

A machine learning model using clinical notes to identify physician fatigue

Corresponding Author: Professor Chenhao Tan

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

The paper uses data from 129,228 emergency room visits to train a model that can identify notes written by fatigued physicians (defined as those who worked 5 or more of the prior 7 days). The authors evaluate the model at detecting this label in a held out test set of 20,543 patients (and the same 60 physicians). The paper then ties model predictions to worse decision-making through the yield of testing for heart attack. The authors also note that the difference in fatigue detected for Black and Hispanic vs. White patients is larger than overnight vs. daytime differences. Finally, the authors note that fatigued notes are more simple, and connect this to LLM-generated note quality.

This paper is interesting, and important. Please see specific comments below.

General Comments:

>> Interestingly, the model also flags notes written in other high-fatigue settings that were not initially labelled (on overnight shifts, and after high patient volumes). I would like to see more about these settings, and whether the model could be used as a general quality detection system. The authors note one specific facets of fatigued notes - that they are simpler/easier to predict. What else?

>> It is interesting that model predictions correlate with worse decision-making through yield of testing for heart attack. I am not sure why only this metric is significantly different (18% lower with each standard deviation increase in model-predicted fatigue). Are there other metrics that were considered, but were not found to be significant?

>> The authors also note that the difference in predicted fatigue for Black and Hispanic vs. White patients is larger than the predicted fatigue for overnight vs. daytime shifts. Can this be tied to any other metric other than the predicted fatigue itself? For instance, are these high-fatigue notes the same ones that lead to poor heart attack testing decisions?

Weaknesses:

>> The model's reported predictive AUC is 0.6. This is not generally considered a well-calibrated model in health prediction settings. Can the authors comment on why this lower predictive quality is reasonable for this work?

>> The authors note that "Our analytic strategy rests on the assumption that the text of notes written today should be statistically uncorrelated with how many days the physician has worked over the past week, except via the direct effect of prior workload on the physician's fatigue state while writing the note. This assumption could be violated if patient characteristics on a given shift differ as a function of prior workload for any reason."

It's not just patient characteristics that are important, but also physician busyness. For example, maybe the fifth day a physician works in a row tends to be the busiest (in terms of number of patients that show up at the ED). The authors check for balance in patient covariates, but I do not think they actually look at the number of patients evaluated. This seems like an easy thing to check, so they should make the case that physician workload (measured in patients/day) is uncorrelated with physician days worked in a week. If they are correlated, the analytical strategy is not valid.

>> The authors check for balance in chief complaints and demographics to check that patient characteristics do not differ as a function of prior workload. It seems like severity is also important; is it feasible to check balance in triage scores?

>> The authors comment that "We emphasize an important, but counterintuitive, aspect of our analytic strategy: while our model is trained to predict a physician's prior workload—our goal is explicitly not to perfectly predict this variable. Instead, we wish to learn a general model of how fatigue affects the text of notes, by training the model to distinguish notes written by physicians we believe to be more (high-workload) vs. less (low-workload) fatigued. In fact, we believe our model predictions are a better measure of a physician's 'true' fatigue when writing a given note than the actual training label (prior workload) itself. We provide the intuition for this argument here, and further explanation and a formal proof in Appendix A; we also provide a range of empirical tests in the following two Sections."

This claim seems controversial to me. They justify it by saying "If fatigue affects notes in the same way—whether it is caused by heavy workload, idiosyncratically poor sleep the night before, etc.—a model that learns to predict workload will learn something about the general way fatigue affects note text." However, that's only true if the differences they detect are caused by fatigue, is difficult to just assume - it should be shown somehow.

>> The study is done in one health center for two years (2010 - 2012), which limits its potential transferability to other settings and the generalization of the findings. Can the authors comment on this?

>> The evaluated test set is also split by patient ID, but not by physician ID. How well does the model do in notes from unseen physicians? This is an important test of whether the model is overfit.

>> The demographic breakdown of the physicians is not noted can this be added and tested for significance?

>> The authors use a small set of tie note predictability (which is a feature of high-fatigue notes) to LLM generated notes, which are also highly predictable and therefore present as more fatigued. The authors do not dig into the features/quality of LLM-generated notes, which is an important area for more work.

