疾病风险动态预测模型方法前沿进展与精准预防

宋雨昕[1],叶倩[2],赵盟生[2],张隆垚[2],魏永越[1,3,4*]

1. 北京大学公众健康与重大疫情防控战略研究中心,北京 100191
2. 南京医科大学公共卫生学院生物统计学系,南京 211166
3. 北京大学公共卫生学院流行病与卫生统计学系,北京 100191
4. 重大疾病流行病学教育部重点实验室(北京大学),北京 100191

摘要　动态疾病风险预测模型将是精确预防策略的核心,在过去20年中,以精准预防为目的的疾病风险预测模型研究呈现快速增长的态势。目前广泛应用的模型未能充分考虑预测因子随时间变化对疾病风险的影响(静态模型),校准漂移不可避免。综述了动态风险预测模型建模方法,得出如下认识:随着医疗健康大数据的互联互通和共享共用的不断推进,统计学和人工智能新方法的不断涌现,如何挖掘出更丰富的预测因子、识别出更准确的作用模式、开发更符合生物医学背景和实际场景的具有可解释性的疾病风险预测模型,赋能共病共防、异病同防,最终实现个体化多疾病谱的精准预防,将是未来的预测模型方法学研究的重点方向。

关键词　预测模型;静态模型;动态预测;精准预防

在21世纪的医疗健康领域,疾病风险预测模型的发展如同一股不可阻挡的潮流,对健康管理和疾病防控的模式有着重要的影响。疾病风险预测模型,是基于模型利用特定时间的个人信息或暴露情况来估计未来发生某一健康事件的风险概率,是精准预防的基础,也被广泛应用于临床决策、预后评估等临床实践中[1]。例如,ORISK2模型可估计个体在未来10年内患心血管疾病的风险,若风险超过10%,则将考虑使用他汀类药物进行治疗[2]。随着医疗健康大数据领域的快速发展、互联互通机制

收稿日期:2024-04-17;修回日期:2024-06-07
基金项目:国家自然科学基金面上项目(81973142)
作者简介:宋雨昕,博士研究生,研究方向为疾病动态风险预测模型统计方法,电子信箱:songpku2023@bjmu.edu.cn;叶倩(共同第一作者),硕士研究生,研究方向为非独立数据统计分析方法,电子信箱:yeqian@stu.njmu.edu.cn;魏永越(通信作者),研究员,研究方向为健康医疗大数据统计分析理论方法与应用、疾病风险和预后预测模型理论方法与应用,电子信箱:ywei@pku.edu.cn
引用格式:宋雨昕, 叶倩, 赵盟生, 等. 疾病风险动态预测模型方法前沿进展与精准预防[J]. 科技导报, 2024, 42(12): 75-91;
doi:10.3981/j.issn.1000-7857.2024.05.00543

的逐渐完善、大型人群队列资源的公开共享以及预测模型研究相关规范的提出[3-5],以精准预防为驱动的疾病风险预测模型研究在过去20年里呈现快速发展的态势。这些模型综合了环境因素、人口学特征、生活方式、遗传学等多维度数据,通过不同方法(算法)为个体提供疾病风险的预测。预测的目的不仅是提前感知疾病的风险,更是在疾病萌芽之时就进行精准预防,以避免疾病的发生。

但是,目前大多数的疾病风险预测模型依赖的变量值是个体某一时间点的变量值,简称静态模型。随着时间的推移,个人特征、暴露情况、临床实践,甚至整个医疗系统都在发生变化,这意味着基于静态数据的预测会发生偏倚,甚至失效,这种现象称为校准漂移(calibration drift),是疾病风险预测模型实践中的主要缺陷之一[6]。如何充分利用日益丰富、实时变化的信息,解析变化趋势对健康结局的影响,构建更为精准的动态风险预测模型(动态模型),是当前模型方法研究的重点,是疾病精准防控的迫切需要[7-8]。

1 预测因子类型

预测模型的构建,首先是预测因子的筛选。候选预测因子类型多样(图1),包括但不限于:群体特征(如环境暴露、气候变化、社会经济因素、公共卫生政策等),个体宏观特征(如人口统计学特征、人体测量、生命体征、个人病史等),物理检查特征(如X射线、超声检查、核磁共振、心电图等)和个体微观特征(如实验室测量指标,基因、转录组学,蛋白质组学,代谢组学等)。

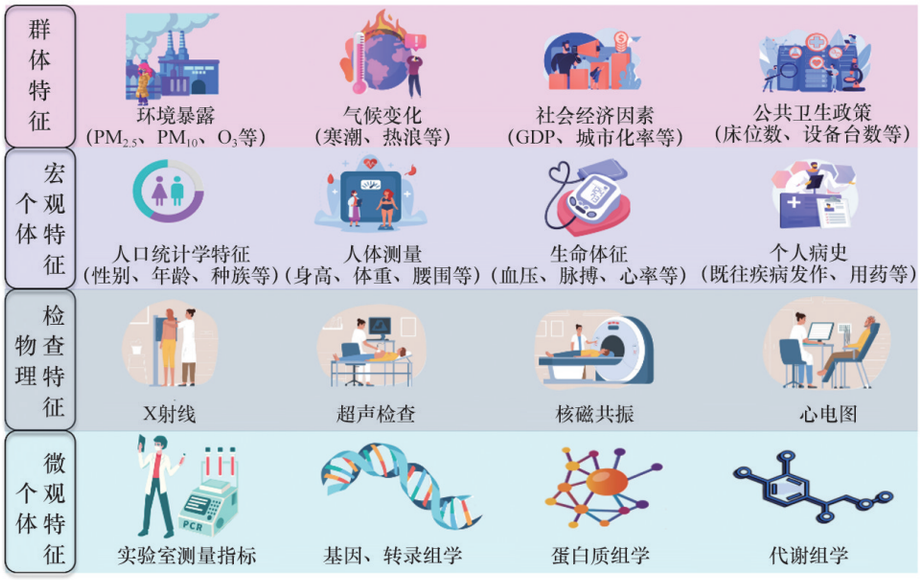


图1 预测因子分类

以肺癌为例,为了深入了解当前肺癌风险预测模型研究现状,严格按照系统评价标准(preferred reporting items for systematic reviews and meta-analyses,PRISMA)流程,筛选出22篇研究论文,其中21篇包括至少经过1次外部前瞻性队列验证的模型,并报道了模型在验证集中的准确性和校准度指标,预测因子累计有55种[9]。包括吸烟暴露情况、人口学因素、环境暴露、个人及亲属肿瘤史、其他疾病和症状史、功能学和实验室测量指标等。其中14个模型仅纳入了吸烟、年龄等传统流行病学因素,另7个模型增加了肺功能或胆红素等实验室指标。其中,$PLCO_{m2012}$模型在美国吸烟(或曾经吸烟)人群中,其6年肺癌风险预测准确度(用AUC(曲线下面积)评价)达79.7%,在欧洲、澳大利亚、

加拿大、巴西等国家和地区人群中被广泛验证[10-11]。Wang等[12]基于中国肺癌筛查项目,针对中国吸烟和不吸烟人群分别构建了肺癌预测模型(China NCC-LC_{m2021}),外部验证集中的AUC分别是75.2%和67.3%。Pan等[9]基于类似的15项易获取的宏观变量,采用XGBoost机器学习算法构建的网络本体语言(web ontology language,OWL)模型,在英国生物银行(United Kingdom Biobank,UKB)人群中预测其8年内新发肺癌的准确度达85%,校准度略优于现有模型。飞速发展的组学(omics)技术,可为风险预测提供更多的分子水平的因素。Shen等[13]基于19个脱氧核糖核酸(deoxyribonucleic acid,DNA)区域上的遗传变异位点(genetic variant),构建了适合于中国人群的肺癌多基因遗传风险评分(polygenic risk score,PRS),可有效地进行人群风险分层。近期多项研究提示克隆性造血(Clonal hematopoiesis,CH)相关基因突变与肺癌风险相关[14],可用于风险预测。Irajizad等[15]利用4个蛋白标志物和$PLCO_{m2012}$模型,将AUC提升了5%。

从上述文献研究可见,在肺癌风险预测模型领域,成熟的模型皆为静态模型,在长期风险预测中的校准偏倚是需要关注的;其他疾病预测模型研究现状与此类似。随着医疗健康大数据领域的快速发展,越来越多的预测因子拥有多次重复测量数据,指标的纵向趋势将是一类极具潜力的疾病风险预测因子。近期一项于新西兰开展的新型冠状病毒所致急性呼吸窘迫综合征(acute respiratory distress syndrome,ARDS)人群预后研究中,用前4天呼吸机参数识别出通气率(ventilatory ratio)轨迹的2个亚型,其28天死亡风险相差近1倍[16]。Haines等[17]构建了尿酸肌酐比这一新指标,发现其轨迹和重症医学病房(intensive care unit,ICU)创伤患者的肌溶解密切相关。基于重症监护室脓毒症患者的血小板计数动态监测数据,识别出具有不同纵向变化趋势的亚组,可提升患者预后预测效果[18]。

融合指标的纵向趋势所构建的动态预测模型为个体层面的精准风险评估提供了契机,为实现个体化的动态、精准预防提供了重要的模型支撑。

2　建模方法

队列研究中的重复调查、人群的定期健康监测、多时点的组学数据检测,产生大量重复测量的具有潜在疾病风险预测价值的自变量。从统计学角度审视,自变量重复测量数据具有以下特点:同一自变量的重复测量之间高度相关,不同自变量之间相关,不同自变量间的关系模式可能随时间变化,不同自变量的不同时间点之间亦可能相关;不同自变量的重复测量时间点和频次不一致等。针对这一特殊数据结构,在预测模型构建过程中,现有分析方法可分为传统回归模型、纵向趋势模型、联合模型、界标模型、动态贝叶斯网络模型等5类。

