

Interpreting the All-Future DeepHealth Model

all-future DeepHealth 模型的解释边界

摘要

The all-future DeepHealth model should be interpreted as a history-conditioned incident disease and mortality risk model. At a query time t , the hidden state $h(t)$ summarizes the observed history up to that time, and the model estimates future first-occurrence risks for diseases in the model vocabulary, together with future mortality risk. The model does not directly estimate current clinical disease burden, organ damage, frailty, disease severity, recurrence risk, or disease-specific weights. Those quantities require labels, mappings, weights, or supervision that are not part of the present model.

all-future DeepHealth 模型应被解释为一个基于既往轨迹的未来新发疾病和死亡风险模型。在查询时刻 t ，隐含状态 $h(t)$ 汇总了截至该时刻的已观测历史；模型输出的是模型词表内各疾病的未来首次发生风险，以及未来死亡风险。当前模型并不直接估计当前临床疾病负担、器官损伤、衰弱程度、疾病严重度、复发风险或疾病特异性权重。这些量都需要当前模型之外的标签、映射、权重或额外监督。

1 English

1.1 Model Object

For individual i , let $\mathcal{H}_i(t)$ denote the observed history up to query time t . The all-future model produces

$$h_i(t) = f_{\theta}(\mathcal{H}_i(t)),$$

and, for each disease token d in the modeled disease vocabulary $\mathcal{D}$, estimates a future first-occurrence risk over a horizon τ :

$$p_{i,d}(t, \tau) = P_{\theta}(T_{i,d} \in (t, t + \tau] \mid h_i(t)),$$

where $T_{i,d}$ is the first observed occurrence time of disease d . If the model includes a death endpoint, it also estimates

$$p_{i,\text{death}}(t, \tau) = P_{\theta}(T_{i,\text{death}} \in (t, t + \tau] \mid h_i(t)).$$

The disease sequence is a first-occurrence sequence. Therefore, from the input history itself we know

$$m_{i,d}(t) = \mathbf{1}\{T_{i,d} \leq t\}.$$

This historical indicator is not learned by the model; it is read directly from the event history.

1.2 Masking Already Occurred Diseases

For a disease that has already occurred before or at t , the model output for that disease should not be interpreted as recurrence risk or current disease activity. For future incident disease summaries, already occurred diseases should be masked:

$$p_{i,d}^{\text{new}}(t, \tau) = [1 - m_{i,d}(t)]p_{i,d}(t, \tau).$$

1.3 Directly Supported Quantities

The current model directly supports the following quantities.

Disease-specific future first-occurrence risk. For each modeled disease d ,

$$p_{i,d}^{\text{new}}(t, \tau)$$

is the estimated risk that disease d newly appears within the next τ years, conditional on the history summarized by $h_i(t)$.

Future mortality risk.

$$p_{i,\text{death}}(t, \tau)$$

is the estimated probability of death within the next τ years. Death is a terminal endpoint and should not be treated as an ordinary disease burden weight.

Future incident disease risk in a specified set. For any analyst-specified subset of disease tokens $G \subseteq \mathcal{D}$, the model can summarize future incident risk within that set:

$$R_i^G(t, \tau) = 1 - \prod_{d \in G} [1 - p_{i,d}^{\text{new}}(t, \tau)].$$

This quantity answers: “What is the model-estimated probability that at least one not-yet-observed disease in G first occurs within the next τ years?” It does not include death. Mortality risk should be reported separately as $p_{i,\text{death}}(t, \tau)$.

If G is called an organ system, the disease-to-organ grouping is external to the model and should not be described as a learned organ score.

