13s	57.6 ± 20.6	< 0.01	56.0 ± 22.8	< 0.01	57.4 ± 21.4	< 0.01
	Logical memory (immediate)	8.7 ± 2.9	0.05	7.3 ± 3.9	< 0.05	6.7 ± 2.8	< 0.05
	Logical memory (delay)	10.7 ± 4.9	< 0.01	10.1 ± 4.5	NS	9.3 ± 4.6	NS
	Laboratory reports (time to sort)	330.1 ± 117.3	< 0.01	392.2 ± 171.7	< 0.05	363.5 ± 170.6	< 0.01
	Laboratory reports (errors)	4.33 ± 2.1	NS	4.6 ± 1.6	NS	4.5 ± 2.8	NS
	Electrocardiographic diagnosis	12.6 ± 2.2	NS	12.2 ± 3.0	NS	11.5 ± 3.7	NS
Wesnes et al 1997 Level IV N = 10	Measure	Test score, mean ± SEM				P-value	
		Off-duty		On-duty			
	Vigilance sensitivity index	0.97 ± 0.11		0.92 ± 0.2		0.02	
	Overall speed on attentional tasks (ms)	1155 ± 30		1210 ± 30		0.05	

NS, not significant; SD, standard deviation; SEM, standard error of the mean

*There were also no significant differences when only the 12 people who completed both arms of the study were used for comparison

Deaconson et al (1989) stated that there were no differences within cohort or overall median performance scores between the sleep-deprived and non sleep-deprived groups for cognitive tests (Table 9). When comparing the reported mean scores, participants performed better in the sleep-deprived state than in the non sleep-deprived state for two of the four tests ($p < 0.05$). There was a learning effect (improvement of tested performance) during the study for both sleep-deprived and non sleep-deprived states. When the results

of performance by each resident on each test was analysed using Mann-Whitney *U* test, there were no differences in performance between the sleep-deprived and non sleep-deprived residents. Further analysis of the correlation between sleep parameters (total sleep and longest interval of uninterrupted sleep) and performance of each component of the psychometric test battery identified changes in performance on some tests but only trivial effects due to sleep.

Deary & Tait (1987) reported significant decrements in three of the eight measured tests when participants were on-call and off-duty ($p \leq 0.05$); and in four of the eight measured tests when participants were off-duty compared with waiting ($p \leq 0.05$) (Table 9). One way ANOVA for repeated measures found significant differences between subject effects for the serial 13s test ($p < 0.01$), logical memory (immediate) ($p < 0.01$), logical memory (delay) ($p < 0.01$), laboratory reports (sorting time) ($p < 0.01$), and laboratory reports (sorting errors) ($p < 0.05$).

Wesnes et al (1997) reported adverse effects for overall attentional speed (calculated from the three attentional tasks) ($p = 0.05$) and vigilance sensitivity ($p = 0.02$) (Table 9). Impaired performance was correlated with the number of hours the participants had worked on Sunday nights ($p < 0.05$). Vigilance sensitivity was positively correlated with the amount of sleep during the weekend ($p = 0.05$); it was negatively correlated with the number of times paged ($p = 0.004$) and the total amount of weekend duty ($p = 0.02$).

Summary of cognitive performance results

- RCT evidence (one study) indicated that sleep deprivation had no effects on factual recall and concentration.
- One non-randomised comparative study reported no differences within or between residents in relation to clear thinking, judgement, memory and learning when residents were acutely fatigued.
- Evidence from three case series studies suggested that were some variations in cognitive performance when participants were tired, but only for some variables in some studies, or only for certain individuals. One study indicated that there was a learning effect for both sleep-deprived and non-sleep deprived states.

Psychomotor skill performance

Simulated performance

There were 11 studies that assessed performance using simulation-based methods when a participant was rested and/or fatigued (Deaconson et al 1988; DeMaria et al 2005; Eastridge et al 2003; Grantcharov et al 2001; Jakubowicz et al 2005; Kahol et al 2008; Leff et al 2008; Light et al 1989; Reznick & Folse 1987; Taffinder et al 1998; Uchal et al 2005). [Please note that although Deaconson et al (1988), Light et al (1989), and Reznick & Folse (1987) appear within these results, their main findings have been categorised under cognitive performance outcomes].

Overall performance scores

Three studies provided details regarding overall performance of simulated skills after being on-call (DeMaria et al 2005; Jakubowicz et al 2005; Kahol et al 2008).

Non-randomised comparative studies

DeMaria et al (2005) reported significant improvements in overall performance for all participants in four of six tasks after being on-call ($p < 0.05$) (Table 10). For all participants, 17/58 (29%) of the tested parameters statistically improved when tested post-call ($p < 0.05$). One parameter statistically deteriorated. When examined separately, junior residents significantly improved in one task ($p < 0.05$). These residents statistically improved in 12/58 (21%) of the tested parameters ($p < 0.05$).

Jakubowicz et al (2005) reported no significant differences in overall score between pre-call and post-call performance (Table 10). When the trials within each condition was analysed separately, there was a significant difference post-call between trials one and two ($p = 0.05$) indicating a learning effect.

Kahol et al (2008) reported that for tasks that combine both psychomotor and cognitive skills, night call led to significantly decreased surgical proficiency (specific data reported in next sections).

Table 10. Simulated performance - overall scores

DeMaria et al 2005 Level III-2	MIST-VR	N	Statistically significant improvement in post-call results from baseline					
			Task 1	Task 2	Task 3	Task 4	Task 5	Task 6
	Total score* (all)	17	< 0.05	< 0.05	< 0.05	< 0.05	NS	NS
	Total score (junior residents only)	10	NS	NS	< 0.05	NS	NS	NS
	MIST-VR	N	Number of tested parameters that					
			Statistically improved ($p < 0.05$)			Statistically deteriorated ($p < 0.05$)		
	All participants	17	17/58 (29.3%)			1/58 (1.7%)		
	Junior residents only	10	12/58 (20.7%)			1/58 (1.7%)		
Jakubowicz et al 2005 Level III-2	ESS	N	Pre-call†		Post-call†		P-value	
	Overall score, mean	8	80.2		80.4		0.92	
	ESS	N	Pre-call†			Post-call†		
	Overall score, mean	8	Trial 1	Trial 2	P-value	Trial 1	Trial 2	P-value
			78.1	82.3	0.34	78.8	82.1	0.05

ESS, Endoscopic Sinus Surgery Simulator; MIST-VR, Minimally Invasive Surgery Trainer - Virtual Reality; NS, not significant
* Total score is the sum of the individual component scores. Parameters were weighted to reflect their importance.
† Measures of variability not clearly recorded.

Performance time

Of the studies that assessed performance using simulation-based methods before and after call, all reported time outcomes.

Randomised controlled trials

Reznick and Folse (1987) reported combined speed and accuracy scores for the manual dexterity simulated wound task. Split plot ANOVA revealed no significant condition effect, group membership effect or learning effect for either of the two scores. An analysis of all 21 subjects revealed no significant difference in the two experimental conditions. The

indices used for these tests correlated significantly in both the sleep-deprived condition ($p < 0.001$) and the non sleep-deprived condition ($p < 0.001$).

Uchal et al (2005) reported no significant differences in operating time of suturing in a pelvic trainer after participants were assessed after a day of work (assessed between 1600 h and 1700 h after work) and post-call (assessed between 0800 h and 0900 h after a night on-call) (Table 11).

Non-randomised comparative studies

Eastridge et al (2003) found no differences in completion time for the six simulated tasks combined (Table 11).

Grantcahrov et al (2001) reported significant increases in time to complete five of the six tasks ($p \leq 0.006$ for all) (Table 11).

Jakubowicz et al (2005) reported that overall there were no significant differences in performance time and diathermy time, but when assessed separately, there were significant differences in these times in trials 1 and 2 of the post-call assessment ($p = 0.08$ and 0.07 respectively)(Table 11).

Kahol et al (2008) reported a decrease in the time it took participants to perform the simulated tasks post-call ($p < 0.0002$). A sub analysis was performed to study the differences between pre-call and post-call performance for exercises that primarily involved psychomotor skill versus exercises those that involved a combination of psychomotor and cognitive skills. There was no difference in time elapsed pre-call and post-call (Table 11).

DeMaria et al (2005) reported statistically significant improvement for three of the six tasks post-call when all participants were combined ($p < 0.05$ for all). When junior residents were analysed alone, there was a significant improvement for the third task only ($p < 0.05$) (Table 11).

Taffinder et al (1998) compared a night of undisturbed sleep (control) with a sham night on-call and a night with no sleep. Repeated-measures ANOVA for a Latin square showed a significant linear trend across the sleep conditions compared with baseline for total time ($p = 0.009$). Surgeons awake all night took 14% longer to complete the tasks than those who had a full night's sleep (Table 11).

Leff et al (2008) reported the results of simulation-based assessment after consecutive nightshifts compared with baseline results. Increased time was required to perform core skills (CS) 1 after the first night shift ($p = 0.002$). Increased time was required to perform CS 2 after nightshifts one, two and three ($p \leq 0.048$) (Table 11).

Table 11. Simulated performance - pre-call and post-call performance time differences

Uchal et al 2005 Level II	Pelvic trainer	N	Post-work (n = 32)			Post-call (n = 32)			P-value			
	Pelvic trainer, Operating time, median s	64	365			381			NS			
Jakubowicz et al 2005 Level III-2	ESS	N	Pre-call*			Post-call*			P-value			
	ESS, overall time, mean s	8	574.1			535.4			NS			
	ESS, dissection time, mean s	8	419.3			390.3			NS			
	ESS	N	Pre-call*			Post-call*						
			Trial 1	Trial 2	P-value	Trial 1	Trial 2	P-value				
	ESS, overall time, mean s	8	563.6	584.6	NS	568.9	501.9	0.08				
	ESS, dissection time, mean s	8	402.4	436.3	NS	423.4	357.3	0.07				
Kahol et al 2008 Level III-2	ProMIS, FLS	N	Pre-call			Post-call†			P-value			
	ProMIS, FLS, time elapsed, units not reported	37	NR			Decreased			< 0.0002			
Eastridge et al 2003 Level III-2	SCMIS, MIST-VR	N	Pre-call			Post-call			P-value			
	MIST-VR, total time, mean s ± SEM	35	65 ± 0.3			74 ± 0.3			NS			
Grantcharov et al 2001 Level III-2	MIST-VR	N	Pre-call	Post-call	Pre-call	Post-call	Pre-call	Post-call	Pre-call	Post-call	Pre-call	Post-call	Pre-call	Post-call	Pre-call	Post-call
			Task 1		Task 2		Task 3		Task 4		Task 5		Task 6	
	MIST-VR, total time, median s (range)	14	5 (4-9)	8 (5-20)	7 (5-14)	9 (5-14)	6 (4-11)	8 (4-21)	7 (5-12)	8 (6-14)	15 (12-18)	18 (14-23)	18 (13-22)	24 (16-30)
	P-value		0.006		NS		0.004		0.003		0.003		0.002	
DeMaria et al 2005 Level III-2	MIST-VR	N	Statistically significant improvement in post-call results from baseline											
			Task 1		Task 2		Task 3		Task 4		Task 5		Task 6	
	MIST-VR, time (all), units not reported	17	< 0.05		NS		< 0.05		< 0.05		NS		NS	
	MIST-VR, time (jr residents only), units not reported	10	NS		NS		< 0.05		NS		NS		NS	
Taffinder et al 1998 Level III-2	ICSAD, MIST-VR	N	Control			Sham on-call			No Sleep			
	Time, mean ± SE	6	-1.0 ± 0.4			-0.5 ± 0.4			0.5 ± 0.5			
	P-value (compared with control)‡		-			NR§			NR‖			
Leff et al 2008 Level III-2	MIST-VR	N	Base-line	Night 1	Night 2	Night 3	Night 4	Night 5	Night 6	Night 7	Follow up	
	CS 1 Median total time, s ± SD	21	33 ± 4.6	42 ± 8.4	34 ± 7.5	35 ± 5.5	32 ± 7.4	33 ± 5.0	33 ± 6.9	32 ± 10.0	32 ± 5.3	
	P-value (compared with baseline)		-	0.002§	NS	NS	NS	NS	NS	NS	NS	
	CS 2 Median total time, s ± SD	21	36 ± 9.7	40 ± 10.2	39 ± 9.4	39 ± 4.2	37 ± 5.2	38 ± 3.3	37 ± 9.8	37 ± 8.6	37 ± 11.4	
	P-value (compared with baseline)		-	0.014§	0.048§	0.032§	NS	NS	NS	NS	NS	

CS, core skills; ICSAD, Imperial College Surgical Assessment Device; MIST-VR, Minimally Invasive Surgery Trainer – Virtual Reality; NS, not significant; SD, standard deviation; SEM, standard error of the mean

* Measures of variability were not clearly reported.