>> The authors state that "Here we investigate an implicit, previously unsuspected form of inequality in clinical care: differences in measured fatigue detected in notes, as a function of racial and ethnic group. We build on studies demonstrating subtle linguistic bias in other settings, from police stops to political speeches (...)" However, at least two of the works subsequently cited are set in a health context, which is misleading given the section quoted.

<https://www.healthaffairs.org/doi/full/10.1377/hlthaff.2021.01423>

<https://dl.acm.org/doi/10.1145/3514094.3534203>

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)

This work entitled "Clinical Notes Reveal Physician Fatigue" is highly relevant, well-done, and original. However, there are a few things I've identified that I believe would significantly strengthen the manuscript:

While the race/ethnicity finding is especially interesting, I suspect it could be driven at least in part by socioeconomic correlation with types of insurance. Different payer classes can require different documentation burden and as the result stands, to me it is not clear whether this is an issue of bias or potentially an artifact of the payer system. Conversely, patients without coverage or who know they are likely to incur costs they can't afford may delay seeking care.

It would strengthen the result to condition this analysis on additional information: insurance, method of arrival to ED, whether the patient was admitted, length of stay, whether the patient was transferred to ICU, died, or other measures of severity. Especially in today's climate, analyses pointing at potential biases should be ironclad or they have the potential to do more harm than good. I would encourage the authors to explore these additional covariates and to perform sensitivity analyses to increase the likelihood that they have identified a potential causal covariate and not one with an easily identified confounder.

For this audience, the metric perplexity needs a stronger introduction and justification for why it is a reasonable choice (beyond that it is widely used in NLP). Ideally, something that would be intuitive to clinicians.

Source code should be provided as part of the review process, otherwise it is impossible to say that the code is peer-reviewed and should be noted as such in the publication.

(Remarks on code availability)

Please make code available or specify in the manuscript that it was not made available for peer review.

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

The authors have answered most of the questions, and added analysis and context that has significantly improved the paper's clarity and contribution.

The only concerns I have remaining are related to the evaluated test set being split by patient ID, but not by physician ID. While the original sample has a small number of doctors, the added MIMIC dataset is larger. However, I did not see a section of the MIMIC analysis dealing with this either. While it is very reassuring to see the generalizability of the findings to MIMIC notes, the generalizability with respect to unseen physicians in any dataset is still unknown.

If there is a reason that this analysis cannot be done, I suggest that it be stated as a limitation of the study and findings.

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)

The authors have fully addressed my feedback.

(Remarks on code availability)

Version 3:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

The authors have answered all of my questions, and addressed all of my concerns.

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)

The authors have answered my concerns.

I would suggest they adjust the paths in code to be relative paths, to make sure there is no private information in the source code (I didn't see any) and to push the code to Github and an archival service like zenodo/figshare etc.

(Remarks on code availability)

The code is comprehensive and relatively easy to follow. It would be helpful to provide sample data so that others can see the input data format.

Open Access This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.

The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

Comments and suggestions from reviewer 1:

Re. Interestingly, the model also flags notes written in other high-fatigue settings that were not initially labelled (on overnight shifts, and after high patient volumes). I would like to see more about these settings, and whether the model could be used as a general quality detection system. The authors note one specific facets of fatigued notes - that they are simpler/easier to predict. What else?

Thanks, we also found this very interesting. We have broadened our discussion of other facets of the notes (beyond perplexity) in Section 6. This highlights facts such as:

“the readability of notes, based on the complexity of sentences and words, is lower when physicians are fatigued.”

“Fatigued physicians tend to use fewer words in the insight category, and more words in the certain category.”

To further address reviewer 1’s suggestion, we included a full table of these features in the appendix (Table 14), and expanded the discussion in the appendix to add our interpretation.

Re. It is interesting that model predictions correlate with worse decision-making through yield of testing for heart attack. I am not sure why only this metric is significantly different (18% lower with each standard deviation increase in model-predicted fatigue). Are there other metrics that were considered, but were not found to be significant?