1.4 Model Attribution to Predicted Mortality Risk

The model can also be queried for a model-internal attribution of historical disease sets to predicted mortality risk. For a disease set G , define the original mortality risk as

$$p_{i,\text{death}}^{\text{orig}}(t, \tau),$$

and the risk after deleting historical disease tokens in G from the input history as

$$p_{i,\text{death}}^{(-G)}(t, \tau).$$

On the probability scale, the attribution is

$$\Delta p_i^G(t, \tau) = p_{i,\text{death}}^{\text{orig}}(t, \tau) - p_{i,\text{death}}^{(-G)}(t, \tau).$$

A more stable primary scale is the cumulative-hazard scale:

$$\Lambda_{i,\text{death}}(t, \tau) = -\log [1 - p_{i,\text{death}}(t, \tau)],$$

$$\Delta \Lambda_i^G(t, \tau) = \Lambda_{i,\text{death}}^{\text{orig}}(t, \tau) - \Lambda_{i,\text{death}}^{(-G)}(t, \tau).$$

This quantity should be described as model attribution to predicted mortality risk. It is not a causal contribution, not an organ damage score, and not a clinical disease-burden weight. Because diseases can interact within the trajectory model, deleting an organ/system as a whole is generally more interpretable than summing single-disease attributions.

1.5 Historical Counts Are Not Model-Derived Burden

One may count observed historical diseases:

$$B_i^{\text{history}}(t) = \sum_{d \in \mathcal{D}} m_{i,d}(t).$$

This quantity is a direct count from the input sequence. It does not require the model and should not be presented as a model-derived disease burden score. Without disease severity labels or disease weights, it treats all disease tokens equally.

1.6 What the Current Model Does Not Estimate

The current model does not directly estimate:

- current clinical disease burden;
- organ damage, organ age, or organ functional reserve;
- frailty or frailty weights;

- severity of a newly diagnosed disease;
- recurrence risk after first occurrence;
- relative clinical importance of different disease tokens.

These interpretations require additional labels, mappings, weights, or model training objectives. Using the present all-future model to claim these quantities would be over-interpretation.

1.7 Post-Onset Prognosis for the Same New Disease

For a disease d that newly occurs at time $T_{i,d}$, the model cannot infer the clinical severity of that disease itself. However, after the disease has entered the history, one may query the model again and compare subsequent future risks:

$$p_{i,e}^{\text{new}}(T_{i,d}, \tau), \quad e \neq d, \quad p_{i,\text{death}}(T_{i,d}, \tau).$$

This supports a prognosis-oriented statement: the same incident diagnosis may be followed by different future disease and mortality risk profiles in people with different prior trajectories. It should not be described as direct disease severity.

1.8 Future Extension with Reliable Recurrence Data

The above interpretation is constrained by the first-occurrence nature of the current disease sequence. UK Biobank does not provide a reliable longitudinal record of disease recurrence, relapse, repeated admissions, treatment escalation, or episode-level severity for the modeled disease tokens. Therefore, the present model cannot be used to estimate recurrence risk or ongoing disease activity after first onset.

If reliable recurrence or repeated-event data were available, one could define a different modeling target. Let $N_{i,d}(t)$ be the counting process for all episodes of disease d , not only its first occurrence. A recurrence-aware model could estimate the future increment

$$P_{\theta}(N_{i,d}(t + \tau) - N_{i,d}(t) > 0 \mid h_i(t)),$$

or the expected number of future episodes

$$E_{\theta}[N_{i,d}(t + \tau) - N_{i,d}(t) \mid h_i(t)].$$

For individuals with $m_{i,d}(t) = 1$, this would support a genuine post-onset interpretation of future recurrence or disease activity. For individuals with $m_{i,d}(t) = 0$, it would remain an incident disease risk.

With such data, the model could also separate three clinically different quantities:

- incident risk: the first future onset of a disease not yet observed;
- recurrence or repeated-event risk: future episodes after a disease has already occurred;
- mortality risk: a terminal endpoint that competes with future non-fatal events.

If recurrence episodes were linked to reliable episode-level severity labels, such as hospitalization intensity, treatment escalation, or validated severity grades, then a further supervised model could learn severity-aware prognosis. These extensions would require new data and new training targets; they are not available from the current all-future first-occurrence model.

