牛钢：虚拟临床试验“喂电子小人吃电子药”

牛钢 中科计算技术西部研究院研究员、图灵－达尔文实验室主任

去年ChatGPT爆火，我们也希望大模型能为医学领域提供更好的帮助，但最终发现，ChatGPT生成的是内容，但不一定是科学洞见。于是从2023年下半年开始，我们逐渐转向开发面向科学的人工智能模型（AI for science, AI4S）。而AI4S生成的不仅是新内容，也是新洞见。

人工智能怎么理解疾病

AI4S主要通过文献挖掘、多组学数据整合、医学影像分析、电子病历挖掘、生理学建模、患者异质性分析和因果推断等手段获取知识和规律并作出判断，从而理解疾病。从根本上说，我们需要的就是知识驱动型AI和数据驱动型AI。

什么是知识驱动型AI？以治疗狂犬病为例，假设一个人被感染病毒的狗咬伤后感染了狂犬病毒，又没有及时打疫苗，AI如何解决这个问题？

首先，AI会在除疫苗之外的所有狂犬病相关文献中挖掘知识颗粒，即特定研究方向的结构化多维信息复合体，然后再将每个知识颗粒用特定文献训练成可以感知特定内容的神经元。这些神经元组成超脑后，就可以把对内容的感知转化为可以解决问题的认知。超脑通过阅读狂犬病的临床病例继续训练，形成世界上最懂狂犬病的认知AI模型，再用该模型分别阅读所有已上市药物的说明书和相关文献，之后给每个药物打分。分数越高，越可能应用到当前这种临床场景，患者越可能从这些老药中直接获益。

如果我们不想让AI完成所有工作，希望专家参与到理解狂犬病机制和选择用药方案的过程中，该模型还可以搭建结构化的狂犬病知识库。专家通过知识库，在很短时间内就能理解疾病并确定方案，不需要再看上万篇文献，这不仅弥补了人脑的局限性，也提高了效率。

从真实患者到数字孪生

建立患者和健康人数字孪生是数据驱动的AI4S在人类健康上最关键的任务。数字孪生既能助力药物研发，促进精准用药、人群差异化、生产质控、药物重定位、靶点发现、药物组合、虚拟临床试验、分子发现和定量药理；也能助力临床医学，推动个性化医疗、精准预防、高效早诊、手术规划、治疗方案、多学科决策、远程医疗、健康管理和虚拟试药。

建立真实世界人类的数字孪生的基本思想主要有四点。第一，从真实世界采集的人类数据包含人与人之间特定方面的差异信息，基于这种差异信息要能在数亿人中精确定位特定人类个体；第二，基于上述数据提取的多维特征继承差异信息且排除噪声；第三，基于多维特征为每个人构建特定用途的数字孪生模型；第四，建立包含大规模自然人群的数字孪生“元宇宙”作为全新的健康基础设施，为了满足不同临床或保健需求，可以建立不同应用，实现不同功能，例如临床诊疗和新药研发。

基于这个模型，我们可以输入个人数据预测其生理、病理或药代动力学特征；输入患者个人疾病信息预测该患者的病因、病理、潜在预后、治疗方案及潜在靶点；输入药物及靶点信息预测潜在适应证，输出临床试验方案等。

需要注意的是，由于生成式模型本身依赖于数据的统计分布和变量之间的条件概率，因此需要进行巨量数据训练。然而，人类疾病数据天生就是“小数据”，尤其是罕见病。即便是癌症与自身免疫性疾病，也存在因病理复杂、疾病机制异质性强而导致的每个亚型数据并不多的问题。在这种情况下，盲目建立和使用大模型，对于在真实临床场景解决真实问题的作用就非常有限。

要解决这些问题，就要回到第一性原理，从最有价值的数据出发，建立能够针对小样本的AI4S模型。从疾病发生的底层逻辑来说，理解人类进化的方法是“第一性”的。而从数据角度来说，组学数据是“第一性”的，而组学数据内部DNA数据是“第一性”的。因此，谁能基于人类基因组DNA序列信息读出每个人更多的机制性定量信息，谁就能做出更好的数字孪生。

