基于电子健康记录的乳腺癌表型识别自然语言处理算法跨机构评估

Sicheng Zhou，理学硕士，健康信息学研究所，明尼苏达大学，美国明尼苏达州明尼阿波利斯；Nan Wang，统计学院，明尼苏达大学，美国明尼苏达州明尼阿波利斯；Liwei Wang，医学博士、哲学博士，人工智能与信息学研究部，梅奥诊所，美国明尼苏达州罗切斯特；Ju Sun，哲学博士，计算机科学与工程系，明尼苏达大学，美国明尼苏达州明尼阿波利斯；Anne Blaes，医学系，明尼苏达大学，美国明尼苏达州明尼阿波利斯；Hongfang Liu，哲学博士，人工智能与信息学研究部，梅奥诊所，美国明尼苏达州罗切斯特；Rui Zhang，哲学博士，计算健康科学部、外科系，明尼苏达大学，美国明尼苏达州明尼阿波利斯

摘要

研究目的：在模型开发过程中，临床大语言模型的泛化能力通常被忽视。本研究通过乳腺癌表型提取任务，评估基于 BERT 的临床自然语言处理模型在不同临床环境中的泛化能力。

材料与方法：我们从明尼苏达大学和梅奥诊所的电子健康记录中收集了两个乳腺癌患者临床语料库，并按照相同的指南进行标注。我们开发了三种类型的自然语言处理模型（即条件随机场、双向长短期记忆网络和 CancerBERT），用于从临床文本中提取癌症表型。我们采用不同的学习策略（模型迁移与本地训练），在不同测试集上评估模型的泛化能力。我们还评估了实体覆盖率得分与模型性能之间的关联。

结果：我们分别在 UMN 和梅奥诊所对 200 份和 161 份临床文档进行了人工标注。研究发现，与整体语料库相比，这两个研究机构的语料库在目标实体方面具有更高的相似性。在来自两个临床机构的独立测试集以及置换测试集上，CancerBERT模型的表现最佳。在一个研究机构开发并在另一个研究机构进一步微调的 CancerBERT模型，与使用本地数据开发的模型相比，表现合理（micro-F1：0.925 对 0.932）。

结论：结果表明，在三种临床自然语言处理模型中，CancerBERT 模型的学习能力和泛化能力最佳。研究发现，模型的泛化能力与语料库之间目标实体的相似性相关。

引言

随着电子健康记录（EHR）的普及率不断提高，人们开发了越来越多的自然语言处理（NLP）方法，以利用临床文本支持临床决策和临床研究 [1]。这些方法广泛应用于实时癌症病例识别 [2], 、医疗处方分类 [3], 以及从电子健康记录中自动提取心力衰竭信息 [4]等场景。已开发的自然语言处理方法主要可分为符号方法和统计方法 [5]。早期，符号方法在临床领域占据主导地位，因为它们能够满足许多临床应用的基本信息需求，而且不需要大量标注数据（这类数据需要投入大量人力）[6]。此外，其结果也易于解释 [6]。符号方法的一项主要缺点是可移植性不足 [1]。近年来，受益于大语言模型（基于 Transformer 的模型 [7]）的发展，以及临床领域标注数据的增加，统计方法迅速发展，并在不同的临床自然语言处理任务中取得了令人瞩目的性能。然而，模型性能与对这些大语言模型泛化能力的理解之间仍存在巨大差距 [8]。成功部署的临床自然语言处理系统通常基于单一医疗机构的数据开发；若应用于不同医疗机构，其性能可能有所差异[9]。不同医疗机构采用的电子健康记录平台、临床文档规则和惯例各不相同，即使记录高度相似的医疗事件，也可能产生不同的临床文本。临床语言变异既包括句法变异，也包括语义变异 [10], ，而这些变异可能在整个临床语料库中累积。在单一医疗机构开发的自然语言处理模型，其泛化能力会受到不同临床语料库间这些变异的何种影响，仍是一个重要的研究问题。

