TITLE PAGE

Title: The effect of sleep deprivation on objective and subjective measures of facial appearance

Running head: Sleep deprivation and facial appearance

Benjamin C. Holding [1]

Tina Sundelin [1, 2, 3]

Patrick Cairns [4]

David I. Perrett [4]

John Axelsson [1, 3]

Corresponding author:

Benjamin C. Holding Department of Clinical Neuroscience, Karolinska Institutet, 171 77 Stockholm, Sweden

Email: benjamin.holding@ki.se

Phone: +46 (0)707 546911

Author Affiliations:

[1] Department of Clinical Neuroscience, Karolinska Institutet, Sweden

[2] Department of Psychology, New York University, USA

[3] Stress Research Institute, Stockholm University, Sweden

[4], School of Psychology and Neuroscience, University of St. Andrews, United Kingdom

Manuscript word count: 4274

Reference count: 31

Conflict of interests for all authors: None

Authorship contribution: BCH and TS conducted the data collection. BCH led the write-up of the manuscript and conducted the statistical analysis. PC and DIP analysed the facial photographs and provided expertise regarding spectrophotometry. TS and JA conceptualised, designed and planned the data collection. All authors contributed to the drafting process and have approved the final manuscript.

SUMMARY

The faces of people who are sleep deprived are perceived by others as looking paler, less healthy, and less attractive compared to when well-rested. However, there is little research using objective measures to investigate sleep-loss-related changes in facial appearance. We aimed to assess the effects of sleep deprivation on skin colour, eye-openness, mouth-curvature, and periorbital darkness using objective measures, as well as to replicate previous findings for subjective ratings. We also investigated the extent to which these facial features predicted ratings of fatigue by others and could classify the sleep condition of the person. Subjects (N =181) were randomised to one night of total sleep deprivation or a night of normal sleep (8-9 hours in bed). The following day (at approximately 14:00) facial photographs were taken and, in a subset (N =141), skin colour was measured using spectrophotometry. A separate set of participants (N = 63) later rated the photographs in terms of health, paleness, and fatigue. The photographs were also digitally analysed with respect to eye-openness, mouth-curvature, and periorbital darkness. Bayesian linear modelling revealed that neither sleep deprivation nor the subjects' sleepiness were related to differences in any facial variable. Similarly, there was no difference in subjective ratings between the groups. Decreased skin yellowness, less eye-openness, downward mouth-curvature, and periorbital darkness all predicted increased fatigue ratings by others. However, the combination of appearance variables could not accurately classify sleep condition. These findings have implications for both face-to-face and computerised visual assessment of sleep loss and fatigue.

Keywords:

Sleep loss, Experimental psychology, Perception, Face, Skin, Health

INTRODUCTION

Humans are incredibly apt at extracting social information from the faces of others. Indeed, seeing a face for only 100ms is enough for lasting judgements to begin forming about the person (Willis and Todorov, 2006). A predisposition to reading faces may have significant evolutionary benefits. For example, we are able to predict whether someone is ill or not just by observing facial photographs (Axelsson et al., 2018), likely because of changes in skin colour and facial expression (Henderson et al., 2016). Judgements of ill-health can subsequently lead to avoidance behaviours, hypothesised to slow the spread of disease (Park et al., 2013).

Humans may also be sensitive to sleep-loss-related facial cues. In a within-subjects study, one night of total sleep deprivation led participants to be rated as looking less healthy, less attractive, and more tired, compared to when they were well rested (Axelsson et al., 2010). Recently, it was also shown that these effects generalise to individuals who were sleep-restricted (4 hours of sleep for 2 days; Sundelin et al., 2017). Moreover, when comparing facial photographs of patients with obstructive sleep apnea before and after treatment, patients post-treatment were rated as looking more alert, attractive, and youthful (Chervin et al., 2013). However, since the measurements were taken up to 4 months following treatment, it is difficult to know whether the effects relate specifically to decreased sleep loss or a general improvement in health due to the treatment.

Previous studies have also explored the specific facial cues that may lead to these changes in subjective judgement. For example, faces of sleep-deprived people were more likely to be rated as having hanging eyelids, red eyes, swollen eyes, dark circles under the eyes, pale skin, wrinkles/fine lines, and droopy corners of the mouth (Sundelin et al., 2013). Evidence of dark

circles under the eyes was not observed in a separate study comparing good to poor habitual sleepers, but poor sleepers did score higher on a skin aging index (Oyetakin-White et al., 2015). Relatedly, eye-openness and mouth-curvature have been objectively measured following sleep restriction, showing a decrease in eye-openness, but no difference in mouth-curvature (Talamas et al., 2016).

Our increasing understanding about how sleep loss affects facial appearance opens future possibilities for creating algorithms that can automatically detect facial cues of fatigue. The rapid development of facial recognition technologies (Owen, 2018), means that soon fatigue recognition may need nothing more than a smartphone. Indeed, analysis of facial cues such as skin colour and shape of the mouth corners have already been successfully utilised to identify fatigue and sleep disorders (Chen et al., 2015; Espinoza-Cuadros et al., 2015; Peng et al., 2017). Yet, since little empirical evidence currently exists, the best features on which to base automatic sleep loss and fatigue recognition remains an open question.

Overall, evidence suggests that the effects of sleep loss are detectable in human faces. However, since the majority of the data come from subjective ratings, the extent to which such ratings parallel objective facial features is unknown. Attempting to identify specific facial cues effected by sleep loss via subjective ratings can be problematic due to the amount of shared variance with other (potentially confounding) environmental cues. The Brunswik lens model (Brunswik, 1956; Jones, 2018) proposes measuring along two dimensions: 1) *cue validity*, measuring whether features are a valid indicator of the underlying state and 2) *cue utilisation*, whether the cues are utilised by observers to formulate judgements. In this study, we investigate cue validity by measuring the effect of one night of total sleep deprivation on four objective aspects of appearance: general facial skin colour, eye-openness, mouth-

curvature, and periorbital skin darkness (i.e. dark patches underneath the eyes). These specific features were chosen since they represent features that were most widely affected by sleep deprivation when measured via subjective ratings (Sundelin et al., 2013). We investigate cue utilisation, by examining the extent to which these objectively measured facial cues predict subjective ratings of fatigue by others.

