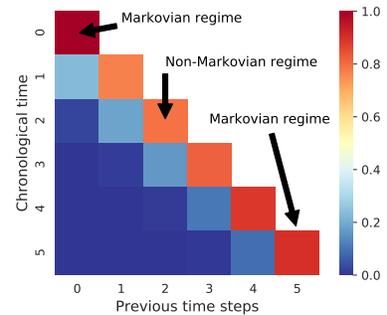


1 [[**Reviewer 1**]] Thank you for your feedback. We will fix the reference formatting in the final manuscript. ■ **Statistical**
2 **significance:** In Table 2, we have already included the 95% confidence intervals obtained via stratified cross-validation.
3 In the final manuscript, we will include the p -values and confidence intervals for the results in Figures 5 and 6 as well.
4 ■ **Summary section:** Thank you for this suggestion. To address this comment, we will move Subsection 2.4 (lines 145
5 - 159) towards the end of Section 3, and add a summary of the inference algorithm to this Subsection.

6 [[**Reviewer 2**]] Thank you for the valuable comments and suggestions. ■ **Minor questions:** Number of states were
7 selected based on the BIC criterion, and the selected number conformed with existing clinical guidelines (please
8 refer to response “Disease phenotypes” for Reviewer 3). Model parameters were initialized via K -means clustering
9 applied to the unrolled time-series data. ■ **Inference-related questions:** Simpler inference algorithms recover emission
10 distributions similar to our algorithm but do a poorer job in recovering state dynamics (transition probabilities and
11 attention weights). Consequently, the predictive accuracy of mean-field and Markovian inference were better than
12 HMMs but slightly worse than RNNs. Accuracy results for both inference strategies will be added to the supplementary.

■ **Originality & baselines:** To emphasize the difference between our model and variants of HMMs, we have collected new results for extra baselines including a 4th-order HMM (4-HMM), a Hidden Semi-Markov model (HSMM), and a factorial HMM (FHMM). Among these, the most competitive was the 4-HMM, but its (predictive) performance was still significantly worse than the RNN benchmark. The reason our model outperforms HMM variants is that it adapts its state dynamics in a non-stationary fashion via time-varying attention weights. The Figure above depicts one patient’s trajectory by visualizing the attention weights allocated to previous states as they are being updated each time step. As we can see, the model exhibits Markovian dynamics towards the beginning and the end of the trajectory, and non-Markovian dynamics in



14 the intermediate steps. Each patient will have their own attention profile (and hence their own contextualized state
15 dynamics) based on their individual clinical observations. This enriches the discriminative power of our model
16 compared to HMM variants that entail stationary/fixed dynamics. ■ **Rao-Blackwellization:** Given the distributional
17 specification in Section 2 and an RNN estimator $RNN(x, z)$ of the posterior state distribution, it can be shown
18 that our inference network is equivalent to the conditional expectation $E[RNN(x, z) | T(x, z)]$ with respect to the
19 sufficient statistic $T(x, z)$. The sufficient statistic in our model is the attention weight sequence, i.e., $T(x, z) = \alpha$.
20 The proof of this equivalence follows from the seminal work on Rao-Blackwellized particle filters in [R2]; we will
21 provide the corresponding derivation for our model in the supplementary material. As pointed out by the reviewer,
22 Rao-Blackwellized estimators reduce the estimation variance: we will demonstrate that our algorithm minimizes
23 the estimation variance by annotating the cross-validation variance of all estimators in Figure 4. ■ **Uncertainty:**
24 Reliability/stability of phenotype inferences were judged by the variance of model parameters obtained via 5-fold
25 cross-validation: all findings were statistically significant (refer to response “Statistical significance” for Reviewer 1).
26 Posterior model uncertainty can be easily quantified via test-time Monte Carlo dropout applied to the inference network
27 as in [R1], without any modifications to our algorithm. A fully Bayesian approach with priors over model parameters is
28 also possible. Detailed discussion on uncertainty quantification will be added to the final manuscript.

29 [[**Reviewer 3**]] Thank you for the helpful comments and suggestions. In the final manuscript, we will fix the typo
30 in line 183 (Q6) and add a supplementary Section with details on the factorization in line 177 (Q4). ■ **Irregularly**
31 **spaced visits (Q1):** Our framework can be straightforwardly extended to the continuous-time (CT) setup by replacing
32 the discrete-time RNNs in the model and inference networks with CT RNN models such as *phased LSTMs* [R3]. Since
33 phased LSTMs also handle asynchronous observations (See [R3]), the inference network can be trained to update the
34 posterior state distribution based on partial observations at arbitrary time steps, emulating a “memoryful CT HMM”. We
35 have already implemented this variant of our model based on phased LSTMs — a related discussion will be added to
36 the supplementary. ■ **Disease phenotypes (Q2, Q3 and Q8):** The model does not need to be restricted to the number
37 of states for existing phenotype schemes (Q1). In general, model selection criteria — such as the Bayesian information
38 criterion (BIC) — would be used to select the number of states in a data-driven fashion. In our experiment, the number
39 of states selected by BIC matched that of existing phenotypic scheme. As it is the case in all unsupervised setups, the
40 meaningfulness of the states cannot be guaranteed and will have to be judged by experts (Q2). Backward transitions in
41 CF are possible through prescriptions of antibiotics. Monotonic progression (e.g., in Kidney diseases) can be easily
42 encoded in our model by setting the relevant transition parameters to 0 (Q8). ■ **NNL (Q7):** In the final manuscript, we
43 will replace the in-sample NNL (Figure 4) with the held-out NNL. Note that in all of the 3 cases in Figure 4, the model
44 is fixed and only the inference strategy is changed. The NNL displayed in Figure 4 is the true LL and not the ELBO.

45 **References:** [R1] Y. Gal, et al. "Dropout as a Bayesian approximation: Insights and applications", ICML, 2015. [R2] A. Doucet,
46 et al. "Rao-Blackwellised particle filtering for dynamic Bayesian networks." UAI, 2000. [R3] D. Neil, et al. "Phased LSTM:
47 Accelerating recurrent network training for long or event-based sequences." NeurIPS, 2016.