很少有研究探讨临床自然语言处理模型的泛化能力。例如，Sohn 等人评估了一种基于规则的自然语言处理系统的可移植性，该系统能够从临床队列中识别哮喘患者 [1]。该自然语言处理系统基于单一医院的队列开发，并使用外部队列进行评估；由于临床文档存在差异，该系统的性能显著下降 [1]。另一项研究开发了一种基于规则的自然语言处理系统，用于识别有胰腺癌家族史的患者，并使用来自两个临床机构的数据对其进行了测试 [11]。研究发现，对于非常具体的自然语言处理任务，只要规则保持简单且可根据新数据进行更新，基于规则的系统就具有可移植性。Liu 等人 [12] 评估了该自然语言处理系统在不同研究机构中检测吸烟状态的表现。该系统由机器学习模块和基于规则的模块组成。研究发现，要使自然语言处理系统具备可移植性，需要付出适度的努力，包括标注更多数据以进一步训练机器学习模块，以及根据新数据添加新规则。这些研究主要聚焦于评估基于规则的方法和传统机器学习方法的泛化能力；目前，尚缺乏探讨基于 Transformer 的模型在临床领域泛化能力的研究 [8]。

基于 Transformer 的模型在各种临床自然语言处理任务中表现出色，但开发这些模型需要大量计算资源和人力进行数据标注。此外，由于隐私问题，标注后的临床数据在大多数情况下无法共享，这给开发可泛化模型带来了障碍。在临床自然语言处理领域，了解基于 Transformer 的模型的泛化能力至关重要；如果模型

能够在不同临床机构之间实现泛化并具备可移植性，就可以避免训练类似模型，从而节省大量人力和计算资源。近期，我们开发了一种癌症专用语言模型 CancerBERT，旨在利用从明尼苏达大学 M Health Fairview（UMN）收集的乳腺癌患者临床文本，识别八种乳腺癌表型 [13]。CancerBERT 基于基于变换器的双向编码器表示（BERT）语言模型开发 [14], , 并在乳腺癌表型提取任务中表现出色。然而，CancerBERT模型仅使用来自 UMN 的电子健康记录语料库进行开发和评估。本研究旨在通过评估CancerBERT模型在从另一家机构——梅奥诊所（MC）收集的语料库上的表现，探究其泛化能力。我们还评估并比较了其他基准模型，包括条件随机场（CRF）模型和双向长短期记忆网络-条件随机场（BiLSTM-CRF）模型。我们通过一项临床信息抽取任务评估了这些模型的泛化能力。研究发现，模型的泛化能力与不同语料库中目标概念的相似性相关。本研究的贡献包括：

1. 我们 评估了语料库异质性对基于 BERT 的自然语言处理模型泛化能力的影响 .

2. 我们还构建了一个置换数据集，用于分析模型的稳健性。

3. 我们比较了在临床机构之间迁移模型的两种策略，即：i）直接迁移与 ii）持续微调。

方法与材料

研究概述

本研究已获 UMN 和 MC 的机构审查委员会 (IRB) 批准。**图 1**展示了本研究的流程。我们从两家机构（UMN 和 MC）的电子健康记录中收集了乳腺癌临床文本。MC 团队遵循 UMN 团队先前制定的同一标注指南 [13]，对其语料库进行了标注。基于 UMN 语料库开发的 $\text{CancerBERT}_{\text{UMN}}$ 模型，以及作为基准的 CRF_{UMN} 和双向长短期记忆网络$_{\text{UMN}}$ 模型，最初旨在从临床文本中提取八种乳腺癌表型（即激素受体类型、激素受体状态、肿瘤大小、肿瘤部位、癌症分级、组织学类型、癌症侧别和癌症分期*)*。我们使用来自 MC 的数据，对在 UMN 网站训练的模型进行了外部性能评估。此外，我们比较了自然语言处理模型的两种迁移学习策略：（1）在 MC 网站上继续训练（MC 修订模型）；（2）仅在 MC 网站上进行本地训练（MC 模型）。此外，我们还开展了两项额外实验，以探究与作为基线模型的传统双向长短期记忆网络-条件随机场和条件随机场模型相比，基于BERT的模型在模型鲁棒性方面的优势。