We hypothesised that sleep deprivation would lead to increased facial paleness, decreased eye-openness, and increased periorbital darkness. Although not found after partial sleep restriction (Talamas et al., 2016), we expected sleep deprivation to also lead to negative mouth-curvature (more droopy corners of the mouth), since this was previously reported using subjective ratings in a study of total sleep deprivation (Sundelin et al., 2013). Additionally, we attempted to replicate the previous findings in ratings of decreased health, and increased paleness and fatigue of sleep-deprived faces. Cue utilisation was examined in an exploratory manner. Finally, we used a machine learning approach to investigate to what extent the combination of all measured facial appearance variables could accurately classify a persons' sleep state.

METHOD

Participants

Subjects

181 healthy individuals (103 female; mean age = 25.39 years, standard deviation (SD) = 6.49 years) took part in a study where subjects were randomised (while keeping an equal number of participants in each condition) into either one night of total sleep deprivation (N = 91) or a night of normal sleep at home (N = 90). Subjects completed an online screening, with exclusion criteria including health problems, poor habitual sleep, or a sleep need outside of 7–9 hours per night, consuming more than 4 cups of coffee per day or being a current smoker (699 potential subjects were excluded at this point). One further subject was removed due to being outside of our age range requirement (18-45) at the time of testing. This upper age limitation was set to reduce any confounding effect of the changing sleep need seen in older adults. The complete screening criteria can be found as a supplement to a previous publication (Holding et al., 2019). The study was approved by the Stockholm Regional Ethical Review Board (no. 2014/1766-32). All subjects gave written informed consent and received financial compensation for participation (non-sleep deprived 800SEK, sleep-deprived 1500SEK).

Raters

63 additional individuals (32 female; mean age = 23.37 years, SD = 3.95 years) were recruited to rate the photographs of the subjects. These raters were obtained from the student population in Stockholm, Sweden. There were no exclusion criteria apart from being required to understand Swedish. Each rater gave written informed consent and was compensated with a cinema ticket.

Materials and Procedure

Each subject was instructed to keep a daily sleep diary for three days before the test day, as well as wear an actigraph (GeneActiv Sleep, Activinsights, Kimbolton, UK or MotionWatch 8, CamNtech, Cambridge, UK) on their non-dominant wrist. Subjects were given instructions to be in bed for 8-9 hours each night, turn off the lights at 23:00 ± 60 minutes, get up at 07:00 ± 60 minutes, and get 8 hours of sleep in that time. Subjects were asked to avoid naps for 4 days before the test day, to abstain from alcohol, and not drink caffeinated drinks later than the morning of the day before the test day.

After three days of this baseline sleep, i.e. on the day before the test day, subjects were informed of which condition they had been randomly assigned to. Those in the sleep-deprivation condition were required to come to the lab at 22:00 that night, and those in the well-rested control condition were instructed to sleep one more night at home and arrive at 10:00 the following day.

During sleep deprivation, subjects were kept in a light-controlled sleep lab and free to choose their activities (e.g. study, use their mobile phone, or watch a film). A research assistant was present at all times to ensure that the subject remained awake. Low-sugar snacks were provided if the subject was hungry, and a 15-minute morning walk was taken to reduce the difference in light-exposure and activity compared to what well-rested control participants may experience while travelling from their home to the lab.

At approximately 14:00 during the test day, subjects were photographed and rated their subjective sleepiness on the Karolinska Sleepiness Scale (KSS; Åkerstedt and Gillberg, 1990) Subjects were not permitted to wear makeup, jewellery, or glasses. All subjects wore an

identical dark-blue t-shirt and for subjects with longer hair, hair bands/clips were used to prevent hair from covering the face. While being photographed, subjects were instructed to sit comfortably, look straight into the camera, and relax their face. As a rule, ten photographs were taken, and after removing poor quality photographs (e.g. subject's eyes were closed), the most representative photograph of each subject was chosen by a person not involved in the study, and blind to the conditions. Photographs were taken using a digital camera (D90, Nikon Corporation, Tokyo, Japan; settings: f-stop = 11, shutter speed = 1/125, ISO = 200) with a white backdrop, professional flash lighting, and a white-balance card.

Following the photo session, skin colour was measured at three facial locations (left/right cheek, and forehead) using a spectrophotometer (CM-700d, Konica Minolta Inc, Tokyo, Japan). This was done only for the last 141 subjects of the study, due to a delay in obtaining the equipment. Colour was assessed using the CIELAB system in three dimensions – dark-light (L*), green-red (a*), and blue-yellow (b*). Higher values represent greater intensity of the second specified colour (i.e. higher a* value indicates greater skin redness). The CIELAB system has been used in previous studies investigating facial appearance, and is a well-validated system for measuring skin colour in humans (Stephen et al., 2009; Tan and Stephen, 2013).

Rating procedure

On a separate occasion, the selected photographs of the subjects' faces were shown to the raters. The images were presented using E-Prime (Version 3.0, Psychology Software Tools Inc., Sharpsburg, USA) on a laptop monitor (HP ZBook, 17.3", Hewlett-Packard Inc., Palo Alto, USA). The instructions were to rate the photographs according to one's first impression and to think of the photographs as if they had been taken five minutes ago. Subsequently, a trial block with four faces was presented to ensure that the raters understood the instructions.