2.1　传统回归模型

当数据集较为稳定且关系为线性的情况下,或研究的重点是估计特定变量对结果的影响时,传统回归模型是一个好的选择。常见分析方法如下。

1)多重回归模型(multivariable regression model)。有研究将所有重复测量的自变量同时放入模型。

2)通径分析模型(path analysis model)。基于重复测量自变量间的时序关系,绘制包含自变量、因变量及其他协变量的有向无环图[19];实为在多重回归基础上,进一步分解和估计间接效应。以上2种方法都要求重复测量的时间点一致,且将面临共线性(collinearity)问题而致使系数估计有偏。

3)条件回归模型(conditional regression model)。某个时间点的自变量与其前序变量之间存在高相关性,可以用前序变量对其进行回归

$$x_p = b_0 + \sum_{j=1}^{p-1} b_j x_j + e \quad (1)$$

式中,x_p为时间点p自变量的观测值,b_j为p前面的$p-1$个时间点的自变量观测值x_j的偏回归系数,b_0为截距,e为残差。

将回归的残差放入后续评估疾病发生风险的结局模型中(如Cox模型)。该方法一定程度上克服了共线性问题,但要求重复测量的时间点要一致,大大地限制了其应用范围。

2.2 纵向趋势模型

纵向趋势模型主要通过重复测量自变量的建模以获得典型特征,如生长曲线参数或潜在轨迹分类,进而将其和疾病结局建立结局模型。它可以识别总体中的异质性,即使在群体中存在不同的发展轨迹,也能够识别并分析这些差异,并且模型对于测量时间点的要求相对灵活,不需要严格的一致性,这使得纵向趋势模型可以适应各种不同的研究设计。当数据随时间变化,并且需要考虑时间依赖性时,可考虑纵向趋势模型。常见分析方法如下。

1)典型特征回归模型(representative characteristics regression model,RCRM)。用"最佳"测量值代替所有重复测量值,或对各受试者的重复测量自变量估算多种描述性统计量。Leffondré等[20]从重复测量的自变量中提取均值、极差、变异、最大变化值、平均变化值、线性回归系数等27种特征作为因子分析的输入信息,以获得具有不同趋势特征的亚组。该方法忽略了时间点的预测价值,且信息有所压缩,损失效能。

2)生长曲线参数回归模型(growth curve parameters regression,GCPR)。对每位研究对象的重复测量的时间作为自变量和其测量值进行回归,构建每位研究对象的测量值"生长曲线"(growth curve),体现曲线特征的参数估计值(如回归系数)将被纳入后续和结局的分析。模型通过参数化的方式,可以准确描述生长过程中的关键特征和阶段,但重复测量次数较少情况下,拟合存在困难。

3)潜在生长曲线模型(latent growth curve model,LGCM)。该方法为建立在结构方程模型(structural equation models,SEM)框架下,将基线自变量和自变量的变化由2类潜变量α(截距因子)和β(斜率因子)估计α上所有测量值的因子负荷是一致的,可解释为估计的基线水平。β上测量值的因子负荷与时间相关联,反映自变量单位时间的变化。图2中展示了等距时间间隔下,时间函数为线性的4次测量模型。α与β之间的双箭头表示2个因子之间的相关性,用于说明个体变化的截距和变化斜率之间的关系,ε代表残差。该方法可以估计和分析个体间在增长轨迹上的差异,因子相互独立,解决了自变量共线性问题。但该方法要求重复测量的时间点要统一,且不兼容缺失数据。

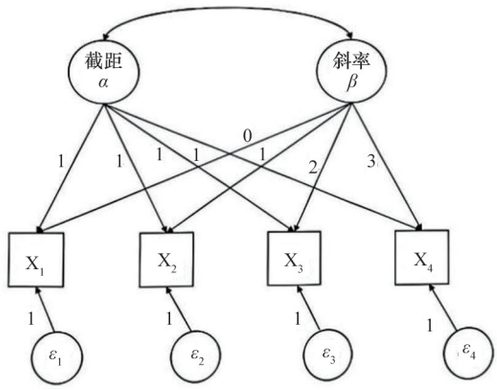


图2 潜在生长曲线模型(LGCM)原理示意

4)混合效应模型(mixed effects model,MEM)。Laird等[21]于1982年提出在线性模型中引入随机效应,用于处理个体水平在不同时间上重复测量值之间相关性的问题,进而构造自变量随时间变化的纵向趋势。在混合效应模型中,重复测量被视为"1水平"变量,聚集在被视为"2水平"变量的个体中,因此也称为多水平模型(multilevel model)。

若对于每个个体建立各自测量值与时间的一次回归方程,则表示每个个体的发展趋势可以用简单的线性关系来描述

$$y_i(t)=\left(\beta_0+b_{i0}\right)+\left(\beta_1+b_{i1}\right)x_i(t)+\varepsilon_i(t) \quad (2)$$

式中,$y_i(t)$表示个体i在时刻t的观测值,$x_i(t)$表示时间;β_0和β_1分别表示总体截距和斜率的均值,由于对于每个个体具有相同的取值,因此也称为固定系数或固定效应(fixed effect);b_{i0}和b_{i1}分别表示个体截距和斜率与总体均值之间的差异,对于每个个体具有特定的取值,因此也称为随机系数或随机效应(random effect);$\varepsilon_i(t)$为残差项。

该模型可以对重复测量时间点施加样条函数,以实现随时间的非线性变化趋势。

5)潜分类增长模型(latent class growth model,LCGM)。在传统的LGCM中假设群体是同质的,群体内所有个体享有相同总体趋势。但在群体中存在异质性时,上述显然是不成立的[22]。Nagin于2005年提出潜类别增长模型,也称作组轨迹模型

(group-based trajectory modeling,GBTM),通过潜类别将研究对象分为几个具有不同增长趋势的组。

潜分类增长模型需要考虑2个关键方面:类别分配模型和增长曲线模型。

假设有N个个体,每个个体在T个时间点上都有观测数据,将这些个体分成K个潜在类别(latent classes),其中g=1,2…,G。令$y_i(t)$表示第i个个体在第t个时间点上的观测值,c_i表示第i个个体所属的潜在类别。使用多项式逻辑回归(multinomial logistic regression)来建立类别分配模型。

$$P\left(c_i = g \middle| X_i\right) = \frac{\exp\left(\alpha_g + \beta_g X_i\right)}{\sum_{j=1}^{G}\exp\left(\alpha_j + \beta_j X_i\right)} \quad (3)$$

式中,α_g是第g个类别的截距,β_g是与个体特征X_i相关的系数。

对于每个潜在类别g,建立增长曲线模型来描述个体特征随时间变化的情况,可以是线性、非线性或其他形式的增长。例如,简单的线性增长模型可以表示为

$$y_i(t)\Big|_{c_i=g} = \beta_{0g} + \beta_{1g}x_i(t) + \varepsilon_i(t) \quad (4)$$

式中,β_{0g}为类别特定的截距,β_{1g}是类别特定的斜率,$\varepsilon_i(t)$是误差项。

该模型可以处理样本中的异质性,即不同个体或子群体可能遵循不同的增长模式,但尚未考虑同一组内的随机效应[23]。R包lavaan(latent variable analysis)可以用于此类分析。

潜在轨迹模型帮助研究者识别出具有相似发展轨迹的患者群体,为精准医疗提供理论基础。模型还可以用来评估治疗干预的长期效果,通过比较不同患者群体的发展轨迹,判断特定治疗方法的有效性。但由于它们是基于无监督学习的,可能会存在个体化分类倾向,特别是在样本量较小或变量多样性较大的情况下,所以潜在变量的选择和模型结构的设定需要仔细考虑,以确保模型的解释性和应用价值。

6)潜分类混合效应模型(latent class mixture modeling,LCMM)。为了考虑组内随机效应,Muthén等[24]将潜变量分析与混合效应模型融合,提出了LCMM模型。该方法是当前应用最为广泛的方法之一,模型假设总体中存在有限个潜在亚组,可以用于探索异质人群中纵向测量指标随时间变化的趋势。每个亚组中,包括类别特定的固定效应和随机效应。

$$y_i(t)\Big|_{c_i=g} = \beta x_{1i}(t) + \upsilon_g x_{2i}(t) + b_{ig}z_i(t) + \varepsilon_i(t) \quad (5)$$

式中,$x_{1i}(t)$、$x_{2i}(t)$和$z_i(t)$为协变量向量,$x_{1i}(t)$与跨类别固定效应β相关,$x_{2i}(t)$则具有特定于类别的固定效应υ_g,b_{ig}仍为随机效应,$\varepsilon_i(t)$是误差项。R包Lcmm可以用于此类分析。

图3(a)和(c)分别展示了LCMM和LCGM中个体测量值随年龄的增长模式,图3(b)和(d)分别对应于某个时刻LCMM和LCGM分布的截面图[25-26]。从图中可以明显看出这2个模型的区别:在图3(a)和(c)中,较粗线条代表了类别组内个体的平均增长曲线,每条粗线表示一个类别组内的平均变化轨迹。在图3(a)中,除粗线外,还有表示类别组内个体差异的增长曲线(细线),这意味着LCMM允许同一类别内个体拥有相似但不完全相同的增长轨迹,即同一潜在类别内个体之间允许存在方差变异[27]。

