We thank the reviewers for their insightful and positive reviews, finding our work well motivated (R4), addressing a clear gap in existing research (R4) by proposing a creative/novel (R3) and intuitive (R2, R3, R4) method with great potential (R3). Along with expanded discussion, we have also addressed minor comments in the updated draft. To summarize, we have proposed a new method, called FIT, that uses KL-divergence to assign instantaneous feature importance for time-series observations, accounting for temporal data shift. FIT shows promising results on complex simulated time-series models as well as two tasks on real healthcare data.

Subset Selection/Importance (R2, R4): Assigning importance to a subset of features is a novel property of FIT and we have added additional experiments and discussions regarding this in our paper. For example, based on R4's suggestion, we identified subsets of correlated features using hierarchical clustering on Spearman correlations for MIMIC and used FIT to evaluate the scores assigned to these subsets (Table 1). Of course, existing methods such as greedy sub-modular search, hill climbing (R2) and modern stochastic search can also be easily used with FIT to find the optimal subset.

Generative model quality (R1, R4): As suggested by R4, we now compare the performance of our generator with simpler approaches for approximating the conditional, such as carry-forward or mean imputation (Table 2). FIT is flexible to the choice of any generator, however, modelling proper conditional distribution is important when time-series data shows significant shifts where carry-forward and mean imputation will result in noisy scores. We have added this discussion and results to the appendix. To demonstrate the quality of the conditional generator, we have also added the likelihood plots, which show that the generator is not overfitting (R1).

Instantaneous attribution (R1): Based on R1's suggestion, we have added the following to the draft: "Instantaneous attribution is valuable to understand the additive information of a *new* observation, particularly for real-time predictions. For example, when managing sepsis in an ICU, instantaneous changes are likely to drive model prediction." We also highlight how FIT may be extended to non-instanteous attribution: "Though out of scope for this work, our method is extendable to non-instantaneous attributions. This requires: 1) evaluating temporal shift (with appropriate delays, e.g. by binning epochs over time); 2) a conditional generator that models distribution over multiple time-steps. Such modifications to the generator are also useful when *gradual shifts* like spikes and trends occur in the data (R2, R4). For explanations of models used for longer term disease management, like chronic conditions, we would suggest using multi-step predictions." Finally, in the logical AND example (R2), we note that all methods will fail when used for instantaneous attribution. This is because the score itself from FO and other methods is biased due to issues of vanishing gradients common in RNNs (Ismail et al. NeurIPS2019). Similarly, no guarantees exist for RETAIN to assign equal importance to both $x_{t-1,i}$ and $x_{t,i}$.

Subject matter expert (SME) evaluation (R2, R3): We asked a clinical collaborator (SME) to annotate important observations over time and we evaluated FIT scores against these. High positive FIT scores were correlated with time-points the clinician identified as important in their decision making. Figure 1 shows an example of such annotations (in red) for 2 different signals from 2 individuals. The clinician determined that patient 1 (top) was tachycardic towards the end (hour 32) and the FIT scores for Heart rate highlight this time point clearly. As R3 suggested, we have also added visualizations for MIMIC experiments along with the clinical insight of the SME.

High-dimensional and binary data (R1): We agree that high dimensionality of the feature-space can increase sample-complexity of estimating a full covariance matrix. We have included the following discussion addressing this: "For high-dimensional data, low-rank approximations can be considered in practice that will reliably model desirable dependencies efficiently. Binary as well as heterogeneous data-types can be incorporated with recent advances in heterogeneous data modeling using recurrent models (e.g. Liu et al. AAAI 2018). For more complex data, FIT can use other conditional generators such as GAIN and Imputation-GANs."

Expand discussion on insights (R2, R4): We have significantly expanded discussion for added insights. R4: The main difference between FIT and other counterfactual methods [5,12] is that we use these counterfactuals to estimate temporal shift, while [5,12] assess perturbations in model output. Also, the counterfactuals sampled using our generator marginalize over complement of the target set. Note that such explanations do not provide causal insights but help understand the predictive mechanism of a model.

Subset	AUROC drop
S1	0.007±0.000
S2	0.005±0.002
S3	0.004±0.003
S4	0.004±0.002
S5	0.011±0.015

Table 1: Subset perf. drop on MIMIC

Generator	AUROC	AUPRC
Conditional	**0.72±0.01**	**0.15±0.00**
Carry-forward	0.53±0.00	0.03±0.00
Mean Imp	0.48±0.004	0.03±0.00

Table 2: Generator quality

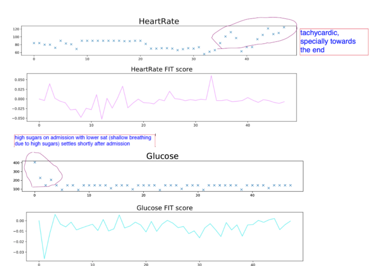

Figure 1: Clinical (SME) evaluation