The 181 facial photographs were then presented one at a time in a random order in three blocks, each block concerning one rating type (i.e. of fatigue, health, or paleness). Each face was thus presented three times, once in each block. Due to a programming error, while the fatigue and health blocks were in a randomised order (blocks 1 and 2) between raters, the paleness block was always last (block 3). The task was self-paced, but with a time limit of 5 seconds per rating. Subjects were allowed to take a short break between each block. The session took approximately 30 minutes to complete.

The faces were rated on seven-point scales pertaining to fatigue (in Swedish: trötthet; "How fatigued is this person?", 1 - very fatigued, to 7 - very alert), health (in Swedish: hälsa; "How is this person's health?", 1 - very bad health, to 7 - very good health), and paleness (in Swedish: blekhet; "How pale is this person's skin?", 1 - not pale at all, to 7 - very pale). Participants were left to make their own interpretation of each word. To make for easier interpretation, the fatigue item was reverse scored in the analysis so that higher score represents greater fatigue.

Image measurements

Psychomorph version 6 (http://users.aber.ac.uk/bpt/jpsychomorph) and Webmorph (DOI: 10.5281/zenodo.1073696) were used to define facial landmarks and construct illustrative facial averages of sleep-deprived versus well-rested participants (Figure 1) and the highest versus lowest fatigue-rated faces (Figure 2). Following previously used methodology (Talamas et al., 2016), the degree of eye-openness of each face was measured by dividing the vertical distance from the centre of the pupil to the top eyelid by the width of the eye inner canthus to outer canthus. Higher values represent greater eye-openness. Mouth-curvature was measured by subtracting the height of the mouth's centre from the mean height of the left and right corners of the mouth, and then dividing by the width of the mouth (Talamas et al.,

2016). This gives a continuum of data where negative values represent downturned corners of the mouth and positive values represent upturned corners. The colour of the faces in each photograph was analysed from cropped patches from the forehead and left/right cheek (see Figure S1). This was done to increase accuracy for the association between ratings and skin colour, as the colours in the images (and hence what the raters viewed) may vary slightly from the spectrophotometer data. Periorbital darkness was measured using the photographs and analysing the lightness of cropped patches from below the left and right eyes (see Figure S2). Since we were interested in the difference relative to the subject's skin colour, we subtracted this value from the lightness of the forehead region (same as the forehead region shown in Figure S1). A higher value represents that the periorbital area is darker relative to the forehead.

[Suggested insert point for Figure 1 and Figure 2]

Statistical procedure

All data were analysed using statistical procedures with R (R Core Team., 2016). The data and statistical code can be viewed online (DOI: 10.5281/zenodo.1414101). Bayesian statistical models were estimated using Markov Chain Monte Carlo sampling using 10000 iterations (1000 warmup) in 15 chains. A prior distribution (representing expectations under the null hypothesis) was set on the possible effects. This was always zero, with a standard deviation of a quarter of the highest possible score. Priors on eyelid-openness, mouth-curvature, and periorbital darkness were set using a quarter of the range between the smallest and largest observed values. All other parameters used the default non-informative priors of the brms package (Bürkner, 2017; Kruschke and Meredith, 2018).

To analyse the effect of sleep deprivation on skin colour, cross-classified multilevel models were used, as the data contains both repeated measures (within-rater and within-face) and between-subjects (sleep deprivation vs well-rested controls) measurements. When analysing subjective ratings of health, fatigue, and paleness of the images as the response variables, ordinal models were used since Likert-type scales cannot be assumed to be interval data (Liddell and Kruschke, 2018). Multiple regression modelling was used to investigate the effects of sleep deprivation on eye-openness, mouth-curvature, and periorbital darkness as well as to analyse the predictors of subjective fatigue ratings. Identical analyses were also run for self-reported sleepiness scores, replacing subjects' condition as the independent variable.

The fitted models provide a point effect estimate (the posterior median) and a probability distribution of plausible values (posterior distribution). Uncertainty around this point was quantified using the 95% Highest Density Interval (95%HDI; 2.5% and 97.5% quantiles of the posterior) representing the most credible estimated values. A Bayesian p-value, p^{MCMC}, was also calculated. This is defined as two times the probability that the parameter value is less than or greater than zero, using the smaller of these probabilities (Hadfield, 2010). Parameter estimates were considered statistically "significant" if both (a) the 95% HDI did not overlap zero and (b) $p^{MCMC} < .05$.

In order to elucidate the meaningfulness of an effect, a 'region of practical equivalence' (ROPE) was included. ROPE is a range of values close to zero representing an effect that is too small to be meaningful. Effect sizes of less than Cohen's d = .2 have been used in previous research as practically equivalent to zero (Pedersen et al., 2018), and correspond to the threshold conventionally associated with "small" effects (Cohen, 1988). However, there is no hard-and-fast rule, and the data can be reanalysed using one's own preferred ROPEs. To construct the upper and lower boundaries of the ROPE for each analysis, raw difference score

values corresponding to ± .2 standard deviation were calculated (Kruschke, 2018). The rule used in this study was that if 95% of the posterior distribution lies outside of the ROPE, the effect is *meaningfully* non-zero, and therefore the null hypothesis can be rejected. Similarly, if 95% of the posterior distribution lies within the ROPE, any probable effect is too small to be meaningful and the null hypothesis can be accepted. Any other state suggests that more data are needed to accept or reject hypotheses at our specified level of certainty (95% probability). To allow for inter-interpretability between outcomes, all continuous predictors were centred and scaled so that 1 unit of the predictor represents a change across 2 standard deviations (95% of the data).

