arXiv:2304.05010v2 [stat.AP] 8 Aug 2023

利用图表示学习刻画家族信息对疾病风险的个性化影响

Sophie Wharrie SOPHIE.WHARRIE@AALTO.FI
芬兰阿尔托大学计算机科学系

ZhiyuYang ZHIYU.YANG@HELSINKI.FI芬兰
分子医学研究所，*FIMM*，赫尔辛基大学，芬兰

Andrea Ganna
AGANNA@BROADINSTITUTE.ORG芬兰分子医学研究所，*FIMM*，赫尔辛基大学，芬兰；麻省理工学院和哈佛大学博德研究所，美国马萨诸塞州剑桥；麻省总医院，美国马萨诸塞州剑桥

塞缪尔·卡斯基 SAMUEL.KASKI@AALTO.FI
芬兰阿尔托大学计算机科学系；英国曼彻斯特大学计算机科学系

摘要

家族史被视为许多疾病的风险因素，因为它隐含地反映了共同的遗传、环境和生活方式因素。芬兰覆盖多代人的全国性电子健康记录（EHR）系统，为研究整个家庭病史的连通网络带来了新的机遇。在本研究中，我们提出一种基于图的深度学习方法，用于学习可解释的监督式表征，揭示每位家庭成员的纵向病史如何影响患者的疾病风险。我们证明，与针对芬兰全国性电子健康记录系统设计的临床启发式基线方法和深度学习基线方法相比，该方法有助于预测5种复杂疾病表型的十年疾病发病情况；该系统涵盖700万名个体及其三级亲属。通过运用图可解释性技术，我们展示了基于图的方法能够通过识别对预测至关重要的亲属和特征，更个性化地建模家族信息与疾病风险。

1. 引言

家族史是患者易患某些健康风险的公认指标。然而，对于病因复杂的常见慢性疾病，例如冠状动脉疾病和2型糖尿病，评估家族史会因多种遗传、环境和生活方式因素之间的相互作用而变得复杂（Wilson等人，2009；Valdez等人，2010）。

近年来，电子健康记录系统的日益普及，使得开发深度学习方法来研究患者的详细病史成为可能。除了单个患者的电子健康记录外，覆盖多代人的全国性和区域级电子健康记录系统，为研究由电子健康记录构成的连通网络带来了新的机遇。

© 2023年 S. Wharrie, Z. Yang, A. Ganna & S. Kaski。

针对整个家庭。我们内部的数据集包含芬兰超过700万人的三十年全国电子健康记录数据，使得能够针对最远至三级亲属开展一种新型深度学习研究。

既有研究指出，在利用临床和生物学数据预测各类疾病风险的机器学习模型中，纳入家族史特征是有益的（Behravan 等人，2020年；Kinreich 等人，2021年）。然而，为表格数据开发的机器学习方法并未明确建模家族史内在的信息性几何结构，而这种结构本身就有助于疾病预测。具体而言，遗传关系的亲疏、罹患某种疾病的亲属人数或相关风险因素（可能有很多），以及亲属确诊这些疾病时的年龄，都是重要信息；我们认为，采用几何深度学习方法能够更恰当地对这些信息进行建模。此外，电子健康记录包含数千个可能具有信息价值的变量，需要借助新的机器学习工具，以整个家庭为尺度进行分析。

在本研究中，我们将利用电子健康记录中的家族史进行疾病风险预测，重新表述为一个图建模问题。图是一种表现力丰富的数据结构，可用于表示每位家庭成员个体层面的高维纵向电子健康记录数据；而遗传亲缘关系的连通结构则编码了先验知识，体现不同类型的家庭成员对于预测个体疾病风险的重要性。我们的主要贡献概括如下：

- 我们介绍了一种可扩展的、疾病无关的机器学习工具，该工具基于图神经网络（图神经网络）和长短期记忆网络（长短期记忆网络），用于学习监督式表征，以根据家族信息预测患者的疾病风险；
- 通过对全国范围电子健康记录数据的研究，预测5种复杂疾病的十年疾病发病情况，我们表明，与受临床启发的基线方法和家族史深度学习基线方法相比，基于图的方法能够更好地预测多因素疾病；
- 我们使用图可解释性技术证明，图神经网络-长短期记忆网络嵌入能够识别家庭成员的医疗特征（从 1500 多个特征中），这些特征比流行病学基线识别的特征更适合预测疾病风险。

我们分享为这项工作开发的软件[1] 以及一个用于模拟私有数据集[2]关键特性的合成数据集。

Generali（忠利保险集团）医疗保健领域中关于机器学习的宝贵见解

家族史长期以来一直被视为许多疾病的风险因素，但在实践中，由于缺乏合适的数据收集和解读工具，这些信息在医疗保健中往往未得到充分利用（Feero 等人，2008年；Ginsburg 等人，2019年）。在这项工作中，我们将电子健康记录建模为扩展家庭病史的连通网络，并学习监督式图表示，以预测疾病的发病。我们希望读者从这项工作中获得的可泛化的机器学习洞见是，对家族史背后的几何结构进行显式建模，可以提供一种信息丰富的表示

1. github.com/dsgelab/family-EHR-graphs2.
https://www.finregistry.fi/

家庭内部影响复杂疾病的共同遗传、环境和生活方式因素。我们在这项工作中开发的软件工具可方便地扩展到为大型家族结构建模数千个变量。我们认为，这项工作的临床相关性体现在：(1) 证明了在个人风险预测中，来自家庭成员且超出一级亲属范围的疾病史，以及除目标疾病之外的健康史的价值；以及 (2) 将图神经网络作为一种整合大量信息与复杂家族关系的新方法，可用于改善临床环境中的家族史采集。

2. 相关工作

2.1. 家族史与疾病预测

我们的工作提出了一种新的机器学习框架，用于建模家族史在疾病风险中的作用。临床医生会定期询问患者的家族史，以便依据既有知识、临床指南和风险评估工具辅助临床决策（Wilson 等人，2009；Valdez等人，2010；Ginsburg 等人，2019年）。人们已开发出多种统计方法，用于建模家族史以及疾病的遗传效应。在有遗传数据可用时，可以使用多基因风险评分方法量化疾病的遗传风险（Lewis 和 Vassos，2020年；Li 等人，2020年）。BLUP（最佳线性无偏预测）是一种广泛用于动物育种的表型预测方法，通常将系谱信息作为线性混合模型中的随机效应进行建模，并结合其他固定效应预测因子构建预测结果（Wang 等人，2018年）。由于缺乏大量系谱数据，这种方法在人群中的应用受到限制。机器学习方法已被用于将不同类型的临床数据和生物学数据与家族史的表格型（非图）特征相结合（Behravan 等人，2020年；Kinreich 等人，2021年）。然而，据我们所知，我们首次将患者的家族史表示为涵盖三级亲属的电子健康记录图，并采用几何深度学习方法对其进行建模。

2.2. 医疗保健中的图神经网络

在本研究中，我们使用图神经网络（图神经网络）将家族信息建模为图数据结构。图表示学习方法（例如图神经网络）已在生物医学和医疗保健领域中，针对具有图结构数据的多项机器学习任务取得领先性能，包括对蛋白质-蛋白质相互作用、药物-药物相互作用和疾病关联进行建模（Li 等人，2022年）。在临床环境中，图神经网络已应用于表示为图数据结构的电子健康记录，例如连接与医疗就诊相关的实验室检查和药物（Mao 等人，2022年；Wu等人，2019年）。其他研究已将图神经网络应用于电子健康记录数据，用于基于患者相似性构建的图，其中节点表示个体患者，边表示患者之间的相似性。例如，Parisot 等人（2018年）根据从影像数据中得到的人口统计学特征和表型来衡量相似性，而Tong 等人（2022年）则考虑相似的诊断结果。我们根据家族关系定义患者之间的相似性，并研究一种图分类任务，其中每位患者的家族图（包括三级亲属以内）用于预测二元疾病结局。

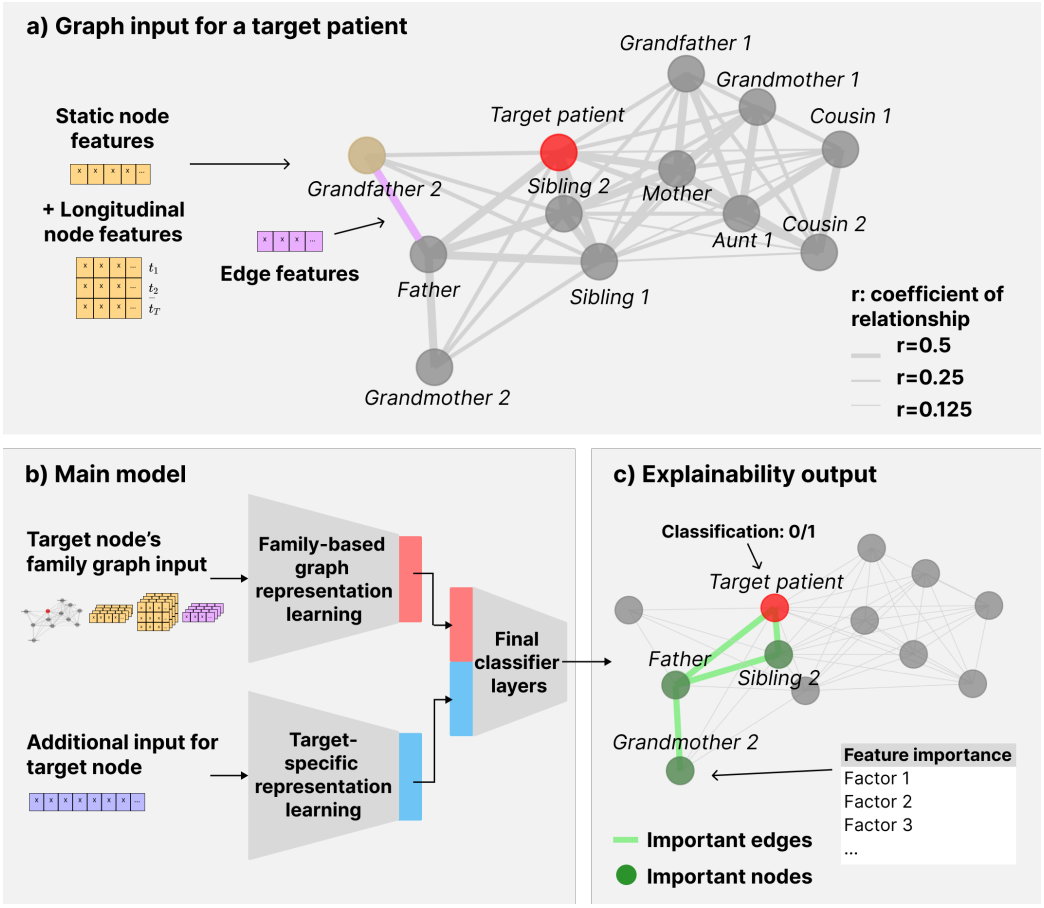


图1：建模方法概览。a）图输入包含（最多三代）家庭成员的电子健康记录，作为节点特征，以及遗传关系的描述信息，作为边特征；b）机器学习模型联合学习家庭（图）表征和患者特异性表征，并在最终层中将二者拼接起来，以分类患者结局；c）图可解释性分析识别出模型在预测患者风险时认为重要的亲属（节点）以及风险因素/共病（节点特征）。