† Raw data difficult to extract from data supplied in study. Time to complete task improved post-call.

‡ Repeated-measures ANOVA for a Latin square showed a significant linear trend across the sleep conditions compared with baseline for total time (p = 0.009).

§ Worse than baseline.

‖ Surgeons awake all night took 14% longer to complete the tasks than those who had a full night's sleep, and also showed increased stress and decreased arousal, which paralleled the decrease in operative dexterity.

Performance errors

Six studies reported performance errors when a surgeon was assessed pre-and post-call (Uchal et al 2005; DeMaria et al 2005; Eastridge et al 2003; Grantcharov et al 2001; Kahol et al 2008; Leff et al 2008; Taffinder et al 1998). Reported performance error data are shown in Table 12.

Randomised controlled trials

Uchal et al (2005) reported no significant differences in error outcomes pre- and post-call.

Non-randomised comparative studies

Eastridge et al (2003) reported a significant increase in errors post-call ($p < 0.001$).

DeMaria et al (2005) reported that one task significantly improved for all residents post-call ($p < 0.05$), and that for junior residents, two tasks significantly improved, and one significantly deteriorated ($p < 0.05$ for all).

Grantcharov et al (2001) reported significant increases in errors for two of the six tasks ($p \leq 0.01$).

In Kahol et al (2008), residents were grouped according to experience levels, and their pre-call and post-call performance data were compared using ANOVA. Performance of PGY 1 – 3 residents was consistently worse for cognitive errors compared with that of PGY 4 – 5 residents ($p < 10e\text{-}6$). An analysis was performed to determine the correlation between measures of proficiency and fatigue (self-reported on a scale of 0 (least fatigued) to 10 (most fatigued). The normalised rating was obtained by taking the weighted average of the three ratings in the questionnaire, with the current rating of fatigue rating being given twice the weight of the minimum and maximum fatigue ratings. The normalised ratings for pre-call and post-call sessions were subtracted to obtain changes in fatigue caused by night call. With increased fatigue ratings, residents tended to make more errors. Similar results were seen with reported hours of sleep and cognitive errors (more cognitive errors as sleep decreased). A correlation coefficient of -0.89 was reported.

Leff et al (2008) reported significantly more errors than at baseline after the first night shift for CS 1 only ($p = 0.025$).

Taffinder et al (1998) reported that the control group made the least errors, followed by the on-call group and then the no sleep group. There was a significant linear trend across the sleep conditions compared with baseline for error score ($p = 0.009$).

Table 12. Simulated performance - pre-call and post-call performance errors

Uchal et al 2005 Level II	Pelvic trainer	N	Post-work (n = 32)			Post-call (n = 32)			P-value			
	Accuracy error, median mm	64	0.5			1.0			0.39			
	Tissue damage, median mm	64	2.2			2.2			1.0			
	Leak rates (%)	64	65.6			56.3			0.61			
Eastridge et al 2003 Level III-2	SCMIS, MIST-VR	N	Pre-call			Post-call			P-value			
	Errors, mean ± SEM	35	5.9 ± 0.3			11.8 ± 0.6			< 0.001			
DeMaria et al 2005 Level III-2	MIST-VR	N	Post-call results*									
			Acquire			Transfer			Traversal		Withdraw	
	Error, all residents	17	NS			NS			Improved		NS	
	Error (junior residents only)	10	Deteriorated			Improved			Improved		NS	
			Diathermy			Manipulation						
	Error, all residents	17	NS			NS						
	Error (junior residents only)	10	NS			NS						
Grantcharov et al 2001 Level III-2	MIST-VR		Task 1		Task 2		Task 3		Task 4		Task 5	
			Before	After	Before	After	Before	After	Before	After	Before	After
	Errors, median (range)	14	1 (0-3)	1(0-6)	0 (0-2)	1 (0-5)	2 (0-5)	3 (0-7)	0 (0 2)	1 (0-3)	1 (0-3)	1 (0-7)
	P-value		0.013		0.44		0.35		0.72		0.061	
			Task 6									
			Before	After								
	Errors, median (range)	14	1 (0-5)	4 (0-12)								
	P-value		0.005									
Kahol et al 2008 Level III-2	ProMIS, FLS	37	Post-call									
	Cognitive errors for PGY 1 – 3	NR	NR†									
	Cognitive errors for PGY 4 – 5	NR	NR†									
	P-value		< 10e-6									
Leff et al 2008 Level III-2	MIST-VR		Base-line	Night 1	Night 2	Night 3	Night 4	Night 5	Night 6	Night 7	Follow up	
	CS 1, total error score	21	86 (4.5)	100 (12.3)	83 (25.1)	67 (21.8)	67 (30.8)	71 (23.3)	78 (26.0)	84 (21.7)	64 (22.1)	
	P-value (compared with baseline)		-	0.025	0.896	0.984	0.825	0.422	0.811	0.629	0.777	
	CS 2, total error score	21	4 (1.7)	7 (3.6)	5 (2.8)	4 (3.3)	5 (2.8)	3 (4.5)	5 (4.4)	4 (0.6)	4 (4.0)	
	P-value (compared with baseline)		-	0.511	0.736	0.585	0.716	0.938	0.775	0.376	0.455	
Taffinder et al 1998 Level III-2	ICSAD, MIST-VR		Control			Sham on-call			No sleep			
	Errors, mean ± SE	6	-0.4 ± 0.1			-0.1 ± 0.1			0.3 ± 0.1			
	P-value (compared with baseline)		Repeated-measures ANOVA for a Latin square showed a significant linear trend across the sleep conditions compared with baseline for error score (p = 0.009).									

CS, core skills; ICSAD, Imperial College Surgical Assessment Device; MIST-VR, Minimally Invasive Surgery Trainer – Virtual Reality; NS, not significant; SD, standard deviation; SEM, standard error of the mean

* Improved, significantly improved post-call ($p < 0.05$); Deteriorated, significantly deteriorated post-call ($p < 0.05$).

† Raw data difficult to extract from data supplied in study. With increased fatigue ratings, residents tended to make more errors.

Performance of movements

Six studies reported outcomes related to the performance of hand movements (DeMaria et al 2005; Grantcharov et al 2001; Eastridge et al 2003; Leff et al 2008; Kahol et al 2008; Uchal et al 2005).

Randomised controlled trials

Uchal et al (2005) reported no significant differences in movements after a night on-call (Table 13).

Non-randomised comparative studies

DeMaria et al (2005) reported significant improvements in economy of movement for up to two tasks for both the left and right hands for all of the participants as well as for the junior residents alone post-call ($p < 0.05$ for all), although no actual data were reported.

Grantcharov et al (2001) reported significantly more unnecessary movements for two of six tasks after a night shift ($p \leq 0.008$) (Table 13).

Eastridge et al (2005) reported no differences in economy of motion for either hand after a night on-call, although no actual data were reported.

Leff et al (2008) reported significant deterioration in total economy for CS 2 (from baseline) after the first and third nightshift ($p \leq 0.02$) (Table 13).

Kahol et al (2008) reported significant decreases in proficiency and smoothness post-call ($p \leq 0.02$). A sub analysis was conducted to study the differences between pre-call and post-call performance for exercises that primarily involved psychomotor skill versus exercises that involved a combination of psychomotor and cognitive skills. Results indicated that the difference between pre-call and post-call psychomotor task performance was not statistically significant, but there was a significant difference in cognitive task performance for gesture proficiency ($p < 0.002$), and tool movement smoothness ($p < 0.04$), but not for hand movement smoothness. In addition to this, residents were grouped according to experience levels, and their pre-call and post-call performance data were compared using ANOVA. Performance of PGY 1 – 3 residents was consistently worse for gesture proficiency ($p < 10e\text{-}9$), hand movement smoothness ($p < 1e\text{-}10$), and tool movement smoothness ($p < 10e\text{-}5$), compared with that of PGY 4 – 5 residents.

Table 13. Simulated performance - pre-call and post-call performance of movements

Uchal et al 2005 Level II	Pelvic trainer	Post-work (n = 32)				Post-call (n = 32)				P-value			
	Goal directed actions, median n	32.5				33.5				0.63			
	Non-goal directed actions, median n	0.31				0.56				0.4			
Grantcharov et al 2001 Level III-2	MIST-VR	Task 1		Task 2		Task 3		Task 4		Task 5		Task 6	
		Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post
	Unnecessary movements, median (range)	4 (3-6)	5 (3-13)	5 (3-9)	4 (4-8)	5 (4-8)	7 (3-14)	4 (3-5)	4 (3-4)	8 (4-10)	9 (5-22)	6 (4-8)	8 (4-16)
	P-value	0.56		0.78		0.27		0.93		0.008		0.004	
Leff et al 2008 Level III-2	MIST-VR	Base-line	Night 1	Night 2	Night 3	Night 4	Night 5	Night 6	Night 7	Follow up			
	CS 1, total economy, median (SD)	3 (0.9)	4 (0.9)	4 (0.3)	4 (0.3)	4 (0.2)	3 (0.4)	3 (0.7)	4 (0.5)	3 (1.0)			
	P-value (compared with baseline)	-	0.101	0.094	0.341	0.090	0.549	1.000	0.420	0.958			
	CS 2, total economy, median (SD)	4 (1.0)	5 (1.4)	4 (1.1)	4 (0.8)	4 (1.0)	4 (0.9)	4 (1.4)	4 (1.2)	4 (1.4)			
	P-value (compared with baseline)	-	0.009	0.178	0.024	0.122	0.131	0.313	0.940	0.303			

CS, core skills; MIST-VR, Minimally Invasive Surgery Trainer – Virtual Reality; SD, standard deviation.

Manual dexterity

Three studies reported outcomes for manual dexterity when participants were tested in rested and unrested states (Light et al 1989; Deaconson et al 1989; Reznick & Folse 1987).

Randomised controlled trials

Reznick & Folse (1987) reported combined speed and accuracy scores (the total number of errors divided by the total number of suture passes) for a manual dexterity simulated wound task. Split-plot ANOVA revealed no significant condition effect, group membership effect, or learning effect, for either the two scores (ie sleep deprived and non-sleep deprived). An analysis of all 21 subjects revealed no significant difference in the two experimental conditions. The indices used in these tests correlated significantly in both the sleep-deprived ($p < 0.001$) and non sleep-deprived condition ($p < 0.001$).

Analysis of the mean scores of the four subtasks (dominant hand, non-dominant hand, both hands, and complex assembly) of the Purdue pegboard of the 12 subjects that participated in both the sleep-deprived and non sleep-deprived conditions revealed that subjects performed significantly better in the non sleep-deprived condition using the dominant hand ($p = 0.006$). In addition, participants demonstrated a significant learning effect, performing better during the second testing episode compared with the first ($p = 0.0001$). Using the non-dominant hand, participants demonstrated a significant learning effect ($p = 0.01$), but there was no difference between the sleep-deprived and non sleep-deprived conditions ($p = 0.89$).

Working with both hands simultaneously, there was no significant effect of condition, learning, or group membership. A significant effect was seen in the complex assembly task ($p = 0.002$), but no effect of condition or group membership was seen.

When results in all subjects were analysed using the unpaired t-test, no significant differences were seen in performance on any of the four subtasks between the sleep-deprived and non sleep-deprived conditions.