Finally, we tested if the combination of all appearance variables could be used to accurately classify someone as being sleep deprived or well-rested. Following a procedure typically used in machine learning (Yarkoni and Westfall, 2017), the data was randomly split into a training set (70% of the data) and a testing set (30% of the data), while keeping an equal proportion of each sleep condition within both sets. Using the caret package (Kuhn, 2008) in R, a model was first fitted using a logistic regression with 10-fold cross-validation using the training dataset. The predictions made by this model was then tested against the withheld testing data set. Our measure of predictive success is the Area Under the Curve (AUC) of the Receiver Operating Curve and it's 95% confidence interval. The AUC represents the efficiency of a model, incorporating both the sensitivity and specificity of the predictions. The maximum value for the AUC is 1.0, indicating perfect classification (100% sensitive, and 100% specificity). An AUC of 0.5 indicates no ability to classify (50% sensitive and 50% specificity).

RESULTS

Descriptive statistics for objective skin colour, other facial features, and subjective rater scores for sleep-deprived and well-rested participants can be seen in Table S1. Visualisations of the correlations between spectrophotometry colour values can be seen in Figure S2.

Effect of sleep deprivation on self-reported sleepiness

Subjects in the sleep deprivation condition reported higher sleepiness ratings (Mean = 5.14, SD = 1.94) compared to the well-rested group (Mean = 3.33, SD = 1.38).

Effect of sleep deprivation on spectrophotometrically-measured skin colour and other facial features

None of the three skin-colour variables, nor mouth-curvature, eye-openness, or periorbital darkness, showed any noticeable difference between conditions, as represented by a 95%HDI that overlapped zero. The null hypothesis that no meaningful effect exists for lightness or redness was accepted, since over 95% of the posterior distribution was within the ROPE. This criterion was not reached for yellowness, eye-openness, mouth-curvature, or periorbital darkness meaning that we cannot exclude that a small effect may exist. However, since 76-88% of values were within the ROPE (representing 76-88% probability) it is unlikely that there was any clear difference between the two conditions. Full results can be seen in Table 1. The average faces of sleep-deprived versus well-rested participants, for men and women respectively, are illustrated in Figure 1.

[Suggested insert point for Table 1]

Effect of sleep deprivation on rater judgments of fatigue, health, and paleness

None of the facial ratings showed a significant difference between conditions (see Table 2). The ROPE contained the majority of the posterior distribution, suggesting higher probability of the null hypothesis (76-87%), however it did not reach the 95% probability threshold for accepting the null hypothesis.

[Suggested insert point for Table 2]

Association between self-reported sleepiness, spectrophotometrically-measured skin colour, and other facial features

Skin-colour, eye-openness, mouth-curvature, and periorbital darkness did not show a significant association with the subjects' own sleepiness ratings (see Table 3). The ROPE contained between 22-77% of the posterior distribution, which does not reach the 95% probability threshold for accepting the null hypothesis. Full results can be seen in Table 3.

[Suggested insert point for Table 3]

Association between self-reported sleepiness on other-rated fatigue, health, and paleness

None of the three factors rated in the facial images were significantly predicted by the subjects' own sleepiness (see Table 4). The probability that the effect was within the ROPE was between 53-70% which again does not reach the 95% probability threshold.

[Suggested insert point for Table 4]

What predicts perceived fatigue?

In a multiple regression model, less eye-openness, less facial yellowness (measured from the images rather than the spectrophotometry), less mouth-curvature, and greater periorbital

darkness, all predicted increases in how fatigued the faces were rated on average (see Table 5 and Figure 3). Additionally, for these four variables, the posterior distributions were found to be almost entirely outside of the ROPE, providing support that these represent meaningful effects. Facial redness and lightness did not show any association with perceived fatigue. To illustrate the most and least fatigued-appearing subjects, averaged images were made for the most fatigued-looking versus least fatigued-looking faces (10 subjects in each image), for men and women respectively, Figure 2.

[suggested insert point for Table 5 and Figure 3]

Sleep state classification

The logistic model based on the training data set had a final AUC of 0.56 (95% CI: 0.46-0.66) representing very low ability to classify sleep condition. When used with the withheld testing dataset the AUC was 0.57 (95% CI: 0.46-0.66) again demonstrating very low ability of the model (and thus the facial cues) to classify sleep condition.

DISCUSSION

This study investigated the effect of sleep deprivation on both objective and subjective measures of facial appearance. It was predicted that sleep deprivation would lead to objectively increased facial paleness, less eye-openness, negative mouth-curvature, and increased periorbital darkness, as well as subjective ratings of looking more pale, less healthy, and more tired. However, the results do not appear to support these hypotheses. Neither sleep deprivation nor subjects' self-reported sleepiness showed any relationship with spectrophotometry-measured lightness, redness, or yellowness. This is a surprising result considering the previous finding that sleep-deprived individuals were subjectively rated as

appearing more pale (Sundelin et al., 2013). However, consistent with the data from the spectrophotometry, we did not observe any distinct differences in how faces were rated with respect to fatigue, health, or paleness. This distinguishes the current study from previous ones finding associations between sleep loss and appearance (Axelsson et al., 2010; Sundelin et al., 2017). In addition, we did not observe any changes in eye-openness, mouth-curvature or periorbital darkness after sleep deprivation, or in relation to self-reported sleepiness of the photographed subjects. This again is different to previous findings, especially regarding eye-openness which has been shown to be decreased following sleep restriction (Talamas et al., 2016).