Unfortunately, quality measurement in the ED is a challenging topic with a long history in health services research and no clear consensus on validated metrics (please see two references below; we note that one member of our author team [ZO] is a practicing emergency physician who has done several years of research in this area). Briefly, most measures of quality in the ER focus on aggregate measures like crowding which do not apply to individual patients, our focus here; others focus on the likelihood of an individual patient revisiting the ER after discharge, but this can have many causes besides poor decision quality.

Measuring the quality of decision making, particularly at the individual patient level, is thus quite difficult. The heart attack testing metric we use has the disadvantage of being narrowly focused on heart attack, and thus only applies to a small fraction of patients who are tested for heart attack. Thus statistical power is diminished and we can only look at one facet of decision quality. On the other hand, the advantage of this metric is that it is unambiguous, has been previously validated, and data are available; we find significant effects despite the small sample size.

For more information on ED quality measurement see:

Quality Measurement In The Emergency Department: Past And Future. Jeremiah D. Schuur, Renee Y. Hsia, Helen Burstin, Michael J. Schull, and Jesse M. Pines.

<https://www.healthaffairs.org/doi/full/10.1377/hlthaff.2013.0730>

Evaluation of emergency department performance – a systematic review on recommended performance and quality-in-care measures. Christian Michel Sørup, Peter Jacobsen, and Jakob Lundager Forberg.

<https://link-springer-com.proxy.uchicago.edu/article/10.1186/1757-7241-21-62>

Re. The authors also note that the difference in predicted fatigue for Black and Hispanic vs. White patients is larger than the predicted fatigue for overnight vs. daytime shifts. Can this be tied to any other metric other than the predicted fatigue itself? For instance, are these high-fatigue notes the same ones that lead to poor heart attack testing decisions?

We continue to find the difference in predicted fatigue for Black and Hispanic vs. White patients intriguing, and have no clear explanation as to why. That said, it seems worth presenting, even in the absence of an explanation, as it's an important phenomenon that deserves further study. To answer the narrow question above, yes, fatigue is correlated to both race and poor heart attack testing decisions, as shown in Table 2 and Table 3. Recall we controlled for race in Table 2.

Re. The model's reported predictive AUC is 0.6. This is not generally considered a well-calibrated model in health prediction settings. Can the authors comment on why this lower predictive quality is reasonable for this work?

While this would indeed be a low AUC in a traditional prediction exercise—where the quantity of interest is the exact variable the model is predicting—we lay out in Appendix A why this is not the case. Here, the training label is not the actual quantity of interest: we do not actually measure fatigue, but rather a randomly assigned proxy for fatigue in the form of prior physician workload. We do not have accurate measures of fatigue readily available and may never have that. So the AUC is not intrinsically meaningful.

The advantage of this setting is that, in the setting of as-good-as-random assignment, the only mechanism that the number of days worked in the past seven days (Y) can affect the notes (X) is through fatigue. We conducted extensive tests to rule out other factors. However, it is impossible to completely validate this assumption. Similar to studies based on causal graphs (see *The Book of Why* by Judea Pearl and Dana Mackenzie and a recent review article, *The Use of Structural Models in Econometrics* by Hamish Low and Costas Meghir), this assumption is required for the validity of our study. Please see a full discussion in Appendix A.

We examine the validity through the correlation with other plausible measures of fatigue in Section 3. We also present more meaningful predictive features in the text and refer to those in Section 6.

Re. The authors note that “Our analytic strategy rests on the assumption that the text of notes written today should be statistically uncorrelated with how many days the physician has worked over the past week, except via the direct effect of prior workload on the physician's fatigue state

while writing the note. This assumption could be violated if patient characteristics on a given shift differ as a function of prior workload for any reason."

It's not just patient characteristics that are important, but also physician busyness. For example, maybe the fifth day a physician works in a row tends to be the busiest (in terms of number of patients that show up at the ED). The authors check for balance in patient covariates, but I do not think they actually look at the number of patients evaluated. This seems like an easy thing to check, so they should make the case that physician workload (measured in patients/day) is uncorrelated with physician days worked in a week. If they are correlated, the analytical strategy is not valid.