Non-randomised comparative studies

Light et al (1989) reported that when residents were in the acute sleep-deprived state, there were no differences within or between resident levels on the functional testing in repetitive skills, continuous tasks, although no data were provided. In the manual dexterity test (pegboard), interns (PGY 1) showed significantly altered ability to perform with both dominant and non-dominant hands when sleep-deprived ($p < 0.05$). This difference was not seen at PGY 2 – 5 levels (results not shown).

Case series (pre-test/post-test outcomes)

Deaconson et al (1989) reported a significant difference in performance on the Purdue Pegboard between sleep-deprived and not sleep-deprived participants ($p < 0.05$). When the results of performance by each resident on each test was analysed using Mann-Whitney *U* test, there were no differences in performance between the sleep-deprived and non sleep-deprived residents.

Summary of psychomotor skill performance results

- RCT evidence (two studies) reported no significant differences in psychomotor skills between rested and unrested groups.
- Non-randomised comparative studies (one study) and case series studies (eight studies) provided more mixed data: for performance time, hand movements and manual dexterity, approximately half of the studies found no significant differences or improvements between the rested and fatigued states post-call, while the other half reported decrements in performance when participants were fatigued. Errors were more likely to occur post-call. Surgical residents with less surgical training/experience appeared to be more affected by sleep deprivation than more senior residents.

Ongoing and unpublished trials

Searches of the Clinical Trials Database, NHS CRD, NHS HTA, Current Controlled Trials and the National Research Register did not identify any unpublished studies.

5. Discussion

Limitations of the evidence

The aim of this systematic review was to determine whether fatigued surgeons or trainee surgeons have different performance outcomes compared with non-fatigued surgeons or trainee surgeons; and to determine the impact of fatigue on the cost of surgery and surgical training

No economic evaluations were found using the search criteria that were developed. This may in part be due to the fact that studies regarding the effect of work hour restrictions were excluded from the review, and because the economic implications of fatigue are not easily characterised. These limitations have resulted in one of the aims of the review not being fully realised, and provides scope for further investigation. These further investigations should include economic evaluations specific to the Australian context to determine the financial effect that fatigue has on the cost of surgery and surgical training. In addition to this, further investigations could include the potential medico-legal risks associated with long working hours.

In relation to performance outcomes, the quality and quantity of the available evidence limited the conclusions that could be drawn. Only two RCTs were included for review (Reznick & Folse 1987; Uchal et al 2005), with the remaining studies being non-randomised comparative studies and case series. Uchal et al (2005) compared different groups of rested and fatigued surgeons, while most other studies compared the same surgeons in both rested and fatigued states. Most of the studies were conducted in single centres, which may have reduced the number of potentially confounding variables within each study.

There was no consensus in what constituted a rested and unrested state in the included studies. In most cases, these states were referred to as on-call/not on-call, fatigued/rested or sleep-deprived/not sleep-deprived. It cannot be definitively stated that participants in the rested states were actually rested, or that the individuals in either the rested/unrested groups were not suffering chronic fatigue at baseline. It was unclear from most of the studies how much *normal* sleep the rested groups had received in the week prior to the study, as opposed to only the night before assessment. The effect of acute sleep deprivation on top of chronic partial sleep loss was not considered.

The amount of time that participants slept was reported in 11 of the 20 studies. In general, the rested states had approximately six to seven hours of sleep the night before the assessment, while the unrested groups had approximately three hours of sleep. Only six studies reported significant differences in sleep time between the rested and unrested states. The methods used to calculate sleep were not uniform, with some studies asking participants to self-report the time they slept, or required participants to complete sleep logs or sleep diaries.

The number of participants in the studies was generally low, with six studies having less than 20 participants (Sawyer et al 1999; Wesnes et al 1997; DeMaria et al 2005; Grantcharov et al 2001; Jakubowicz et al 2005; Taffinder et al 1998). These small sample sizes limited the statistical power of the studies.

The included studies were selected on the basis that participants were surgeons or training surgeons, but in some cases this was difficult to determine from abstracts or inclusion/exclusion criteria. Many participants in the included studies were volunteers, which could indicate a pre-study willingness to participate or an obligation to do so, which may have affected performance and introduced bias. In addition to this, residents vulnerable to sleep loss may not have agreed to participate in the studies.

The level of training of the participants was diverse: although most participants were surgical residents at different levels of training, some studies also included fellowship trained surgeons (Schieman et al 2008), board certified surgeons (Uchal et al 2005), attending surgeons (Ellman et al 2004) and general surgery physicians (DeMaria et al 2005). However, because most participants were residents in training, results may not be extrapolatable to surgeons of higher experience levels. It was unclear from many studies whether there were differences in baseline characteristics of the study participants.

The methods of assessment of performance varied greatly between studies. Although these methods have been broadly grouped into categories, there was still a large amount of variation in the outcome measures within each group. Studies that used surveys or relied on self-reporting of variables may have resulted in bias as there may have been differences in attitudes or reporting between responders and nonresponders. The time-of-day of assessment differed between studies, with some studies assessing participants in the morning (Deaconson et al 1988; Leff et al 2008; Wesnes et al 1997; Eastridge et al 2003), while others assessed them in the afternoon (Deary & Tait 1987), or a combination of both (Taffinder et al 1998; Uchal et al 2005; DeMaria et al 2005). In some cases the assessment time was unclear (Grantcharov et al 2001; Jakubowicz et al 2005; Kahol et al 2008; Light et al 1989). There is evidence that performance can vary depending on whether participants are assessed during a circadian upswing (in the morning) or during a circadian low (late in the day) (Schmidt et al 2007), so these differences may have impacted the findings.

The duration of assessments was not evenly balanced between studies, or within categories. Overall, only six studies reported how long assessments took (Deaconson et al 1988; Deary & Tait 1987; Eastridge et al 2003; Leff et al 2008; Light et al 1989; Reznick & Folse 1987). In the cognitive performance category, five studies reported the length of assessment. The assessments in Deaconson et al (1998) were reported to take approximately half an hour, while the assessments in Reznick & Folse (1987), and Light et al (1989) were reported to take approximately one hour. These variations in time are likely to be related to the number and complexity of tests, and the number of times each participant was required to be

assessed, and may have impacted on the results depending on a person's alertness at the time of assessment.

Performance outcomes

Clinical performance

Clinical performance was measured by five studies (Ellman et al 2005; Ellman et al 2004; Haynes et al 1995; Sawyer et al 1999; Schieman et al 2008). Three of these studies reported no differences whether the surgeon had been on-call or not the night before the assessed operation. One study reported that when residents operated on a not on-call day, complications were 45% more likely when the resident had been on-call the day before (Haynes et al 1995). Another study found that being on-call every second night resulted in greater subjective fatigue, stress, and satisfaction, and less operating time than being on-call every fourth night, and that errors were related to on-call fatigue (Sawyer et al 1999). This study did not measure performance objectively, and relied on residents to report subjective questions using visual analogue scales and submit data. Subsequently, response rates were just over 50%, and may have been subject to attrition bias.

Surgical mortality, either as a direct result of the operation, or because of the occurrence of a complication was not considered by one study (Haynes et al 1995), and procedures that were cancelled because a surgeon felt too tired, or surgeon absenteeism were not taken into account by any study. In addition to this, these studies did not consider the potential for sleep-deprived surgeons to alter their work loads to take into account how tired they were (such as only doing standard, uncomplicated surgeries on days where they knew they were tired).

Correlating complication rates and mortality with on-call status is difficult as there are an overwhelming number of potential confounding factors which may have affected outcomes. Patient co-morbidities, anatomical and procedural complexities, the call status and experience of residents who acted as assistants, and interactions of staff in the operating room may all contribute to overall patient outcomes. There is a need to control for such factors, as well as for differences in surgery type, complexity of disease, patient status, and other factors that could affect outcomes and skew the results.

Academic performance

Two studies examined academic scores when residents had been on-call or not on-call the night before the ABSITE examination (Minion et al 2007; Stone et al 2000). Call status was not found to be a significant factor in academic outcome. Sleep time the night before the exam or during a night on-call was not reported. It can be argued that both groups may have been equally sleep-deprived because students who were not on-call would have studied the night before the examination. Studies of this nature are problematic because no baseline differences between groups, or other demographic information that could have affected the outcomes were reported,

resulting in any important differences not being able to be ascertained. In addition to this, one of the studies (Minion et al 2007), did not take into account variations in the difficulty of the examination from one year to the next, or institutional variations in results.

Cognitive performance

Five studies were included in this category (Deaconson et al 1988; Deary & Tait 1987; Light et al 1989; Reznick & Folse 1987; Wesnes et al 1997). Results regarding psychological outcomes were mixed, limiting the ability to draw firm conclusions. RCT evidence indicated that sleep deprivation had no effects on factual recall and concentration. In other studies, differences in performance were attributable to between-subject differences (Deaconson et al 1989) or decrements in performance by junior residents (Light et al 1989). The results of Light et al (1989) may provide evidence that the effect of experience and practice may mitigate the effects of fatigue.

Five of the six studies (including the single RCT) in this category were published more than 10 years ago. Changes in surgical training curricula, study methodologies and assessment tools over this time may mean that these results are no longer applicable to the current surgical training environment. In addition to this, the grouping of these studies into one category and summarising the results may not be appropriate because of study differences.

Psychomotor skill performance

Eight studies were included in this category (and three additional studies were included from the cognitive performance category because they provided some psychomotor results). The RCTs reported no significant differences between rested and unrested groups (Reznick & Folse 1987; Uchal et al 2005). A change between the two groups in Uchal et al (2005) (if apparent) may not have been detected because the participants were assessed at different times of day: the post-call group was assessed during a circadian upswing (0800 – 0900 h), while the post-work group was assessed at a circadian low (1600 – 1700 h). The other studies provided more mixed data. For performance time and hand movements, half of the studies found no significant differences or improvements between the rested and fatigued states post-call, while the other half reported decrements in performance. Errors were more likely to occur post-call.

There were large variations in the length of time participants were trained and assessed. The end-points of training were often ill-defined and were not consistent between studies, making it difficult to comment on the skill level at the end of the training. Only two studies reported training to a predetermined level of proficiency before assessment (Jakubowicz et al 2005; Leff et al 2008). It can be argued that the short duration of simulation training may have resulted in an inability to show differences in skill level when a person was rested or not. Learning or practice effects, although acknowledged by five studies (DeMaria et al 2005; Eastridge et al 2003; Jakubowicz et al 2005; Kahol et al 2008; Leff et al 2008; Taffinder et al 1998),

may have skewed the results in those studies that did not account for it in the results. Leff et al (2008) showed the effects of adaptation over time. The performance times initially increased over the first few nightshifts, but then remained stable at baseline levels over the course of the week, indicating that surgeons can adapt and prepare for a known task even when sleep-deprived.

Most of the studies examined particular aspects of surgical performance, with many studies acknowledging that more factors would contribute to a safe and effective surgeon. The development of technical skills is only one part of surgical training, and no single parameter measured in a simulator can by itself demonstrate that a trainee has acquired an expert level of proficiency or competence (Ahlberg et al 2005). A good example of this is performance time, which was measured by many of the included studies. The measurement of this variable is unlikely to be a clinically significant result as it does not give any indication of the quality of the task performed, and caution should be taken when interpreting it without any additional objective quality data. A single study demonstrated that performance time improved post-call, suggesting that fatigue may have no detrimental effects on performance. Further examination of the results indicated, however, that this improvement was at the detriment of errors and the performance of movements (Kahol et al 2008), suggesting that surgeons are not immune to the effects of sleep loss. This highlights the issue of using surrogate markers as measures of clinical performance.

Some of the manual tasks included in the assessments of the studies may have had questionable content validity for surgeons: tasks such as peg transfers on poles, threading pipe cleaner onto rubber pipe, passing balls through hoops, report sorting, and counting backwards may not have had high clinical relevance to the participants which may have affected performance. Challenge may heighten focus during surgery (a surgeon confronted with a complex clinical task is likely to be more alert than when performing a simulated task in low risk environment), raising doubts about the use of simulators as a surrogate discriminator for fatigue in clinical performance.