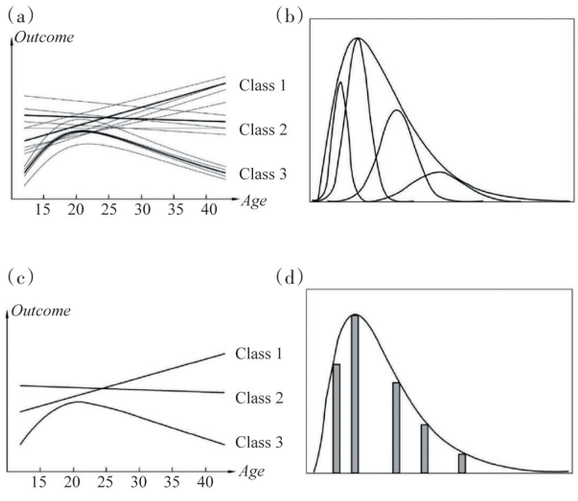


图3　LCMM与LCGM的增长模式图和某时刻测量值分布图

7)多元LCMM(multivariate LCMM,MVLCMM)。多元LCMM模型通过共享随机效应部分,以

推断其共同的潜类别,为多元纵向数据分析提供了一种可能。它可以处理不同类型的数据,包括连续的、非高斯的和有序的结果。然而,若重复测量自变量种类较多、趋势差异较大时,仅用有限个数的潜类别来综合描述,其信息损失严重,预测效果必然受影响。R包Lcmm可以用于此类分析。

相比横断面分析,纵向分析不仅关注个体之间的差异,还关注指标随时间的变化轨迹。随着大型队列研究和电子病例系统的不断发展,医学研究中个体层面的数据变得越来越丰富。这些数据不仅包括了各种临床指标的重复测量,还记录了病人的临床结局和发生时间。在对拥有纵向数据的患者进行生存分析时,除了关注结局状态和生存时间外,还需要考虑自变量的时间依赖性、测量误差,以及因结局发生导致的纵向数据非随机缺失问题。

在建立预测模型时,希望能同时关注患者结局和自变量测量值的变化,以期获得更准确的预后预测信息。因此,在研究死亡率的关联时,考虑风险因素随时间的变化具有很大的吸引力[28],传统统计学模型在分析此类数据时由于处理参与者失访和未观测指标相关引起的非随机缺失时存在局限性[29],因此有学者提出整合纵向数据与生存数据进行联合建模(joint modeling,JM)的方法[30-31]。

2.3 联合模型

联合模型用于同时处理纵向数据(重复测量数据)和生存数据(时间到事件数据)。1997年,Wulfsohn等[31]首次提出了用于处理重复测量和生存结局数据的联合模型。最初,联合模型主要应用于艾滋病研究[32],但随着时间的推移,它已被广泛用于其他临床研究领域,包括肾病、癌症等[33-34]。2023年,Zhang等[35]发现多变量联合模型可以高效地预测精神病高危人群发病情况,受试者工作特征AUC达到0.9,优于静态模型(AUC=0.6)。

目前已发表的大多数文章中,常使用共享随机效应联合模型[36-38]。此外,也有少数研究从潜在类别联合模型[39-40]、功能模型[41-42]和加性模型等方面探讨了联合模型[43-44]。本文简单回顾共享随机效应模型和潜在类别联合模型的模型结构及关联的方法。

2.3.1 共享随机效应模型

共享随机效应模型(shared random-effect model,SREM)是一种利用线性混合效应模型对服从多元正态分布的纵向资料进行建模的方法。其核心假设是纵向过程和生存过程之间存在共享的随机效应,并且能够同时解释纵向过程和生存过程之间的关联,以及纵向过程中重复测量之间的相关性[45]。SREM主要由2部分组成:一是刻画重复测量轨迹的纵向子模型,二是描述生存过程的生存子模型。

1)纵向子模型。纵向子模型通常是一个线性混合效应模型,用于描述个体随时间变化的测量结果。

$$y_i(t)=m_i(t)+\varepsilon_i(t)=X_i(t)\beta+z_i(t)b_i+\varepsilon_i(t) \quad (6)$$

式中,$y_i(t)$是第i个受试者在时间点t上的纵向观测值;β是固定效应;b_i是随机效应;$X_i(t)$为协变量;$\varepsilon_i(t)$是测量误差项,与b_i无关,且服从均值为0、方差为σ^2的正态分布。

2)生存子模型。生存子模型则是一个比例风险模型,用于描述时间到事件的数据。假设受试者i事件发生的风险取决于时间点t处标志物的真实值$m_i(t)$,则有

$$\begin{aligned} &h_i\left(t\middle|M_i(t),\boldsymbol{w}_i\right)\\ &=\frac{\lim\limits_{s\to 0}\Pr\left\{t\leqslant T_i^*<t+s\middle|T_i^*\geqslant t,M_i(t),\boldsymbol{w}_i\right\}}{s} \quad (7)\\ &=h_0(t)\exp\left\{\gamma^T\boldsymbol{\omega}_i+\alpha m_i(t)\right\},t>0 \end{aligned}$$

式中,$M_i(t)=\{m_i(s),0\leqslant s<t\}$,表示直至时间点$t$的真实纵向过程;$h_0(\cdot)$是基线风险函数;$\boldsymbol{\omega}_i$是基线协变量的向量;$\gamma$是回归系数。

3)共享随机效应建模。基于2个纵向子模型,可以将SREM总结成一种标准的表述形式:即在经典的生存模型中加入代表了纵向过程的$m_i(t)$作为协变量,$m_i(t)$前的系数α量化了纵向指标水平与结局发生风险之间关联。在纵向子模型中,当假设存在简单的随机截距和随机斜率结构时,这种参数化更有意义。在这种情况下,随机效应表示了个体特定的偏差,即与平均截距和平均斜率相比,某些个体可能具有较低或较高的截距,或者在其纵向

轨迹上显示出较陡或较缓的增加或减少[45]。图4展示了SREM的方法原理,其核心思想是假设纵向过程与之间存在共享的随机效应,能够同时解释重复测量数据和时间至事件之间的关联,以及纵向过程中重复测量之间的相关性。

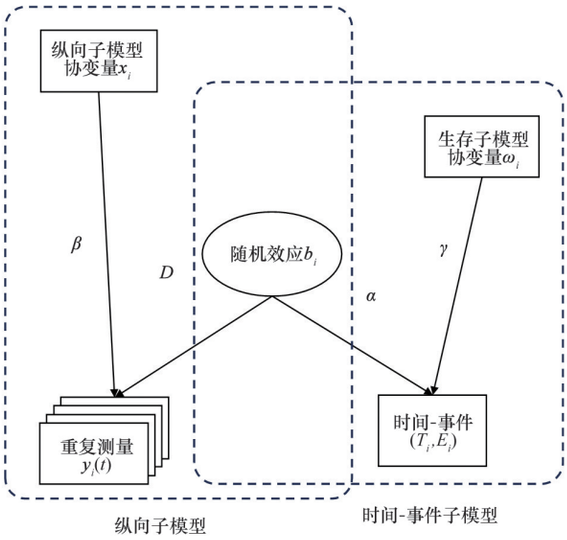


图4　SREM示意

2.3.2　潜在类别联合模型

共享随机效应模型假设是种群同质的,所有个体都遵循一个单一的平均轨迹,且具有共同基线风险。潜在类别联合模型(joint latent class model,JLCM)认为这种假设可能与实际的高度异质性群体情况不符[22,46]。其假设总体可以被划分为有限个子群,每个子群内部的个体遵循相同的平均轨迹和共同基线风险,并且假设在给定潜在类别的情况下,标志物和事件发生的时间是条件独立的[47-48]。

因此,在描述个体的标志物特征与事件风险之前,需提前定义总体内存在的潜在子群。若总体样本量为N,则每位受试者$i(i=1,2,...,N)$的潜在类可以通过一个分类的潜在变量c_i来定义,如果受试者i属于潜在类别$g(g=1,2,...,G)$,则$c_i=g$,这里的c_i变量是潜在的、不可观察得到的。个体属于潜在类别g的概率可以通过协变量向量X_{pi}使用多项式logistic回归模型来进行计算

$$P\left(c_i=g\middle|X_{pi}\right)=\pi_{ig}=\frac{e^{\xi_{0g}+X_{pi}^{\top}\xi_{1g}}}{\sum_{l=1}^{G}e^{\xi_{0l}+X_{pi}^{\top}\xi_{1l}}} \tag{8}$$

式中,ξ_{0g}是潜在类g的截距,ξ_{1g}是与时间无关的协变量向量X_{pi}的类特定参数向量。并且有

$$\prod_{g=1}^{G}\pi_{ig}=1 \tag{9}$$

1)类别特定的纵向子模型。在JLCM中,由于其假设"每个子群内部的个体遵循相同的平均轨迹和共同基线风险",因此,其纵向子模型和生存子模型额外有了"类别特定"的特征。给定潜在类别g的情况下,患者i的纵向标志物在测量时间$t_j(j=1,2,...,J)$的重复测量$y_i(t)=(y_i(t1),...,y_i(t_j),...,y_i(t_J))$,也可以被描述为一个线性混合模型

$$\begin{aligned}y_i(t)\big|_{c_i=g}&=y_{ig}^{*}(t)+\varepsilon_i(t)=x_{L1i}^{\top}(t)\beta\\&+x_{L2i}^{\top}(t)\upsilon_g+z_i^{\top}(t)b_{ig}+\varepsilon_i(t)\end{aligned} \tag{10}$$

与SREM不同的是,JLCM将之前定义的固定效应协变量分解为$x_{L1i}^{\top}(t)$和$x_{L2i}^{\top}(t)$,其中,$x_{L1i}^{\top}(t)$与跨类别固定效应β相关,$x_{L2i}^{\top}(t)$则具有特定于类别的固定效应υ_g,b_i仍为随机效应。

