To the editor and both referees, thank you for the constructive feedback. I have now revised the manuscript to your concerns. In particular, I have clarified the novelty of the current manuscript, have added more details, and have clarified the conclusions. My detailed responses are below (preceded by ">>>"), and changes in the main text are highlighted.

Reviewer #1: The paper is interesting and meaningful.
Overall it is a good work, needing some improvements.
>>> Thank you.

The author should justify why ZeitZeiger is a machine learning technique.
>>> I have more clearly explained ZeitZeiger as a supervised learning method in the Background. I have also left the more detailed description in the Methods.

The author should clearly state the advancements in this paper with respect to their paper ref (31) in Nucleic Acids Research (2016).
>>> I have added text in the Background to clarify the advancements in the current paper. In short, our NAR paper was based on data from mice, whereas here I apply ZeitZeiger to data from humans.

The phrase " ... periodic variables, i.e. variables that are continuous and bounded and for which the maximum value is equivalent to the minimum value" should be clarified; what means "maximum value equivalent to minimum value". Are they the same ? What means equivalent ?
>>> I added an example to clarify this: the angle in polar coordinates between 0 and 2π. In this paper, I focus on another periodic variable: time of day.

In Fig. 2 some correlation coefficient should be computed, because the dispersion of data around the periodic curve is rather high. If a linear regression is made, would the correlation coefficient be very different ?
>>> The dispersion around the curve is indeed high. If the dispersion were lower, prediction accuracy would be higher. Because circadian time is periodic (CT0 is the same as CT24), however, linear regression is inappropriate. Instead, I now mention in the Fig. 2 caption the signal-to-noise ratios (SNRs) for the two SPCs. As described in the Methods, the SNR describes the strength of circadian rhythmicity, and corresponds to the peak-to-trough amplitude of the spline fit divided by the root mean squared error of the spline fit.

Reviewer #2: General
This is an important clinical and research problem -and I commend the author for taking it up.

From what I can tell the methods are well described and well reasoned. In large part the methods build directly off the author's recent work

Thus my primary suggestions are in the exposition of the results (and in discussion). I think more emphasis needs to be placed on describing the accuracy of the predictor during

perturbations (as opposed to "unrealistic"/"non-clinical" control conditions). These are now relegated to the supplement. Moreover I think there are a few claims that could be better tested (or explained) All in all I think there is much value to this work and my concerns can be relatively easily addressed.
>>> Thanks for your feedback.

General Comment
As a reader it was difficult to keep track of text that just referenced experiments by GEO accession #'s (maybe that's just me) I would suggested first describing the experiments briefly and then referring them by either some short identifier "mistimed sleep experiment" or the first author.. (Archer et al)"
(at a minimum I think you need to describe the protocols a little more)
>>> Thanks for the suggestion. I have now added more detail to Table 1 and cleaned up how I refer to the datasets throughout the text and in the caption for Fig. 3.

Methods
Small question?
In combining datasets for training - how did you handle genes that were represented by more that 1 probe set?
>>> This is handled by the MetaPredict package (Hughey and Butte 2015). In short, for processed microarray data (all three of the datasets in this study), each gene's expression is calculated as the median over all probe sets that map to that Entrez Gene ID.

In looking briefly at the methods of the papers this data comes from - it looks like the subjects were asked to stay on a relatively constant schedule - mirroring the baseline protocol schedule - in the weeks prior to the study. (Or in the first weak of the study -prior to sample collection). So why is any further adjustment to circadian phase needed (based on location)?.
I would think you might be better off keeping things relative to "Lights on" How much does this change your results?
>>> My understanding of the studies is that the subjects were asked to follow a consistent sleep schedule prior to the study, but the schedule could vary from one individual to another. As you allude to and as I state in the Methods, in each dataset, the phase of each subject's circadian clock should be based primarily on the natural light-dark cycle. Given that I lack individual-level data for DLMO, I believe using the external time of day (i.e., normalizing to the sunrise time) is the proper method.

>>> I would argue that the time of "lights on" is arbitrary and should not be used to standardize time of day. To give an extreme example, imagine a study had each person following his/her own sleep schedule before coming into the lab, but on the first day in the lab, the lights didn't turn on until noon.