Other considerations

Conducting well designed RCTs to determine the effect of fatigue on performance in surgery is difficult because it would be unethical to have intentionally sleep-deprived surgeons operating on patients. As has been done, an alternative method is to measure performance via simulated methods. Unfortunately simulation does not always provide a realistic environment, with participants not performing as they would in an actual clinical situation. The relationship between surrogate markers for performance (eg time of simulators, cognitive performance) to actual clinical performance is currently unclear and it is an area that deserves further study. Understanding exactly which surrogate markers, if any, have a high correlation with actual clinical performance will determine which of these is useful to use in studies concerning sleep deprivation and fatigue.

Studies indicate that neurobehavioural impairment may be similar when people suffer

short-term sleep loss (recent 24 hour complete sleep loss) or chronic partial sleep deprivation (less than six hours of sleep per night on average for at least one week) (Linde & Bergstorm 1992; Polzella 1975), suggesting a reason why there were no differences between groups in many of the studies. In addition to this, some individuals may react differently to sleep deprivation than others, and different operations may evoke different levels of energy or arousal. It has been shown that performance, subjective alertness, and sleepiness vary between individuals, and are related to individual peak periods of circadian arousal and the time-of-day that testing occurs (Schmidt et al 2007). There is a belief that some processes are more vulnerable to variations in circadian arousal level than others: highly practised responses being less susceptible across the day, with all other responses being vulnerable to the time-of-day effect during normal day/night conditions since they require a certain degree of control over stimuli and responses (Schmidt et al 2007).

Greater decrements in performance were seen in junior residents compared with senior residents indicating that years of training may affect changes in perfor mance when tired. Jabukowicz et al (2005) state that in general, the more skilled, experienced or knowledgeable a person is, the less likely loss of sleep is to affect his or her performance. These authors further state that in all individuals fatigue is more likely to affect cognition, regardless of experience, skill, or knowledge. Fatigue is most likely to impact negatively on attention, decision-making, and information retrieval from long term memory.

Many factors contribute to performance, and it is acknowledged that more than just the number of hours slept or time of assessment would determine outcomes. Other factors, such as the cognitive skills of anatomical recognition, decision making, judgement, leadership and communication are also vital to the performance of a surgeon. Other factors such as caffeine, naps, food intake, physical exertion and stress also affect an individual's performance, which in most cases were not controlled in the studies. Factors influencing quality surgical care are multifactorial, and sound judgement and compassion may be at least as important as technical skill (Eastridge et al 2003).

There is a notion that surgeons self-select themselves into surgery because they know that they can work when tired, and those who are incapable of performing when sleep-deprived enter other specialties (Ellman et al 2005; Veasey et al 2002). If this is indeed the case, then some studies may have favoured participants better adapted to sleep deprivation if those participants who could not cope well with sleep deprivation were lost to follow-up.

This review did not include studies examining changes in performance before and after work hour restrictions, and instead focussed on performance directly related to sleep. Investigations into the effects of work hour restrictions warrant further attention as they could provide insight into the effects of shorter shifts, quality of life of surgeons, and clinical outcomes for patients.

Employers have a duty of care to be aware of workplace risks, and to prevent the

chance of accident or injury. Forward rather than backward shift rotation, education about good sleep hygiene, and strategic napping before or during shifts may reduce fatigue and improve performance (Jha et al 2001). Taking a risk management approach (QLD Health 2009) and educating employees about the effects of fatigue and teaching them practical strategies to improve sleep quality and increase alertness could potentially also reduce accidents and reduce the risk of potential litigation.

6. Conclusions and Recommendations

The aim of this systematic review was to determine whether fatigued surgeons or trainee surgeons have different performance outcomes compared with non-fatigued surgeons or trainee surgeons, and to investigate the impact of fatigue on the cost of surgery and surgical training. Many studies used surrogate markers to measure performance, although the relationship between these markers to actual clinical performance is unclear. Variations in results were in some cases attributable to the level of training of participants, and between-subject differences. The search strategy did not identify any economic evaluations, resulting in an inability to comment on the financial effect of fatigue on surgery and surgical training.

Work hour restrictions were implemented to reduce the level of fatigue, and to provide greater patient safety under the premise that less sleep leads to decrements in performance. This review highlights that a paucity of evidence exists in this field. The nature of surgery and the surgical workforce limits the types of studies that can be conducted. It is technically difficult, if not impossible to produce a placebo control condition for fatigue in a surgical setting. The absence of evidence of an effect is not necessarily evidence of an absence of an effect, which leaves the questions in relation to the effect of fatigue on surgeons largely unanswered. A lack of evidence should not provide a rationale for not taking action, as this may be interpreted as a prejudice that rationalises the maintenance of the status quo. Therefore, it may be warranted to extrapolate evidence from other settings or industries which do have a large evidence base to the surgical setting and allow this research to guide policy and decision making processes.

We acknowledge that it would be beneficial to compare the results of this systematic review with data from professions other than the field of surgery. A systematic assessment of fatigue in other professions, such as aeronautics, transport, military and shift workers, was beyond the scope of this current assessment but, where available, have generally demonstrated similar findings to this review, although individual reports written within these industries do suggest detrimental effects of fatigue on performance.

Classification and Recommendations

The evidence-base in this review is rated as poor. The studies included were of variable quality, differed in study design, and many used surrogate markers to assess performance, resulting in an inability to draw solid conclusions.

Clinical and Research Recommendations

It is recommended that further studies be conducted into the effect of sleep

deprivation on performance in the surgical setting, specifically:

- The identification of surrogate markers, if any, to actual clinical performance
    - The strength of the relationship between these surrogate markers (eg time of simulators, cognitive performance) and actual clinical performance
- The development of a clearer definition of fatigue and its relationship to sleep deprivation
- The impact of fatigue on the cost of surgery and surgical training.
- The development of common numerical values for acute and chronic sleep deprivation
- The effect of acute sleep deprivation on performance
- The effects of acute sleep deprivation on top of chronic partial sleep loss on performance
- Comparisons of sleep-deprived surgeons with those who have had at least one week of normal sleep
- Comparisons of performance at difference times of day to assess outcomes at different circadian points
- Comparisons of performance of inexperienced surgeons with experienced surgeons with respect to fatigue and sleep loss

Acknowledgements

The ASERNIP-S project is funded by the Australian Commonwealth Department of Health and Ageing and the South Australian Department of Health.

References

Accreditation Council for Graduate Medical Education (ACGME). 2007. http://www.acgme.org/acWebsite/dutyHours/dh_ComProgrRequirmentsDutyHours0707.pdf [accessed 9 March 2009].

Accreditation Council for Graduate Medical Education (ACGME). Report of the ACGME Work Group on Resident Duty Hours and the Learning Environment. 2002. http://www.acgme.org/acWebsite/dutyHours/dh_wkgroupreport611.pdf [accessed 9 March 2009].

Aeschbach D, Sher L, Postolache TT, Matthews JR, Jackson MA, Wehr TA. A longer biological night in long sleepers than in short sleepers. *Journal of Clinical Endocrinology & Metabolism* 2003; 88(1): 26-30.

Ahlberg G, Hultcrantz R, Jaramillo E, Lindblom A, Arvidsson D. Virtual reality colonoscopy simulation: a compulsory practice for the future colonoscopist? *Endoscopy* 2005; 37(12): 1198-1204.

Akerstedt T. Shift work and disturbed sleep/wakefulness. *Occupational Medicine* 2003; 53(2): 89-94.

Australian Medical Association (AMA). National Code of Practice – Hours of Work, Shiftwork and Rostering for Hospital Doctors. 2005. http://www.ama.com.au/web.nsf/doc/WEEN-6V76DZ/$file/National_Code_of_Practice_January_2005.pdf [accessed 14 March 2009].

Australian Medical Association (AMA). Risk assessment of junior doctor residents. 2001. http://www.ama.com.au/web.nsf/doc/SHED-5G3C2R/$file/wplace14_ve_shp%20risk%20assess%20audit%20report.pdf [accessed 14 March 2009].

Australian Medical Association (AMA). Safe work hours = safe patients. AMA safe work hours audit 2006. http://www.ama.com.au/web.nsf/doc/WEEN-6UWAVM/$file/Safe_Hours_Report.pdf [accessed 14 March 2009].

Ayas NT, Barger LK, Cade BE, Hashimoto DM, Rosner B, Cronin JW, Speizer FE, Czeisler CA. Extended work duration and the risk of self-reported percutaneous injuries in interns. *Journal of the American Medical Association* 2006; 6; 296(9): 1055-1062.

Barger LK, Cade BE, Ayas NT, Cronin JW, Rosner B, Speizer FE, Czeisler CA; Harvard Work Hours, Health, and Safety Group. Extended work shifts and the risk of motor vehicle crashes among interns. *New England Journal of Medicine* 2005; 13; 352(2): 125-134.

Cook DJ, Mulrow CD, Haynes BR. Systematic reviews: synthesis of best evidence

for clinical decisions. *Annals of Internal Medicine* 1997; 126(5): 376–380.

Dawson D & Reid K. Fatigue, alcohol and performance impairment. *Nature* 1997; 388: 235.

Dawson D & Zee P. Work hours and reducing fatigue-related risk: good research vs good policy. *Journal of the American Medical Association* 1995; 294(9): 1104-1106.

DeMaria EJ, McBride CL, Broderick TJ, Kaplan BJ. Night call does not impair learning of laparoscopic skills. *Surgical Innovation* 2005; 12(2): 145-149.

Deaconson TF, O'Hair DP, Levy MF, Lee MB, Schueneman AL, Codon RE. Sleep deprivation and resident performance. *Journal of the American Medical Association* 1988; 260(12): 1721-1727.

Deary IJ & Tait R. Effects of sleep disruption on cognitive performance and mood in medical house officers. *British medical journal (Clinical research ed.)* 1987; 295(6612): 1513-1516.

Eastridge BJ, Hamilton EC, O'Keefe GE, Rege RV, Valentine RJ, Jones DJ, Tesfay S, Thal ER. Effect of sleep deprivation on the performance of simulated laparoscopic surgical skill. *American Journal of Surgery* 2003; 186(2): 169-174.

Ellman PI, Kron IL, Alvis JS, Tache-Leon C, Maxey TS, Reece TB, Peeler BB, Kern JA, Tribble CG. Acute sleep deprivation in the thoracic surgical resident does not affect operative outcomes. *Annals of Thoracic Surgery* 2005; 80(1): 60-65.

Ellman PI, Law MG, Tache-Leon C, Reece TB, Maxey TS, Peeler BB, Kern JA, Tribble CG, Kron IL. Sleep deprivation does not affect operative results in cardiac surgery. *Annals of Thoracic Surgery* 2004; 78(3): 906-911.

Fisman DN, Harris AD, Rubin M, Sorock GS, Mittleman MA. Fatigue increases the risk of injury from sharp devices in medical trainees: results from a case-crossover study. *Infection Control and Hospital Epidemiology* 2007; 28(1):10-7.

Fletcher KE, Davis SQ, Underwood W, Mangrulkar RS, McMahon LF Jr, Saint S. Systematic review: effects of resident work hours on patient safety. *Annals of Internal Medicine* 2004; 141(11):851-857.

Folkard S, Lombardi DA, Tucker PT. Shiftwork: Safety, sleepiness and sleep. *Industrial Health* 2005; 43(1): 20-23.

Frank AL. Injuries related to shiftwork. *American Journal of Preventive Medicine* 2000; 18(4 Supp 1): 33-36.

Gander P, Purnell H, Garden A, Woodward A. Work patterns and fatigue-related risk among junior doctors. *Occupational Environmental Medicine* 2007; 64: 733-738.

Grantcharov TP, Bardram L, Funch-Jensen P, Rosenberg J. Laparoscopic performance after one night on call in a surgical department: prospective study. *British Medical Journal* 2001; 323(7323): 1222-1223.

Haynes DF, Schwedler M, Dyslin DC, Rice JC, Kerstein MD. Are postoperative complications related to resident sleep deprivation? *Southern Medical Journal* 1995; 88(3): 283-289.

Hedman C, Pohjasvaara T, Tolonen U, Suhonen-Malm AS, Myllyla VV. Effects of pregnancy on mothers' sleep. *Sleep Medicine* 2002; 3(1): 37-42.