2)类别特定的生存子模型。依旧使用比例风险模型对患者发生事件的时间T_i进行描述

$$h_i\left(t\middle|c_i=g;\varsigma_g\right)=h_{0g}\left(t;\varsigma_g\right)\exp\left(\gamma_g^{\top}w_{ig}\right) \tag{11}$$

这里的类特定参数向量γ_g用于描述协变量向量w_{ig}和风险与时间之间的关系。与SREM不同,JLCM不再将纵向过程作为协变量放入生存子模型,而是假设了类特定基线风险h_{0g}与类特定协变量w_{ig}及其参数γ_g,这些假设在不同类内可以不尽相同。

3)潜在类别个数的确定。进行JLCM分析时,建模之前需要确定潜在类别的数量。这个过程通常从假设仅存在一个类别开始,此时拟合的模型称为零模型或独立模型;然后逐步增加潜在类别的数量,重新建模并计算各个模型的参数;最后,根据拟合评价指标对模型进行比较,以确定最佳模型[49-50]。常用的评价标准包括AIC(Akaike information criterion)、BIC(Bayesian information criterion)及样本矫正BIC(sample size adjusted Bayesian information criterion,SA-BIC),这些准则分数越低表

示模型拟合效果越好[47]。另一个常用的度量指标是“熵”(entropy),用于衡量模型分类的质量,熵接近1表示良好的分类[51]。需要注意的是,这些拟合度量标准可能存在差异,因此当它们分别达到最佳值时,不同的评价指标可能对应不同的模型。

4)后验分类。确定最佳类数模型后,可以通过贝叶斯定理计算个体潜在类别归属的后验概率[52]为

$$\widehat{\pi}_{ig}^{Y,T} = P\left(c_i = g \middle| Y_i,\left(T_i,E_i\right);\widehat{\theta}_G\right) = \frac{\pi_{ig}P\left(Y_i \middle| c_i = g;\theta_G\right)P\left(T_i,E_i \middle| c_i = g;\theta_G\right)}{\sum_{l=1}^{G}\pi_{il}P\left(Y_i \middle| c_i = l;\theta_G\right)P\left(T_i,E_i \middle| c_i = g;\theta_G\right)} \tag{12}$$

$\widehat{\pi}_{ig}^{Y,T}$具体解释为具有某些特征的个体i属于潜在类别g的概率,既可以作为个体分组依据,也可以用于评价模型拟合质量。若个体对某一类别的后验概率越高,表示其分类结果越可靠,模型拟合效果越优[53]。通常情况下,个体潜在类别的最大后验概率大于0.8被认为是较为理想的情况[54]。

5)潜在类联合建模。JLCM的核心思想是将个体划分为不同的潜在类别,并在每个类别内部建立混合效应模型。这种模型允许在潜在类别内对个体进行分组,以捕捉不同类别间的异质性。通过这种方式,JLCM能够在考虑潜在类别之间的差异性的同时,对每个类别内部的数据进行建模,从而更准确地描述个体间的差异和共享的特征。图5展示了JLCM方法原理,该方法的主要思想是将数据分解为不同的潜在类别,并在每个类别内部建立模型。通过潜在类别连接纵向过程和生存过程,从而在患者亚群内部的纵向标志物轨迹特征和事件发生风险之间建立关联,以更好地理解数据内在的结构和模式。

目前,联合建模的形式已拓展到单纵向指标和单时间-事件结局、单纵向指标和多时间-事件结局(也称竞争风险)[46,55]、多纵向指标和单时间-事件结局[48,56]以及两者都不唯一[57-58]的情况。

2.3.3 联合模型的动态预测

联合模型的动态预测意味着:(1)利用整个纵向历史数据;(2)在每次有新数据可用时更新预测[59]。与传统生存分析模型中仅利用患者基线或特定时间点测量的变量信息不同,联合模型的动态预测利用了整个纵向测量的历史数据。在贝叶斯框架下,可以针对生存或纵向结果推导出特定于个体的预测[55]。

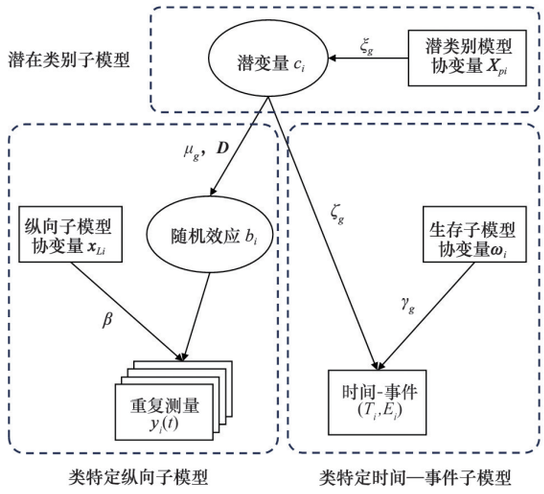


图5 JLCM示意

1)共享随机效应模型动态预测。具体来讲,SREM是从目标人群$D_n=\{T_i,\delta_i,y_i;i=1,...,n\}$的样本中拟合的,希望为来自相同人群的新个体$k$推导出预测,该个体提供了一组纵向测量$Y_j(t)=\{y_j(s);0\leqslant s\leqslant t\}$,并具有基线协变量向量$\boldsymbol{w}_j$。值得注意的是,若个体标志物的测量已记录到$t$,意味着该个体在此在时间点$t$之前都存活,因此,需要关注的是给定到$t$时刻存活的条件下特定于个体的预测。即对已存活到$t$时刻的新个体,需要关注的是其至少能够继续存活$\Delta t$,即存活到$t+\Delta t$时刻的概率[45],则

$$\pi_j^{\mathrm{SREM}}\left(t+\Delta t \middle| t\right) = P\left(T_j^* \geqslant t+\Delta t \middle| T_j^* > t, \mathrm{y}_i(t), \boldsymbol{w}_j, D_n; \boldsymbol{\theta}^*\right) \tag{13}$$

式中,$\boldsymbol{\theta}^*$表示参数的真实值。由$\pi_j^{\mathrm{SREM}}\left(t+\Delta t \middle| t\right)$的定义可知,$\pi_j^{\mathrm{SREM}}\left(t+\Delta t \middle| t\right)$具有时间动态性。

当患者k在时间$t'>t$记录新的信息时,可以更新得到$\pi_j^{\mathrm{SREM}}\left(t'+\Delta t \middle| t'\right)$,从而以时间动态的方式进行预测[60]。

2）潜在类别模型动态预测。与SREM类似，对来自相同总体$D_n=\{T_i,\delta_i,y_i;i=1,...,n\}$，已存活到$t$，具有所有其他协变量$\boldsymbol{X}_i$，并提供了一组纵向测量$y_j(t)=\{y_j(s);0\leqslant s\leqslant t\}$的新个体，潜在类别模型也可以推导出该个体至少存活到$t+\Delta t$时刻的条件后验概率[61]

$$\pi_j^{\mathrm{JLCM}}(t+\Delta t\mid t)=P\left(T_j^*\leqslant t+\Delta t\mid T_j^*\geqslant t,y_i(t),\boldsymbol{X}_j,D_n;\boldsymbol{\theta}^*\right)\tag{14}$$

对于式(13)与式(14)真实参数$\boldsymbol{\theta}^*$，可以使用$\{\boldsymbol{\theta}\mid D_n\}$以近似估计$N\{\boldsymbol{\theta}^*,\mathrm{var}(\boldsymbol{\theta}^*)\}$，并使用马尔可夫链蒙特卡罗(Markov chain Monte Carlo algorithm，MCMC)方法计算该个体事件发生的条件概率分布的2.5%和97.5%百分位数[61-62]。

SREM中将纵向过程作为生存过程的协变量以达到联合建模的目的；而在JLCM中，个体的纵向测量不再作为与时间有关的函数带入生存模型中，而是通过潜类别这一变量体现，因而式(13)与式(14)在计算求解上稍有不同[63]，本文不做赘述。