2 中文

2.1 模型对象

对于个体 i , 令 $\mathcal{H}_i(t)$ 表示查询时刻 t 之前已经观测到的历史。all-future 模型产生

$$h_i(t) = f_{\theta}(\mathcal{H}_i(t)),$$

并且对模型疾病词表 $\mathcal{D}$ 中的每个疾病 d , 估计未来 τ 年内的首次发生风险:

$$p_{i,d}(t, \tau) = P_{\theta}(T_{i,d} \in (t, t + \tau] \mid h_i(t)),$$

其中 $T_{i,d}$ 是疾病 d 的首次观测发生时间。如果模型包含死亡终点, 则同时估计

$$p_{i,\text{death}}(t, \tau) = P_{\theta}(T_{i,\text{death}} \in (t, t + \tau] \mid h_i(t)).$$

当前疾病序列是 first-occurrence 序列。因此, 从输入历史本身即可得到

$$m_{i,d}(t) = \mathbf{1}\{T_{i,d} \leq t\}.$$

这个历史发生指示量不是模型学出来的, 而是直接从事件历史中读出的。

2.2 已经发生疾病的 Mask

如果某个疾病在 t 之前或 t 时已经发生, 那么该疾病对应的模型输出不应解释为复发风险, 也不应解释为当前疾病活动程度。在汇总未来新发疾病风险时, 应对已经发生过的疾病进行 mask:

$$p_{i,d}^{\text{new}}(t, \tau) = [1 - m_{i,d}(t)]p_{i,d}(t, \tau).$$

2.3 当前模型直接支持的量

当前模型直接支持以下几类量。

疾病层面的未来首次发生风险。对每一个模型内疾病 d ,

$$p_{i,d}^{\text{new}}(t, \tau)$$

表示在 $h_i(t)$ 所总结的历史条件下, 疾病 d 在未来 τ 年内新发生的风险。

未来死亡风险。

$$p_{i,\text{death}}(t, \tau)$$

表示未来 τ 年内死亡的概率。死亡是终末结局, 不应作为普通疾病负担权重加入疾病负担求和。

指定疾病集合内的未来新发风险。对于任意由分析者预先指定的疾病 token 集合 $G \subseteq \mathcal{D}$, 模型可以汇总该集合内的未来新发风险:

$$R_i^G(t, \tau) = 1 - \prod_{d \in G} [1 - p_{i,d}^{\text{new}}(t, \tau)].$$

这个量回答的是: 模型估计该个体在未来 τ 年内, 至少新发生一个 G 中尚未发生疾病的概率是多少。它不包含死亡; 死亡风险应单独报告为 $p_{i,\text{death}}(t, \tau)$ 。

如果将 G 称为某个器官系统, 那么疾病到器官的分组来自模型外部, 不应表述为模型学到的器官评分。

2.4 死亡风险预测的模型归因

模型还可以用于计算历史疾病集合对死亡风险预测的模型内部归因。对于疾病集合 G , 令原始死亡风险为

$$p_{i,\text{death}}^{\text{orig}}(t, \tau),$$

将输入历史中属于 G 的历史疾病 token 删除后, 再次查询模型, 得到

$$p_{i,\text{death}}^{(-G)}(t, \tau).$$

在概率尺度上, 归因可以写为

$$\Delta p_i^G(t, \tau) = p_{i,\text{death}}^{\text{orig}}(t, \tau) - p_{i,\text{death}}^{(-G)}(t, \tau).$$

更推荐的主尺度是累计 hazard:

$$\Lambda_{i,\text{death}}(t, \tau) = -\log [1 - p_{i,\text{death}}(t, \tau)],$$

$$\Delta \Lambda_i^G(t, \tau) = \Lambda_{i,\text{death}}^{\text{orig}}(t, \tau) - \Lambda_{i,\text{death}}^{(-G)}(t, \tau).$$

这个量应表述为疾病集合对死亡风险预测的模型归因。它不是因果贡献, 不是器官损伤评分, 也不是临床疾病负担权重。由于轨迹模型中疾病之间可能存在交互, 整体删除一个器官系统通常比逐个疾病归因后相加更容易解释。