Higgins JPT & Green S (eds). Cochrane Handbook for Systematic Reviews of Interventions Version 5.0.0 [updated February 2008]. The Cochrane Collaboration, 2008. Available from www.cochrane.org/resources/handbook

Holley DC, Sundaram B, Wood DK. Shift work and aviation safety. *Clinics in Occupational and Environmental Medicine* 2003; 3(2): 231-262.

Jakubowicz DM, Price EM, Glassman HJ, Gallagher AJ, Mandava N, Ralph WP, Fried MP. Effects of a twenty-four hour call period on resident performance during simulated endoscopic sinus surgery in an accreditation council for graduate medical education-compliant training program. *Laryngoscope* 2005; 115(1): 143-146.

Jenni OG, Iglowstein I, Benz C, Largo RH. Percentile curves for sleep duration in the first 16 years of life. *Tagliche Praxis* 2004; 45(1): 127-135.

Jensen S & Given BA. Fatigue affecting family caregivers of cancer patients. *Cancer Nursing.* 1991; 14(4): 181-187.

Jha A, Duncan BW, Bates DW. Fatigue, sleepiness, and medical errors. Making health care safer: A critical analysis of patient safety practices. Chapter 46. Evidence Report/Technology Assessment: Number 43. AHRQ Publication No. 01-E058, July 2001. Agency for Healthcare Research and Quality, Rockville, MD. http://www.ahrq.gov/clinic/ptsafety/ [accessed 7 March 2009].

Kahol K, Leyba MJ, Deka M, Deka V, Mayes S, Smith M, Ferrara JJ, Panchanathan S. Effect of fatigue on psychomotor and cognitive skills. *American Journal of Surgery* 2008; 195(2): 195-204.

Kripke DF, Garfinkel L, Wingard DL, Klauber MR, Marler MR. Mortality associated with sleep duration and insomnia. *Archives of General Psychiatry* 2002; 59:131-136.

Kuhn G. Circadian rhythm, shift work, and emergency medicine. *Annals of Emergency Medicine* 2001; 37(1):88-98.

Laine C, Goldman L, Soukup JR, Hayes JG. The impact of a regulation restricting medical house staff working hours on the quality of patient care. *Journal of the American Medical Association* 1993; 269(3): 374-378.

Lal SKL, Craig A. A critical review of the psychophysiology of driver fatigue. *Biological Psychology* 2001; 55(3): 173-194.

Landrigan CP, Rothschild JM, Cronin JW, Kaushal R, Burdick E, Katz JT, Lilly CM, Stone PH, Lockley SW, Bates DW, Czeisler CA. Effect of reducing interns' work hours on serious medical errors in intensive care units. *New England Journal of Medicine* 2004; 351(18):1838-1848.

Leff D, Aggarwal D, Rana M, Nakhjavani B, Purkayastha S, Khullar V, Darzi A. Laparoscopic skills suffer on the first shift of sequential night shifts: program directors beware and residents prepare. *Annals of Surgery* 2008; 247(3): 530-539.

Leonard C, Fanning N, Attwood J, Buckley M. The effect of fatigue, sleep deprivation and erroneous working hours on the physical and mental well-being of pre-registration house officers. *Irish Journal of Medical Science* 1998; 167: 22-25.

Light AI, Sun JH, McCool C, Thompson L, Heaton S, Bartle EJ. The effects of acute sleep deprivation on level of resident training. *Current Surgery* 1989; 46(1): 29-30.

Linde L & Bergstrom M. The effect of one night without sleep on problem-solving and immediate recall. *Psychological Research* 1992; 54(2): 127-36.

Lockley SW, Cronin JW, Evans EE, Cade BE, Lee CJ, Landrigan CP, Rothschild JM, Katz JT, Lilly CM, Stone PH, Aeschbach D, Czeisler CA. Harvard Work Hours, Health and Safety Group. Effect of reducing interns' weekly work hours on sleep and attentional failures. *New England Journal of Medicine* 2004; 28; 351(18): 1829-1837.

Minion D, Plymale M, Donnelly M, Endean E. The effect of prior night call status on the American Board of Surgery in-training examination scores: eight years of data from a single institution. *Journal of Surgical Education* 2007; 64(6): 416-419.

National Health and Medical Research Council (NHMRC). How to use the Evidence: assessment and application of scientific evidence, p8. Canberra: NHMRC. 2000.

Nuckols TK & Escarce JJ. Residency work-hours reform. A cost analysis including preventable adverse events. *Journal of General Internal Medicine* 2005; 20(10); 873.

Philibert I. Sleep loss and performance in residents and nonphysicians: A meta-analytic examination. *Sleep* 2005; 28(11): 1392-1402.

Piper B. Fatigue. In: Carrieri-Kohlman V, Lindsey A, West C (Eds). *Pathophysiological phenomena in nursing: human responses to illness* (2nd Edn). Philadelphia: Saunders. pp 279-302.

Polzella DJ. Effects of sleep-deprivation on short-term recognition memory. *Journal of Experimental Psychology: Human Learning and Memory* 1975; 104: 194-200.

QLD Health. Medical fatigue risk management. Human resources policy January 2009. http://www.health.qld.gov.au/hrpolicies/other/i_1.pdf [accessed 15 April 2009].

Reznick RK & Folse JR. Effect of sleep deprivation on the performance of surgical residents. *American Journal of Surgery* 1987; 154(5): 520-525.

Royal Australasian College of Surgeons (RACS). Standards for safe working hours and conditions for fellows, surgical trainees and international medical graduates. RACS. 2007.

Sawyer RG, Tribble CG, Newberg DS, Pruett TL, Minasi JS. Intern call schedules and their relationship to sleep, operating room participation, stress, and satisfaction. *Surgery* 1999; 126(2): 337-342.

Schenarts PJ, Anderson Schenarts KD, Rotondo MF. Myths and realities of the 80-hour work week. *Current Surgery* 2006; 63(4):269-274.

Schieman C, MacLean AR, Buie WD, Rudmik LR, Ghali WA, Dixon E. Does surgeon fatigue influence outcomes after anterior resection for rectal cancer? *American Journal of Surgery* 2008; 195(5): 684-687.

Schmidt C, Collette F, Cajochen C, Peigneux P. A time to think: circadian rhythms in human cognition. C*ognitive Neuropsychology* 2007; 24(7): 755-789.

Stone MD, Doyle J, Bosch RJ, Bothe A, Steele G. Effect of resident call status on ABSITE performance. *Surgery* 2000; 128(3): 465-471.

Strunk CL, Bailey BJ, Scott BA, et al Resident work hours and working environment in otolaryngology: analysis of daily activity and resident perception. *Journal of the American Medical Association* 1991; 266: 1371-1374.

Taffinder NJ, McManus IC, Gul Y, Russell RCG, Darzi A. Effect of sleep deprivation on surgeons' dexterity on laparoscopy simulator. *Lancet* 1998; 352(10): 1191.

Tiesinga LJ, Dassen TWN, Halfens RJD. Fatigue: a summary of the definitions, dimensions, and indicators. *Nursing Diagnosis* 1996; 7(2): 51–62.

Uchal M, Tjugum J, Martinsen E, Qiu X, Bergamaschi R. The impact of sleep deprivation on product quality and procedure effectiveness in a laparoscopic physical simulator: a randomized controlled trial. *American Journal of Surgery* 2005; 189(6): 753-757.

UK Department of Health. What is the European Working Time Directive? Updated 9 August 2007. http://www.dh.gov.uk/en/Managingyourorganisation/Humanresourcesandt

raining/Modernisingworkforceplanninghome/Europeanworkingtimedirective/DH_077304 [accessed 14 March 2009].

van Dongen HP, Maislin G, Mullington JM, Dinges DF. The cumulative cost of additional wakefulness: dose-response effects on neurobehavioral functions and sleep physiology from chronic sleep restriction and total sleep deprivation. *Sleep* 2003; 26(2): 117-126.

Veasey S, Rosen R, Barzansky B, Rosen I, Owens J. Sleep loss and fatigue in residency training - A reappraisal. *Journal of the American Medical Association* 2002; 288(9): 1116-1124.

Weatherby BA, Rudd JN, Ervin TB, Stafford PR, Norris BL. The effect of resident work hour regulations on orthopaedic surgical education. *Journal of Surgical Orthopaedic Advances* 2007; 16(1): 19-22.

Weinger MB, Ancoli-Israel S. Sleep deprivation and clinical performance. *Journal of the American Medical Association* 2002; 287(8): 955-957.

Wesnes KA, Walker MB, Walker LG, Heys SD, White L, Warren R, Eremin O. Cognitive performance and mood after a weekend on-call in a surgical unit. *British Journal of Surgery* 1997; 84: 493-495.

Whang EE, Perez A, Ito H, Mello MM, Ashley SW, Zinner MJ. Work hours reform: perceptions and desires of contemporary surgical residents. *Journal of the American College of Surgeons* 2003; 197: 624-630.

APPENDIX A – EXCLUDED STUDIES

Appendix A - Excluded studies

The following articles were excluded from the methodological assessment as outlined in the methods section of the review.

Excluded Studies

Study	Reason for exclusion
Bartle EJ, Sun JH, Thompson L, Light AI, McCool C, Heaton S. The effects of acute sleep deprivation during residency training. *Surgery* 1988; 104(2): 311-316.	Study unavailable
Browne BJ, Van Susteren T, Onsager DR, Simpson D, Salaymeh B, Condon RE. Influence of sleep deprivation on learning among surgical house staff and medical students. *Surgery* 1994; 115(5): 604-610.	Study unavailable
Bunch WH, Dvonch VM, Storr CL, Baldwin DC, Jr., Hughes PH. The stresses of the surgical residency. *Journal of Surgical Research* 1992; 53(3): 268-271.	Not fatigue/sleep deprivation specific
Chung RS and Ahmed N. How surgical residents spend their training time - The effect of a goal-oriented work style on efficiency and work satisfaction. *Archives of Surgery* 2007; 142(3): 249-252.	No rested/unrested groups
Folse R and DaRosa DA. The balance between service and education in surgical residency training. *Current Surgery* 1989; 46(3): 193-202.	No rested/unrested groups
Goldstein MJ, Kim E, Widmann WD, Hardy MA. A 360degrees evaluation of a night-float system for general surgery: A response to mandated work-hours reduction. *Current Surgery* 2004; 61(5): 445-451.	No baseline data
Jensen A, Milner R, Fisher C, Gaughan J, Rolandelli R, Grewal H. Short-term sleep deficits do not adversely affect acquisition of laparoscopic skills in a laboratory setting. *Surgical Endoscopy and Other Interventional Techniques* 2004; 18(6): 948-953.	No baseline data
Kocher HM, Warwick J, Al Ghnaniem R, Patel AG. Surgical dexterity after a 'night out on the town'. *ANZ Journal of Surgery* 2006; 76(3): 110-112.	Sleep parameters not work related
Lee DTY, Chan SWW, Kwok SPY. Introduction of night shift call system for surgical trainees: a prospective self-controlled trial. *Medical Education* 2003; 37(6): 495-499.	No rested/unrested groups
Woodrow SI, Park J, Murray BJ, Wang C, Bernstein M, Reznick RK, Hamstra SJ. Differences in the perceived impact of sleep deprivation among surgical and non-surgical residents. *Medical Education* 2008; 42(5): 459-467.	No intervention/exposure