图6为R包JMBayes2中通过共享随机模型将多个纵向趋势模型同时纳入联合估计进行动态预测的实例[64]。

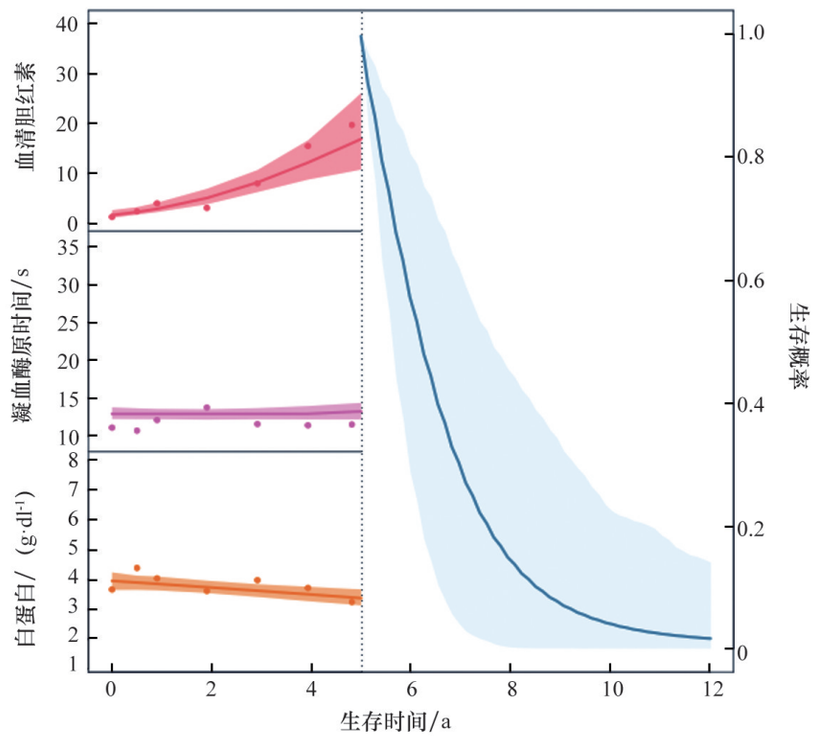


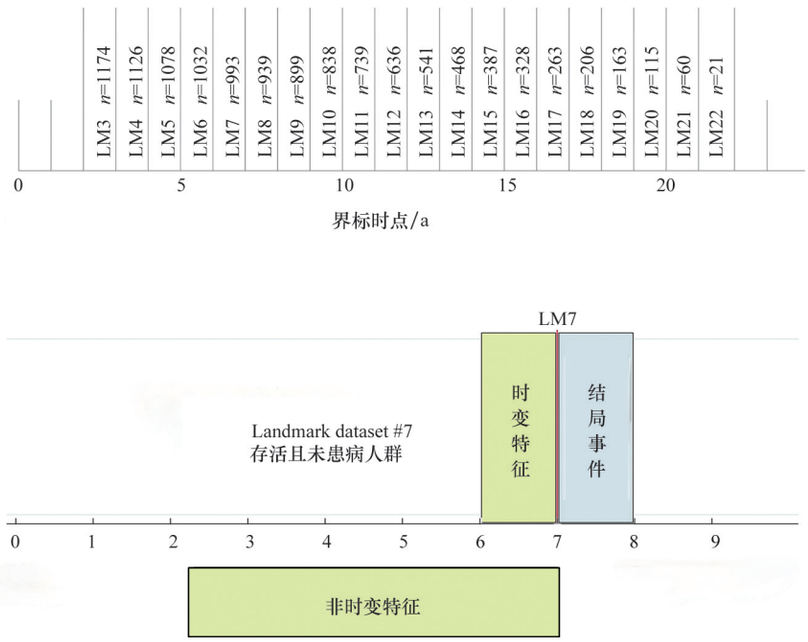
图6　共享随机效应模型多纵向指标和单时间-事件结局动态预测示例

联合模型提供了一种强大的工具，用于分析在医学随访研究中常见的复杂数据类型，当需要同时分析时间到事件数据(如生存时间)和重复测量数据(如生物标志物水平)时，联合模型是合适的。它允许研究者探索变量随时间的变化如何影响事件的风险，能够正确处理噪声和未能完全观察到的时依性协变量信息，从而无偏估计纵向过程和生存过程之间的关系[65]。然而实际情况下的假设和参数估计可能更加复杂，尤其是在分析大型数据集时，可能需要除MCMC和极大似然法以外的更高效的参数估计方法，这增加了对模型的计算需求[66]。

2.4　界标模型

界标模型(landmarking model)的发展始于2004年，Zheng等[67]提出将界标模型用于生存分析

的动态预测。它通过在特定的时间点(即界标时点)对风险进行评估,来预测未来某一时间段内事件发生的概率,在界标时点之前已经死亡或发生结局事件的人群将被排除并忽略时点之后的特征变化。界标时点可以选择一个或多个,并在每个时点构建一个风险人群集合构成一个超预测数据集来进行生存分析(图7)。通过组合多个预测模型,形成一个综合的超级预测模型(super prediction model),风险比率随着界标时间的推移而稳定变化,从而提升条件生存概率估计的精确度和合理性。



在第3年至第22年之间,以每年为间隔选择界标时点(LMs),并为每个时点创建单独的数据集;以数据集7为例,该数据集包括所有在7年后仍存活且未患病的人群,时间变化特征的值均按发病后7年的情况记录

图7 超预测数据集示例

界标超级预测模型的基本原理可以用以下数学表达式来描述。选择界标时点$t_{LM1},\cdots,t_{LMn}$,通过删失创建每个时点的预测数据集,并将一系列数据集合并为超级预测数据集。构建Cox比例风险模型

$$h\left(t\middle|X,t_{LM}\right)=h_0\left(t\middle|t_{LM}\right)\exp\left(X\beta_{LM}\left(t_{LM}\right)\right)\quad t\geqslant t_{LM}\ (15)$$

式(15)中,$\beta_{LM}(t_{LM})=f(t_{LM})\boldsymbol{\theta}$,$f(t_{LM})$是一组光滑函数,多采用的是多项式基函数的线性组合形式,如$\beta_{LM}=\beta_0+\beta_1 t_{LM}+\beta_2 t_{LM}^2$,$\boldsymbol{\theta}$是参数向量。

界标模型已经在囊性纤维化、宫颈癌、慢性肾功能疾病等多个临床领域得到应用[68-70]。该模型能够在每个界标时点根据最新的协变量信息来更新风险预测,从而提供动态的风险评估。这种模型的优势在于其简单的结构、易于实现的特性及高计算效率,使其性能与更为复杂的联合模型相近。它适用于需要在特定时间点更新预测的情况。但该方法目前缺乏统一标准来设定界标时间点,通常根据研究的实际情况进行设置,这可能导致预测结果受数据驱动的影响[71],并且界标模型不考虑测量误差,可能会因为观察值和真实值之间的差异而导致偏倚。

2.5 动态贝叶斯网络模型

贝叶斯网络(Bayesian network,BN)作为一种新兴的概率图模型(probabilistic graphical model),通过有向无环图(directed acyclic graph,DAG)表示生物医学因果关联。在贝叶斯网络中,条件概率仅取决于有向无环图中父节点(parent nodes)项,即

$$p\left(\mathrm{X}\middle|\mathrm{G},\Theta\right)=\prod_{j=1}^{p}p\left(X_j\middle|X_{pa_G(j)},\theta_j\right)\tag{16}$$

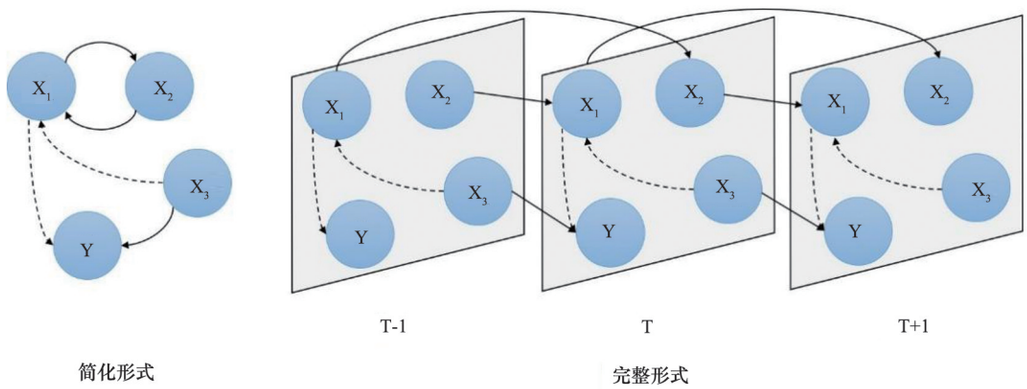
式(16)中,$pa_G(j)=\{i\in V:(i,j)\in E(G)\}$是节点$j$在图$G$中的父节点集,$\Theta=(\theta_1,\theta_2,\cdots,\theta_p)$是所需参数。

然而,传统的贝叶斯网络并不擅长处理时间信息。动态贝叶斯网络(dynamic Bayesian network,DBN)是基于时间序列数据集构建的贝叶斯网络,将不同时间点上的随机变量区别开来,作为不同的随机变量,对处理动态系统具有较强的优势。在动态贝叶斯网络中,模型具有若干个离散时间片(time slices),其中每个时间片的每个变量均被表示为一个节点,由此组成的有向无环图被称作动态贝叶斯网络。该模型存在以下几个常用假设:(1)模型具有一阶马尔可夫性,即变量i在t时刻的取值,依赖于父节点在$t-1$时刻的取值,而与$t-1$时刻前的状态完全独立;(2)模型是静态的,即模型中的关联模式及参数不随时间变化而改变。为方便描述,可以将动态贝叶斯网络模型拆解成初始时间片网络(G_0)和转移网络($G_{\rightarrow}$)2部分,其条件概率为

$$\begin{aligned}p\left(\mathrm{X}\middle|\mathrm{G},\Theta\right)=&\prod_{j=1}^{p}p\left(X_{j,1}\middle|\theta_j^{(0)}\right)\\&\prod_{t=2}^{T}p\left(X_{j,t}\middle|X_{pa_G(j),t-1},\theta_j\right)\end{aligned}\tag{17}$$

式(17)中,θ_j表示在$G_{\rightarrow}$中的参数,$\theta_j^{(0)}$表示在G_0中的参数,$\theta_j=\left(\theta_j^{(0)},\theta_j\right)_{j=1,\cdots,p}$。

因此,在一个简单动态贝叶斯网络中,每个变量都依赖于前一时间点的变量子集,该子集由图G中变量的父节点集给出(图8)。此外,更通用的动态贝叶斯网络模型允许同一时间片内部的有向边,以刻画模型中变量在时间尺度上同期的影响。



实线代表相邻时间片之间的有向边,虚线代表同一时间片内部的有向边

图8　动态贝叶斯网络示意

在重复测量数据的预测模型构建中,动态贝叶斯网络具有独特的优势。第一,网络模型相较于传统回归模型能够考虑自变量间的复杂关联模式,并且基于条件概率来描述复杂关联的不确定性,以降低过拟合风险,提高了模型的泛化能力;除此之外,相较于其他网络模型如神经网络,动态贝叶斯网络模型并不是一个“黑匣子”,其可以提供因果层面的疾病通路信息。第二,根据变量间的条件依赖关系,可以方便地处理数据缺失问题,提高了模型对缺失数据的敏感性。第三,模型综合应用了现有信息及历史信息,将当前变量与历史变量的影响结合,共同用于预测模型构建,提高了模型预测的准确性。