2.5 历史计数不是模型派生的疾病负担

可以计算历史已经发生过多少个模型内疾病：

$$B_i^{\text{history}}(t) = \sum_{d \in \mathcal{D}} m_{i,d}(t).$$

但这个量只是从输入序列直接计数，不需要模型，因此不应表述为模型派生的疾病负担评分。在没有疾病严重度标签或疾病权重时，它默认所有疾病 token 等价。

2.6 当前模型不能估计什么

当前模型不能直接估计：

- 当前临床疾病负担；
- 器官损伤、器官年龄或器官功能储备；
- 衰弱程度或衰弱权重；
- 某个新发疾病本身的严重程度；
- 首次发生后的复发风险；
- 不同疾病 token 之间的相对临床重要性。

这些解释都需要额外标签、映射、权重或新的训练目标。用当前 all-future 模型直接声称这些量，属于过分解读。

2.7 同一新发疾病后的预后差异

对于在 $T_{i,d}$ 时刻新发生的疾病 d ，模型不能推断这个疾病本身的临床严重程度。但是，当该疾病已经进入历史之后，可以再次查询模型，并比较之后的未来风险：

$$p_{i,e}^{\text{new}}(T_{i,d}, \tau), \quad e \neq d, \quad p_{i,\text{death}}(T_{i,d}, \tau).$$

这支持一种预后层面的表述：同一个新发诊断出现在不同既往轨迹的人身上，可能对应不同的后续新病和死亡风险结构。但这不应被表述为模型直接判断了疾病严重度。

2.8 如果有可靠复发数据，未来可以做什么

上述解释受到当前 first-occurrence 疾病序列的限制。UK Biobank 并没有为模型词表中的疾病提供可靠的纵向复发、复燃、重复住院、治疗升级或 episode 层面严重程度记录。因此，当前模型不能用于估计首次发病后的复发风险或持续疾病活动。

如果未来有可靠的复发或重复事件数据，可以定义一个不同的建模目标。令 $N_{i,d}(t)$ 表示疾病 d 的所有事件计数过程，而不只是首次发生。一个考虑复发的模型可以估计未来事件增量：

$$P_{\theta}(N_{i,d}(t + \tau) - N_{i,d}(t) > 0 \mid h_i(t)),$$

或者估计未来事件次数的期望：

$$E_{\theta}[N_{i,d}(t + \tau) - N_{i,d}(t) \mid h_i(t)].$$

对于 $m_{i,d}(t) = 1$ 的个体，这才可以支持真正的发病后复发风险或疾病活动解释；对于 $m_{i,d}(t) = 0$ 的个体，它仍然对应未来首次发病风险。

在这样的数据条件下，模型可以区分三个临床上不同的量：

- 新发风险：尚未发生疾病的未来首次发生风险；
- 复发或重复事件风险：疾病已经发生后的未来 episode 风险；
- 死亡风险：与非致死事件竞争的终末结局风险。

如果复发 episode 还带有可靠的 episode 层面严重程度标签，例如住院强度、治疗升级或经过验证的严重程度分级，那么还可以进一步训练有监督的严重程度相关预后模型。但这些扩展都需要新的数据和新的训练目标，并不是当前 all-future first-occurrence 模型已经具备的能力。

3 Recommended Wording

English. The all-future model is a history-conditioned incident disease and mortality risk model. Its outputs support separate reporting of future mortality risk and future incident disease risk over analyst-specified disease sets. Historical disease sets may be ablated to obtain model attribution to predicted mortality risk, but this is not causal contribution and does not directly quantify current disease burden, organ damage, frailty, or disease severity.

中文。 all-future 模型是基于既往轨迹的未来新发疾病和死亡风险模型。它的输出可以支持分别报告未来死亡风险，以及指定疾病集合内的未来新发疾病风险。通过删除历史疾病集合可以得到其对死亡风险预测的模型归因，但这不是因果贡献，也不能直接量化当前疾病负担、器官损伤、衰弱程度或疾病严重度。