A key difference of this study compared to previous ones is the between-subjects design, i.e. the raters were presented with just one image of each subject, *either* sleep deprived *or* well-rested. Previous experimental studies have focused on differences within subjects, always including two photographs of the same subject, one from each sleep condition. It seems possible that the effects of sleep loss on facial appearance is most easily distinguished if the observer has information about what the individual looks like well-rested. However, this might not explain why objectively measured eye-openness, which was previously found to have a medium effect size decrease following sleep restriction (Talamas et al., 2016), was not seen. One possibility is that the greater inter-individual variance in our sample compared to previous within-subject studies is what is obscuring the effects. However, given the substantially larger sample size, it should be possible to observe effects nonetheless. An alternative explanation is that because previous studies took photographs on two occasions, this may prime subjects (vis-à-vis demand characteristics (Nichols and Maner, 2008)) to look more or less tired at the second photo opportunity, as the reason for taking photographs might be more apparent then. This could, for example, take the form that subjects try to force their

eyes more open. Additionally, despite the findings of previous studies being consistent, they have all been relatively small (23-25 subjects), making the likelihood of unreliable results higher. A benefit of using Bayesian statistics along with the 95%HDI metric is that it is possible to discount certain effect sizes. For example, regarding subjective ratings of paleness in sleep-deprived faces, the boundary the 95%HDI was 0.53, suggesting that this is the largest unstandardised effect (on the 7-point scale) that is credible at the 95% probability threshold. This makes it clear that while no large between-subjects effects are realistic, it does not rule out smaller ones that may be hidden in the variance.

An interesting aspect is that reduced skin yellowness, eye-openness, mouth-curvature, and periorbital darkness were all related to perceived fatigue, despite not being noticeably affected by sleep deprivation or associated with sleepiness. This suggests that despite having low cue validity, these features in fact have significant cue utilization. Raters clearly used these cues to evaluate fatigue in the photographs. This process could be an example of overgeneralisation of learnt associations, which are common during impression formation (Zebrowitz and Montepare, 2008). These overgeneralisations may be learnt from aquantances who show the effects very distinctly, or conceivably from television and film.

A key implication of our findings regards the development of automatic fatigue-detection systems. Researchers are working to produce artificial intelligence systems to detect sleep loss and fatigue using facial cues similar to those investigated in this study (Gu and Ji, 2004; Peng et al., 2017; Vural et al., 2007). The current study indicates vulnerabilities in this methodology, since it appears that commonly extracted features may have high cue utility but in fact have lower cue validity. This could mean that predictions of sleep loss become based on incorrect assumptions regarding the physical effect of sleep loss on facial appearance. We

also underline the problems of using facial data from a single point in time. Automatic fatigue recognition systems would probably need to focus on within-person changes to be effective. On a more individual level, the results imply that while we are likely poor at judging whether strangers are fatigued based on a quick look, we may have more success with friends or colleagues, by virtue of knowing how they appear when alert.

There are a number of limitations that should be considered as well as directions for future research. Firstly, we did not explicitly ask raters whether they could identify who was sleep deprived and who was not. While there was no association between subjects' self-rated sleepiness and how others rated their fatigue, it is possible that with a more direct question, others may be able to distinguish who is sleep deprived. Secondly, we did not stratify the results based on ethnicity or skin pigmentation (two trained coders classified over 85% of the faces as Caucasian), though we attempted to statistically control for this by including the other two spectrophotometer colour outcomes as covariates in each colour model. Nonetheless, future studies may benefit from investigating whether the effects of sleep loss on appearance varies depending on ethnicity. Thirdly, the rater sample was not stratified by insomnia symptoms, which previous evidence has shown to alter how individuals perceive tiredness in others (Akram et al., 2017). Fourthly, our measures of appearance focused on specific areas of the face and were only analysed using static images. An interesting future direction would be to learn more about how dynamic movements of the face act as cues of sleep loss to influence social judgements and impression formation. A handful of studies have shown that cues such as blink duration and body posture are associated with self-reported sleepiness (Anund et al., 2013; Ingre et al., 2006). Future studies analysing biological motion of sleep-deprived and fatigued people can give additional information on the validity of the presented cues, as well as indicating others. Finally, while the between-subjects design of the

data allowed us to collect a large sample, the results suggest that any possible effects must be smaller than the variance in our sample allows us reliably to test for. Future studies will likely benefit from following participants over time in order to assess relative changes in appearance.

Overall, we find that skin colour, eye-openness, mouth-curvature, and periorbital darkness were not impacted by one night of total sleep deprivation in a between-subjects design. These features nonetheless appear important for social signalling of fatigue as they predicted how individuals were rated. The results have implications for artificial-intelligence applications that attempt to identify fatigue through facial features by highlighting that accurate classification of fatigue and sleep loss through a single static image is problematic. We suggest that automatic recognition systems focused on detecting sleep-deprived or fatigued individuals should use data on changes within individuals.

ACKNOWLEDGEMENTS
This study was funded by the Swedish Research Council, FORTE (Swedish Research Council for Health, Working Life and Welfare), and The Swedish Foundation for Humanities and Social Sciences. Thank you to Lisa Debruine for help with face delineation.

DISCLOSURE STATEMENT

Financial Disclosure: None

Non-financial Disclosure: None

Table 1. Effects of sleep deprivation on objective measures of skin colour and facial features