例如,Chen等[72]基于重症监护数据中的时依性生理指标,如体温、血压、白细胞计数、血糖、血清肌酐等以及入院基线数据,利用动态贝叶斯网络,构建了评估包括器官功能障碍,如肾脏、肝脏、心血管和血液功能障碍和死亡风险的动态预测模型。Marini等[73]使用1型糖尿病患者的时依性体格指标(如腰臀比、体重指数)、时依性生化指标(糖化血红

蛋白、低密度脂蛋白、高密度脂蛋白、甘油三酯等),以及静态指标(性别、年龄、治疗方式等)构建了用于长期模拟1型糖尿病临床并发症(心血管疾病、肾病)的动态贝叶斯网络模型,其模拟时间跨度超过15年,测试数据误差低于10%。Orphanou等[74]基于患者的既往病史、低密度脂蛋白等生化指标的重复测量以及吸烟等基线信息,使用动态贝叶斯网络框架,结合时间抽象技术,成功构建了患者首次冠心病发作的风险预测模型。

然而,使用动态贝叶斯网络在构建预测模型时也存在问题,如模型中的时间片之间要求离散且均匀,这削弱了模型的灵活性,但目前已经提出了应用于连续时间变量的贝叶斯网络模型;在算法层面,基于条件独立关系构建的贝叶斯网络模型计算复杂度一般为指数级,目前尚无法满足高维预测因子(如组学数据)的建模需求。建模方法总结见表1。

表1 针对重复测量自变量的5类建模方法的优势、局限和应用场景

方法	优势	局限	应用场景
传统回归模型	计算简单; 适用于二分类结局	面临自变量共线性问题; 受限于重复测量时间点要一致; 未考虑结局	数据集较为稳定且关系线性; 估计特定变量对结果的影响
纵向趋势模型	识别异质性和个体差异; 灵活的数据要求	忽略潜类别的不确定性; 不可同时纳入多个自变量; 未考虑结局	数据随时间变化并且需要考虑时间依赖性
联合模型	无偏估计纵向过程和生存过程之间的关系	计算复杂、依赖算力	同时分析时间到事件数据和重复测量数据
界标模型	结构简单、易于实现、高计算效率	缺乏统一标准来设定界标时间点; 忽略测量误差	关注在特定时间点之后的风险预测
动态贝叶斯网络	降低过拟合风险; 方便处理数据缺失问题; 综合应用现有信息及历史信息	时间片之间要求离散且均匀; 指数级算法复杂度	考虑自变量间的复杂关联模式

3 展望

在互联互通的医疗健康大数据的背景下,动态预测模型将成为精准预防的基石。这些模型能够综合多源、动态数据,实时更新疾病风险评估,为医疗决策提供实时、精准科学依据。

3.1 互联互通的医疗健康大数据是预测模型研究的基础

医疗健康领域的大数据储存了关于居民健康、疾病进程、预防措施及其效果,以及医疗服务使用情况的纵向数据,成为维护公众健康的重要数据宝库和证据基础。据国家卫生健康委员会发布的《2022年中国卫生健康统计年鉴》,截至2021年,中国现有医院36570所,基层卫生医疗机构977790所;其中三级医院3275所,二级医院10848所,一级医院12649所,未定级医院9798所[75]。平均每家医院每年新增约500 Tb的医疗健康相关数据,则全国每年将产生约16 Eb(17282 Pb)的数据。对这些大数据进行及时且深入的科学分析,将极大地促进中国医疗卫生行业的进步,并为维护国民健康及生命安全提供坚实的科学支撑。

当前,中国在医疗健康大数据的应用与共享方面尚未建立起完整的机制。政府、医院、研究机构和医疗企业等成为了这些数据的主要管理者,而数据共享通常限于这些实体之间,缺少向更广泛的专业团体和个人开放的共享原则和标准。此外,对于医疗保健大数据的分析和利用还不够系统和有序。目前的诊疗和医保数据以及互联互通的医疗数据,还未经过科学的严格处理,包括数据清洗、整合、完善和筛选,以达到科研所需的质量。没有经过精确处理的数据无法产生可靠的证据。同时,还缺少针对国家医疗健康的重大需求,对医疗大数据进行组

织化、系统化和深入的分析,以产生高质量的科学证据。这是目前面临的关键问题,也是利用大数据实现其潜力的核心目标。

3.2　更全面的预测因子是模型提升的关键

精准医疗代表了一种医疗革新,它根据个人的特定特征(遗传、环境暴露和生活方式等)来定制治疗方案和干预措施。这种治疗方法以患者自身的信息为基础,指导医疗诊断与疾病预测过程[76]。实现精准医疗的关键之一是采纳更全面的预测因子,这种全面性确保了模型能够考虑所有可能的影响因素,从而提高预测的准确性和可靠性。全面的因子能够提供更丰富的数据维度,从而增强模型对疾病发展动态的捕捉能力。在海量的医疗健康大数据下,挖掘关键指标的动态趋势成为可能。众所周知,病变是一个动态的过程,除了基线水平的预测因子,动态变化本身也构成了预测因子的一个类别。对于大多数预测因子,其动态变化趋势或幅度应具有重要的预测价值。指标的动态变化模式以及对结局的作用模式,将是一类极具潜力的预测因子,对疾病风险预测指标体系和模型具有重要的补充价值。近期有数项研究显示随访过程中的吸烟暴露、体重指数、饮酒等指标重复测量的纵向趋势与肺癌发生风险存在关联[77-79]。此外,这些预测因子之间的非线性、高维度、高阶交互作用是提升模型性能的关键。人工智能技术的优势在于能够揭示和利用这些复杂的高维度交互效应,从而增强预测模型的能力。因此,指标趋势特征的挖掘不仅是预测模型的一个重要组成部分,更是赋能这些模型以实现高效预测的关键步骤。

3.3　人工智能将赋能预测模型研究

机器学习和人工智能技术的飞速发展,使得构建更加精准和复杂的预测模型成为可能。Jarrett等[80]运用卷积神经网络提出Match-Net算法,利用ADNI(Alzheimer's disease neuroimaging initiative)研究的1737名患者的纵向数据,展现了阿尔茨海默病的动态预测能力,AUROC值达到了0.89。Li等[81]基于函数型主成分分析提取术后12个月内CEA、CA19-9以及CA125的纵向变化信息,并将其纳入随机生存森林模型进行生存预测;模型验证显示,纳入CEA、CA19-9和CA125的围手术期测量信息后,结直肠癌预后预测模型的预测性能改善。

然而,尽管机器学习和AI技术在预测模型研究中表现出了巨大的潜力,但也面临着众多挑战。这些局限性恰恰提示了未来医学人工智能发展方向。机器学习和AI模型的性能在很大程度上依赖于数据的质量和数量。如果训练数据存在偏差或错误,模型的预测结果很可能会受到影响[82]。同时模型需要大量的数据来进行训练,但数据隐私和安全问题也变得越来越突出,需要采取相应的措施保护用户数据。在对安全性要求极高的应用场景中,AI模型的"黑箱"特性必须被严格审查和评估[83]。机器学习和AI模型可解释性存在局限性,这给大范围应用带来了困扰。解释性AI(explainable AI,XAI)是未来的重要研究方向,一定程度上可协助人类理解AI模型的运行机制,这在医学场景下更为重要。

3.4　预测模型构建须遵循规范

在预测模型的发展应用日益广泛的背景下,学界已经认识到有必要建立一套相关标准来确保这些模型构建和报告的规范性和科学性。为此制定了《针对个体预后或诊断的多因素预测模型透明报告规范》(TRIPOD指南)[3]。TRIPOD指南作为一套报告规范,其目的在于增强已发表模型研究设计、分析、结果的透明度。虽然预测模型的技术不断进步,但对众多模型进行的严格评估显示,未遵循报告规范的模型研究不在少数[84]。模型的外部验证,甚至第三方团队的独立验证,是评价模型的表现和泛化能力的必要手段[85]。提倡外部验证试验的样本量应经估算以满足统计学基本要求[86]。在验证过程中,还应充分展示亚组和分层分析结果,以充分评估模型的稳健性和适用人群[87]。预测模型的开发,应当设计在先,报告需遵循规范,独立验证不可或缺。

3.5　动态预测模型是未来趋势

当前在疾病预防和临床实践中应用的模型以"静态模型"为主。医疗健康大数据提供了更丰富的个体水平的重复采集的健康相关指标,为"动态模型"的发展提供了必要的数据基础。动态预测模

型的发展,将有助于更早、更及时地识别疾病风险,赋能精准预防。中国在这一领域的方法学和应用研究有待加强。

综上所述,动态预测模型在精准预防中扮演着重要角色。随着医疗健康大数据的互联互通和共享共用机制的不断完善,统计学和人工智能新方法的不断涌现,挖掘出更丰富的预测因子、识别出更准确的作用模式、将生物医学规律融入重复测量的预测因子变化轨迹识别过程、开发更符合生物医学背景和实际场景的且具有可解释性的疾病风险预测模型等内容,将是预测模型方法学研究的重点方向,以期赋能共病共防、异病同防,最终实现个体化多疾病谱的精准预防。

参考文献(References)

[1] Grant S W, Collins G S, Nashef S A M. Statistical Primer: Developing and validating a risk prediction model[J]. European Journal of Cardio-Thoracic Surgery, 2018, 54(2): 203-208.