APPENDIX B – STUDY PROFILE TABLES

Appendix B – Study profile tables

Author	Intervention	Study Design	Study population	Inclusion/Exclusion Criteria
DeMaria et al 2005 Location Department of Surgery, Medical College of Virginia and Virginia Commonwealth University, Richmond USA	Objective: To evaluate performance of surgeons on laparoscopic simulator before and after a night on-call. All participants Participants recorded • Amount of sleep before testing session • Longest period of uninterrupted sleep • Day time sleeping (naps) • Assessed their own level of fatigue on a scale of well rested, rested, partially rested, and not rested. Training Participants were instructed how to use MIST-VR and allowed to practice all tasks. Intervention Call Assessment Tested on all 6 parameters pre call (time not specified) and post call (late in the day post-call, usually about 34 hours before testing). Difference between pre and post call was calculated. Device MIST-VR	Case series (pre-test/post-test outcomes) Prospective, single centre Level of evidence: IV Lost to follow-up: 30 surgeons performed pre call assessments, 13 did not complete post call assessments and data were not analysed. Study period: not reported Operator details: not reported Outcome measures and validity: MIST-VR • Acquire • Transfer • Traversal • Withdraw • Diathermy • Manipulation Performance of dominant and non-dominant hand measured: • Economy of motion • Time to complete task • Errors by each hand/foot • Total time • Total number of errors	Sample size: n = 17 (all right hand dominant) • General surgery residents n = 16 o intern and PGY 2: n = 10 o senior residents PGY3-5: n = 6. Residents had varying laparoscopic experience. • General surgery attending n = 1 No subject had any sessions with the MIST-VR in the previous year. Mean age: not reported Gender (M/F): 12/4	Inclusion: general surgery residents and general surgery physician Exclusion: not reported

Author	Intervention	Study Design	Study population	Inclusion/Exclusion Criteria
Deaconson et al 1988 Location Department of Surgery, Medical College, Milwaukee, Wisconsin USA	Objective: to determine whether sleep deprivation by residents in an every-other-night call schedule affects cognitive or motor performance over an extended period All participants Three cohorts. Residents tested only once. Financial incentives given for participation. Subjects were told their performance scores would be publicised to peers. Intervention On-call every other night Assessment Tested each morning, for 18 or 19 days	Case series (pre-test/post-test outcomes) Single centre, serial repeated measures Level of evidence: IV Lost to follow-up: not reported Power calculation: study designed to detect a difference as small as 10% or less for each of the 5 psychometric tests. When comparing sleep-deprived and non sleep-deprived participants, there was a 90% chance (power equals 0.90) of detecting a difference for each of the 5 variables when there was a 10% difference between the 2 sleep states. Study period: August – November 1987 Operator details: no feedback given on performance during study Outcome measures and validity: • Sleep diary • Fatigue on a scale of 1, rested to 10, significant difficulty staying awake • Motivation on a scale of 1, hostile about test situation to 10, strong motivation to excel • Paced Auditory Serial Addition Test • Trail-Making Test • Grammatical Reasoning Test • Modified version of Minnesota Paper Form Board Test • Purdue Pegboard	Sample size: N = 26 • PGY 1: n = 7 • PGY 2: n = 7 • PGY 3: n = 5 • PGY 4: n = 4 • PGY 5: n = 3 Age: 26 – 35 years Gender (M/F): 25/1 Dominant hand (R/L): 23/3	Inclusion: each cohort comprised all residents assigned to the trauma, vascular and cardiothoracic services Exclusion: not reported

Author	Intervention	Study Design	Study population	Inclusion/Exclusion Criteria
Deary & Tait 1987 Location Department of Psychology, University of Edinburgh, Edinburgh Scotland	Objective: to determine the effect of sleep loss and long hours on performance Intervention Call • Night off-duty – previous evening and night spent at home • Night on-call – previous evening spent attending calls but no new admissions taken • Night spent admitting emergency cases (waiting) – previous evening and night spent admitting new emergency cases to wards and looking after patients already in wards Assessment Each subject tested 3 times Testing performed between 1400 h and 1700 h on the day after each condition	Case series (pre-test/post-test outcomes) Prospective, single centre Level of evidence: IV Lost to follow-up: not reported Study period: February and mid April Operator details: not reported Outcome measures and validity: • Nowlis mood adjective checklist • Digit span • Serial 13s test • Logical memory – immediate and delay • Information processing • Electrocardiogram assessments • Questionnaire	Sample size: n = 12 Median age: not reported Gender (M/F): 7/5	Inclusion: All house officers at a teaching hospital Exclusion: not reported

Author	Intervention	Study Design	Study population	Inclusion/Exclusion Criteria

Eastridge et al 2003
 	
Location
Southwestern Centre for Minimally Invasive Surgery and the Division of Burns, Trauma and Surgical Care Centre, Department of Surgery, University of Texas Southwestern Medical Centre, Dallas, Texas
 	
USA
 	
Funding provided in part by Southwestern Centre for Minimally Invasive Surgery as supported in party by an educational grant from the United States Surgical Corporation, a division of Tyco Healthcare Group
	Objective: To determine the impact of acute sleep deprivation cased by on-call on the performance of simulated laparoscopic skills
	
	All participants
	All instructed on how to perform tasks.
	6 tasks of progressive complexity. Each task (except task 3) performed twice with each hand before task is complete.
	
	Intervention
	Call
	
	Assessment
	Tested 3 times during the month according to:
	• Pre call (morning before scheduled 24 h in-house call) (RESTED)
	• On-call (the morning of scheduled 24 h in-house call) (RESTED)
	• Post call (morning after scheduled 24 h in-house call (ACUTELY SLEEP-DEPRIVED)
	Tested between 0800 and 1100 h
	Subjects served as own controls
	Caffeine intake not controlled.
	Interval between testing sessions varied in length and testing performed in random order to minimise learning effect.
	
	Device
	MIST-VR - valid (Taffinder et al 1998; Hamilton et al 2002)
		Case series (pre-test/post-test outcomes)
		Prospective, single centre
		
		Level of evidence: IV
		
		Lost to follow-up: not reported
		
		Study period: May – June 2001
		
		Outcome measures and validity:
		• Completed a questionnaire at the beginning of each testing session:
		  • Basic demographics
		  • Number of hours slept in the preceding 24 hours
		  • Quality of sleep (number of interruptions)
		  • Number of hours worked in the previous 7 days
		  • Subjective levels of fatigue (1 none to 1-exhausted)
		• MIST-VR
		  • Time to perform task
		  • Number of errors
		  • Economy of motion
		  • Economy of diathermy
			Sample size: n = 35
			Surgery residents
			• PGY1 n = 13
			• PGY2 n = 9
			• PGY3 n = 8
			• PGY4 n = 2
			• PGY5 n = 3
			Residents had variable prior experience on the simulator.
			
			Average age: 28 years (range 24 – 33)
			
			Gender (M/F): 26/9
				Inclusion: surgery residents
				
				Exclusion: not reported

- Taffinder N, Russell RC, McManus IC, et al. An objective assessment of laparoscopic psychomotor skills: the effect of a training course on performance. Surgical Endoscopy 1998; 12: 493.
- Hamilton EC, Scott DJ, Fleming JB, et al. Comparison of video trainer and virtual reality training systems on acquisition of laparoscopic skills. Surgical Endoscopy 2002; 16: 406-411.

Author	Intervention	Study Design	Study population	Inclusion/Exclusion Criteria

Ellman et al 2005
 	
Location
University of Virginia, Department of Cardiovascular Surgery, Charlottesville, Virginia
 	
USA
	Objective: to determine whether sleep deprivation in cardiac surgical residents affects patient outcomes
	
	Exposure
	Thoracic surgical resident sleep-deprived if he/she preformed a case that started between 2300 h and 0500, or ended a case between 2300 h and 0730 h.
	If the resident performed a subsequent case within the next 24 hours that case was considered a sleep-deprived case, while all other cases were considered not sleep-deprived cases.
	
	Assessment
	Complications recorded prospectively and compared with resident call status.
		Non-randomised comparative
		Single centre, retrospective
		
		Level of evidence: III-2
		
		Lost to follow-up: not reported
		
		Study period: Jan 1994 – March 2004
		
		Operator details: not reported
		
		Outcome measures and validity:
		• In-hospital mortality rates of coronary artery bypass graft (CABG) operations, valve operations, combined CABG-valve operations and 'other' cardiac cases
		• Operative efficiency:
		  • Cardiopulmonary bypass times and cross clamp times
		  • Total in-hospital length of stay after operation
		• Whether patient had received blood products
		• Operative, neurologic, renal, infectious and pulmonary complications
			Sample size: n = 7323
			• SD: 229
			• NSD: 7094
			
			Mean patient age (years):
			• SD: 64.1 ± 0.8
			• NSD: 63.4 ± 0.1
			
			Gender (M/F):
			• SD: 153/76
			• NSD: 4966/2128
			
			Groups were well matched and there were no significant differences in age, sex, race or the operations performed
				Inclusion: cases performed by thoracic surgical residents
				
				Exclusion: non cardiac thoracic cases

Author	Intervention	Study Design	Study population	Inclusion/Exclusion Criteria

Ellman et al 2004
 	
Location
Department of Cardiovascular Surgery, Charlottesville, Virginia
 	
USA
	Objective: to determine whether sleep deprivation in cardiac surgeons affects patient outcomes
	
	Exposure
	Attending cardiac surgeon sleep-deprived if he/she preformed a case that started between 1000 h and 0500, or ended a case between 2300 h and 0730 h. If the surgeon performed a subsequent case within the next 24 hours that case was considered a sleep-deprived case, while all other cases were considered not sleep-deprived cases.
	
	Assessment
	Complications recorded prospectively and compared with resident call status.
		Non-randomised comparative
		Single centre, retrospective
		
		Level of evidence: III-2
		
		Lost to follow-up: not reported
		
		Study period: Jan 1994 – April 2003
		
		Operator details: not reported
		
		Outcome measures and validity:
		• In-hospital mortality rates
		• Operative efficiency:
		  • Cardiopulmonary bypass times and cross clamp times
		  • Total in-hospital length of stay after operation
		• Whether patient had received blood products
		Operative, neurologic, renal, infectious and pulmonary complications
			Sample size: n = 6751 cases
			• SD surgeons: 339
			• NSD surgeons : 6412
			
			Mean patient age (years):
			• SD surgeons: 63.4 ± 0.7
			• NSD surgeons: 63.5 ± 0.1
			
			Gender (M/F):
			• SD surgeons: 237/102
			• NSD surgeons: 4488/1924
			
			Groups were well matched and there were no significant differences in age, sex, race or the operations performed
				Inclusion: cases performed by attending cardiac surgeons
				
				Exclusion: not reported

Author	Intervention	Study Design	Study population	Inclusion/Exclusion Criteria

Grantcharov et al 2001
 	
Location
Department of Surgical Gastroenterology L, Aarhus University, Kommunehospitalet, Aarhus
 	
Denmark
	Objective: to determine whether 1 night on-call in a surgical department would adversely affect the surgeon's performance on simulated laparoscopic tasks.
	
	Training
	All participants received identical pre training by performing 9 repetitions of 6 tasks.
	
	Exposure
	Call
	Night shift: 1530 h – 0900 h
	
	Assessment
	Skills of 10th repetition assessed:
	• during normal day time working hours (time not specified)
	• again at 0930 h after a night on-call with impaired sleep
	The period between the 1st and the 10th repetition was predetermined to be no longer than 1 month.
	