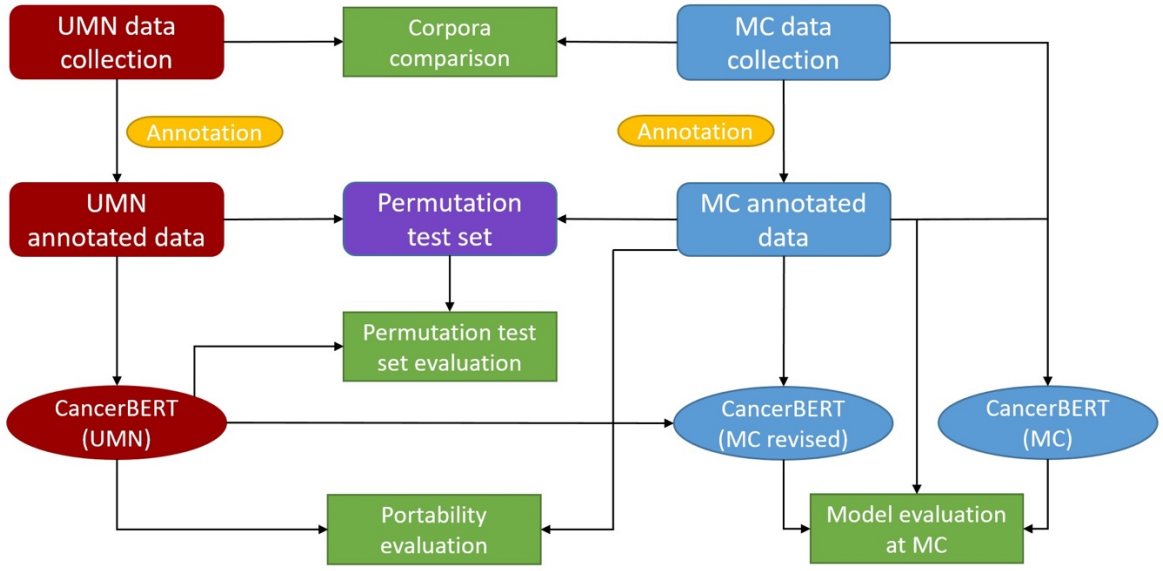


图 1。研究流程。数据从 UMN 和 MC 收集并进行标注。UMN 模型在 MC 数据上进行了外部评估。随后，UMN 模型在 MC 数据上进一步修订，并作为比较对象进行评估。

数据来源

本研究使用的数据来自两家临床机构的电子健康记录，即 UMN 临床数据仓库和 MC。UMN 的电子健康记录数据包含 2001 年至 2018 年间 17,970 名乳腺癌患者的健康记录。MC 的电子健康记录数据包含 2000 年至 2020 年间 54,050 名乳腺癌患者的记录。

人工标注与语料库比较

标注方案已在先前的研究中介绍 [13], 我们使用相同的方案对从 MC 电子健康记录中提取的临床文本进行了标注。我们随机抽取了 100 名乳腺癌患者，并为每位患者随机抽取一份诊断日期后 90 天内的临床记录和一份病理报告进行标注。最终纳入 81 份临床记录和 80 份病理报告。两名标注员共同完成了 10% 文本的标注，标注者间一致性（IAA）以科恩卡帕系数评估，结果为 0.80。随后，一名标注员完成了所有标注任务。

为更好地了解不同部位的语料库对自然语言处理方法性能的影响，我们比较了两个部位标注的临床文本的差异。我们汇总并比较了基本的语料库统计数据，例如句子数量、句子的平均长度和词元的唯一数量。此外，我们将两个语料库表示为归一化的词频-逆文档频率（TF-IDF）向量，并计算其余弦相似度 [1]，以衡量两个语料库之间的相似度。每个语料库向量都包含该语料库中唯一术语的平均 TF-IDF 值。术语 t 的平均 TF-IDF 值定义如公式 1 所示，其中 N 是语料库中的文档总数，$TF_i(t)$ 是词元 t 在第 i 个文档中的出现频率除以

第 *i* 个文档中的词元总数，而 *IDF(t)* 则是 *log(N)* 除以包含词元 *t* 的文档数量。

$$TF - IDF(t) = \sum_{i=1}^{N} TF_i(t) * \frac{IDF(t)}{N} \qquad (1)$$

除了语料库层面的相似度外，我们还计算了两个语料库中各乳腺癌表型类别的相似度（表型相似度）。同样，每种乳腺癌表型均由一个归一化的词频-逆文档频率向量表示，该向量包含各乳腺癌表型类别下唯一术语的平均词频-逆文档频率值。我们针对两个语料库中的同一乳腺癌表型类别计算了余弦相似度。