We checked the correlation between busyness of a day and days worked in a week (workload) controlled for physician and time, and found no significant effect (p-value = 0.673). We have added this to Appendix Section E.

Re. The authors check for balance in chief complaints and demographics to check that patient characteristics do not differ as a function of prior workload. It seems like severity is also important; is it feasible to check balance in triage scores?

We don't have the triage score but we do have length of stay as a signal of patient severity. Length of stay can be seen as a measure of severity of the patient. Thus, we checked for balance in length of stay, and found no correlation between length of stay and the prior workload (days worked in past week) where the p-value is 0.31. We added this as a control variable in all the tables in the main paper.

Re. The authors comment that "We emphasize an important, but counterintuitive, aspect of our analytic strategy: while our model is trained to predict a physician's prior workload— our goal is explicitly not to perfectly predict this variable. Instead, we wish to learn a general model of how fatigue affects the text of notes, by training the model to distinguish notes written by physicians we believe to be more (high-workload) vs. less (low-workload) fatigued. In fact, we believe our model predictions are a better measure of a physician's 'true' fatigue when writing a given note than the actual training label (prior workload) itself. We provide the intuition for this argument here, and further explanation and a formal proof in Appendix A; we also provide a range of empirical tests in the following two Sections."

This claim seems controversial to me. They justify it by saying "If fatigue affects notes in the same way—whether it is caused by heavy workload, idiosyncratically poor sleep the night before, etc.— a model that learns to predict workload will learn something about the general way fatigue affects note text." However, that's only true if the differences they detect are caused by fatigue, is difficult to just assume - it should be shown somehow.

We have rewritten to emphasize that this is a critical assumption. While we tried to falsify it in many ways and could not, it is impossible to prove directly, similar to other causal assumptions commonly made in the literature (e.g., see *The Book of Why* by Judea Pearl and Dana

Mackenzie and a recent review article, The Use of Structural Models in Econometrics by Hamish Low and Costas Meghir).

Re. The study is done in one health center for two years (2010 - 2012), which limits its potential transferability to other settings and the generalization of the findings. Can the authors comment on this?

Thank you, generalizability is of course a major concern for us as well. A major problem with health records, and free-text notes in particular, is how hard it is to find freely accessible datasets, because of privacy concerns. We were lucky to have access to the data we use, but there are few datasets available online, and most publicly-available datasets lack the meaningful variables necessary for comparison with our paper, such as patient race, physician ID, and time of admission.

That said, we appreciate the prompt to think about this a bit more carefully, and were able to find one dataset in which to test the generalizability of our approach. The MIMIC-III data at first glance does not contain physician workload information, as dates are only internally consistent for the same patient to protect patient privacy. More details can be found at <https://mimic.mit.edu/docs/iii/about/time/#date-shifting>. But unlike i2b2 (the standard datasets of notes), we are able to infer which notes in MIMIC-III were written on an overnight shift.

We thus apply the fatigue prediction model trained with our internal dataset. While statistical note features are easy to obtain, we need to fine-tune a language model specifically for MIMIC-III notes. After collecting the necessary note features, we were able to obtain predicted fatigue scores on MIMIC-III notes using the fatigue model trained with our internal dataset. In this case, we are able to replicate our result with respect to notes written in overnight shifts. We have included this in the main paper (section 3.1) and added more details in the appendix.

It is useful to note that MIMIC-III includes notes from the ICU, which is substantially different from our datasets from the emergency department. The fact that the observation about overnight shifts generalizes provides evidence that our work holds promise for wide applicability in healthcare. The lack of comparable datasets highlights the uniqueness of our dataset and the potential value of our study.

Re. he evaluated test set is also split by patient ID, but not by physician ID. How well does the model do in notes from unseen physicians? This is an important test of whether the model is overfit.

We hope that our analysis with the generalizability of the findings addresses the concerns. As shown in the next response, the sample size is very small with respect to physicians, so we could not evaluate generalizability to unseen physicians in this dataset.

Re. The demographic breakdown of the physicians is not noted can this be added and tested for significance?