	Estimate (posterior median)	95% HDI		p^{MCMC}	Probability effect is within ROPE	Null hypothesis decision
		Low	High			
Model 1. Lightness						
Fixed effects:						
Intercept	63.43	62.98	63.86	-	-	-
Sleep deprivation	0.06	-0.56	0.69	.85	95.66%	Accept
Redness	-7.55	-7.94	-7.18	< .001	0%	Reject
Yellowness	-2.88	-3.33	-2.43	< .001	0%	Reject
Random effects:						
Face ID (intercept)	1.90	1.68	2.15	-	-	-
Model 2. Redness						
Fixed effects:						
Intercept	12.48	12.22	12.74	-	-	-
Sleep deprivation	-0.03	-0.41	0.33	.87	96.63%	Accept
Lightness	-6.82	-7.15	-6.49	< .001	0%	Reject
Yellowness	-3.20	-3.50	-2.89	< .001	0%	Reject
Random effects:						
Face ID (intercept)	1.12	0.99	1.26	-	-	-
Model 3. Yellowness						
Fixed effects:						
Intercept	16.76	16.39	17.13	-	-	-
Sleep deprivation	-0.04	-0.56	0.50	.89	88.49%	Undecided
Lightness	-3.60	-4.17	-3.05	< .001	0%	Reject
Redness	-4.34	-4.75	-3.91	< .001	0%	Reject
Random effects:						
Face ID (intercept)	1.62	1.42	1.82	-	-	-
Model 4. Mouth-curvature						
Intercept	0.00	-0.01	0.02	-	-	-
Sleep deprivation	0.00	-0.01	0.01	.89	82.45%	Undecided
Model 5. Eye-openness						
Intercept	0.20	0.19	0.20	-	-	-
Sleep deprivation	0.00	-0.00	0.01	.60	76.16%	Undecided
Model 6. Periorbital darkness						
Intercept	7.45	6.81	8.12	-	-	-
Sleep deprivation	-0.15	-1.06	0.79	.75	80.00%	Undecided

Note. Results of six models showing changes in skin colour and facial features. Sleep deprivation was dummy coded and the intercept refers to the well-rested control condition. All continuous predictors are centred and scaled such that 1 unit of the predictor represents a change of 2 standard deviations (95% of the data). Random effect estimates represent the standard deviation. All values are on the latent scale. HDI = Highest density interval; p^{MCMC} = Bayesian p value; ROPE = region of practical equivalence around zero (defined as less that Cohen's d = 0.2). Specific ROPEs are as follows: Model 1 = 0±0.66; Model 2. = 0±0.41; Model 3 = 0±0.43; Model 4 = 0±0.01; Model 5 = 0±0.01; Model 6 = 0±0.62.

Table 2. Effects of sleep deprivation on rater-perceived fatigue, health, and paleness

	Estimate (posterior median)	95% HDI		p^{MCMC}	Probability effect is within ROPE	Null hypothesis decision
		Low	High			
Model 1. Fatigue						
Fixed effects:						
Sleep deprivation	-0.10	-0.41	0.22	.54	86.68%	Undecided
Random effects:						
Face ID (intercept)	1.06	0.94	1.17	-	-	-
Rater ID (intercept)	0.80	0.66	0.96	-	-	-
Rater ID (slope of sleep deprivation)	0.07	0.00	0.17	-	-	-
Model 2. Health						
Fixed effects:						
Sleep deprivation	0.10	-0.19	0.39	.50	87.25%	Undecided
Random effects:						
Face ID (intercept)	0.97	0.87	1.08	-	-	-
Rater ID (intercept)	1.27	1.04	1.51	-	-	-
Rater ID (slope of sleep deprivation)	0.10	0.00	0.20	-	-	-
Model 3. Paleness						
Fixed effects:						
Sleep deprivation	0.20	-0.11	0.53	.21	76.48%	Undecided
Random effects:						
Face ID (intercept)	1.07	0.95	1.19	-	-	-
Rater ID (intercept)	1.20	0.98	1.43	-	-	-
Rater ID (slope of sleep deprivation)	0.15	0.01	0.26	-	-	-

Note. Results of three ordinal multilevel models predicting changes in subjective ratings. The same table including ordinal thresholds (intercepts) can be found in Table S2. Random effect estimates represent the standard deviation. All values are on the latent scale. HDI = Highest density interval; p^{MCMC} = Bayesian p value; ROPE = region of practical equivalence around zero (defined as less that Cohen's d = 0.2; Model 1: ±0.28; Model 2: ±0.27; Model 3: ±0.32).

Table 3. Association between self-reported sleepiness and objective measures of skin colour and facial features

	Estimate (posterior median)	95% HDI		p^{MCMC}	Probability effect is within ROPE	Null hypothesis decision
		Low	High			
Model 1. Lightness						
Fixed effects:						
Intercept	63.41	63.09	63.73	-	-	-
Sleepiness	-0.11	-1.16	0.94	.84	76.61%	Undecided
Redness	-7.65	-8.04	-7.27	< .001	0%	Reject
Yellowness	-3.05	-3.51	-2.61	< .001	0%	Reject
Random effects:						
Face ID (intercept)	1.88	1.65	2.12	-	-	-
Model 2. Redness						
Fixed effects:						
Intercept	12.48	12.30	12.67	-	-	-
Sleepiness	0.01	-0.67	0.69	.99	76.31%	Undecided
Lightness	-6.77	-7.10	-6.43	< .001	0%	Reject
Yellowness	-3.27	-3.58	-2.98	< .001	0%	Reject
Random effects:						
Face ID (intercept)	1.13	0.99	1.27	-	-	-
Model 3. Yellowness						
Fixed effects:						
Intercept	16.74	16.47	17.01	-	-	-
Sleepiness	0.33	-0.58	1.25	.48	53.28%	Undecided
Lightness	-3.80	-4.36	-3.24	< .001	0%	Reject
Redness	-4.53	-4.96	-4.10	< .001	0%	Reject
Random effects:						
Face ID (intercept)	1.61	1.41	1.81	-	-	-
Model 4. Mouth-curvature						
Intercept	0.01	0.00	0.02	-	-	-
Sleepiness	-0.01	-0.03	0.02	.55	40.81%	Undecided
Model 5. Eye-openness						
Intercept	0.20	0.20	0.20	-	-	-
Sleepiness	-0.00	-0.01	0.01	.91	46.71%	Undecided
Model 6. Periorbital darkness						
Intercept	7.20	6.69	7.66	-	-	-
Sleepiness	-1.28	-3.07	0.64	.18	22.69%	Undecided

Note. Results of six models predicting changes in skin colour and facial features. All continuous predictors are centred and scaled such that 1 unit of the predictor represents a change of 2 standard deviations (95% of the data). Random effect estimates represent the standard deviation. All values are on the latent scale. HDI = Highest density interval; p^{MCMC} = Bayesian p value; ROPE = region of practical equivalence around zero (defined as less that Cohen's d = 0.2). Specific ROPEs are as follows: Model 1 = 0±0.65; Model 2 = 0±0.41; Model 3 = 0±0.43; Model 4 = 0±0.01; Model 5 = 0±0.01; Model 6 = 0±0.61.