3. 方法

本文开发的建模方法概览见图 1。以下各节将介绍我们如何设计一个机器学习系统，该系统可为患者预测二元疾病结局，并解释哪些亲属及其医疗特征对模型得出预测结果最为重要。

3.1. 预备知识与符号

为了研究家族信息对疾病风险的影响，我们考虑这样一个二元分类问题：基于先前观察到的电子健康记录数据预测未来的健康结局。我们

将我们用于预测健康结局的研究队列称为目标样本，记为集合 P_{target}。我们还定义了一个更广泛的患者集合，包括目标样本及其亲属，我们称之为图样本，并记为 P_{graph}。

目标患者及其家庭成员的数据，以及他们之间遗传关系的性质，可以使用图数据结构进行编码，如图1a所示。图通常定义为 $\mathcal{G} := (V, E)$，其中 V 是一组顶点（或节点），E 是一组称为边的顶点对。在机器学习中，将图定义为 $\mathcal{G} := (\boldsymbol{X}, (\boldsymbol{I}, \boldsymbol{E}))$ 更为方便，其中 $\boldsymbol{X} \in \mathbb{R}^{N \times F}$ 是一个节点特征矩阵，包含 F 个特征和 N 个节点，而 $(\boldsymbol{I}, \boldsymbol{E})$ 是一个包含 E 条边信息的元组，其中 $\boldsymbol{I} \in \mathbb{N}^{2 \times E}$ 包含边对的索引，$\boldsymbol{E} \in \mathbb{R}^{E \times D}$ 是一个边特征矩阵，包含 D 个特征和 E 条边（Fey和Lenssen，2019年）。在本文中，我们考虑形式为 $\mathcal{D} := \{(\mathcal{G}_i, \tilde{\boldsymbol{X}}_i, \boldsymbol{y}_i)\}_{i=1}^{|P_{target}|}$ 的数据集，其目标是根据目标患者的家族图 $\mathcal{G}_i = (\boldsymbol{X}_i, (\boldsymbol{I}_i, \boldsymbol{E}_i))$ 及其自身的特征集 $\boldsymbol{X}_i$，预测该患者的二元结局 $\boldsymbol{y}_i \in \{0, 1\}$。

3.2. 构建家族图

几何深度学习算法中的图可视为几何先验，为学习算法提供归纳偏置（Bronstein等人，2021年）。从生物学角度来看，一级亲属共享的DNA多于二级或三级亲属，而且一级亲属也更有可能共享环境条件和生活方式因素。除了遗传关系的亲疏程度外，亲属病症的确诊年龄以及是否存在相关风险因素等家族史的其他属性，在研究复杂疾病时也具有参考价值（Valdez等人，2010）。

基于这些假设，我们构建了一个家族图 $\mathcal{G}_i = (\boldsymbol{X}_i, (\boldsymbol{I}_i, \boldsymbol{E}_i))$，其中包含目标患者 $i \in P_{target}$ 及其在 P_{graph} 中的一级、二级和三级亲属的节点。对于表示 $\boldsymbol{I}_i$ 遗传亲缘关系的图边，我们在合成数据实验中对三种不同配置进行了实证比较（详细结果见附录6图A）：（1）父母/子女边（类似家谱或谱系图）；（2）父母/子女边，并额外添加将每个节点直接连接到目标节点的边；（3）连接所有遗传相关节点的边。这种做法源于机器学习的视角：为了提高图神经网络的有效性，图通常会采用虚拟节点以及连接两跳邻居的边等策略进行扩充（Wu等人，2022年）。边特征 $\boldsymbol{E}_i$ 包括亲缘系数、r；对于支持多个边特征的图神经网络层，我们还添加了二元指示变量，用于描述相对于目标节点的遗传关系类型（父母/子女、兄弟姐妹/祖父母等）。节点特征包括静态特征（年龄、性别）和纵向特征（按年份汇总的诊断代码）。图1a展示了一个图示例，实验所用特征的更多详情见第4节。

3.3. 学习监督式图表示

我们构建了一种神经网络架构，能够根据患者自身的电子健康记录数据以及其家庭成员电子健康记录的图表示来学习对患者的健康结局进行分类。多个输入头分别对应于目标个体自身的数据、$\boldsymbol{X}_i = (\boldsymbol{X}_{i,\ \text{static}}, \boldsymbol{X}_{i,\ \text{long}})$，以及

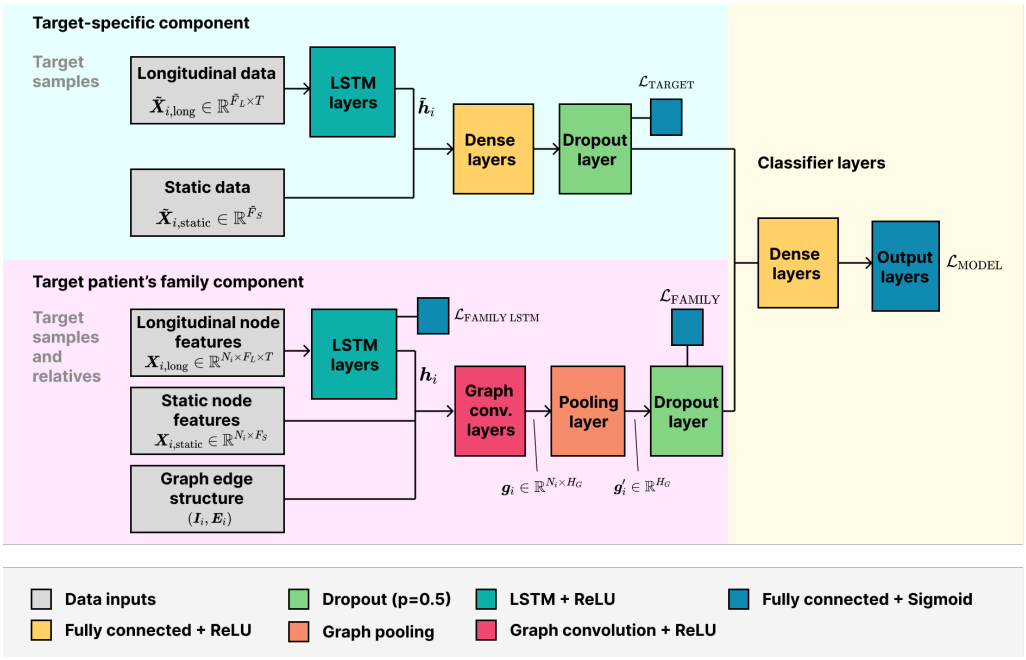


图2：神经网络架构图。

其家族的数据，$\mathcal{G}_i = ((\boldsymbol{X}_{i,\text{ static}}, \boldsymbol{X}_{i,\text{ long}}), (\boldsymbol{I}_i, \boldsymbol{E}_i))$，共同学习目标表征和基于家族的表征，这些表征在最终的分类器层中进行组合，如图 1b 所示。

完整的神经网络架构如图2所示。利用双向长短期记忆层的最终隐藏状态输出，将纵向输入降维为 H_L维嵌入，$\tilde{\boldsymbol{h}}_i \in \mathbb{R}^{H_L}$ 用于 $\tilde{\boldsymbol{X}}_{i,\text{ long}} \in \mathbb{R}^{F_L \times T}$ 以及 $\boldsymbol{h}_i \in \mathbb{R}^{N_i \times H_L}$ 用于 $\boldsymbol{X}_{i,\text{ long}} \in \mathbb{R}^{N_i \times F_L \times T}$；这是一种常用于纵向电子健康记录的架构（Morid 等人，2023年）。与目标患者数据对应的模型组件实际上对应于一个长短期记忆网络：嵌入 $\tilde{\boldsymbol{h}}_i$与静态数据 $\tilde{\boldsymbol{X}}_{i,\text{ static}} \in \mathbb{R}^{F_S}$进行拼接，随后连接堆叠的全连接层和一个丢弃层，以防止过拟合。与目标样本及其亲属对应的家族组件结合了长短期记忆网络层和图神经网络层，类似于Tong 等人提出的方法（2022年）。图卷积层为每个节点学习一个独立的嵌入 $\boldsymbol{g}_i \in \mathbb{R}^{N_i \times H_G}$，其隐藏维度为 H_G，输入来自 $((\boldsymbol{h}_i, \boldsymbol{X}_{i,\text{ static}}), (\boldsymbol{I}_i, \boldsymbol{E}_i))$.