乳腺癌表型提取任务自然语言处理模型的开发

本研究评估了三种类型机器学习模型的泛化能力，即条件随机场、双向长短期记忆网络-条件随机场和 CancerBERT [13] 模型。条件随机场和双向长短期记忆网络-条件随机场是传统机器学习模型，已被证明在临床领域的命名实体识别（NER）任务中有效 [15-17]。对于条件随机场和双向长短期记忆网络基线模型，我们尝试将 GloVe（基于维基百科训练）[18] 和 Word2vec（基于 Google 新闻训练）嵌入 [19] 作为输入。CancerBERT模型基于 BERT语言模型开发，在我们的乳腺癌表型提取任务中，其性能优于条件随机场和双向长短期记忆网络-条件随机场模型 [13]。

乳腺癌表型提取模型的可移植性评估

MC 标注数据被划分为训练集（60%）、开发集（10%）和测试集（30%）。在本研究中，我们评估了以下三组模型在 MC 语料库测试集上的表现。

1. 最初仅使用 UMN 语料库训练的模型 [13]。这些 UMN 模型（条件随机场$_{UMN}$、双向长短期记忆网络-条件随机场$_{UMN,}$ CancerBERT$_{UMN_397}$、CancerBERT$_{UMN_997}$）在 MC 数据（测试集）上进行评估，以测试它们通过从临床文本中提取乳腺癌表型这一任务所展现的可移植性。所有其他模型（例如上一节介绍的 CancerBERTMC$_{_refined}$ 和 CancerBERTMC$_{_397}$ ）也进行了测试，以作比较。2. 在 MC 语料库上继续微调 UMN 模型，包括条件随机场 $_{MC\ 精调,}$ 双向长短期记忆网络-条件随机场$_{MC_refined,}$ 以及 CancerBERT$_{MC_refined}$。随后使用 MC 标注数据进一步微调 CancerBERT$_{UMN_397}$ 模型，以获得 MC 精调的 CancerBERT$_{MC_refined}$ 模型。为精调 UMN 模型，我们使用训练集和开发集微调所有在 UMN 本地训练的模型，从而获得相应的 MC 精调模型。3. 仅使用 MC 语料库训练的 CancerBERT 模型（不使用 UMN 数据）. 按照相同步骤 [13], , 我们开发了另一个 CancerBERT$_{MC_397}$。该模型仅在 MC 语料库（约包含 500 万份临床文档）上进行预训练，并在 MC 数据的训练集和开发集上进行微调，作为基准模型。

乳腺癌表型提取可以表述为命名实体识别任务，并采用表型级别（实体名称）评估。我们使用微平均（对每个样本赋予相同权重）F1 和宏平均（对每个类别赋予相同

权重）F1 作为指标，并按照 i2b2standard[20] 同时采用精确匹配和宽松匹配。F1 分数是精确率和召回率的调和平均值，在数据不平衡时能更好地评估模型性能。所有实验均重复进行 10 次，并采用单因素方差分析检验比较不同模型的性能。

UMN 模型的置换测试集评估

置换数据集用于模拟数据中的变化，并能更好地了解模型在遇到新数据时的泛化能力 [8, 21]。因此，为了探究数据差异如何影响模型的性能，我们手动创建了置换测试集，并在其中评估了双向长短期记忆网络-条件随机场$_{UMN}$、条件随机场$_{UMN}$ 和 CancerBERT$_{UMN}$ 模型。由于实体是命名实体识别任务的核心，因此我们采用了实体置换。为创建置换测试集，我们首先找出梅奥诊所标注数据中所有不在 UMN 标注数据中的唯一实体；然后，从梅奥诊所的唯一实体中随机抽取实体，并用同类别的抽取实体替换 UMN 测试集中的实体。这样一来，置换数据集便包含了目标实体与相关上下文的新组合。在置换测试集上评估模型，可以揭示模型学习实体上下文信息而非记忆实体的能力。所有使用 UMN 数据开发的模型均使用置换测试集进行了评估。评估方案与上一节介绍的模型可移植性评估相同。

基于实体覆盖率（ECR）的模型泛化能力评估

机器学习模型的可移植性通常通过测量模型在不同测试集上的标准指标（例如 F1、精确率、召回率）的变化来评估。然而，这种方法只能提供高层次的见解，无法对可移植性进行细粒度分析 [22]。为从不同角度分析模型的可移植性，我们将先前研究提出的 ECR 改用于我们的命名实体识别任务 [22]。ECR 衡量测试集中的目标实体在训练集中以相同类别出现过的程度。其定义如公式（2）和（3）所示：

$$ECR(e_i) = \begin{cases} 0 & , \quad C = 0 \\ \sum_{k=1}^{K} \frac{\#(e_i^{tr,k})\#(e_i^{te,k})}{C^{tr}} / C^{te}, & otherwise \end{cases} \qquad (2)$$

