Response to Reviewer GE3W

Submission1374

April 12, 2024

1 Response to Reviewer GE3W

Thanks for asking these good questions, some of them are very insightful and will help us improve our manuscript. We provide detailed responses below to each of your questions.

1.1 The core literature anchoring this survey, particularly the review article pivotal to the authors' argument, dates back thirteen years (in line 129). Given the fast-paced advancements in disease modeling, a thirteen-year gap can overlook substantial developments that could critically inform and possibly transform current understanding and methodologies

Thanks for pointing this out. The main purpose of adding this old survey is to introduce different applications of disease progression models, (e.g., in cancer and HIV). In our experiments, we didn't use any methods from this survey as baselines. We also introduced the most recently proposed methods in the remaining part of related work, e.g., IEF was proposed in 2021.

1.2 Can the model handle cases where genetic data is sparse or incomplete, which is a common scenario in real-world clinical datasets?

This is a good question, and a hard question to answer based on our current result. If genetic data is sparse but complete, our method works since the VAE structure is good at inferring posterior from high dimensional space. However, if the genetic data is incomplete, it's very hard to answer. Since missing values is a huge topic, the best conclusion is that if we have a very good imputation method, and the non-missing part of the data is sufficiently informative, our method can still learn from genetic data. Alternatively, if the missing values are scarce, we can assume the data is sparse but complete in the VAE structure.

1.3 How does the model perform when applied to diseases with less genetic influence compared to environmental factors?

One of our motivations of using genetic information is to build a causally correct inference model. According to our factorization, the transition between latent disease state is conditioned on its direct or indirect cause (genetic influence and environmental factors) and its generic enough on those underlying causal factors. The only assumptions of our model is that the latent disease state will be defined by clinical observations that are typically measured by the healthcare system.

However, due to data limitations, we only have access to genetic information, and only designed the structure for genetic influence. For diseases with less genetic influence compared to environmental factors, our model can easily augment the former with genetic data and then build the model without any further modification.

1.4 What are the computational requirements for implementing this model in real-time clinical decision-making processes? Could the authors provide more insight into how the model might be integrated into existing clinical workflows?

Since disease progression modeling is usually designed to model chronic diseases, the application scenario is closer to a longitudinal analysis problem than a time series analysis problem. Thus, in the "real-time clinical decision-making" scenario, after deployment, we don't have to update the model very frequently, i.e., no need for continuous training, since the next visit of the same patient could be several months later. We can update/retrain the model every 3 or 6 months using the most updated dataset.

Given genetic data and historical clinical observations, like labs, vitals, diagnosis, treatment, using our proposed framework, the sampled posterior from inference model $q_\phi(Z_t|Y_t, Z_{t-1})$ indicate what's the patient current disease state (Z_t). Also, given patient's current disease states (Z_t) and genetic data (V), sampling prior from the generative model $p_\theta(Z_{t+1}|Z_t, V)$ can predict the probability of the disease state at next visit.

1.5 How does the model's predictive power vary with the stage of disease progression (e.g., early vs. late-stage chronic kidney disease)?

This is a very good question. For CKD, the disease states vary from 1 to 5, reflecting the increase of severity. One observation from the data is that most patients having CKD didn't recover, although they received treatment. In other words, transition from state 5 to 1 is very rare. We also show this trend in Figure

6, where the disease severity index goes up as the disease progresses. This makes the inference of disease state at late-stage easier than the early stage, because the number of possible transitions is less in late-stage. For example, given a patient at stage 4, transitions 4→4 and 4→5 are predominant. However, for patients at stage 3, transitions 3→3, 3→4, 3→5 are predominant.

Thus, we believe the disease progression model has a better inference power for late-stage than early stage CKD. The biology of the disease also supports this, as late-stage CKD patients transition faster than early stages.

1.6 In line 74 ”disease trajectory patterns are uniform across all patients”: what’s trajectory patterns patterns? Time patterns? Clinical detection patterns? or Functional decay patterns?

Here, the terminology “disease trajectory patterns” is expected to describe the pattern of disease states transition. For example, a patient with fast disease progression will exhibit very different disease states transition probabilities between disease states than a patient having slow progression. In other words, the probability transition matrix for these two patients would be different. In this statement, we highlighted existing methods assuming all patients share the same probability transition matrix. That’s why we use the terminology ”are uniform across all patients.” One of our contributions is our proposed framework allows patients from different genetic groups to have different probability transition matrices.

1.7 Please summarise the advantages of genetic makeups inference and genetics driven state transition modelling respectively

The main advantage of using genetic makeups inference is to learn a latent representation from high-dimensional and noisy genetic information. When jointly inferred with genetics driven state transition, the learned representation can capture the information from target-disease-related genomic makeups. The advantage of genetics driven state transition is to allow the disease progression to be conditioned on these genetic makeups and treatments.

We hope our response address all your concerns. We like the idea of assessing the model’s predictive power when the stage of disease progression varies. If there any other questions or you want to discuss more, we are happy to answer, and we kindly request that you reconsider your assessment based on our response.