图神经网络的卷积层充当传播模块，从邻居节点聚合信息，随后池化层将节点级表征聚合为图级表征（Zhou 等人，2020年）。目前已经为图神经网络开发了各种类型的传播层和池化层，因此在实验中，我们评估模型对各种设计选择的敏感性（图 6）。例如，一个简单的图卷积网络层在每一层 l 中，通过以下操作更新家庭成员 j 的节点嵌入$\boldsymbol{g}_{i,\,j}$:

$$\boldsymbol{g}_{i,j}^{(l+1)} = \sigma(\hat{\boldsymbol{D}}^{-\frac{1}{2}}\hat{\boldsymbol{A}}\hat{\boldsymbol{D}}^{-\frac{1}{2}}\boldsymbol{g}_{i,j}^{(l)}\boldsymbol{W}^{(l)}), \tag{1}$$

其中 $\hat{\boldsymbol{D}}^{-\frac{1}{2}}\hat{\boldsymbol{A}}\hat{\boldsymbol{D}}^{-\frac{1}{2}}$是归一化邻接矩阵[3]， $\boldsymbol{W}^{(l)}$是神经网络的权重矩阵，而 σ 是激活函数，例如 ReLU$(\cdot) = \max(0, \cdot)$ (Kipf 和 Welling，2016年)。直观来说，这相当于对节点 j 及其在图中的邻居节点（亲属）的嵌入进行加权求和。随后通过池化操作，从一组节点嵌入 $\boldsymbol{g}_i \in \mathbb{R}^{N_i \times H_G}$中应用求和或求平均等操作，获得图级嵌入 $\boldsymbol{g}_i' \in \mathbb{R}^{H_G}$。

从神经网络的目标组件和图组件中学习到的表示在最终层中进行组合，以预测二元结局 $\hat{\boldsymbol{y}}_i$。 Tong 等人（2022年）观察到，将长短期记忆网络层与图神经网络结合时，表示学习的质量可能会下降。我们采用其工作中提出的解决方案，并使用以下形式的损失函数

$$\mathcal{L} = \gamma\mathcal{L}_{\text{MODEL}} + \alpha\mathcal{L}_{\text{TARGET}} + \beta\mathcal{L}_{\text{FAMILY}} + \delta\mathcal{L}_{\text{FAMILY LSTM}},$$

其中包含完整模型输出 $\mathcal{L}_{\text{MODEL}}$、特定目标表示学习组件 $\mathcal{L}_{\text{TARGET}}$、图表示学习组件 $\mathcal{L}_{\text{FAMILY}}$ 以及图节点特征的长短期记忆网络嵌入 $\mathcal{L}_{\text{FAMILY 长短期记忆网络}}$ 的个体（加权）二元交叉熵（BCE）损失项，其中 γ， α， β， $\delta \in \mathbb{R}$ 为超参数。每个单独的损失项形式为

$$\mathcal{L}_j = -\frac{1}{N}\sum_{i=1}^{N} w_i[y_i \log(\hat{y}_{i,j}) + (1 - y_i)\log(1 - \hat{y}_{i,j})],$$

其中 $N = |P_{target}|$ 是目标患者的数量，$\hat{y}_{i,\ j}$ 是通过在相应的 $\mathcal{L}_j$ （图 2）所指示的位置应用全连接层和 sigmoid 层获得的，并且$w_i = N/n_i$ 是由于数据集中的类别不平衡而使用的类别权重，其中 n_i 是与个体 i属于同一类别的样本数量。我们的消融研究评估了模型对各种设计选择的敏感性，例如是否使用家庭数据和长短期记忆网络（表 2）。实验中使用的超参数值的完整细节见附录 A ，我们在表 4中报告了 GNN-LSTM 算法和基线方法的计算时间。

3.3.1. Additional considerations for model training

在实践中，我们需要考虑真实医疗数据的特征，这些特征会给机器学习模型带来挑战。对于我们的数据集，主要问题是类别不平衡和家庭成员数据缺失。类别不平衡是因为我们预测的疾病患病率$<$ 10% 存在于所选队列中。我们通过使用类别加权损失函数、针对不平衡分类设计的采样策略、防止过拟合的方法，以及谨慎选择和解读评估指标来缓解这一问题。这些方法的具体细节列于附录A.1中。目标患者家庭成员的医疗记录可能会缺失，因为这些记录不在电子健康记录系统的覆盖范围内。在第4节中，我们将介绍如何处理缺失的家庭成员数据。

3. $\boldsymbol{A}$ 是无向图的邻接矩阵，$\hat{\boldsymbol{A}} = \boldsymbol{A} + \boldsymbol{I}$ 是 $\boldsymbol{A}$ （含自环），$\hat{\boldsymbol{D}}$ 是一个对角矩阵，其中包含每个节点的度

3.4. 可解释性方法

可以对图数据结构，特别是 GNN-LSTM 为每位家庭成员学习的监督式节点嵌入进行分析，以解释模型认为对于预测特定患者疾病风险很重要的亲属及这些亲属的医疗特征，如图1c所示。我们应用模型无关的、基于扰动的可解释性方法 GNNExplainer（Amara等人，2022年；Ying等人，2019年）来估计每个家族图$\boldsymbol{M}_{imp,\ i} \in [0,1]^{N_i}$的归一化节点重要性权重$\boldsymbol{X}_{imp,\ i} \in [0,1]^{N_i \times F'}$以及节点特征重要性权重[4]的软掩码。GNNEx- planer 作为训练后分析方法应用，可通过最大化互信息目标，联合学习节点和节点特征的重要性

$$\max_{\mathcal{G}_{i,S}, \boldsymbol{p}_i} MI(\boldsymbol{y}_i, (\mathcal{G}_{i,S}, \boldsymbol{p}_i)) = H(\boldsymbol{y}_i) - H(\boldsymbol{y}_i \mid \mathcal{G}_i = \mathcal{G}_{i,S}, \boldsymbol{X}_{i,\text{long}} = \boldsymbol{X}_{i,\text{long},S}), \tag{2}$$

其中，$\mathcal{G}_{i,\ S} \subset \mathcal{G}_i$是排除部分节点后的子图，$\boldsymbol{X}_{i,\ \text{long},\ S}$是掩码特征 $\boldsymbol{p}_i \in \{0,1\}^{F'}$对应的节点特征子集。互信息 $MI(\boldsymbol{y}_i,\ (\mathcal{G}_{i,\ S},\ \boldsymbol{p}_i))$量化了当节点和特征被限制为（$\mathcal{G}_{i,\ S}$，$\boldsymbol{p}_i$）时，$\hat{\boldsymbol{y}}_i$ 的预测概率变化，其基本思想是，预测概率发生较大变化表明被排除的节点和特征对预测很重要。对于纵向输入$\boldsymbol{X}_{i,\ \text{long}} \in \mathbb{R}^{N \times F \times T}$，特征掩码针对等价的二维输入$\boldsymbol{X}_{i,\ \text{long}} \in \mathbb{R}^{N \times F'}$进行定义，其中 $F' = F \times T$。节点特征重要性可以通过沿时间维度对$\boldsymbol{X}_{imp,\ i}$ 取平均获得。出于数据隐私方面的原因，我们无法在结果中报告个体层面的患者数据，因此，某一特定类型（父母、兄弟姐妹等）的所有亲属的全局特征重要性，是通过对该类型所有亲属的节点特征重要性取平均，并按节点重要性加权$\boldsymbol{M}_{imp,\ i}$来计算的。

4. 队列

4.1. 队列选择

实验使用 FinRegistry 的数据进行[5]，；FinRegistry 是一个汇总的全国性健康登记数据集。在总计 7,166,416 名可用个体中，我们排除了有移民记录、截至 2010 年 1 月 1 日年龄小于 18 岁，或因数据不完整而缺少性别或出生日期的个体。我们将剩余的 5,305,347 名个体定义为符合建图条件群体，P_{graph}，并将其用作大多数实验的主要队列。除此之外，对于每种选定的疾病，d，我们进一步定义一个目标队列，$P_{target,\ d}$，用于疾病预测任务，该队列由符合建图条件且满足以下标准的个体构成：（1）截至 2010 年 1 月 1 日，在指定年龄时仍然存活（通常为建议进行疾病筛查的年龄范围）；（2）父母双方也都是符合建图条件的个体；（3）截至 2010 年 1 月 1 日，没有与目标疾病相关的医疗记录；（4）不是目标队列中其他任何个体的父母。有关针对特定疾病的队列定义详情，请参见图3 和表5 。为每种疾病定义目标队列后，我们随后提取该队列中个体的符合建图条件的亲属，以构建

4.图可解释性方法通常以边重要性的形式进行表述，但对于本研究所研究的应用而言，计算节点重要性更为合理 5. https://www.finregistry.fi/

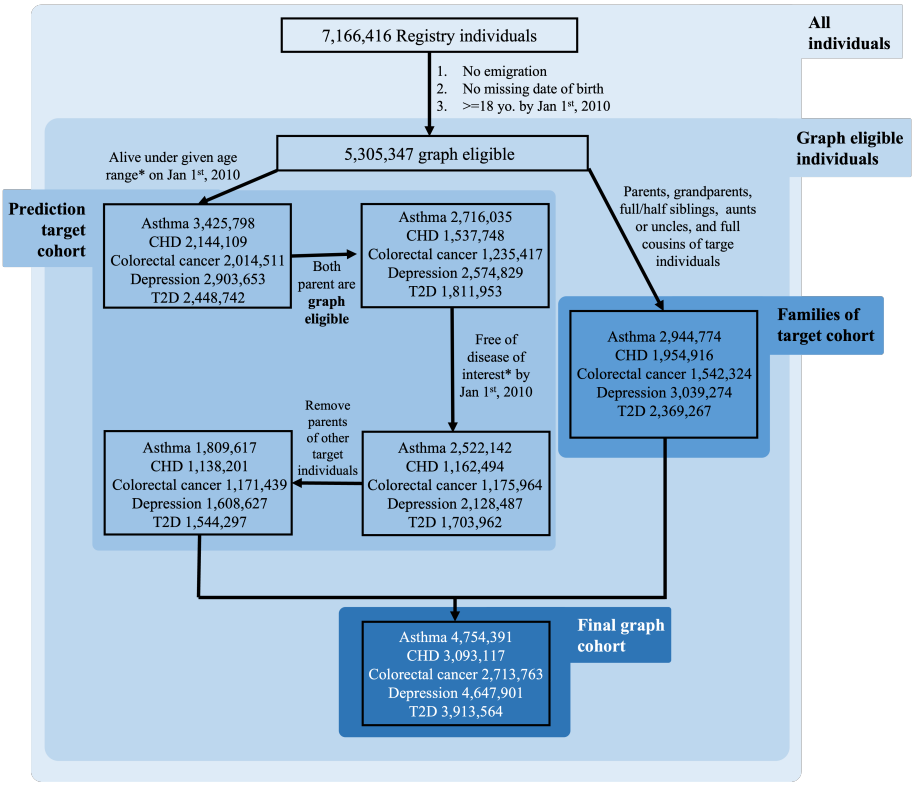


图3：各目标疾病的质量控制与纳入标准。另请参阅表 5 ，其中列出了带星号标注的疾病特异性年龄范围和对照组排除标准。

每位目标患者的家族图。每位目标个体的家庭成员包括其父母、祖父母、亲兄弟姐妹或同父异母／同母异父兄弟姐妹、叔伯姑舅姨以及堂表兄弟姐妹。

根据我们的纳入标准，每种疾病目标队列中的所有个体在其子图中均有父母双方的信息，因此任何目标个体对应的家族图最小规模为 $N_i \geq 3$。此外，截至2010年1月1日，所有目标个体均未患目标疾病。我们的机器学习任务是根据1990年1月1日至2010年1月1日期间观察到的电子健康记录数据，预测他们在2011年1月1日至2021年12月31日期间的疾病发病情况。因此，在预测时间窗内确诊目标疾病的目标患者被定义为病例，其余患者则被定义为对照。