其中 $C^{tr} = \sum \#\mathrm{P}_{k=1}^{K} \quad e_i^{tr\,k}$,Q 和 $C^{te} = \sum \#\mathrm{P}e_i^{te\,k},\mathrm{Q}_{k=1}^{K}$。$e_i$ 表示一个测试实体，$\#(e_i^{tr,k})$ 是训练集中标签为 k 的实体数量，$\#\mathrm{P}e_i^{te\,k},\mathrm{Q}$ 是测试集中标签为 k 的实体数量。ECR 分数介于 0 到 1 之间，表示预测一个实体的难度（从容易到困难）[22]。

在本研究中，我们计算了 UMN 标注数据测试集中实体的 ECR 分数。我们进一步根据实体的 ECR 值将测试集中的实体划分为不同组别，即 0<= ECR <0.33、0.33<= ECR <0.67、0.67 <= ECR<1 和 ECR = 1。我们在 UMN 测试集上评估了 UMN 模型（即仅基于 UMN 数据开发的 CancerBERT$_{UMN_397}$、双向长短期记忆网络-条件随机场$_{UMN}$ 和条件随机场$_{UMN}$ 模型），并探究了 ECR 值与模型在相应实体上的表现之间的关系。

结果

UMN 与梅奥诊所语料库比较

从多个角度比较了 UMN 和梅奥诊所的语料库（**表 1**）。总体而言，UMN 语料库中每位患者对应的临床文档更多（190.5 篇 vs 140.8 篇），且每篇临床文档的平均长度短于梅奥诊所（259 个词元 vs 361 个词元）。对两个部位的临床文本中标注的乳腺癌表型数量进行了统计。括号中的数字表示每种表型实体所包含的唯一术语数。UMN 数据中标注了更多与乳腺癌表型相关的实体。UMN 和梅奥诊所的标注者间一致性评分分别为 0.91 和 0.80。还使用 TF-IDF 计算了两个语料库中表型的相似度。总体而言，乳腺癌表型的相似度分数高于整个语料库（表型平均值：0.9088；整体：0.5411）。

表 ***1****– UMN* 和梅奥诊所的标注病理报告和临床文本统计数据。括号中的数字表示每个表型概念的唯一术语数量。表型相似度指各乳腺癌表型类别在不同语料库之间的余弦相似度。

	UMN	**MC**	**表型相似度**
数量 of 已标注文档	200	200	NA
句子数量	10452	6165	NA
词元数量	266079	110,980	NA
唯一词元数量	9151	5735	NA
平均值 数量 of 句子中的词元	25.5	18.0	NA
激素受体类型	1673 (29)	417 (18)	0.9923
激素受体状态	436 (14)	178 (12)	0.9815
肿瘤大小	540 (305)	393 (262)	0.8076
肿瘤部位	329 (173)	187 (135)	0.7399
癌症分级	271 (15)	234 (23)	0.8810
癌症侧别	1192 (4)	1846 (10)	0.9833
癌症分期	173 (38)	207 (55)	0.9965
组织学类型	1070 (95)	726 (77)	0.8880

基于机器学习的模型和 BERT模型的可移植性评估

两种经典命名实体识别模型（即条件随机场和双向长短期记忆网络-条件随机场）的可移植性评估结果（严格匹配和宽松匹配的 F1 分数）见**表2**。模型以两种词嵌入作为输入（GloVe 维基百科 60 亿词 [18]和 Word2Vec Google News [19]），并使用 MC 测试集进行评估。结果展示了直接从 UMN 网站获取的模型（条件随机场$_{UMN}$、双向长短期记忆网络-条件随机场$_{UMN}$），以及在 MC数据上进一步微调的模型（条件随机场$_{MC_refined}$、双向长短期记忆网络-条件随机场$_{MC_refined}$）的性能。

表*2– CRF* 和双向长短期记忆网络-条件随机场模型的可移植性评估结果（严格匹配（宽松匹配）*F1* 分数）。请注意，下标为*"UMN"*的模型表示仅在 *UMN* 语料库上训练的模型，而下标为*"MC_refined"*的模型表示在 *MC* 语料库上进行连续微调的 *UMN* 模型。