>>> The other issue is more practical: for GSE39445 (sleep restriction study) and GSE48113 (forced-desynchrony study), the time of "lights on" is not provided and is not entirely obvious

from the publication. So one would have to resort to simply using clock time, which, in my opinion, is much less defensible and generalizable than standardizing clock time by sunrise time.

Results
FIG1 - As Is, should highlight in figure (or legend) that this only includes "control" data without sleep/wake perturbation (Same for Figs 3 and 4 I think).
>>> I now mention in the captions for Figs. 1, 2, 4, and 5 that the analysis is based only on control samples.

But, I think that clinically - the difficulty in estimating someone's circadian phase is less of an issue when they live on a "perfectly" regimented schedule as was seemed to be requested of the subjects in the weeks prior to the study start. It would seem that the performance during the sleep/wake perturbations is probably the most relevant for clinical populations (ie Sup Figure 3) . I would ask that (for figs 1 3, and 4) you showed both control and sleep perturbation prediction errors in the main figures.
>>> Your point is well taken. I have now moved the analysis of the perturbation data to the main text (Fig. 3) and adjusted the Results and Discussion accordingly. After much thought, however, I have decided not to include the perturbation data in the analysis for Figs. 1-2 and 4-5. Here is my reasoning:

>>> Even in the control samples, my ground truth for circadian time is murky, and it becomes murkier still in the perturbation samples. If I had detailed DLMO data for each subject in each condition, I would feel more comfortable training a predictor using the control and perturbation samples together in the main analysis. Even that, however, would be assuming that the central clock and the blood clock stay in phase during the perturbations, which my analysis now shows may not be true. Thus, analyzing the control and perturbation samples together could be introducing a systematic bias. Given the limitations in the data, I believe the control samples are an appropriate place to start, and I would strongly prefer to not include the perturbation data in the main analyses.

Similarly the results section in the abstract should reference performance both in the setting of perturbations and control conditions (or combined performance).
>>> I have now mentioned the results of analyzing the sleep perturbations in the abstract.

You say that that a "single night of complete sleep deprivation" appeared to induce a phase delay of 2.1 hours" I think you mean that the perturbation resulted in a systematic bias in your estimation. It is possible that this was a true phase delay. I would think it is also possible this just reflects an error in attempting to generalize from the control data. I thought that the classic experiments that supported "the 2 process model of sleep" suggested that acute sleep perturbations (without light) does not markedly shift phase? Do I misunderstand this? This should be clarified.

>>> Although this could be generalization error or bias, I am inclined to believe it is a real phase delay. In the sleep deprivation study (GSE56931), the lights were on during the entire sleep deprivation period (this is stated in their supplement). Thus, a phase delay of ~2 h would be consistent with previous studies in humans (Khalsa et al. 2003; Hilaire et al. 2012). I have now also calculated the phase difference between control and perturbation conditions for each clock gene (Figure S4), which also is suggestive of a phase delay. I have added text to clarify these points.

>>> This exemplifies why I am reluctant to include the perturbation samples in the main analysis (especially without individual-level DLMO): because doing so makes a strong assumption that the perturbation does not affect the relationship between external time and the circadian clock in the blood.

Similarly you write that the 'four 28-h days (forced desynchorny protocol in GSE48113)..Induced a phase delay of 2 hrs". You later describe your ability to estimate DLMO in that paper. Thus, if I understand correctly, there is DLMO data to test your hypothesis that there was a 2 hours phase delay after the desynchrony (perhaps the author's of the original study did this?) Was a 2hrs phase delay notable in those results? If that data is available you should compare. If there is no corresponding delay in DLMO - you should be clear in saying that the estimator has a bias (that may reflect a true phase shift in blood cells?) compared to central rhythms.
>>> I apologize for not explaining this sufficiently. In each condition, GSE48113 provided DLMO as an average over all subjects (see their Fig. 1 caption). Average DLMO was ~1 h later (so a delay) in the “out of phase” samples. As you mention, this is why I am reluctant to rely too much on this aggregate value for DLMO: it reflects the central clock but not the clock in blood cells. I have added text in the Results and the Discussion around this issue (see also Table S1).