4.2. 疾病与特征选择

我们选择五种常见疾病作为预测目标：成人哮喘、结直肠癌、冠心病、抑郁症和自杀，以及2型糖尿病。这些疾病之所以被选中，是因为它们具有以下特点：(1) 在人群中患病率较高；(2) 影响较大，且全球疾病负担（Roth，2018年）；(3) 遗传影响相对较大；(4) 遗传结构存在差异（Han 等人，2020年；Wray 等人，2018年；Nelson 等人，2017年；Mahajan 等人，2018年；Law 等人，2019年）。模型特征由年龄、性别和家族史的静态特征，以及纵向电子健康记录特征组成。家族史特征的构建方式为

作为上述亲属的类别特征，时间截至2010年1月1日。例如，目标个体表亲的疾病史为：若其任一表亲在2010年1月1日之前被诊断出患有该疾病，则为1，否则为0。如上所述，所有目标个体的疾病史都包含父母双方的信息。对于其他符合条件的亲属，我们使用额外的指示变量对其是否可用进行编码，因为家庭成员缺失本身可能具有信息价值；也就是说，除了表亲的疾病史外，还会增加一个特征：对于至少有一位表亲的目标个体，该特征为1，否则为0。对于纵向电子健康记录特征，在数据集中3000个疾病终点中[6]，我们选择发生次数达到$> 0.1\% \times$ 目标队列样本量的终点，从而得到 ~ 1500 个特征。我们仅考虑1990年1月1日至2010年1月1日期间的医疗记录，以避免目标队列与目标个体家庭成员之间的重叠导致标签泄漏。终点数据按年汇总为二元特征，用于表示该终点是否在该年被报告。

4.3. 合成数据集

由于我们使用的是私有数据集，为了使结果可复现，我们还创建并发布了合成的家系和表型数据。使用责任阈值模型（Neale，2005年；Holst 等人，2016年），我们假设个体的疾病易感性 L 由以下几个部分构成：

$$L = L_{Herr} + L_{Age} + L_{Sex} + \epsilon,$$

其中 L_{Herr} 从广义上讲，是涵盖任何类型家族影响的可遗传风险，包括可从父母处获得的遗传力和可遗传环境风险；$L_{Age} = \beta_{age} \cdot Age$ 以及 $L_{Sex} = \beta_{sex} \cdot Gender$ 是个体年龄和性别带来的影响，这些影响可以推广到任何固定效应因素，而 ϵ 是表型上的个体特异性噪声。具有超过某一截断阈值的易感性的个体会被赋予疾病标签。合成数据通过从一个具有随机分配的可遗传风险的祖源群体开始，并假设随机交配以生成下一代来创建。重复此过程，直到达到目标代数。有关我们的合成数据算法的完整细节，请参阅附录 B 。

5. 实验

我们使用真实患者数据进行预测性能、消融研究和可解释性实验，并使用合成数据集进一步开展消融研究和分析。预测性能对比实验采用附录中所述的训练设置 A，并报告测试数据集的指标：受试者工作特征曲线下面积、精确率-召回率曲线下面积和马修斯相关系数（马修斯相关系数），这些指标是针对类别不平衡的医学数据集所推荐的（Hicks 等人，2022年；Saito 和 Rehmsmeier，2015年）。对于真实数据实验，我们使用 MC 随机失活方法（Gal 和 Ghahramani，2016年）计算不确定性估计，并报告 3 个样本的均值和 95 % 置信区间；对于合成数据实验，则对合成数据集的 5 个随机样本的结果取平均。

6. 详情请见 https://risteys.finregistry.fi/

5.1. 与临床基线方法的比较

基于图的方法与两种临床基线方法进行比较，用于预测疾病发病 $\boldsymbol{y}_i$，这些方法近似于在临床环境中收集家族史的一般建议（美国医学会、2004年；Walker 等人、1990年；Shaukat 等人、2021年；检测专家组 等人，2001年；Handelsman 等人，2015年）：

- **基于规则的临床基线（B1）**：基于规则的方法，如果至少一位（最高至三级）亲属有该疾病史，则将患者标记为高风险（$\hat{\boldsymbol{y}}_i = 1$）；

- **基于多层感知机的临床基线（B2）**：采用 2 个线性层的多层感知机（MLP），并使用年龄、性别和家族史静态特征，定义见第4.2节。

基于图的模型无法与不使用纵向数据的基线方法直接比较，因此我们还在实验中纳入了一个不含纵向特征的简化模型版本。我们将这个非 LSTM 变体简称为**图神经网络（G1）**，并将第3.3 节所述的完整模型称为**GNN-LSTM（G2）**。表1报告了这 4 种模型在 5 种疾病表型上的结果。我们发现，基于图的方法始终优于基线方法。图神经网络（G1）的表现优于针对家族史的相应多层感知机临床基线（B2），这表明仅使用图神经网络模型本身就能提升预测性能。GNN-LSTM 模型还纳入了各家庭成员的详细病史，在除结直肠癌之外的所有疾病上均表现最佳。由于已知癌症在我们研究的疾病中遗传性最低，我们推测，使用图神经网络对家族信息建模能为遗传性更强的疾病带来最大益处，并在第5.3节中利用合成数据进一步研究这一问题。

Table 1: Comparison of predictive performance for GNN and GNN-LSTM models with clinical baselines. Reports the average value for each metric and 95% CI of the uncertainty estimates in brackets (except for the rule-based approach), with best results indicated in bold.

Endpoint	**Model**	**AUC-ROC**	**AUC-PRC**	**MCC**
Coronary heart disease	B1: Clinical baseline binary	0.573	0.039	0.052
	B2: Clinical baseline MLP	0.710 (0.7097 - 0.71031)	0.073 (0.0721 - 0.0739)	0.114 (0.1071 - 0.1209)
	G1: GNN	0.720 (0.7198 - 0.7202)	0.079 (0.0789 - 0.0791)	0.125 (0.0590 - 0.191)
	G2: GNN-LSTM	**0.775 (0.7747 - 0.7754)**	**0.126 (0.1249 - 0.1271)**	**0.169 (0.0800 - 0.2580)**
Type 2 diabetes	B1: Clinical baseline binary	0.582	0.075	0.085
	B2: Clinical baseline MLP	0.662 (0.6615 - 0.6625)	0.111 (0.1108 - 0.1112)	0.114 (0.0530 - 0.1750)
	G1: GNN	0.673 (0.6727 - 0.6733)	0.117 (0.1168 - 0.1172)	0.122 (0.0590 - 0.1850)
	G2: GNN-LSTM	**0.700 (0.6994 - 0.7006)**	**0.127 (0.1263 - 0.1277)**	**0.141 (0.0700 - 0.2120)**
Depression	B1: Clinical baseline binary	0.557	0.072	0.056
	B2: Clinical baseline MLP	0.639 (0.6387 - 0.6393)	0.105 (0.1043 - 0.1057)	0.102 (0.0450 - 0.1590)
	G1: GNN	0.642 (0.6419 - 0.6421)	0.112 (0.1119 - 0.1121)	0.111 (0.0550 - 0.1670)
	G2: GNN-LSTM	**0.655 (0.6548 - 0.6552)**	**0.119 (0.1183 - 0.1197)**	**0.116 (0.0570 - 0.1750)**
Asthma	B1: Clinical baseline binary	0.542	0.013	0.019
	B2: Clinical baseline MLP	0.635 (0.6347 - 0.6352)	0.020 (0.0195 - 0.0205)	0.031 (0.0140 - 0.0480)
	G1: GNN	0.647 (0.6467 - 0.6473)	0.021 (0.0204 - 0.0216)	0.039 (0.0210 - 0.0570)
	G2: GNN-LSTM	**0.666 (0.6657 - 0.6661)**	**0.023 (0.0228 - 0.0232)**	**0.042 (0.0418 - 0.0422)**
Colorectal cancer	B1: Clinical baseline binary	0.524	0.009	0.019
	B2: Clinical baseline MLP	0.650 (0.6497 - 0.6503)	0.015 (0.0146 - 0.0154)	0.032 (0.0150 - 0.0490)
	G1: GNN	**0.654 (0.6534 - 0.6546)**	**0.016 (0.0160 - 0.0160)**	**0.035 (0.0170 - 0.0530)**
	G2: GNN-LSTM	0.653 (0.6522 - 0.6538)	**0.016 (0.0158 - 0.0162)**	0.031 (0.0200 - 0.0420)

5.2. 消融研究

GNN-LSTM 的结果可能对纵向电子健康记录数据机器学习方法的开发具有更广泛的参考价值，因此我们进一步通过消融研究检验这种方法。我们以一个包含 2 层**年龄和性别特征多层感知机（A1）**的基线为起点，比较以下模型：