Word 嵌入	GloVe 维基百科 60 亿词		GloVe 维基百科 60 亿词		Word2vec Google News		Word2vec Google News	
模型	条件随机场UMN 随机场	条件 MC 精调	双向长短期记忆网络 -CRF UMN	双向长短期记忆网络 -CRF MC 精调	条件随机场UMN 随机场	条件 MC 精调	双向长短期记忆网络 -CRF UMN	双向长短期记忆网络 -CRF MC 精调
激素受体类型	0.939 (0.941)	0.944 **(0.954)**	0.925 (0.925)	0.943 **(0.954)**	0.945 (0.948)	**0.948** (0.951)	0.939 (0.939)	0.918 (0.929)
激素受体状态	0.527 (0.527)	0.542 (0.542)	0.794 (0.794)	**0.876*** **(0.876*)**	0.529 (0.529)	0.491 (0.491)	0.867 (0867)	0.837 (0.837)
肿瘤大小	0.224 (0.294)	0.694 (0.749)	0.392 (0.509)	**0.711*** **(0.813*)**	0.244 (0.350)	0.663 (0.724)	0.367 (0.421)	0.592 (0.738)
肿瘤部位	0.053 (0.136)	0.303 (0.600)	0.266 (0.400)	0.472 (0.758)	0.044 (0.128)	0.321 (0.596)	0.205 (0.323)	**0.361*** **(0.668*)**
肿瘤分级	0.647 (0.681)	0.881 **(0.925*)**	0.903 (0.903)	**0.916*** (0.916)	0.647 (0.685)	0.861 (0.913)	0.849 (0.849)	0.881 (0.881)
肿瘤侧别	0.846 (0.846)	0.935 (0.935)	0.882 (0.882)	**0.952*** **(0.952*)**	0.853 (0.853)	0.934 (0.934)	0.874 (0.874)	0.930 (0.930)
癌症分期	0.773 (0.773)	0.868 (0.868)	0.578 (0.578)	**0.891*** **(0.891*)**	0.632 (0.632)	0.873 (0.873)	0.593 (0.593)	0.838 (0.838)
组织学 type	0.829 (0.896)	**0.937** (0.964)	0.847 (0.917)	0.934 (0.959)	0.845 (0.907)	0.934 (**0.965**)	0.839 (0.921)	0.899 (0.917)
宏平均 F1 平均值	0.605 (0.637)	0.763 (0.817)	0.698 (0.738)	**0.837*** **(0.889*)**	0.593 (0.629)	0.753 (0.806)	0.691 (0.723)	0.782 (0.842)
F1 微平均 平均值	0.803 (0.822)	0.883 (0.905)	0.785 (0.814)	**0.893*** **(0.922*)**	0.804 (0.825)	0.880 (0.903)	0.777 (0.802)	0.853 （0.885）

注：分数为 *10* 次运行后的平均值，粗体文本表示最高性能，* 表示在统计学上高于其他方法（置信区间：*0.95*）。

基于BERT的模型的可移植性评估结果（严格匹配（宽松匹配）F1 分数）见**表3**。评估了在 UMN 开发的 CancerBERT模型（$CancerBERT_{UMN}$），以及两个基准基于BERT的模型（BERT-large Origin [23] 和 BlueBERT [14]）。$CancerBERT_{UMN}$ 模型有两个变体：一个在词汇表中包含 997 个基于词频的自定义词（$CancerBERT_{UMN_997}$），另一个包含 397 个基于知识的自定义词（$CancerBERT_{UMN_397}$）。每个 UMN 模型（**表3**，*UMN*列）均直接

在 MC 测试集上进行了评估。随后，使用 MC数据对这些模型进一步微调，得到 MC 精调模型（**表3**，*MC* 精调列），并再次进行评估。此外，还评估了仅使用 MC 语料库训练的 CancerBERT$_{MC_397}$ 模型作为对照，以全面了解采用不同迁移学习策略开发的模型的表现。

表*3*– 基于*BERT*的模型在 *MC* 测试数据集上的可移植性评估结果（严格匹配（宽松匹配）*F1* 分数）。*“UMN”*列包含仅在 *UMN* 语料库上训练的模型，*“MC_refined”*列包含在 *MC* 语料库上进行微调的模型。*“仅 MC”*列表示仅在 *MC* 语料库上训练的模型。