[2] Hippisley-Cox J, Coupland C, Robson J, et al. Derivation, validation, and evaluation of a new QRISK model to estimate lifetime risk of cardiovascular disease: Cohort study using QResearch database[J]. BMJ, 2010, 341: c6624.

[3] Collins G S, Reitsma J B, Altman D G, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement. The TRIPOD Group[J]. Circulation, 2015, 131(2): 211-219.

[4] Patzer R E, Kaji A H, Fong Y. TRIPOD reporting guidelines for diagnostic and prognostic studies[J]. JAMA Surgery, 2021, 156(7): 675-676.

[5] Moons K G M, Wolff R F, Riley R D, et al. PROBAST: A tool to assess risk of bias and applicability of prediction model studies: Explanation and elaboration[J]. Annals of Internal Medicine, 2019, 170(1): W1-W33.

[6] Davis S E, Lasko T A, Chen G H, et al. Calibration drift in regression and machine learning models for acute kidney injury[J]. Journal of the American Medical Informatics Association, 2017, 24(6): 1052-1061.

[7] Greene T, Li L. From static to dynamic risk prediction: Time is everything[J]. American Journal of Kidney Diseases, 2017, 69(4): 492-494.

[8] Tangri N, Inker L A, Hiebert B, et al. A dynamic predictive model for progression of CKD[J]. American Journal of Kidney Diseases, 2017, 69(4): 514-520.

[9] Pan Z C, Zhang R Y, Shen S P, et al. OWL: An optimized and independently validated machine learning prediction model for lung cancer screening based on the UK Biobank, PLCO, and NLST populations[J]. EBioMedicine, 2023, 88: 104443.

[10] Davis A M, Cifu A S. Lung cancer screening[J]. JAMA, 2014, 312(12): 1248.

[11] Tammemägi M C, Ruparel M, Tremblay A, et al. USPSTF2013 versus $PLCO_{m2012}$ lung cancer screening eligibility criteria (International Lung Screening Trial): Interim analysis of a prospective cohort study[J]. The Lancet Oncology, 2022, 23(1): 138-148.

[12] Wang F, Tan F W, Shen S P, et al. Risk-stratified approach for never- and ever-smokers in lung cancer screening: A prospective cohort study in China[J]. American Journal of Respiratory and Critical Care Medicine, 2023, 207(1): 77-88.

[13] Shen H B, Zhu M, Wang C. Precision oncology of lung cancer: Genetic and genomic differences in Chinese population[J]. NPJ Precision Oncology, 2019, 3: 14.

[14] Hong W, Li A, Liu Y H, et al. Clonal hematopoiesis mutations in patients with lung cancer are associated with lung cancer risk factors[J]. Cancer Research, 2022, 82(2): 199-209.

[15] Irajizad E, Fahrmann J F, Marsh T, et al. Mortality benefit of a blood-based biomarker panel for lung cancer on the basis of the prostate, lung, colorectal, and ovarian cohort[J]. Journal of Clinical Oncology, 2023, 41(27): 4360-4368.

[16] Bos L D J, Sjoding M, Sinha P, et al. Longitudinal respiratory subphenotypes in patients with COVID-19-related acute respiratory distress syndrome: Results from three observational cohorts[J]. The Lancet Respiratory Medicine, 2021, 9(12): 1377-1386.

[17] Haines R W, Zolfaghari P, Wan Y Z, et al. Elevated urea-to-creatinine ratio provides a biochemical signature of muscle catabolism and persistent critical illness after major trauma[J]. Intensive Care Medicine, 2019, 45(12): 1718-1731.

[18] Ye Q, Wang X, Xu X S, et al. Serial platelet count as a dynamic prediction marker of hospital mortality among septic patients[J]. Burns & Trauma, 2024,12: tkae016.

[19] Tu Y K, Tilling K, Sterne J A C, et al. A critical evaluation of statistical approaches to examining the role of growth trajectories in the developmental origins of health and disease[J]. International Journal of Epidemiology, 2013, 42(5): 1327-1339.

[20] Leffondré K, Abrahamowicz M, Regeasse A, et al. Statistical measures were proposed for identifying longitudinal patterns of change in quantitative health indicators[J]. Journal of Clinical Epidemiology, 2004, 57(10): 1049-1062.

[21] Laird N M, Ware J H. Random-effects models for longi-

tudinal data[J]. Biometrics, 1982, 38(4): 963-974.
[22] Nguena Nguefack H L, Pagé M G, Katz J, et al. Trajectory modelling techniques useful to epidemiological research: A comparative narrative review of approaches[J]. Clinical Epidemiology, 2020, 12: 1205-1222.
[23] Thurston R C, Chang Y F, Kline C E, et al. Trajectories of sleep over midlife and incident cardiovascular disease events in the study of women's health across the nation [J]. Circulation, 2024, 149(7): 545-555.
[24] Muthén B, Shedden K. Finite mixture modeling with mixture outcomes using the EM algorithm[J]. Biometrics, 1999, 55(2): 463-469.
[25] Feldman B J, Masyn K E, Conger R D. New approaches to studying problem behaviors: A comparison of methods for modeling longitudinal, categorical adolescent drinking data[J]. Developmental Psychology, 2009, 45(3): 652-676.
[26] Muthén B. Latent variable hybrids: Overview of old and new models[J]. Advances in latent variable mixture models, 2008, 1: 1-24.
[27] Muthén B, Asparouhov T. Growth mixture modeling with non-normal distributions[J]. Statistics in Medicine, 2015, 34(6): 1041-1058.
[28] Wei Y H. Review for Dynamic Prediction in Clinical Survival Analysis[J/OL]. [2023-11-27]. https://arxiv.org/abs/2311.15743.
[29] Little R J A, Rubin D B. Nonignorable missing-data models[M]. Hoboken: John Wiley & Sons, Inc., 2014.
[30] Wu L, Liu W, Yi G Y, et al. Analysis of longitudinal and survival data: Joint modeling, inference methods, and issues[J]. Journal of Probability and Statistics, 2012, 2012: 1-17.
[31] Wulfsohn M S, Tsiatis A A. A joint model for survival and longitudinal data measured with error[J]. Biometrics, 1997, 53(1): 330-339.
[32] Tsiatis A A, DeGruttola V, Wulfsohn M S. Modeling the relationship of survival to longitudinal data measured with error. applications to survival and CD4 counts in patients with AIDS[J]. Journal of the American Statistical Association, 1995, 90(429): 27.
[33] Parr H, Hall E, Porta N. Joint models for dynamic prediction in localised prostate cancer: A literature review [J]. BMC Medical Research Methodology, 2022, 22(1): 245.
[34] Chesnaye N C, Tripepi G, Dekker F W, et al. An introduction to joint models-applications in nephrology[J]. Clinical Kidney Journal, 2020, 13(2): 143-149.
[35] Zhang T H, Tang X C, Zhang Y, et al. Multivariate joint models for the dynamic prediction of psychosis in individuals with clinical high risk[J]. Asian Journal of Psychiatry, 2023, 81: 103468.
[36] Hennessey V, Novelo L L, Li J, et al. A Bayesian joint model for longitudinal DAS28 scores and competing risk informative drop out in a rheumatoid arthritis clinical trial[J/OL]. [2018-01-25]. https://arxiv.org/abs/1801.08628.
[37] Chen M H, Ibrahim J G, Sinha D. A new joint model for longitudinal and survival data with a cure fraction[J]. Journal of Multivariate Analysis, 2004, 91(1): 18-34.
[38] Chi Y Y, Ibrahim J G. Bayesian approaches to joint longitudinal and survival models accommodating both zero and nonzero cure fractions[J]. Statistica Sinica, 2007, 17: 445-462.
[39] Andrinopoulou E R, Nasserinejad K, Szczesniak R, et al. Integrating latent classes in the Bayesian shared parameter joint model of longitudinal and survival outcomes[J]. Statistical Methods in Medical Research, 2020, 29(11): 3294-3307.
[40] Garre F G, Zwinderman A H, Geskus R B, et al. A joint latent class changepoint model to improve the prediction of time to graft failure[J]. Journal of the Royal Statistical Society Series A: Statistics in Society, 2008, 171(1): 299-308.
[41] Li K, Luo S. Dynamic predictions in Bayesian functional joint models for longitudinal and time-to-event data: An application to Alzheimer's disease[J]. Statistical Methods in Medical Research, 2019, 28(2): 327-342.
[42] Li K, Luo S. Bayesian functional joint models for multivariate longitudinal and time-to-event data[J]. Computational Statistics & Data Analysis, 2019, 129: 14-29.
[43] Köhler M, Umlauf N, Greven S. Nonlinear association structures in flexible Bayesian additive joint models[J]. Statistics in Medicine, 2018, 37(30): 4771-4788.
[44] Köhler M, Umlauf N, Beyerlein A, et al. Flexible Bayesian additive joint models with an application to type 1 diabetes research[J]. Biometrical Journal Biometrische Zeitschrift, 2017, 59(6): 1144-1165.
[45] Rizopoulos D. Joint models for longitudinal and time-to-event data: With applications in R[M]. Boca Raton: CRC Press, 2012.
[46] Huang X, Li G, Elashoff R M, et al. A general joint model for longitudinal measurements and competing risks survival data with heterogeneous random effects[J]. Lifetime Data Analysis, 2011, 17(1): 80-100.
[47] Herle M, Micali N, Abdulkadir M, et al. Identifying typical trajectories in longitudinal data: Modelling strategies and interpretations[J]. European Journal of Epidemiology, 2020, 35(3): 205-222.
[48] Proust-Lima C, Dartigues J F, Jacqmin-Gadda H. Joint modeling of repeated multivariate cognitive measures and competing risks of dementia and death: A latent process and latent class approach[J]. Statistics in Medicine, 2016, 35(3): 382-398.
[49] 邱皓政. 潜在类别模型的原理与技术[M]. 北京: 教育科学出版社, 2008.