- **年龄、性别和家族史多层感知机（A2，与 B2 相同）**：添加家族史特征，定义见第 4.2 节；
- **年龄、性别和图连通性多层感知机（A3）**：添加拓扑信息，使用一个 8 维 Node2Vec 嵌入（Grover 和 Leskovec，2016年），该嵌入是根据整个数据集的连通家族网络为每位患者生成的；
- **年龄、性别和纵向电子健康记录数据长短期记忆网络（A4）**：添加纵向电子健康记录数据（针对目标患者），采用单层双向长短期记忆网络架构，即与 GNN-LSTM 模型的目标特异性组件类似。神经网络架构的选择基于一篇近期综述，该综述指出，单层长短期记忆网络常用于患者深度时间序列预测，且双向变体的表现始终优于单向变体（Morid 等人，2023年）；
- **年龄、性别、家族史和纵向电子健康记录数据LSTM（A5）**LSTM模型（A4），但还将家族史作为静态特征纳入其中。将家族史特征添加到LSTM中，与使用GNN-LSTM对此进行建模有所不同，因此，该模型有助于我们了解使用图神经网络的益处（如果有的话）。

与 **GNN-LSTM（AG，与 G2 相同）** 相比，这些模型的结果见表2。与年龄和性别基线（A1）相比，性能提升主要来自加入家族史（A2）和纵向电子健康记录数据（A4），而非单独的拓扑信息（A3）。当使用 GNN-LSTM 方法（AG）将所有这些信息整合到单个模型中时，其整体表现通常优于各个单独的方法（A1-5），尤其是在冠心病和2型糖尿病等高度多因素疾病上。GNN-LSTM（AG）的表现优于包含家族史静态特征的长短期记忆网络（A5），这表明图更适合作为家族信息的模型。然而，对于哮喘和结直肠癌，不包含家族史的长短期记忆网络（A4）获得了比 GNN-LSTM（AG）更高的马修斯相关系数，这可能表明对于某些疾病，仅基于遗传亲缘关系的几何先验并不足够（例如，由于非遗传效应更强）。由于难以区分真实电子健康记录数据中的遗传效应和环境效应，我们在第5.3节使用合成数据单独研究遗传力的影响。

Table 2: Ablation studies examining how different aspects modeled by the GNN-LSTM contribute to the performance: covariates age and sex (A1), family history (A2), family relatedness (A3), and detailed medical histories (A4). A GNN-LSTM (AG) is a graph-based model for this information, which is a conceptually different approach to adding features for family history to an LSTM model (A5). Reports the average value for each metric and 95% CI of the uncertainty estimates in brackets, with best results indicated in bold.

Endpoint	**Model**	**AUC-ROC**	**AUC-PRC**	**MCC**
Coronary heart disease	A1: Age and sex (MLP)	0.696 (0.6956 - 0.6964)	0.066 (0.0656 - 0.0665)	0.093 (0.0820 - 0.1040)
	A2: Age, sex, family history (MLP)	0.710 (0.7097 - 0.7103)	0.073 (0.0721 - 0.0739)	0.114 (0.1071 - 0.1209)
	A3: Age, sex, graph connectivity (MLP)	0.696 (0.6956 - 0.6964)	0.067 (0.0669 - 0.0671)	0.091 (0.0818 - 0.1002)
	A4: Age, sex, EHR (LSTM)	0.763 (0.7625 - 0.7635)	0.105 (0.1044 - 0.1056)	0.148 (0.1250 - 0.1710)
	A5: Age, sex, EHR, family history (LSTM)	0.771 (0.7706 - 0.7714)	0.123 (0.1228 - 0.1232)	0.164 (0.1580 - 0.1700)
	AG: GNN-LSTM	**0.775 (0.7746 - 0.7754)**	**0.126 (0.1249 - 0.1271)**	**0.169 (0.0800 - 0.2580)**
Type 2 diabetes	A1: Age and sex (MLP)	0.617 (0.6166 - 0.6174)	0.085 (0.0848 - 0.0853)	0.072 (0.0330 - 0.1110)
	A2: Age, sex, family history (MLP)	0.662 (0.6615 - 0.6625)	0.111 (0.1108 - 0.1112)	0.114 (0.0530 - 0.1750)
	A3: Age, sex, graph connectivity (MLP)	0.619 (0.6187 - 0.6193)	0.087 (0.0870 - 0.0870)	0.073 (0.0340 - 0.1120)
	A4: Age, sex, EHR (LSTM)	0.675 (0.6745 - 0.6755)	0.110 (0.1091 - 0.1109)	0.111 (0.0520 - 0.1700)
	A5: Age, sex, EHR, family history (LSTM)	0.689 (0.6884 - 0.6896)	0.121 (0.1200 - 0.1220)	0.127 (0.0600 - 0.1940)
	AG: GNN-LSTM	**0.700 (0.6994 - 0.7006)**	**0.127 (0.1263 - 0.1277)**	**0.141 (0.0700 - 0.2120)**
Depression	A1: Age and sex (MLP)	0.632 (0.6318 - 0.6323)	0.102 (0.1018 - 0.1022)	0.097 (0.0450 - 0.1490)
	A2: Age, sex, family history (MLP)	0.639 (0.6387 - 0.6393)	0.105 (0.1043 - 0.1057)	0.102 (0.0450 - 0.1590)
	A3: Age, sex, graph connectivity (MLP)	0.633 (0.6328 - 0.6332)	0.102 (0.1017 - 0.1023)	0.098 (0.0460 - 0.1500)
	A4: Age, sex, EHR (LSTM)	0.647 (0.6467 - 0.6473)	0.113 (0.1121 - 0.1139)	0.108 (0.0510 - 0.1650)
	A5: Age, sex, EHR, family history (LSTM)	0.651 (0.6508 - 0.6512)	**0.119 (0.1183 - 0.1197)**	0.111 (0.0520 - 0.1700)
	AG: GNN-LSTM	**0.655 (0.6548 - 0.6552)**	**0.119 (0.1183 - 0.1197)**	**0.116 (0.0570 - 0.1750)**
Asthma	A1: Age and sex (MLP)	0.625 (0.6249 - 0.6251)	0.018 (0.0179 - 0.0181)	0.034 (0.0160 - 0.0520)
	A2: Age, sex, family history (MLP)	0.635 (0.6348 - 0.6352)	0.020 (0.0195 - 0.0205)	0.031 (0.0140 - 0.0480)
	A3: Age, sex, graph connectivity (MLP)	0.622 (0.6217 - 0.6223)	0.017 (0.0168 - 0.0172)	0.034 (0.0160 - 0.0520)
	A4: Age, sex, EHR (LSTM)	0.653 (0.6526 - 0.6534)	0.021 (0.0209 - 0.0211)	**0.043 (0.0200 - 0.0660)**
	A5: Age, sex, EHR, family history (LSTM)	0.664 (0.6636 - 0.6645)	0.022 (0.0219 - 0.0222)	0.037 (0.0170 - 0.0570)
	AG: GNN-LSTM	**0.666 (0.6659 - 0.6661)**	**0.023 (0.0228 - 0.0232)**	0.042 (0.0418 - 0.0422)
Colorectal cancer	A1: Age and sex (MLP)	0.651 (0.6505 - 0.6515)	0.015 (0.0149 - 0.0151)	0.026 (0.0120 - 0.0400)
	A2: Age, sex, family history (MLP)	0.650 (0.6497 - 0.6503)	0.015 (0.0146 - 0.0154)	0.032 (0.0150 - 0.0490)
	A3: Age, sex, graph connectivity (MLP)	0.645 (0.6447 - 0.6453)	0.015 (0.0150 - 0.0150)	0.034 (0.0160 - 0.0520)
	A4: Age, sex, EHR (LSTM)	0.652 (0.6519 - 0.6521)	0.015 (0.0149 - 0.0152)	**0.035 (0.0160 - 0.0540)**
	A5: Age, sex, EHR, family history (LSTM)	0.646 (0.6456 - 0.6464)	0.015 (0.0149 - 0.0151)	0.027 (0.0130 - 0.0410)
	AG: GNN-LSTM	**0.653 (0.6522 - 0.6538)**	**0.016 (0.0158 - 0.0162)**	0.031 (0.0200 - 0.0420)

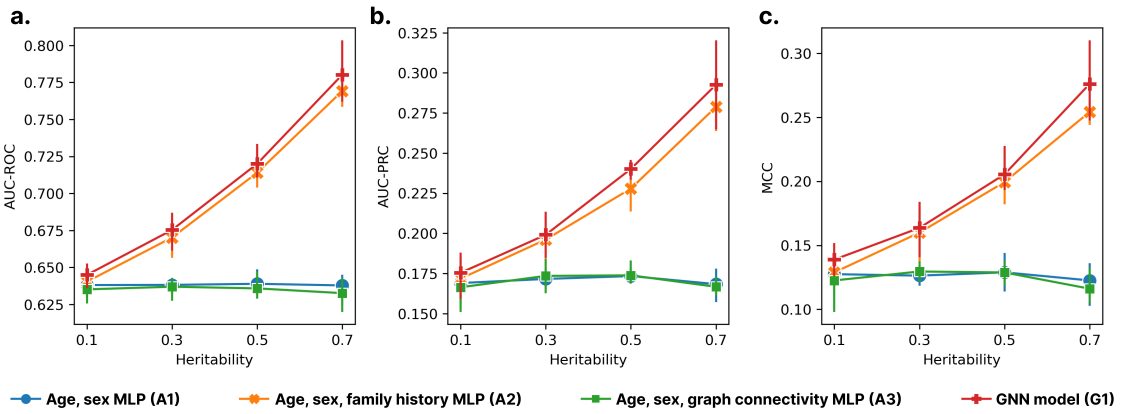


图4：使用以不同程度遗传力模拟的表型对合成数据进行的消融研究， h。报告了每个数据集的5次模拟的平均值：(a) 受试者工作特征曲线下面积，(b) 精确率-召回率曲线下面积和 (c) 马修斯相关系数 $h = 0.1$,0.3,0.5,0.7，并以误差条绘制。

5.3. 合成数据实验

我们使用第 4.3 节所述算法生成的合成数据集，考察遗传力程度不同的表型， h。消融研究的结果（图 4）表明，对于各种模拟表型 $h = 0.1$，0。3，0。5，0。7，图神经网络方法（G1）始终优于基线方法，并且对于遗传力更高的表型，这种性能提升更加明显。收益主要来自加入家族信息（A2），而使用图神经网络（G1）将其建模为几何先验还能进一步提升性能；这一结果与真实数据中的观察结果（表 2）一致。我们还使用合成数据考察图神经网络模型对各种设计选择的敏感性（图 6）。结果表明，所考察的设计选择对受试者工作特征曲线下面积的影响差异不大。构建家谱的标准方法使用父母/子女边，但我们发现，对于图神经网络，若同时添加从每个节点直接指向目标节点的边，或添加连接所有遗传相关节点的边，性能会更好。受试者工作特征曲线下面积平均分别提升了 0.22% 和 0.28%，但具体提升幅度因图神经网络架构的选择而异。总体而言，对于连接所有遗传相关节点的家族图，采用 k-图神经网络层（Morris 等人，2019年）和目标节点表示（不进行池化）时，受试者工作特征曲线下面积最高。