	Device: MIST-VR (Mentice Medical Simulation, Gothenburg, Sweden)
		Case series (pre-test/post-test outcomes)
		Single centre, prospective
		
		Level of evidence: IV
		
		Blinding: not reported
		
		Lost to follow-up: not reported
		
		Study period: not reported
		
		Operator details: not reported
		
		Outcome measures and validity:
		• MIST-VR
		  • Error of motion
		  • Time of motion
		  • Economy of motion
			Sample size: n = 14
			Median time since graduation: 6 years (1 – 11 years)
			Median number prior cholecystectomies: 0 (0 – 5)
			
			Median age: 34 years (24 – 43 years)
			
			Gender (M/F): 11/3
				Inclusion: surgeons in training in department with total sleep time of less than 3 hours
				
				Exclusion: not reported

Author	Intervention	Study Design	Study population	Inclusion/Exclusion Criteria
Haynes et al 1995 Location Tulane University School of Medicine, Tulane Division of Charity Hospital of Louisianna, New Orleans USA	Objective: retrospective examination of complications after operations and on-call status of surgeon Exposure A resident was considered on-call (subject to sleep deprivation) only when required to remain in house for the duration of the 24 hour call day. No other determinants of sleep deprivation were used. Each surgical procedure was allowed a max or 1 complication. If more than 1, most serious used. Assessment Complications were identified by senior resident and attending surgeon and entered into computer (prospectively). Operating surgeons on-call status at the time of the operation determined by computer correlation. Call status of residents who acted as assistant was not considered. Surgical mortality, either as a direct result of the operation of because of the occurrence of a complication was not considered.	Non-randomised comparative Single centre, retrospective Level of evidence: III-2 Lost to follow-up: not reported Study period: Jan 1985 – April 1988 Operator details: not reported Outcome measures and validity: • Patient complications as entered on computer (not valid) • Resident survey (not valid) Surgical procedures categorised by • Surgeon's on-call status • Level of training (postgraduate year) • Emergency or non emergency status of procedure Survey distributed to all surgical residents who rotated on these services during a 2 month period: • Current service rotation • Number of hours slept during his/her last night on-call in-house • Number of hours usually slept when in-house on-call • Whether lack of sleep ever interfered with performance as a resident	Sample size: n = 6541 surgical cases Of these, n = 351 cases identified with complications Median age: not reported Gender (M/F): not reported	Inclusion: All emergency and elective procedures done by residents on the services of general surgery, vascular surgery, oncologic surgery, paediatric surgery and transplantation surgery Exclusion: Procedures done by attending surgeons or surgical fellows. Services of cardiothoracic and plastic surgeries

Author	Intervention	Study Design	Study population	Inclusion/Exclusion Criteria
Jakubowicz et al 2005 Location Department of Otolaryngology, Montefiore Medical Centre – Albert Einstein College of Medicine, Bronx, New York USA	Objective: To determine whether an endoscopic sinus simulator can measure performance changes before and after a 24 hour on-call period in residents following mandated work hour limitations. Training Trained on novice mode of simulator, 6 or 7 times over a month. This level corresponded to a plateau in the learning curve for this simulator. Training occurred between 0800 and 1300 h. First session lasted approx 2 hour – regimented, interactive training between student and proctor. Other sessions lasted 20 – 40 minutes with less coaching as skills increased. During each trial, subjects answered questions related to number of hours slept, caffeine, alcohol, cigarettes during the previous 24 hours. No training occurred after a 24 hour on-call period. Exposure Call Assessment Each resident served as his or her own control. Tested twice before and twice after a 24 hour on-call. Order between pre call and post call altered to minimise variance attributable to practice effects. Subjects answered same questions related to number of hours slept, caffeine, alcohol, cigarettes during the previous 24 hours. During post call trial subjects completed a sleep log quantifying the number of interruptions, total hours of sleep and fatigue at time of testing. Device: Endoscopic Sinus Simulator (Mentice AB) Virtual reality	Case series (pre-test/post-test outcomes) Prospective, case control cross over trial, single centre Participant selection: not reported Level of evidence: IV Blinding: none Lost to follow-up: not reported Study period: June – September 2003 Operator details: all training performed by one of the study investigators Outcome measures and validity: • Time and accuracy. • Overall score = (ideal time x 1/trial time) x accuracy • A hazard score is a negative score reflecting the % of area of each obstacle dissected. • Overall simulation score = navigation score + injection score + dissection score + hazards score	Sample size: n = 8 (but PGY values reported = 10) • PGY1: n = 4 • PGY2: n = 2 • PGY3: n = 2 • PGY4: n = 2 None had used an endoscope before. Mean age: 35 years (range 28 45) Gender (M/F): 7/1	Inclusion: surgical residents rotating on general surgery service Exclusion: not reported

Author	Intervention	Study Design	Study population	Inclusion/Exclusion Criteria
Kahol et al 2008 Location Simulation and Education Training Centre, Banner Good Samaritan Medical Centre, Phoenix, Arizona USA	Objective: to evaluate the effect on fatigue (pre- and post- call) on residents' ability to accomplish tasks that require both psychomotor and cognitive skills. AI participants Participants completed fatigue questionnaire. 8 measurement sessions (4 pre and 4 post call). Exposure Call Intervention 3 random exercises performed during each session. Each exercise repeated twice. Cyberglove and Polhelmus Liberty Tracker worn on dominant hand. Exercises pre call and post call were not matched to control for learning effect. Tasks Tasks that require both pshychomotor and cognitive decision making skills evaluated on a visiohaptic simulator: • Virtual ring transfer task - part of a validated basic laparoscopic course offered by the ProMIS and FLS simulators. • Sensorimotor coordination exercise • Slow 2-dimensional tracking exercise • 3-dimensional tracking exercise • Orientation exercise • Preparatory attention exercise • Working memory exercise • Visiohaptic transfer Device ProMIS and FLS simulators	Case series (pre-test/post-test outcomes) Prospective, unclear how many centres Level of evidence: IV Lost to follow-up: not reported Study period: not reported Operator details: not reported Outcome measures and validity: Level of fatigue assessed by questionnaire (Behrenz & Monga 1999). • Scale of 1 (least fatigued) to 10 (most fatigued) Laparoscopic proficiency measured by combination of hand and tool movement. Tool and hand movement representative of economy of motion and overall smoothness in execution. Measures of proficiency: • Gesture-level proficiency (construct validity - Kahol et al 2006) • Hand and tool movement smoothness (tool movement valid – Gallagher et al 2001) • Time elapsed • Cognitive errors	Sample size: n = 37 • PGY 1 or 2: n =25 • PGY 3 or higher: n =12 Mean age: not reported Gender (M/F): 18/19	Inclusion: trauma surgery and obstetric/gynecologic residents Exclusion: not reported

- Behrenz K, Monga M. Fatigue in pregnancy: a comparative study. American Journal of Perinatology 1999; 16(4): 185-188.
- Gallagher AG, Richie K, McClure N, et al. Objective psychomotor skills assessment of experienced, junior and novice laparoscopists with virtual reality. World Journal of Surgery 2001; 25: 1478-1483.
- Kahol K, Smith M, Tripathi P, et al. Gesture based hand movement analysis and haptic feedback for surgical training. Poster presentation at the 14th Meeting of medicine meets virtual reality, Ling Beach, CA: Jan 25-28, 2006.

Author	Intervention	Study Design	Study population	Inclusion/Exclusion Criteria
Leff et al 2008 Location Department of Biosurgery and Surgical Technology, Imperial College London, London UK	Objective: To assess the impact of 7 consecutive night shifts on newly acquired surgical performance of junior residents. Training Supervised training 1 week prior to study commencement. Training involved working through a proficiency based VR training curriculum. Participants performed tasks on the easy, medium and hard levels. All tasks repeated in 2 separate sessions on the same day, spaced at last 1 hour apart. This was done to give practice distribution. Subjects were considered sufficiently trained based on their performance for 2 of the most complex tasks (these 2 tasks have shown to be construct valid). A pilot study demonstrated that these criteria were attainable after approx 3 days prior to the start of the night shift duty. All subjects achieved predefined benchmark the day prior to the start of their night shift and aimed to minimise effects of on going learning during the week of nights assessment. Exposure 7 consecutive night shifts General surgical, orthopaedics and O&G residents: 2000 h to 0800 h Emergency medicine residents: either 2000 h – 0700 h or 2100 h – 0800 h. Participants wore pedometer. Assessment Between 0800 h – 1030 h following each night shift. Completed the same 2 higher level tasks on which proficiency based. No feedback given to participants. Follow up testing All participants returned 1 week after initial nightshift duty to perform the 2 simulator tasks during normal duty hours. Device: MIST-VR (Mentice Medical Simulation, Gothenberg, Sweden)	Case series (pre-test/post-test outcomes) Prospective, observational cohort study Level of evidence: IV Blinding: not reported Lost to follow-up: not reported Study period: not reported Operator details: Assessment supervised by a study coordinator. Outcome measures and validity: • VR training curriculum validated (Aggarwal et al 2006) • Assessment metrics measured by simulator • Questionnaire (approx total sleep in last week, hrs daytime sleep before nightshift, length of naps during night shift, quality of sleep, number if interruptions, number caffeinated drinks and cigarettes). • Completed Epworth Sleepiness Scale. • Epworth Sleepiness Scale (developed by Johns 1991; validated by Johns 1992). • Participants recorded number of patients, calls and operations attended.	Sample size: n = 21 • General surgery: n = 7 • Obstetrics and gynaecology: n = 4 • Orthopaedics: n = 2 • Emergency medicine: n = 8 Variable prior laparoscopic experience. Minimal, and no participant had performed one unsupervised. Median age: not reported Gender (M/F): not reported	Inclusion: junior post graduate residents from surgical specialties Exclusion: not reported

- Aggarwal R, Grantcharov T, Moorthy K, et al. A competency-based virtual reality training curriculum for the acquisition of laparoscopic psychomotor skill. *American Journal of Surgery* 2006; 191: 128-133.
- Johns MW. A new method for measuring daytime sleepiness: The Epworth sleepiness scale. *Sleep* 1991; 14: 540-545.
- Johns MW. Reliability and factor analysis of the Epworth sleepiness score. *Sleep* 1992; 15: 376–381.

Author	Intervention	Study Design	Study population	Inclusion/Exclusion Criteria
Light et al 1989 Location Department of Surgery and Psychiatry, University of Colorado Health Sciences Centre, Denver, Colorado USA Extended abstract of a paper presented at 1998 conference	Objective: to determine the effects of sleep deprivation on mood and performance Participants placed in 21 matched pairs. Exposure Rest and fatigue (less than 4 hours sleep) Assessment Each participant took a series of 8 neuropsychologic tests regarded as Test A or Test B, each of which had variation of rest and fatigue. This battery examined abilities in sustained concentration, clear thinking, repetitive skills, verbal attention, memory, judgment, and learning.	Non-randomised comparative Prospective Level of evidence: III-2 Lost to follow-up: not reported Study period: not reported Operator details: not reported Outcome measures and validity: • POMS • Manual dexterity pegboard	Sample size: 42 • Rested n = 21 • Fatigued n = 21 Median age: not reported Gender (M/F): not reported	Inclusion: surgical house staff (residents) Exclusion: not reported

Author	Intervention	Study Design	Study population	Inclusion/Exclusion Criteria
Minion et al 2007 Location Department of Surgery, University of Kentucky College of Medicine, Lexington, Kentucky USA	Objective: to evaluate effect of on-call status on ABSITE scores Exposure Call Assessment ABSITE scores (standardised and percentile ranks) for all general surgery residents from 1999 – 2006 recorded. Compared with on-call status of the night before, and the postgraduate year.	Non-randomised comparative Single centre, retrospective Level of evidence: III-2 Lost to follow-up: not reported Study period: 1999 - 2006 Operator details: not reported Outcome measures and validity: • ABSITE score • On-call status • Postgraduate year • USMLE Step 2 performance	Sample size: n = 282 • (n = 69 on-call) Median age: not reported Gender (M/F): not reported	Inclusion: not reported Exclusion: not reported

Author	Intervention	Study Design	Study population	Inclusion/Exclusion Criteria
Reznick & Folse 1987 Location Department of Surgery, Southern Illinois University School of Medicine, Springfield, Illinois USA	Objective: investigate whether sleep deprivation negatively effects the performance of surgical residents Exposure Acute sleep deprivation defined as less than or equal to 4 hours of sleep in the preceding 24 hours. Residents who had more sleep were considered rested. Assessment • Subjects tested between 0800 h and 1800 h during the working day. • Subjects tested in sleep-deprived state only if it was a result of previous night's duties. No attempt was made to control for caffeine.	Randomised controlled trial Prospective, single centre Level of evidence: II Method of randomisation: random numbers table Blinding: not reported Method of allocation concealment: not reported Intention-to-treat analysis: not reported Lost to follow-up: none (but 9 participants were only tested in the non sleep-deprived condition) Study period: not reported Operator details: not reported Outcome measures and validity: • Factual recall test • Concentration ability task • Manual dexterity task • Manual dexterity task (Purdue Pegboard)	Sample size: n = 21 Mean age: 29 ± 2.5 (range 25 – 34 years) Gender (M/F): not reported	Inclusion: residents in training in department of surgery Exclusion: not reported