实体 type	BERT-large Origin		BlueBERT（PubMed+MIMIC III）		CancerBERT UMN_997		CancerBERT UMN_397		CancerBERT 梅奥诊所_397
	UMN	MC 精炼的	UMN	MC 精炼的	UMN	MC 精炼的	UMN	MC 精炼的	仅 MC
激素受体 type	0.926 (0.935)	0.967 (0.969)	0.897 (0.911)	0.977 (0.977)	0.923 (0.963)	0.984 (0.988)	0.942 (0.947)	0.975 (0.981)	**0.993*** **(0.993*)**
激素受体状态	0.816 (0.816)	0.910 (0.910)	0.897 (0.897)	0.932 (0.932)	0.842 (0.842)	0.901 (0.901)	0.819 (0.819)	0.926 (0.926)	**0.943*** **(0.943*)**
肿瘤 size	0.633 (0.737)	0.837 (0.903)	0.440 (0.648)	0.839 (0.915)	0.595 (0.664)	0.765 (0.813)	0.745 (0.781)	**0.864** (**0.928***)	0.862 (0.907)
肿瘤 site	0.666 (0.739)	0.609 (0760)	0.308 (0.671)	0.590 (0.790)	0.186 (0.742)	**0.733*** (0.792)	0.709 (0.759)	0.601 (0.786)	0.661 (**0.832***)
肿瘤分级	0.827 (0.827)	0.909 (0.922)	0.886 (0. 886)	0.859 (0.939)	0.869 (0.869)	0.891 (0.891)	0.846 (0.846)	**0.927*** **(0.943*)**	0.863 (0.933)
肿瘤侧别	0.896 (0.896)	0.928 (0.928)	0.936 (0.936)	0.954 (0.954)	0.928 (0.928)	0.939 (0.939)	0.903 (0.903)	0.959 (0.959)	**0.962** **(0.962)**
癌症分期	0.774 (0.774)	0.934 (0.934)	0.799 (0.799)	**0.953** (**0.953**)	0.806 (0.806)	0.870 (0.870)	0.829 (0.829)	0.949 (0.949)	0.950 （0.950）
组织学类型	0.793 （0.888）	0.926 （0.950）	0.794 （0.902）	0.931 (0.958)	0.828 (0.923)	0.849 (0.922)	0.815 (0.914)	0.934 (0.965)	**0.950*** **(0.981*)**
宏平均值	0.724 (0.829)	0.874 (0.908)	0.744 (0.831)	0.879 (0.927)	0.747 (0.842)	0.867 (0.889)	0.828 (0.849)	0.892 (0.930)	**0.898*** (**0.932**)
微平均值	0.843 (0.877)	0.905 (0.922)	0.817 (0.876)	0.917 (0.943)	0.829 (0.886)	0.903 (0.925)	0.864 (0.906)	0.925 (0.947)	**0.932*** **(0.952*)**

注：分数为基于 *10* 次运行得出的平均值，粗体文字表示最高性能，* 表示在统计学上高于其他方法（置信区间：*0.95*）。

根据**表2**和**表3**，与在原始 UMN 测试集上评估的模型表现相比，所有模型的表现均出现了不同程度的下降。而在 MC 数据上进一步微调模型后，其表现显著提升。

置换数据集评估

对于每种类型的模型，我们选择表现最佳的模型，在置换数据集上进行评估。**表4**展示了 各模型的表现，以及与其在常规测试集上的表现相比，表现（F1 分数）的变化幅度。结果表明，与其他模型相比，$\text{CancerBERT}_{\text{UMN_397}}$ 模型的性能下降幅度最小，这说明它在处理来自其他机构的新数据时具有较强的鲁棒性。