5.4. 可解释性分析

我们以冠心病（CHD）的 GNN-LSTM 模型为示例，使用第 3.4 节中描述的图可解释性方法，分析模型学习到的图表示。由于整个电子健康记录特征集都没有每位亲属的重要特征真实值，我们将 GNN-LSTM 方法与逻辑回归基线进行比较，后者是流行病学研究中常用的方法。

我们重新训练图神经网络[7]，使用前n 个电子健康记录特征，这些特征通过以下方法为父母识别：(1)GNN-LSTM 可解释性分析，以及 (2) 逻辑回归基线，用于 $n = 20$、50、100。对于GNN-LSTM模型识别出的特征，我们观察到更高的精确率-召回率曲线下面积（平均提升19.5%）（图5a），这表明，基于图的方法更擅长识别对预测子女冠心病风险有用的父母特征。

作为案例研究，我们进一步考察了 GNN-LSTM 为冠心病患者的父代节点学习到的节点嵌入。我们使用 t-SNE 算法对这些嵌入进行可视化，并使用 K均值聚类识别相似父代的群（图 5b.i），图中还展示了每个簇对应的特征重要性5c。我们观察到，父代的既有冠心病风险因素、合并症和治疗方式（例如降压药物、他汀类药物、2型糖尿病及各种心脏病（Escobar，2002年；Goodarzi and Rotter，2020年；Megaet al.，2015年）是其子代 10 年冠心病结局的重要预测因素。图 5b.ii 显示，嵌入能够根据父代的冠心病病史，区分其预测影响；与父代患有冠心病相关的程度因簇而异。然而，尽管这种可解释性分析有助于识别 GNN-LSTM 用于预测疾病的特征，但我们注意到，它并不能直接对某一特征对疾病的影响提供因果解释。

6. 讨论与结论

在本研究中，我们开发了一种新的几何深度学习方法，用于对家族内部高维、纵向医疗史的关系结构进行建模，并学习患者疾病风险的可解释监督式图表示。我们将该方法与用于家族史的（统计）临床基线方法以及用于电子健康记录数据的深度学习基线方法进行了比较，结果表明，在预测患者十年疾病发病（针对5种复杂疾病表型）时，基于图的方法取得了更高的受试者工作特征曲线下面积、精确率-召回率曲线下面积和马修斯相关系数（表1和2）。针对合成数据的消融研究表明，对于遗传性更强的疾病，基于图神经网络的方法在预测性能方面带来了最大的提升（图4）；针对真实数据的消融研究则表明，图神经网络本身促成了这些提升，即与用于家族史的等效非图多层感知机和长短期记忆网络基线方法相比（表1 和2）。这些结果表明，家族史所具有的底层几何结构适合使用基于图的方法进行建模。此外，通过使用可解释性技术，我们展示了基于图的方法如何通过识别对模型预测患者疾病风险最为重要的个体亲属和特征，以个性化的方式刻画家族信息的影响。我们观察到，与使用逻辑回归基线所识别的特征相比，使用GNN-LSTM模型所识别的重要特征重新训练后，图神经网络的精确率-召回率曲线下面积平均提高了19.5%。

6.1. 局限性

我们分析的局限性包括缺乏与解读家族史的疾病特异性方法进行比较，因为我们的目标是提供一项范围更广的疾病无关研究。

7. 对于 GNN 模型，纵向电子健康记录特征汇总为：如果在观察期内至少记录过一次该诊断，则为 1；否则为 0

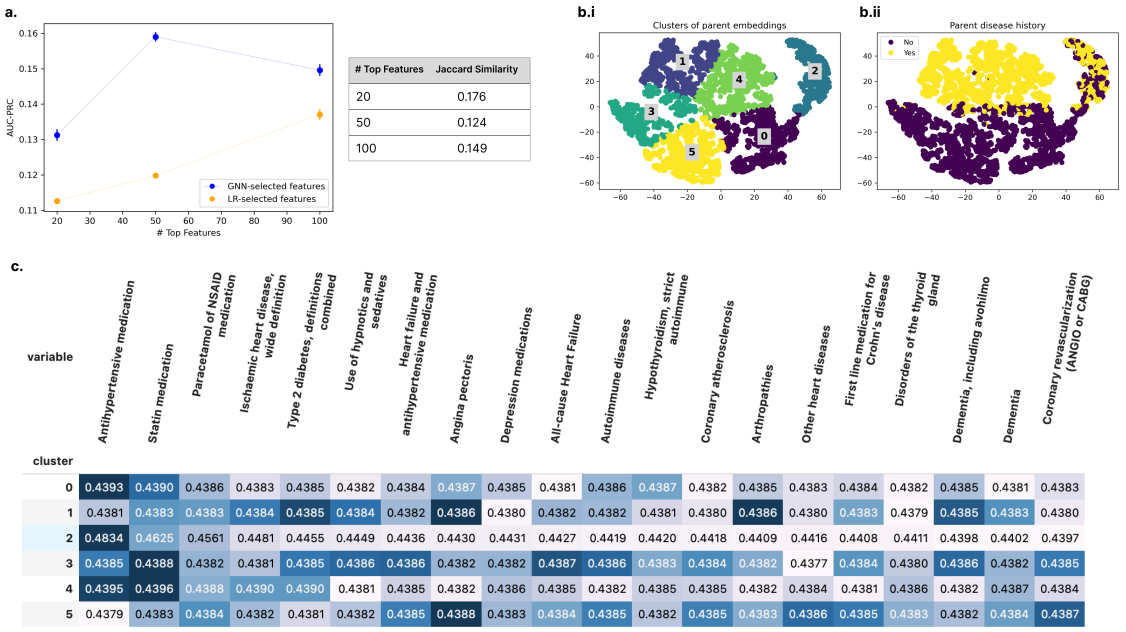


图 5：针对父节点嵌入的 GNN-LSTM 解释，冠心病终点。**a)** 绘制了使用（1）GNN-LSTM 和（2）逻辑回归（LR）基线所识别的前-n 个特征重新训练的图神经网络的 AUC-PRC。同时报告了 GNN 与 LR 所选特征之间的杰卡德相似度；**b.i)** 使用 t-SNE 和 K均值聚类对父节点的图表示进行可视化；**b.ii)** 这些表示清楚地表明，家族史是决定儿童风险的一个因素；**c)** 模型总体识别出的前20个特征在各簇中的特征重要性。

方法。我们承认，任何新的机器学习方法都需要进一步开发，才能在临床环境中投入使用，而且伦理和法律问题可能会影响家庭电子健康记录数据的可用性。如果无法访问家庭电子健康记录，我们建议先使用大规模电子健康记录数据集训练模型，再通过图可解释性分析识别重要特征。在临床环境中，如果能够获取家庭电子健康记录数据，模型便可提供个性化推断；否则，可使用训练数据中的人群平均值，为与患者讨论家族史（涵盖三级亲属）时应关注哪些方面提供一般性指导。

我们的工作重点是考察几何方法是否有助于对家族史进行建模。由于我们的研究是评估几何方法是否适用于家族史的第一步，因此我们将分析限制在现有的图神经网络架构上。未来的研究可以针对这一机器学习问题开发专门的图神经网络层。异质性疾病的类别不平衡和类内变异带来了机器学习方面的挑战，我们使用附录 A.1中所述的技术缓解了这些挑战。我们还采取了措施，以减少队列设计和机器学习中潜在的混杂因素

问题设置，但这意味着真实数据的某些方面并未在我们的分析中得到明确建模，例如缺失家族信息的不确定性，以及家族图中的子节点。我们还注意到，尽管我们的可解释性分析识别出了对预测有用的节点和特征，但这并不能直接转化为因果解释（例如，重要性可能源于相关性）。

6.2. 结论

总体而言，我们认为，本研究开发的机器学习方法为临床应用、流行病学研究和精准医学应用带来了新的机遇。我们展示了基于图的方法如何提供信息量更丰富的表示，以结合家族史预测疾病风险。GNN-LSTM 将家庭电子健康记录建模为连通网络，能够学习针对各位亲属的可解释节点嵌入，并且在复杂疾病的预测性能方面优于受临床启发的基线方法和深度学习基线方法。未来沿这一方向开展的工作将有助于推动个性化家族史评估等临床应用，以及研究亲属关系如何影响健康与疾病的流行病学研究。

致谢

我们感谢 FinRegistry 团队为本研究提供数据，并感谢 CSC – IT科学中心（芬兰）提供计算资源。本研究获得欧盟地平线2020研究与创新计划资助，资助协议编号为101016775。FinRegistry 是芬兰健康与福利研究所（THL）与芬兰分子医学研究所（FIMM）、赫尔辛基大学数据科学遗传流行病学研究组的合作项目。FinRegistry 项目已获得国家健康与福利研究所（THL/1776/6.02.00/2019 及后续修订）、DVV（VRK/5722/2019-2）、芬兰养老金中心（ETK/SUTI 22003）和芬兰统计局（TK-53-1451-19）对数据访问的批准。FinRegistry 项目已获得国家健康与福利研究所的机构审查委员会批准（会议 7/2019）。

参考文献

美国医学会。疾病预防中的家族病史。宣传册，2004年。
URLhttps://www.ama-assn.org/sites/ama-assn.org/files/corp/media-browser/public/genetics/family_病史02_0.pdf。

Kenza Amara、Rex Ying、Zitao Zhang、Zhihao Han、Yinan Shan、Ulrik Brandes、Sebastian Schemm 和 Ce Zhang。Graphframex：迈向图神经网络可解释性方法的系统评估。*arXiv* 预印本 *arXiv:2206.09677*，2022年。

Hamid Behravan、Jaana MHartikainen、Maria Tengstr¨om、Veli-MattiKosma 和Arto Mannermaa。利用相互作用的遗传因素和人口统计学因素以及机器学习预测乳腺癌风险。《科学报告》，10(1):11044，2020年。