Author	Intervention	Study Design	Study population	Inclusion/Exclusion Criteria
Sawyer et al 1999 Location Department of Surgery, University of Virginia, Charlottesville, Virginia USA	Objective: to assess the impact of different call schedules on intern performance and education Exposure Intern schedule determined primarily by number of residents taking call and desires of staff. Long call – overnight, in-hospital duties General schedule patterns: • 2 interns taking long call every other day • 3 interns taking long call ever 3rd day • 2 interns taking long call every 4th night and day call only (short call) every 4th night, with patients cared for every other night by an intern on a separate service (cross over) Generally 14 interns assigned (2 per service) to 3 general surgical services, or the cardiothoracic, transplant, paediatric or vascular surgery services. Kept sleep logs and completed anonymous monthly surveys Errors defined as preventable mistakes that delayed or diminished the care of patients. Assessment Subjective surveys by residents and staff	Case series (pre-test/post-test outcomes) Prospective, observational, single centre Level of evidence: IV Lost to follow-up: • 260/560 weekly sleep/ operative logs not returned • 56/126 monthly questionnaires not returned • 52/207 faculty questionnaires not returned Study period: 1997 - 1998 Operator details: not reported Outcome measures and validity: • Sleep log – call status, cross covering, hours slept, no of cases • Monthly surveys – hours of sleep, service census, frequency of fatigue, sleepless nights, errors, inability to complete work, operating room participation, stress levels • Faculty surveys	Sample size: n = 19 Median age: not reported Gender (M/F): not reported	Inclusion: first year surgical residents Exclusion: • The 5 interns per month assigned to subspecialty services or on holidays were not studied • Non surgery interns rotating on surgery services • Stressful months not used

Author	Intervention	Study Design	Study population	Inclusion/Exclusion Criteria
Schieman et al 2008 Location Department of Surgery, University of Calgary, Tom Baker Cancer Centre, Calgary, Alberta Canada	Objective: to determine whether fatigue adversely impacts surgical outcomes for rectal cancer, with respect to complications and cancer recurrence. Exposure Defined surgeons fatigued if the surgeon billed for clinical work after 2200 h the night before the anterior resection. All other cases were considered nonfatigued. Assessment Cases cross referenced with surgeons billing data. Intraoperative complications included injury to other organs, gross spillage of stool, stapler misfire, or blood loss greater than 1 L. Long term complications included chronic abscesses, incisional hernias, anastomotic strictures, and bowel obstructions requiring hospital admission.	Non-randomised comparative Retrospective, 2 teaching hospitals, blinded chart review Level of evidence: III-2 Blinding: chart reviewers blinded to surgeon fatigue status Lost to follow-up: not reported Study period: 1994 - 2005 Operator details: not reported Outcome measures and validity: cross referencing with billing data is valid Outcomes: Patient characteristics, perioperative complications, long term complications, cancer recurrence graded according to the Clavien Complication Scale (Clavien et al 1992; Dindo et al 2004). Major complications defined as those with Clavien grade 3 or higher (these complications are significant and by definition typically require surgical or radiologic intervention)	Sample size: n = 270 cases • Done by fatigued surgeons n = 22 • Done by nonfatigued surgeons n = 248 Gender (M/F): not reported Groups well matched. No significant differences in respect to age, body mass, American Society of Anaesthesiology class, Charlson comorbidity score, proportion of patients with low-lying tumours, or anastomotic height.	Inclusion: • adult patients operated on by fellowship trained colorectal surgeons Exclusion: • patients undergoing nonrestorative procedures (abdominoperineal resection or local excision) • patients with stage 4 disease • patients with incomplete data (missing hospital records, missing operating room records, missing pathology reports, absence of the recorded follow up etc). • procedures performed by non-colorectal surgeons

- Clavien D, Sanabria JR, Strasberg SM. Proposed classifications of surgery with examples of utility in cholecystectomy. Surgery 1992; 111: 518-26.
- Dindo D, Demartines N, Clavien PA. Classification of surgical complications: a new proposal with evaluation in a cohort of 6336 patients and results of a survey. Annals of Surgery 2004; 240: 205-13.

Author	Intervention	Study Design	Study population	Inclusion/Exclusion Criteria
Stone et al 2000 Location Department of Surgery, Harvard Medical School and Deaconess Hospital and Department of Biostatistics, Harvard School of Public Health, Boston, Mass USA	Objective: to evaluate effect of on-call status on ABSITE scores Exposure Call Assessment ABSITE scores (standard score) for all general surgery residents in 1994 obtained from 15/21 general surgery programs. Compared with on-call status of the night before, the postgraduate year and program track. Program directors were surveyed regarding their use of ABSITE scores in making promotion or dismissal decisions about individual residents. Individual programs' usual call frequency was extracted from self-reports published in Surgery's "Little Red Book".	Non-randomised comparative Multicentre, retrospective Level of evidence: III-2 Blinding: investigators were blinded to resident names Lost to follow-up: • 6 centres did not provide data • Of those that responded 1/15 program directors did not respond to survey Study period: 1994 Operator details: not reported Outcome measures and validity: • ABSITE score • Postgraduate year • Program track • On-call status	Sample size: n = 424 • On-call: n = 70 • Off-call: n = 354 Median age: not reported Gender (M/F): not reported	Inclusion: scores from 21 general surgery programs for general surgery residents Exclusion: not reported

Author	Intervention	Study Design	Study population	Inclusion/Exclusion Criteria
Taffinder et al 1998 Location Minimal Access Surgical Unit, Imperial College School of Medicine. St Mary's London UK	Objective: The effect of sleep deprivation on surgeons' laparoscopic skills using a virtual reality laparoscopic simulator. All participants Surgical dexterity measured using Imperial College Surgical Assessment Device (ICSAD). Pre-trained on the simulator. Standard stress and arousal questionnaire completed after each test. Simulator measured two handed skill with use of foot pedal to simulated electro-coagulation of virtual targets. Practice effects controlled by a Latin square design. Exposure Sleep deprivation Assessment Tested on 6 different nights, with 1 week intervals. • 20 repetitions in the evening (1700 – 1800 h) and identical testing the following morning (0800 – 0900 h) after either: o An undisturbed night (n = 12) o Sham night on-call (disturbed at 000 h, 0300 h, 0600 h) (n = 12) o Night with no sleep (n = 12) Device MIST-VR, Ethicon, UK	Case series (pre-test/post-test outcomes) Single centre, prospective Level of evidence: IV Lost to follow-up: not reported Study period: not reported Operator details: not reported Outcome measures and validity: • Dexterity - Imperial College Surgical Assessment Device (ICSAD) – valid (Taffinder et al 1998). • Standard stress and arousal questionnaire - valid (King et al 1983). • Surgical skills - MIST-VR • Errors • Time to complete task • Stress • Arousal	Sample size: n = 6 Median age: 30 (range 26 -33 years) Median years since qualification: 5 (range 2 – 9) Gender (M/F): not reported	Inclusion: surgeons in training Exclusion: not reported

- King et al. Measurement of stress and arousal: validation of the stress/arousal adjective checklist. British Journal of Psychology 1983; 74: 473-479.
- Taffinder N, Sutton C, Fishwork RJ, McManus IC, Darzi A. Validation of virtual reality to teach and assess psychomotor skills in laparoscopic surgery. In: Westwood J, Hoffman H, Stredney D, Weghorst S (Eds). Technology and Informatics 50: proceedings of medicine meets virtual reality 6 – 1998 Jan 28-31; San Diego, USA. Amsterdam: IOS Press. 1998: 124-130.

Author	Intervention	Study Design	Study population	Inclusion/Exclusion Criteria
Uchal et al 2005 Location Department of Research and Development, Forde Health System, University of Bergen, Forde Norway	Objective: To compare the impact of sleep deprivation after 24 h duty (post call) with that of 8 h work (post work) on product quality and procedure effectiveness in suturing of a perforated ulcer in a laparoscopic physical simulator. All participants Amount of sleep during 24 h prior to testing was self-reported. Prior to performance of tasks, participants familiarised with equipment and tasks. Participants viewed videotape of expert surgeon's performance featuring 11 overall actions. Information given individually on outcome measures used for evaluation. Pretest of all participants at 1600 – 1700 h involving : • Epworth Sleepiness Scale to measure sleepiness • MIST-VR Exposure • Post call study arm: on-call from 0800 h to 0800 h • Post work study arm: worked from 0800 h to 1600 h, with undisturbed night sleep beforehand Assessment Suturing a perforated ulcer - a foam hollow stomach placed in a pelvic trainer • Post call study arm assessed between 0800 h - 0900 h after on-call duty in hospital. • Post work study arm assessed between 1600 h – 1700 h after working day. All tasks video taped, identifying participants by identity numbers. Device: Pelvic Trainer, USSC, Norwalk, CT.	Randomised controlled trial Multicentre, prospective Level of evidence: II Method of randomisation: computer generated random sample. Subjects randomly assigned to either post-call or post-work study arms Blinding: outcome measures evaluated by 2 assessors blinded to participants' identity and arm assignment. Method of allocation concealment: identity numbers given to participants. The generator of assignment was separated from its executor. Intention-to-treat analysis: not reported Power calculation: sample size based on previous study (Bergasmashi & Dicko 2000) which showed that 60 subjects would be needed to detect significant differences in operating time. Power of study was 80%. Parallel block randomisation used. Lost to follow-up: n = 2. Withdrew after the pre-test and were excluded Study period: not reported Operator details: Evaluation of all outcome measures performed by 2 raters and inter-rater reliability (IRR) assessed. IRR defined as extent to which rating of the 6 outcome measures yielded consistent results when carried out by 2 independent raters. Kendall's tau concordance coefficient was 1.0 ($P < 0.0001$) for AR, TD, LR and OT. It was 0.75 ($P = 0.325$) for GDA and 0.77 ($P = 0.305$) for NGDA. Outcome measures and validity: • Assessed by observers from video tapes • Outcome measures such as AE, TD, LR,	Sample size: n = 66 • Surgeons: n = 64 • Nurse controls: n = 64 Mean age (years): Surgeons • Post call: 33 • Post work: 38 Nurse controls • Post call: 35 • Post work: 40 Gender (M/F): Surgeons • Post call: 19/13 • Post work: 24/8 Nurse controls • Post call: 8/24 • Post work: 10/22 Surgeons on the study arms were comparable for age, gender, duration of practice, and ESS and MIST-VR scores. Nurses in the study arms were well matched for age, gender, practice duration, and ESS and MIST-VR scores, but not for hours slept in previous 24 hours.	Inclusion: • Board certified general surgeons, gynecologists, orthopedic surgeons, urologists, vascular surgeons with extensive experience in minimally invasive surgery • Control: registered nurses, registered scrub nurses, midwives Exclusion: not reported

		GDA, NGDA and OT valid • Construct validity was defined as its ability to distinguish between trained (surgeons) and untrained (nurses) subjects.		

• Bergasmashi R, Dicko A. Instruction versus passive observation: a randomised educational research study on laparoscopic suture skills. *Surgical Laparoscopy Endoscopy & Percutaneous Techniques* 2000; 10: 319-322.

Author	Intervention	Study Design	Study population	Inclusion/Exclusion Criteria
Wesnes et al 1997 Location Cognitive Drug Research Ltd, Reading and Behavioural Oncology Unit, Department of Surgery, Surgical Nutrition and Metabolism Unit, University of Aberdeen UK	Objective: to evaluate the emotional and cognitive effects of a weekend on-call in a surgical ward Exposure • On-call: Saturday 6 am to Monday 9 am • Not on-call Assessment Surveys Assessed in counterbalanced design on 4 Monday mornings: • twice after a weekend off-duty • twice after a weekend on-call	Case series (pre-test/post-test outcomes) Prospective, single centre Level of evidence: IV Lost to follow-up: not reported Study period: not reported Operator details: not reported Outcome measures and validity: • Cognitive Drug Research computerised cognitive assessment system o Attention o Working memory o Long term memory • Aberdeen Mood Rating Scale • Diary with weekend activities • Questionnaire on-call or off-call	Sample size: n = 10 Median age: not reported Gender (M/F): not reported	Inclusion: house officers in general surgical unit in 1 hospital Exclusion: not reported