Table 4. Association between self-reported sleepiness and rater-perceived fatigue, health, and paleness

	Estimate (posterior median)	95% HDI		p^{MCMC}	Probability effect is within ROPE	Null hypothesis decision
		Low	*High*			
Model 1. Fatigue						
Fixed effects:						
Sleepiness	0.20	-0.43	0.84	.54	53.23%	Undecided
Random effects:						
Face (intercept)	1.10	0.97	1.23	-	-	-
Rater (intercept)	0.80	0.66	0.95	-	-	-
Rater (Slope of sleepiness)	0.16	0.00	0.34	-	-	-
Model 2. Health						
Fixed effects:						
Sleepiness	-0.02	-0.60	0.55	.93	64.52%	Undecided
Random effects:						
Face (intercept)	0.96	0.85	1.08	-	-	-
Rater (intercept)	1.30	1.07	1.54	-	-	-
Rater (Slope of sleepiness)	0.12	0.00	0.28	-	-	-
Model 3. Paleness						
Fixed effects:						
Sleepiness	0.01	-0.62	0.61	.99	69.98%	Undecided
Random effects:						
Face (intercept)	1.04	0.92	1.17	-	-	-
Rater (intercept)	1.20	0.99	1.43	-	-	-
Rater (Slope of sleepiness)	0.18	0.00	0.38	-	-	-

Note. Results of three ordinal multilevel models predicting changes in subjective ratings. Subjective sleepiness is centred and scaled such that 1 unit of the predictor represents a change of 2 standard deviations (95% of the data). The same table with ordinal thresholds (intercepts) can be found in Table S3. Random effect estimates represent the standard deviation. All values are on the latent scale. HDI = Highest density interval; p^{MCMC} = Bayesian p value; ROPE = region of practical equivalence around zero (defined as less that Cohen's d = 0.2). Specific ROPEs are as follows: Model 1 = 0±0.28; Model 2 = 0±0.27; Model 3 = 0±0.32).

Table 5. Association between facial appearance and rater-perceived fatigue

	Estimate (posterior median)	95% HDI		p^{MCMC}	Probability effect is within ROPE	Null hypothesis decision
		Low	*High*			
Intercept	2.52	2.43	2.60	-	-	-
Lightness	-0.03	-0.45	0.39	.90	46.90	Undecided
Redness	0.06	-0.33	0.45	.77	48.26	Undecided
Yellowness	-0.59	-1.01	-0.19	< .001	1.42	Reject
Mouth-curvature	-0.71	-1.08	-0.35	< .001	0.08	Reject
Eye-openness	-0.73	-1.09	-0.37	< .001	0.06	Reject
Periorbital darkness	0.47	0.10	0.85	.01	3.61	Reject

Note. Results of a multiple regression model predicting fatigue ratings from facial appearance. All continuous predictors are centred and scaled such that 1 unit of the predictor represents a change of 2 standard deviations (95% of the data). All values are on the latent scale. HDI = Highest density interval; p^{MCMC} = Bayesian p value; ROPE = region of practical equivalence around zero (defined as less that Cohen's d = 0.2). ROPE = 0±0.13).

Figure 1. Averaged images of well-rested vs sleep deprived faces
Top – female well-rested (left) vs female sleep deprived (right).
Bottom – male well-rested (left) vs male sleep deprived (right).

Figure 2. Averaged images of faces rated as least fatigued and most fatigued
Top – ten female faces rated least fatigued (left) vs ten female faces rated most fatigued (right).
Bottom – ten male faces rated least fatigued (left) vs ten male faces rated most fatigued (right).

Figure 3. The estimated feature-specific effects on fatigue ratings (as a probability distribution) following an increase of 2 standard deviations. Red band represents region of practical equivalence (ROPE) defined as a less than a small effect (Cohen's $d < 0.2$). Green-filled densities (solid outline) represent that the parameter shows a significant and meaningful effect. Grey-filled (no outline) densities represent that the parameter cannot be excluded from being zero or near zero.