Michael M Bronstein、Joan Bruna、TacoCohen和PetarVeli˘ckovi´c。几何深度学习：网格、群、图、测地线和规范场。*arXiv*预印本 *arXiv:2104.13478*，2021年。DeponkerSarker Depto、Md Mashfiq Rizvee、AimonRahman、Hasib Zunair、M SohelRah-man 和 MRC Mahdy。白血病检测的类别不平衡分类方法量化。《计算机在生物学与医学中的应用》，152:106372，2023年。EEscobar。高血压与冠心病。《人类高血压杂志》，16(1):S61–S63，2002年。检测、评估及治疗成年人高血胆固醇问题专家组等。国家胆固醇教育计划（NCEP）检测、评估及治疗成年人高血胆固醇问题专家组（成人治疗专家组第三版）第三次报告执行摘要。《美国医学会杂志》，285(19):2486–2497，2001年。W Gregory Feero、MaryBeth Bigley和 KristinMBrinner。电子健康记录中家族健康史信息的新标准和增强效用：来自美国健康信息社区家族健康史多利益相关方工作组的更新。《美国医学信息学会杂志》，15(6): 723–728，2008年。Matthias Fey 和 J an Eric Lenssen。使用 PyTorch Geometric进行快速图表示学习。*arXiv*预印本 *arXiv:1903.02428*，2019年。YarinGal和 Zoubin Ghahramani。将随机失活作为贝叶斯近似：表示深度学习中的模型不确定性。在国际机器学习会议，第 1050–1059 页。PMLR，2016年。Geoffrey S Ginsburg、RRyanneWu和 Lori A Orlando。家族健康史：未充分用于可操作的风险评估。《柳叶刀》，394(10198):596–603，2019年。MarkOGoodarzi和Jerome I Rotter。2型糖尿病与冠心病关系中的遗传学见解。《循环研究》，126(11):1526–1548，2020年。AdityaGrover和Jure Leskovec。Node2Vec：面向网络的可扩展特征学习。在第*22*届*ACMSIGKDD*知识发现与数据挖掘国际会议论文集，第 855–864 页，2016年。Yi Han、Qiong Jia、Pedram Shafiei Jahani、Benjamin P Hurrell、Calvin Pan、Pin Huang、Janet Gukasyan、Nicholas C Woodward、Eleazar Eskin、FrankD Gilliland 等。全基因组分析凸显免疫系统通路对哮喘遗传结构的贡献。《自然—通讯》，11(1):1776，2020年。Yehuda Handelsman、ZacharyT Bloomgarden、George Grunberger、Guillermo Umpierrez、Robert S Zimmerman、Timothy S Bailey、Lawrence Blonde、George A Bray、A JayCohen、SamuelDagogo-Jack 等。美国临床内分泌医师协会和美国内分泌学会关于制定糖尿病综合护理计划的临床实践指南——2015年——执行摘要。《内分泌实践》，21(4):413–437，2015年。

Steven A Hicks, Inga Str¨umke,Vajira Thambawita,MalekHammou, Michael A Riegler, P˚al Halvorsen, 和 Sravanthi Parasa。《人工智能医学应用的评估指标》。《科学报告》，12(1):5979，2022年。Klaus K Holst, ThomasHScheike,和 Jacob B Hjelmborg。删失双胞胎数据的责任阈值模型。《计算统计学与数据分析》，93:324–335，2016年。Sivan Kinreich, Jacquelyn L Meyers, Adi Maron-Katz, Chella Kamarajan, Ashwini K Pandey, David B Chorlian, Jian Zhang, Gayathri Pandey, Stacey Subbie-SaenzdeViteri, DanPitti 等。使用具有多模态生物标志物和家族史的纵向数据预测酒精使用障碍风险：一项机器学习研究。《分子精神病学》，26(4):1133–1141，2021年。Thomas N Kipf 和 Max Welling。使用图卷积网络进行半监督分类。*arXiv* 预印本*arXiv:1609.02907*，2016年。Bartosz Krawczyk和Micha˙l Wo´zniak。用于类别不平衡分类的基于 ROC 的移动阈值成本敏感神经网络。在智能数据工程与自动化学习——*IDEAL 2015*年：第*16*届国际会议，波兰弗罗茨瓦夫，*2015*年*10*月*14*日至*16*日，会议论文集*16*，第45–52页。施普普林格，2015年。

Philip J Law、Maria Timofeeva、Ceres Fernandez-Rozadilla、Peter Broderick、James Studd、Juan Fernandez-Tajes、Susan Farrington、Victoria Svinti、Claire Palles、Giulia Orlando等。关联分析确定了31个与结直肠癌易感性相关的新风险位点。《自然—通讯》，10(1):2154，2019年。

Junhyun Lee、Inyeop Lee 和 Jaewoo Kang。自注意力图池化。载于国际机器学习会议，第 3734–3743 页。PMLR，2019年。Cathryn M Lewis 和 Evangelos Vassos。多基因风险评分：从研究工具到临床工具。《基因组医学》，12(1):1–11，2020年。Michelle M Li、Kexin Huang 和 Marinka Zitnik。生物医学与医疗保健中的图表示学习。《自然·生物医学工程》，第 1–17 页，2022年。Ruowang Li、Yong Chen、Marylyn D Ritchie 和 Jason H Moore。利用电子健康记录和多基因风险评分预测疾病风险。《自然 · 遗传学评论》，21(8):493–502，2020年。Chuang Liu、Yibing Zhan、Chang Li、Bo Du、Jia Wu、Wenbin Hu、Tongliang Liu 和 Dacheng Tao。图神经网络的图池化：进展、挑战与机遇。*arXiv*预印本 *arXiv:2204.07321*，2022年。Anubha Mahajan、Daniel Taliun、Matthias Thurner、Neil R Robertson、Jason M Torres、N William Rayner、Anthony J Payne、Valgerdur Steinthorsdottir、Robert A Scott、Niels Grarup 等。利用高密度插补和胰岛特异性表观基因组图谱，将2型糖尿病位点精细定位至单变异分辨率。《自然·遗传学》，50(11):1505–1513，2018年。

Chengsheng Mao、Liang Yao 和 Yuan Luo。Medgcn：通过图卷积网络实现药物推荐和实验室检测插补。《生物医学信息学杂志》，127:104000，2022年。Jessica L Mega、Nathan O Stitziel、J Gustav Smith、Daniel I Chasman、Mark J Caulfield、James J Devlin、Francesco Nordio、Craig L Hyde、Christopher P Cannon、Frank M Sacks 等。遗传风险、冠心病事件与他汀治疗的临床获益：对一级和二级预防试验的分析。《柳叶刀》，385(9984):2264–2271，2015年。Mohammad Amin Morid、Olivia R. Liu Sheng 和 Joseph Dunbar。医疗保健领域中采用深度学习方法进行时间序列预测。《*ACM*管理信息系统汇刊》，14(1)，2023年1月。ISSN 2158-656X。doi: 10.1145/3531326。URLhttps://doi.org/10.1145/3531326。Christopher Morris、Martin Ritzert、Matthias Fey、William L Hamilton、Jan Eric Lenssen、Gaurav Rattan 和 Martin Grohe。魏斯费勒-莱曼走向神经网络：高阶图神经网络。载于*AAAI*人工智能大会论文集，第33卷，第 4602–4609 页，2019年。Ben Neale。易感性阈值模型。《行为科学统计学百科全书》，2005年。Christopher P Nelson、Anuj Goel、Adam S Butterworth、Stavroula Kanoni、Tom R Webb、Eirini Marouli、Lingyao Zeng、Ioanna Ntalla、Florence Y Lai、Jemma C Hopewell 等。基于错误发现率的关联分析揭示了冠状动脉疾病的新位点。《自然·遗传学》，49(9):1385–1391，2017年。Sarah Parisot、Sofia Ira Ktena、Enzo Ferrante、Matthew Lee、Ricardo Guerrero、Ben Glocker 和 Daniel Rueckert。利用图卷积网络进行疾病预测：应用于孤独症谱系障碍和阿尔茨海默病。《医学图像分析》，48:117–130，2018年。G Roth。全球疾病负担协作网络。2017年全球疾病负担研究（GBD 2017）结果。美国西雅图：健康指标与评估研究所（IHME），2018年。《柳叶刀》，392:1736–88，2018年。Takaya Saito 和 Marc Rehmsmeier。在不平衡数据集上评估二分类器时，精确率-召回率曲线比ROC曲线更具信息量。《*PLOS ONE*》，10(3):e0118432，2015年。Aasma Shaukat、Charles J Kahi、Carol A Burke、Linda Rabeneck、Bryan G Sauer 和 Douglas K Rex。ACG临床指南：2021年结直肠癌筛查。美国胃肠病学会官方期刊——*ACG*，116(3):458–479，2021年。Catherine Tong、Emma Rocheteau、Petar Veli˘ckovi´c、Nicholas Lane 和 Pietro Li`o。利用图表示学习预测患者结局。载于疾病监测与大流行病智能：智能疾病检测实践，第 281–293 页。施普林格，2022年。

Rodolfo Valdez、Paula W Yoon、Nadeem Qureshi、Ridgely Fisk Green 和 Muin J Khoury。公共卫生实践中的家族史：用于疾病预防和健康促进的基因组工具。公共卫生年度评论，31:69–87，2010年。

Petar Veli˘ckovi´c、Guillem Cucurull、Arantxa Casanova、Adriana Romero、Pietro Lio 和 Yoshua Bengio。图注意力网络。*arXiv* 预印本 *arXiv:1710.10903*，2017年。

H Kenneth Walker、W Dallas Hall 和 J Willis Hurst。临床方法：病史、体格检查和实验室检查。1990年。

Jiabo Wang、Zhengkui Zhou、Zhe Zhang、Hui Li、Di Liu、Qin Zhang、Peter J Bradbury、EdwardS Buckler 和 ZhiwuZhang。扩展蓝色字母表，以进行适应复杂性状遗传结构的基因组预测。遗传，121(6):648–662，2018年。