Among the 60 doctors in the study, there were 44 males and 16 females. The racial distribution of the participants was as follows: 46 white, 11 Asian, 1 black, and 2 Latino. Due to the small sample size of physicians, we could not include physician demographics in the regression analysis, both to avoid compromising privacy and potentially noisy results.

Re. The authors use a small set of tie note predictability (which is a feature of high-fatigue notes) to LLM generated notes, which are also highly predictable and therefore present as more fatigued. The authors do not dig into the features/quality of LLM-generated notes, which is an important area for more work.

In addition to the perplexity difference mentioned in the paper, we observed that LLM-generated notes usually come with a lower note length and a lower fraction of positive emotion words. In general, notes generated with the sampling decoding method are more similar to human-written notes compared to the greedy decoding method, which has a higher readability and comes with a higher fraction of exclusive words in LIWC, e.g., “but” and “without”. We added the table of feature values from original HPI and LLM-generated HPI with different decoding methods in the Appendix Table 14.

*Re. The authors state that “Here we investigate an implicit, previously unsuspected form of inequality in clinical care: differences in measured fatigue detected in notes, as a function of racial and ethnic group. We build on studies demonstrating subtle linguistic bias in other settings, from police stops to political speeches (...)”
However, at least two of the works subsequently cited are set in a health context, which is misleading given the section quoted.*

Thank you, this was an oversight. We meant to say that measured fatigue from physician notes has not been studied as a form of inequality for racial and ethnic groups. We have updated the wording in the manuscript. We have also added the recommended references.

Comments and suggestions from reviewer 2:

Re. While the race/ethnicity finding is especially interesting, I suspect it could be driven at least in part by socioeconomic correlation with types of insurance. Different payer classes can require different documentation burden and as the result stands, to me it is not clear whether this is an issue of bias or potentially an artifact of the payer system. Conversely, patients without coverage or who know they are likely to incur costs they can't afford may delay seeking care.

It would strengthen the result to condition this analysis on additional information: insurance, method of arrival to ED, whether the patient was admitted, length of stay, whether the patient was transferred to ICU, died, or other measures of severity. Especially in today's climate, analyses pointing at potential biases should be ironclad or they have the potential to do more harm than good. I would encourage the authors to explore these additional covariates and to

perform sensitivity analyses to increase the likelihood that they have identified a potential causal covariate and not one with an easily identified confounder.

Thanks for the suggestion! We take the socioeconomic status of patients into account with the insurance information, e.g., commercial insurance or Medicare, and add additional controls that are available in our dataset to represent the severity of the patient with length of stay (days). The new regression results show that predicted fatigue is still correlated with other fatigue proxies and decision quality on the yield of the test for heart attack. We have updated all the tables in the main paper to reflect this change.

Re. For this audience, the metric perplexity needs a stronger introduction and justification for why it is a reasonable choice (beyond that it is widely used in NLP). Ideally, something that would be intuitive to clinicians.

We use perplexity which captures the average log likelihood of words being generated from a language model. Intuitively, large language models are trained to predict the next word and if they can reliably predict what the next word is, it suggests that the note is very predictable. We added several sentences in the main paper and have a detailed introduction in Appendix Section C.

Re. Source code should be provided as part of the review process, otherwise it is impossible to say that the code is peer-reviewed and should be noted as such in the publication.

We have attached the code base of our experiment in the submission.

Comment from the Reviewer 1

While the original sample has a small number of doctors, the added MIMIC dataset is larger. However, I did not see a section of the MIMIC analysis dealing with this either. While it is very reassuring to see the generalizability of the findings to MIMIC notes, the generalizability with respect to unseen physicians in any dataset is still unknown. If there is a reason that this analysis cannot be done, I suggest that it be stated as a limitation of the study and findings.

We have experimented with splitting by physician IDs on MIMIC-III admission notes, which yields varying results. It suggests that seeing the physician notes in the training set is important for our approach on training the language model to obtain note unpredictability (perplexity). We have added this information in the supplementary materials to clarify this limitation.

Comment from the Reviewer 2

I would suggest they adjust the paths in code to be relative paths, to make sure there is no private information in the source code (I didn't see any) and to push the code to Github and an archival service like zenodo/figshare etc.

We have made the change and released the code on GitHub.