Åkerstedt, T., Gillberg, M. Subjective and Objective Sleepiness in the Active Individual. Int. J. Neurosci., 1990, 52: 29–37.
Akram, U., Sharman, R., Newman, A. Altered Perception of Facially Expressed Tiredness in Insomnia. Perception, 2017, 030100661772524.
Anund, A., Fors, C., Hallvig, D., Åkerstedt, T., Kecklund, G. Observer Rated Sleepiness and Real Road Driving: An Explorative Study. PLoS One, 2013, 8: e64782.
Axelsson, J., Sundelin, T., Ingre, M., Van Someren, E.J.W., Olsson, A., Lekander, M. Beauty sleep: experimental study on the perceived health and attractiveness of sleep deprived people. BMJ, 2010, 341: c6614.
Axelsson, J., Sundelin, T., Olsson, M.J., et al. Identification of acutely sick people and facial cues of sickness. Proc. R. Soc. B Biol. Sci., 2018, 285: 3–9.
Brunswik, E. Perception and the representative design of psychological experiments, 2nd ed. , 1956.
Bürkner, P.-C. brms : An R Package for Bayesian Multilevel Models Using Stan. J. Stat. Softw., 2017, 80: 1–28.
Chen, Y., Liu, W., Zhang, L., Yan, M., Zeng, Y. Hybrid facial image feature extraction and recognition for non-invasive chronic fatigue syndrome diagnosis. Comput. Biol. Med., 2015, 64: 30–39.
Chervin, R.D., Ruzicka, D.L., Vahabzadeh, A., Burns, M.C., Burns, J.W., Buchman, S.R. The Face of Sleepiness: Improvement in Appearance after Treatment of Sleep Apnea. J. Clin. Sleep Med., 2013, 9: 845–852.
Cohen, J. Statistical power analysis for the behavioral sciences. Statistical Power Analysis for the Behavioral Sciences. p. 567 (1988).
Espinoza-Cuadros, F., Fernández-Pozo, R., Toledano, D.T., Alcázar-Ramírez, J.D., López-Gonzalo, E., Hernández-Gómez, L.A. Speech Signal and Facial Image Processing for Obstructive Sleep Apnea Assessment. Comput. Math. Methods Med., 2015, 2015: 1–13.
Gu, H.G.H., Ji, Q.J.Q. An automated face reader for fatigue detection. Sixth IEEE Int. Conf. Autom. Face Gesture Recognition, 2004. Proceedings., 2004, 0–5.
Hadfield, J.D. MCMC methods for multi-respoinse generalized linear mixed models: The MCMCglmm R package. J. Stat. Softw., 2010, 33: 1–22.
Henderson, A.J., Holzleitner, I.J., Talamas, S.N., Perrett, D.I. Perception of health from facial cues. Philos. Trans. R. Soc. B Biol. Sci., 2016, 371: 20150380.
Holding, B.C., Sundelin, T., Lekander, M., Axelsson, J. Sleep deprivation and its effects on communication during individual and collaborative tasks. Sci. Rep., 2019, 9: 3131.
Ingre, M., Åkerstedt, T., Peters, B., Anund, A., Kecklund, G. Subjective sleepiness, simulated driving performance and blink duration: Examining individual differences. J. Sleep Res., 2006, 15: 47–53.
Jones, A.L. The influence of shape and colour cue classes on facial health perception. Evol. Hum. Behav., 2018, 39: 19–29.
Kruschke, J.K. Rejecting or Accepting Parameter Values in Bayesian Estimation. Adv. Methods Pract. Psychol. Sci., 2018, 1: 270–280.
Kruschke, J.K., Meredith, M. BEST: Bayesian estimation supersedes the t test., https://cran.r-project.org/web/packages/BEST/vignettes/BEST.pdf, 2018.
Kuhn, M. caret Package. J. Stat. Softw., 2008,.
Liddell, T.M., Kruschke, J.K. Analyzing ordinal data with metric models: What could possibly go wrong? J. Exp. Soc. Psychol., 2018, 79: 328–348.
Nichols, A.L., Maner, J.K. The good-subject effect: Investigating participant demand characteristics. J. Gen. Psychol., 2008, 135: 151–165.
Owen, D. Should We Be Worried About Computerized Facial Recognition?, https://www.newyorker.com/magazine/2018/12/17/should-we-be-worried-about-computerized-facial-recognition, 2018.
Oyetakin-White, P., Suggs, A., Koo, B., et al. Does poor sleep quality affect skin ageing? Clin. Exp. Dermatol., 2015, 40: 17–22.
Park, J.H., Van Leeuwen, F., Chochorelou, Y. Disease-avoidance processes and stigmatization: Cues of substandard health arouse heightened discomfort with physical contact. J. Soc. Psychol., 2013, 153: 212–228.
Pedersen, E.J., McAuliffe, W.H.B., McCullough, M.E. The unresponsive avenger: More evidence that disinterested third parties do not punish altruistically. J. Exp. Psychol. Gen., 2018, 147: 514–544.
Peng, X., Luo, J., Glenn, C., Chi, L.-K., Zhan, J. Sleep-deprived fatigue pattern analysis using large-scale selfies from social media. Big Data (Big Data), 2017 IEEE International Conference on. pp. 2076–2084. IEEE (2017).
R Core Team. R: A language and environment for statistical computing. , Vienna, Austria , 2016.
Stephen, I.D., Law Smith, M.J., Stirrat, M.R., Perrett, D.I. Facial skin coloration affects perceived health of human faces. Int. J. Primatol., 2009, 30: 845–857.
Sundelin, T., Lekander, M., Kecklund, G., Van Someren, E.J.W., Olsson, A., Axelsson, J. Cues of fatigue: effects of sleep deprivation on facial appearance. Sleep, 2013, 36: 1355–1360.
Sundelin, T., Lekander, M., Sorjonen, K., Axelsson, J. Negative effects of restricted sleep on facial appearance and social appeal. Open Sci., 2017, 4: 160918.

Talamas, S.N., Mavor, K.I., Axelsson, J., Sundelin, T., Perrett, D.I. Eyelid-openness and mouth curvature influence perceived intelligence beyond attractiveness. J. Exp. Psychol. Gen., 2016, 145: 603–620.
Tan, K.W., Stephen, I.D. Colour detection thresholds in faces and colour patches. 2013, 42: 733–741.
Vural, E., Cetin, M., Ercil, A., Littlewort, G., Bartlett, M., Movellan, J. Drowsy Driver Detection Through Facial Movement Analysis. Human–Computer Interaction. pp. 6–18. Springer, Berlin, Heidelberg (2007).
Willis, J., Todorov, A. First impressions: Making up your mind after a 100-ms exposure to a face. Psychol. Sci., 2006, 17: 592–598.
Yarkoni, T., Westfall, J. Choosing Prediction Over Explanation in Psychology: Lessons From Machine Learning. Perspect. Psychol. Sci., 2017, 12: 1100–1122.
Zebrowitz, L.A., Montepare, J.M. Social Psychological Face Perception: Why Appearance Matters. Soc. Personal. Psychol. Compass, 2008, 2: 1497–1517.