Discussion
You comment (p14) that the relatively poorer performance in human as compared to mouse data is a result of "genetic and environmental factors" I would think that another major issue is in the tissue itself. (As you note) your previous paper looked at solid tissues - not blood. As the composition of formed elements in blood greatly varies between individuals - this may be another issue. Brain (the other tissue you mention - but don't show also has I think much weaker rhythms in mice (at least the # of genes identified to cycle).
>>> This is a valid point, which I have added to the Discussion.

2nd you conclude that the reason that the mouse predictor was more likely to include elements of the "core circadian clock" was that it was forced to predict on many tissues. An alternative is that the mouse experiment was on a single inbred strain with little genetic heterogeneity. Here you are searching for genes with a strong circadian signal relative other causes of variance in gene expression. It could simply be that core clock genes are also subject to significant heterogeneity in baseline expression between people - and this limits the predictive power of those genes…

>>> If I understand your argument correctly, this would require that the clock genes suffer from greater heterogeneity in expression than do other (clock-controlled) genes. I would have trouble coming up with an explanation for that. Furthermore, the Zhang et al. 2014 paper showed that the core clock genes are essentially the only ones that show a strong, consistent rhythm in all tissues. In a given tissue, however, there will be other genes whose rhythms are stronger. Anecdotally, when ZeitZeiger is trained on a single tissue (e.g., mouse liver or human brain), it selects several genes that are not part of the clock. I have added text around this issue.

The use of ~2 samples from a patient to determine their phase seems possible (but not ideal) the requirement for 7+ samples from each person (personal predictor) seems completely impractical for a clinical diagnostic test.
>>> Thank you for pushing back on this. I totally agree, given the current technology. However, one could imagine both advances in technology and ultimately combining tissue-based measurements with actigraphy, so that either of these methods (or derivatives thereof) could become more practical. I have revised the text and the title accordingly.

P15 "Our work implies that the forced desynchrony protocol affects not just the central clock…: I believe the whole point of a forced desyncrhony protocol is that the imposed schedule is outside of the range of entrainment of the central clock - and leaves the the central clock relatively unaffected. Do I misunderstand?
>>> This was poorly worded. By “affects”, I meant “causes it to go into free-run”. I have revised this section accordingly.

You begin to address the issue of whether or not the transcripts in question really predict circadian phase - or some other associated feature or clock output which can sometimes be out of phase with other rhythms. (Like sleep/wake, eating, body temperature)

However - I must admit - I did not fully understand your arguments.
(1) For 2/3 of the experiments sleep/wake , body temp, eating, and circadian rhythms are likely aligned in the control condition - having a strong transcriptional signal doesn't necessarily imply specificity for circadian rhythms compared to the others. Please explain what you meant
(2) I do see that GSE39445 (Archer paper) used a constant routine to separate sleep/wake (and probably eating) from circadian. But the "free running" (ie unentertained) clock in GSE48113 would seem to only to be a factor in the perturbation condition - as the baseline condition is designed so that sleep/wake is in phase with circadian. Please explain a bit more what you mean..
>>> You’re right about the control condition in GSE48113, and I apologize that my explanations were unclear. I have now extensively rewritten this paragraph, and added more evidence to support my hypothesis that the predictor reflects the state of the clock.

(3) The authors of GSE48113 emphasized the point that most genes lost cycling in their perturbation. So (just from your exposition) its hard for me to know if that the reason you "get similar results" when train just on GS#48113 is simply because that perturbation data can't be

well fit - and only the baseline data makes a real contribution. So do you get similar results if you train off the perturbation data (only) in GSE48113? I realize this is less data - but even if you just used this group to select the genes you are to train on - it would be a useful check.
>>> Thanks for the suggestion. I have now performed cross-validation on only the perturbation data from GSE48113 (Figure S6), and the results are very similar to Fig. 3 and Figure S5. Given the analysis I'd already done, here are a few more points. First, when training on all samples from GSE48113, the absolute error for "out of phase" samples is not drastically higher than for "in phase" samples (Figure S5C). Second, the two SPCs show circadian variation in both "in phase" and "out of phase" samples (Figure S5D). Third, the SPCs in Figure 2 are shifted from those in Figure S5C. All this suggests to me that the "out of phase" (perturbation) samples are contributing, although certainly gene expression in those samples is noisier (which one could argue is a real manifestation of the clock in blood cells).