“电子药物”的开发

开发电子药物的前提是已经开发出患者和疾病的数字孪生。在此基础上，特定药物也需要建立数字孪生，之后才可以自由开展硅基的虚拟临床试验，探索药物的适用人群、新适应证、潜在耐药原因，以及联合用药方案的理性设计。那么如何建立药物的数字孪生，也就是所谓的“电子药物”？

一种策略是基于靶点和既往同类药物的所有知识，采用知识驱动的AI模型建立电子药物。例如建立知识库后建立真实作用机制（MOA）模型，提取生物标志物、药物敏感或耐药机制等，把这些信息转化为数字化标签，通过非监督方式在患者的数字孪生库中进行标注。标注过程可以看作虚拟临床试验，而标注的统计分布结果就是虚拟临床试验的结果。

第二种策略是利用靶基因的分子生物学与细胞生物学数据建立功能性模型和数字化标签，之后按照第一种策略中的标注和统计方法进行模拟。

第三种策略是通过不同疾病特征间接建立模型。例如CDK4/6抑制剂在Luminal B型乳腺癌治疗上获得成功，而对三阴性乳腺癌效果不佳，那么这个药物的机制可以被两种乳腺癌的差异所代表。如果可以根据特定组学数据将这种差异反映出来，并转化为评分，那么这种评分就能向其他癌症类型推广。以上这些工作完成，只要药物性质没有问题，临床试验想失败都难。

我们和上海市胸科医院教授陆舜合作，对肺腺癌免疫药物一线治疗做了两轮预测。第一轮单独采取肿瘤基因组数据预测，其中有两位患者预测错误。第二轮增加了胚系基因组数据后，所有患者全部预测正确。究其原因，胚系基因组编码了免疫系统先天的抑制状态，因此尽管患者肿瘤并未产生免疫抑制，但是T细胞很难浸润肿瘤组织，导致患者使用免疫药物无效。这表明，只有把胚系基因组和肿瘤基因组结合在一起，才能解释清楚肿瘤的大部分功能。

整体而言，上述电子药物建立的方法不仅可以预测PD-1/PD-L1单抗的疗效，更重要的是找到了PD-1/PD-L1在泛癌种中出现耐药现象的基本规律。摸清这个规律，我们就能明白是肿瘤的哪条信号通路导致了原发耐药，继而研发一个新的药物解决这个问题。利好的消息是，目前这个新药已经在开发当中。

（3～6版由本报记者陈祎琪、张思玮采写整理）

（上接第4版）

陈蕾：大知识模型驱动癫痫“老药新用”

通过语言大模型构建育龄期癫痫女性本体库，为医生提供标准化诊断的同时也促进患者正确了解癫痫诊疗常识。

通过大数据驱动，在癫痫治疗方法上我们也取得了创新突破。例如，国际上首创卵圆孔未闭封堵微创手术，控制耐药癫痫的有效率可达70%；建立从胃肠论治难治性癫痫的药物新方案；提出针对癫痫伴发多囊卵巢综合征的“三早方案”。

这些创新的背后是计算医学对医学高质量创新发展的推动。它的计算方法可用于了解人类疾病，数学、信息学和计算模型可用于为疾病的机制、诊断和治疗提供新见解，从而最终提高治疗效果。

在癫痫领域，运用计算医学构建通用大知识模型，形成多模态异构数据群并建立数字孪生患者数据库，进一步完善育龄期女性癫痫精准诊疗知识库，以药物信息评价认知模型评估药物的疗效－安全性/不良事件等信息，最终发掘“老药”在癫痫治疗中潜在的新用途和新范围。这一思路不仅可以避免新药研发的高昂投入，还能最大化利用现有药物的安全性和可及性优势。