[50] Nylund K L, Asparouhov T, Muthén B O. Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study[J]. Structural Equation Modeling: A Multidisciplinary Journal, 2007, 14(4): 535-569.

[51] Larose C, Harel O, Kordas K, et al. Latent class analysis of incomplete data via an entropy-based criterion[J]. Statistical Methodology, 2016, 32: 107-121.

[52] Han J, Slate E H, Peña E A. Parametric latent class joint model for a longitudinal biomarker and recurrent events[J]. Statistics in Medicine, 2007, 26(29): 5285-5302.

[53] Proust-Lima C, Joly P, Dartigues J F, et al. Joint modelling of multivariate longitudinal outcomes and a time-to-event: A nonlinear latent class approach[J]. Computational Statistics & Data Analysis, 2009, 53(4): 1142-1154.

[54] Beunckens C, Molenberghs G, Verbeke G, et al. A latent-class mixture model for incomplete longitudinal Gaussian data[J]. Biometrics, 2008, 64(1): 96-105.

[55] Andrinopoulou E R, Rizopoulos D, Takkenberg J J,et al. Combined dynamic predictions using joint models of two longitudinal outcomes and competing risk data[J]. Statistical Methods In Medical Research, 2017, 26(4): 1787-801.

[56] Hatfield L A, Boye M E, Carlin B P. Joint modeling of multiple longitudinal patient-reported outcomes and survival[J]. Journal of Biopharmaceutical Statistics, 2011, 21 (5): 971-991.

[57] He B, Luo S. Joint modeling of multivariate longitudinal measurements and survival data with applications to Parkinson′s disease[J]. Statistical Methods in Medical Research, 2016, 25(4): 1346-1358.

[58] Andrinopoulou E R, Rizopoulos D, Takkenberg J J M, et al. Joint modeling of two longitudinal outcomes and competing risk data[J]. Statistics in Medicine, 2014, 33(18): 3167-3178.

[59] Taylor J M G, Yu M G, Sandler H M. Individualized predictions of disease progression following radiation therapy for prostate cancer[J]. Journal of Clinical Oncology, 2005, 23(4): 816-825.

[60] Rizopoulos D, Hatfield L A, Carlin B P, et al. Combining dynamic predictions from joint models for longitudinal and time-to-event data using Bayesian model averaging[J]. Journal of the American Statistical Association, 2014, 109(508): 1385-1397.

[61] Proust-Lima C, Taylor J M G. Development and validation of a dynamic prognostic tool for prostate cancer recurrence using repeated measures of posttreatment PSA: A joint modeling approach[J]. Biostatistics, 2009, 10(3): 535-549.

[62] Rizopoulos D. Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data[J]. Biometrics, 2011, 67(3): 819-829.

[63] Proust-Lima C, Séne M, Taylor J M G, et al. Joint latent class models for longitudinal and time-to-event data: A review[J]. Statistical methods in medical research, 2014, 23(1): 74-90.

[64] Rizopoulos D, Taylor J M G. Optimizing dynamic predictions from joint models using super learning[J]. Statistics in Medicine, 2024, 43(7): 1315-1328.

[65] Barrett J K, Sweeting M J, Wood A M. Dynamic risk prediction for cardiovascular disease: An illustration using the ARIC study[M]//Handbook of Statistics. Amsterdam: Elsevier, 2017: 47-65.

[66] McCrink L M, Marshall A H, Cairns K J. Advances in joint modelling: A review of recent developments with application to the survival of end stage renal disease patients[J]. International Statistical Review, 2013, 81(2): 249-269.

[67] Zheng Y Y, Heagerty P J. Partly conditional survival models for longitudinal data[J]. Biometrics, 2005, 61(2): 379-391.

[68] Keogh R H, Seaman S R, Barrett J K, et al. Dynamic prediction of survival in cystic fibrosis: A landmarking analysis using UK patient registry data[J]. Epidemiology, 2019, 30(1): 29-37.

[69] Yang Z J, Hou Y W, Lyu J J, et al. Dynamic prediction and prognostic analysis of patients with cervical cancer: A landmarking analysis approach[J]. Annals of Epidemiology, 2020, 44: 45-51.

[70] Yao Y, Li L, Astor B, et al. Predicting the risk of a clinical event using longitudinal data: The generalized landmark analysis[J]. BMC Medical Research Methodology, 2023, 23(1): 5.

[71] Bull L M, Lunt M, Martin G P, et al. Harnessing repeated measurements of predictor variables for clinical risk prediction: A review of existing methods[J]. Diagnostic and Prognostic Research, 2020, 4: 9.

[72] Chen Q, Tang B H, Song J Q, et al. Dynamic Bayesian network for predicting physiological changes, organ dysfunctions and mortality risk in critical trauma patients [J]. BMC Medical Informatics and Decision Making, 2022, 22(1): 119.

[73] Marini S, Trifoglio E, Barbarini N, et al. A Dynamic Bayesian Network model for long-term simulation of clinical complications in type 1 diabetes[J]. Journal of Biomedical Informatics, 2015, 57: 369-376.

[74] Orphanou K, Stassopoulou A, Keravnou E. DBN-extended: A dynamic Bayesian network model extended with temporal abstractions for coronary heart disease prognosis[J]. IEEE Journal of Biomedical and Health Informatics, 2016, 20(3): 944-952.

[75] 国家统计局. 2022年中国卫生健康统计年鉴[M]. 北京: 中国统计出版社, 2023.

[76] Allen B. The promise of explainable AI in digital health for precision medicine: A systematic review[J]. Journal of Personalized Medicine, 2024, 14(3): 277.
[77] Luu M N, Han M J, Bui T T, et al. Smoking trajectory and cancer risk: A population-based cohort study[J]. Tobacco Induced Diseases, 2022, 20: 71.
[78] You D F, Wang D H, Wu Y Q, et al. Associations of genetic risk, BMI trajectories, and the risk of non-small cell lung cancer: A population-based cohort study[J]. BMC Medicine, 2022, 20(1): 203.
[79] Bui T T, Han M J, Luu N M, et al. Cancer risk according to alcohol consumption trajectories: A population-based cohort study of 2.8 million Korean men[J]. Journal of Epidemiology, 2023, 33(12): 624-632.
[80] Jarrett D, Yoon J, van der Schaar M. Dynamic prediction in clinical survival analysis using temporal convolutional networks[J]. IEEE Journal of Biomedical and Health Informatics, 2020, 24(2): 424-436.
[81] Li C X, Zhao K, Zhang D F, et al. Prediction models of colorectal cancer prognosis incorporating perioperative longitudinal serum tumor markers: A retrospective longitudinal cohort study[J]. BMC Medicine, 2023, 21(1): 63.
[82] Averbuch T, Sullivan K, Sauer A, et al. Applications of artificial intelligence and machine learning in heart failure[J]. The European Heart Journal-Digital Health, 2022, 3(2): 311-322.
[83] Hunter D J, Holmes C. Where medical statistics meets artificial intelligence[J]. The New England Journal of Medicine, 2023, 389(13): 1211-1219.
[84] Fihn S D, Berlin J A, Haneuse S J P A, et al. Prediction models and clinical outcomes: A call for papers[J]. JAMA Network Open, 2024, 7(4): e249640.
[85] Collins G S, Dhiman P, Ma J, et al. Evaluation of clinical prediction models (part 1): From development to external validation[J]. BMJ, 2024, 384: e074819.
[86] Riley R D, Snell K I E, Archer L, et al. Evaluation of clinical prediction models (part 3): Calculating the sample size required for an external validation study[J]. BMJ, 2024, 384: e074821.
[87] Riley R D, Archer L, Snell K I E, et al. Evaluation of clinical prediction models (part 2): How to undertake an external validation study[J]. BMJ, 2024, 384: e074820.

Disease dynamic risk prediction modeling methods and precision prevention

SONG Yuxin[1], YE Qian[2], ZHAO Mengsheng[2], ZHANG Longyao[2], WEI Yongyue[1,3,4*]

1. Center for Public Health and Epidemic Preparedness & Response, Peking University, Beijing 100191, China
2. Department of Biostatistics, School of Public Health, Nanjing Medical University, Nanjing 211166, China
3. Department of Epidemiology and Biostatistics, School of Public Health, Peking University, Beijing 100191, China
4. Key Laboratory of Epidemiology of Major Diseases (Peking University), Ministry of Education, Beijing 100191, China

Abstract Dynamic disease risk prediction models are essential for precision prevention strategies. Over the last twenty years, there has been a surge in research focused on these models for precision prevention. However, widely used models(static models) often overlook the impact of changes in predictors over time on disease risk, leading to inevitable calibration drift. This paper reviewed modeling methods for dynamic risk prediction models and provided reference for their development. The conclusions are as follows: As healthcare big data becomes more interconnected and shared, and new methods of statistics and artificial intelligence emerge, the challenge lies in discovering richer predictors, in identifying more accurate modes of action, and in creating interpretable disease risk prediction models which align with biomedical contexts and practical scenarios, to enhance common prevention of common diseases and co-prevention of heterogeneous diseases and to achieve precision and personalized prevention across a spectrum of diseases. This will be a crucial focus for future research on predictive modeling methodologies.

Keywords prediction model; static model; dynamic prediction; precision prevention

(责任编辑　王微)