Brenda J Wilson、Nadeem Qureshi、Pasqualina Santaguida、Julian Little、June C Carroll、JudithAllanson和 Parminder Raina。《系统综述：常见疾病风险评估中的家族史》。《内科医学年鉴》，151(12):878–885，2009。

Naomi R Wray、Stephan Ripke、Manuel Mattheisen、Maciej Trzaskowski、Enda M Byrne、Abdel Abdellaoui、Mark J Adams、Esben Agerbo、Tracy M Air、Till MFAndlauer等。全基因组关联分析确定44个风险变异，并细化重度抑郁症的遗传结构。《自然·遗传学》，50(5):668–681，2018年。

Shiwen Wu、Fei Sun、Wentao Zhang、Xu Xie 和 Bin Cui。推荐系统中的图神经网络：综述。*ACM Comput. Surv.*，55(5)，2022年12月。ISSN 0360-0300。doi: 10.1145/3535101。URLhttps://doi.org/10.1145/3535101。

Tong Wu、Yunlong Wang、Yue Wang、Emily Zhao、Yilian Yuan 和 Zhi Yang。基于图的医疗实体嵌入的电子健康记录数据表示学习。*arXiv* 预印本 *arXiv:1910.02574*，2019年。

Zhitao Ying、Dylan Bourgeois、Jiaxuan You、Marinka Zitnik 和 Jure Leskovec。Gnnex-plainer：为图神经网络生成解释。神经信息处理系统进展，32，2019年。

Jie Zhou、Ganqu Cui、Shengding Hu、Zhengyan Zhang、Cheng Yang、Zhiyuan Liu、Lifeng Wang、Changcheng Li 和 Maosong Sun。图神经网络：方法与应用综述。*AI open*，1:57–81，2020年。

Zhi-Hua Zhou 和 Xu-Ying Liu. 使用解决类别不平衡问题的方法训练成本敏感的神经网络。*IEEE Transactions on knowledge and data en-gineering*,18(1):63–77, 2005年。

附录A. 模型训练细节

表3 列出了实验中使用的超参数值和架构设计选项。对于合成数据实验，我们对所列出的所有组合进行了网格搜索，

结果如图 6所示。对于真实数据实验，由于敏感数据计算环境中的计算资源限制，我们采用了合成数据实验中确定的表现最佳架构所对应的参数值。在适用情况下，实验中比较的基线方法使用相同的超参数值。计算复杂度见表4。

表 3：超参数值和架构设计选择。真实数据实验所使用的值以粗体标示。

描述	值
边的定义	父母/子女、包含目标对象的父母/子女、**所有具有遗传关系的对象**
图神经网络层	GCN(Kipf和 Welling，2016年)，**k-图神经网络** (Morris 等，2019年)，GAT (Veli˘ckovi´c 等, 2017年)
池化层	**目标**，总计，　平均值、SAGPool（求和）、SAGPool（平均值）（李等人，2019年；Liu等人，2022年）
SAGPool 比率	0.4, 0.7
优化器	Adam与 $\beta_1 = 0.9, \beta_2 = 0.999$
批大小	250
最大训练轮数	100
耐心（提前停止标准）	5
学习率	0.01，**0.001**
隐藏维度 - 图神经网络单元	$H_G = 20$
隐藏维度 - 长短期记忆网络单元	$H_L = 40$
隐藏维度 - 多层感知机单元	$H_M = 20$
γ	1
α	1
β	1
δ	1
随机失活率	0.5

A.1. 训练期间管理类别不平衡的方法

- 采样策略：通过将少数类（病例）按 70:10:20 的比例划分为训练集、验证集和测试集，来设置训练集/验证集/测试集划分；随后，通过对多数类（对照组）进行欠采样，按训练集中病例与对照组 15:85 的比例，将对照组随机分配到这些集合中。这能够提高训练期间的统计效能；但验证集和测试集采用数据集中的病例与对照组比例，以便在验证和测试时维持符合实际的人群分布。

- 类别加权损失：研究表明，基于损失函数的方法有助于解决健康领域中的类别不平衡分类问题（Depto 等人，2023年）。我们根据类别比例对 BCE 损失项进行加权，以加大对少数类误分类的惩罚。

- 过拟合应对策略：为防止少数类过拟合，采用随机失活层，并在验证集上计算早停标准；如果验证损失开始增加，则停止训练。分类阈值被视为超参数，并通过对验证集应用阈值移动方法进行优化。 vali-

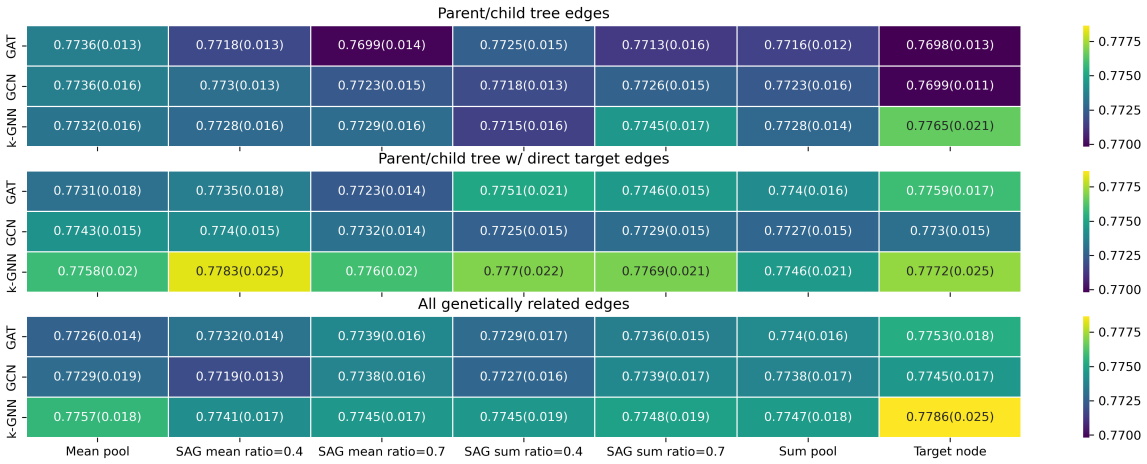


图6：图神经网络设计研究，报告在表中列出的超参数下，对 $h = 0.7$ 合成数据集进行5次试验后计算的受试者工作特征曲线下面积平均值（方括号内为95%置信区间）3。

验证集（Krawczyk 和 Wo´zniak，2015年；Zhou 和 Liu，2005年），基于精确率-召回率曲线。

- 评估指标：类别不平衡可能会影响对广泛使用的指标（如精确率-召回率曲线下面积（Hicks 等人，2022年；Saito 和 Rehmsmeier，2015年）的解读。在测试集上评估性能时会使用多种指标：受试者工作特征曲线下面积、精确率-召回率曲线下面积和马修斯相关系数（MCC）。

表 4：5 种疾病的平均训练时间（及标准差），单位为分钟。模型在单张 NVIDIA A100 GPU 上训练。

模型	平均训练时间
年龄、性别 多层感知机 (A1)	12.98 (5.07)
年龄、性别、家族史 多层感知机 (A2)	19.30（7.12）
年龄、性别、图连通性（A3）	15.56 (7.90)
图神经网络 (G1)	59.88（28.20）
年龄、性别、电子健康记录 长短期记忆网络（A4）	88.92（26.66）
年龄、性别、电子健康记录长短期记忆网络、家族史 (A5)	84.34 (26.37)
GNN-LSTM (G2)	381.90 (165.73)

附录 B. 合成数据算法

在第 4.3 节所述假设的基础上，我们使用以下算法生成合成数据：

1. 为祖源生成（第 0 代）生成个体 i，其

$$L^0_{Herr,i} \sim N(0,h^2), \epsilon^0_i \sim N(0,e^2), Gender^0_i \sim \mathrm{Ber}(0.5), Age^0_i \sim \mathrm{Uni}(0,1),$$

以及其总表型易感性 $L^0_i = L^0_{Herr,\ i} + \beta_{age} \cdot Age^0_i + \beta_{sex} \cdot Gender^0_i + \epsilon^0_i$。$h^2$，$e^2$，$\beta_{age}$ 和 β_{sex} 均为输入参数，用于调整各组成部分对总易感性的贡献。

2. 随后可按如下方式生成第 k 代（$k > 0$）的个体 j

$$\epsilon^k_j \sim N(0,e^2), Gender^k_j \sim \mathrm{Ber}(0.5), Age^k_j \sim \mathrm{Uni}(0,1),$$

以及 $L^k_{Herr,\ j} = c \cdot (\alpha_1 L^{k-1}_{Herr,\ mo} + \alpha_2 L^{k-1}_{Herr,\ fa})$，其中 mo，fa 是从上一代中随机选出的母亲和父亲，并满足以下条件：

(a) 母亲为女性，父亲为男性；(b) 母亲和父亲在三代以内没有亲缘关系（他们不是样本个体，也没有共同的父母或祖父母）

α_1 和 α_2 是父母对其子代易感性的贡献参数，且 $\alpha_1 + \alpha_2 = 1$。根据定义，L_{Herr} 涵盖遗传易感风险和非遗传性父母影响，因此两者不一定相等。$c = \frac{1}{\alpha_1^2 + \alpha_2}$ 是一个常数，用于稳定后续世代的易感性方差。该个体的总表型易感性为 $L^k_j = L^k_{Herr,\ j} + \beta_{age} \cdot Age^k_j + \beta_{sex} \cdot Gender^k_j + \epsilon^k_j$。重复步骤 2，直至达到指定的世代数。为将连续表型易感性映射为二元疾病状态，我们可以假设疾病患病率为 k，并在标准化易感性上设置截断阈值 $\Phi^{-1}(k)$。然后将所有 $L^k_i \geq \Phi^{-1}(k)$ 判定为病例，其余判定为对照。

附录 C：终点定义

表5：终点详情。疾病特定年龄范围和对照组排除标准

关注疾病	目标年龄范围	目标排除条件
成人哮喘	20-70	任何哮喘，慢性阻塞性肺疾病（COPD）及相关医疗报销
冠心病（冠心病）	40-70	任何心血管疾病
结直肠癌	45-75	任何癌症诊断
抑郁症	18-60	任何情绪障碍、自杀未遂、抗抑郁药物购买
II型糖尿病（2型糖尿病）	35-70	I型、II型糖尿病及相关并发症、胰岛素购买