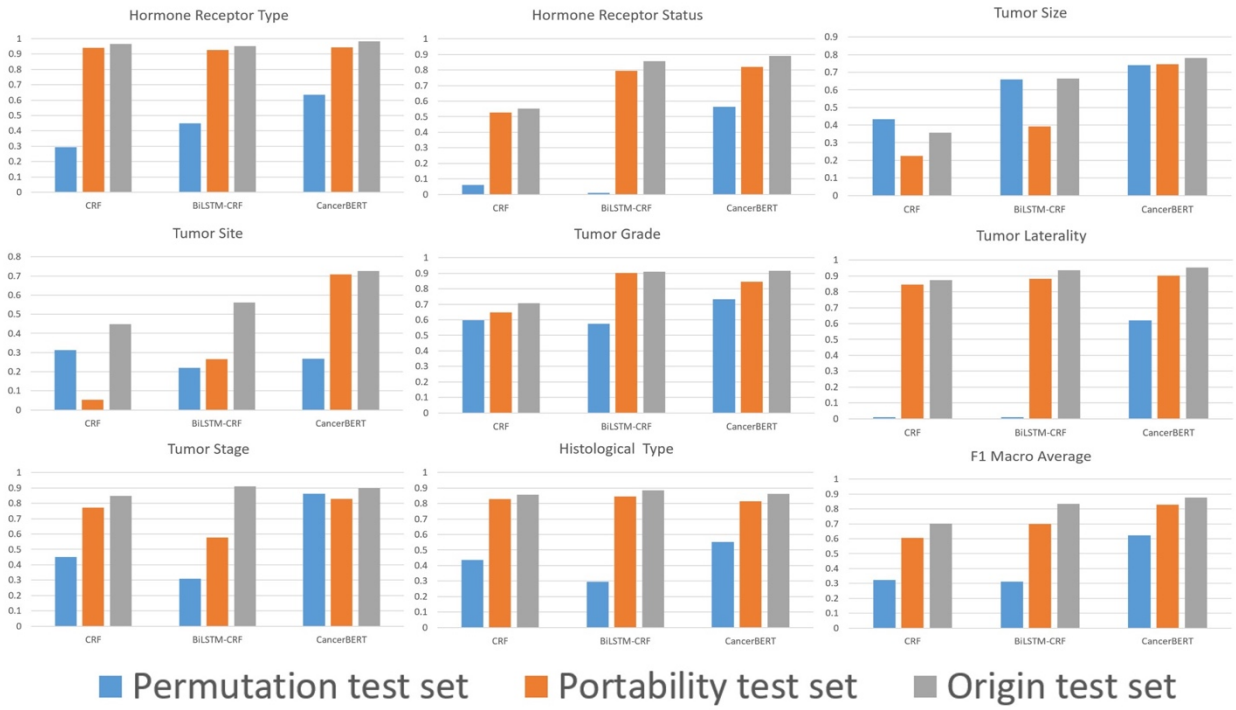
表*4*–置换数据集评估结果

模型 实体	条件随机场 UMN	变化后的 F1 分数	双向长短期记忆网络-条件随机场 UMN	变化后的 F1 分数	CancerBERT UMN_397	变更后的 F1 分数
激素 受体类型	0.295 (0.646)	-0.649 (-0.308)	0.448 (0.559)	-0.495 (-0.395)	0.634 (0.842)	**-0.341*** **(-0.139*)**
激素 受体 状态	0.061 (0.061)	-0.481 (-0.481)	0.000 (0.000)	-0.876 (-0.876)	0.564 (0.564)	**-0.362*** **(-0.362*)**
肿瘤大小	0.434 (0.582)	-0.260 (-0.167)	0.659 (0.832)	**-0.052*** **(0.019*)**	0.742 (0.921)	-0.122 (0.007)
肿瘤部位	0.314 (0.568)	**0.110*** **(-0.018*)**	0.220 (0.615)	-0.252 (-0.143)	0.267 (0.736)	-0.334 (-0.050)
肿瘤分级	0.596 (0.596)	-0.285 (-0.329)	0.575 (0. 575)	-0.341 (-0.341)	0.733 (0.733)	**-0.194*** **(-0.210*)**
肿瘤 侧别	0.001 (0.001)	-0.934 (-0.934)	0.000 (0.000)	-0.952 (-0.952)	0.620 (0.620)	**-0.339*** （**-0.339***）
癌症分期	0.450 （0.450）	-0.418 （-0.418）	0.310 （0.310）	-0.581 （-0.581）	0.863 (0.863)	**-0.086*** **(-0.086*)**
组织学 type	0.435 (0.794)	-0.502 (-0.170)	0.296 (0.718)	-0.638 (-0.241)	0.552 (0.818)	**-0.382*** **(-0.147*)**
宏平均 F1 平均值	0.323 (0.462)	-0.440 (-0.355)	0.313 (0.451)	-0.524 (-0.438)	0.622 (0.761)	**-0.270*** **(-0.169*)**
F1 微平均 平均值	0.354 (0.558)	-0.529 (-0.347)	0.210 (0.328)	-0.683 (-0.594)	0.616 (0.722)	**-0.309*** **(-0.225*)**

注：粗体文本表示 *F1* 分数变化最小。结果为基于 *10* 次运行得出的平均分数，* 表示 *F1* 分数的变化在统计学上低于其他方法（置信区间：*0.95*）。

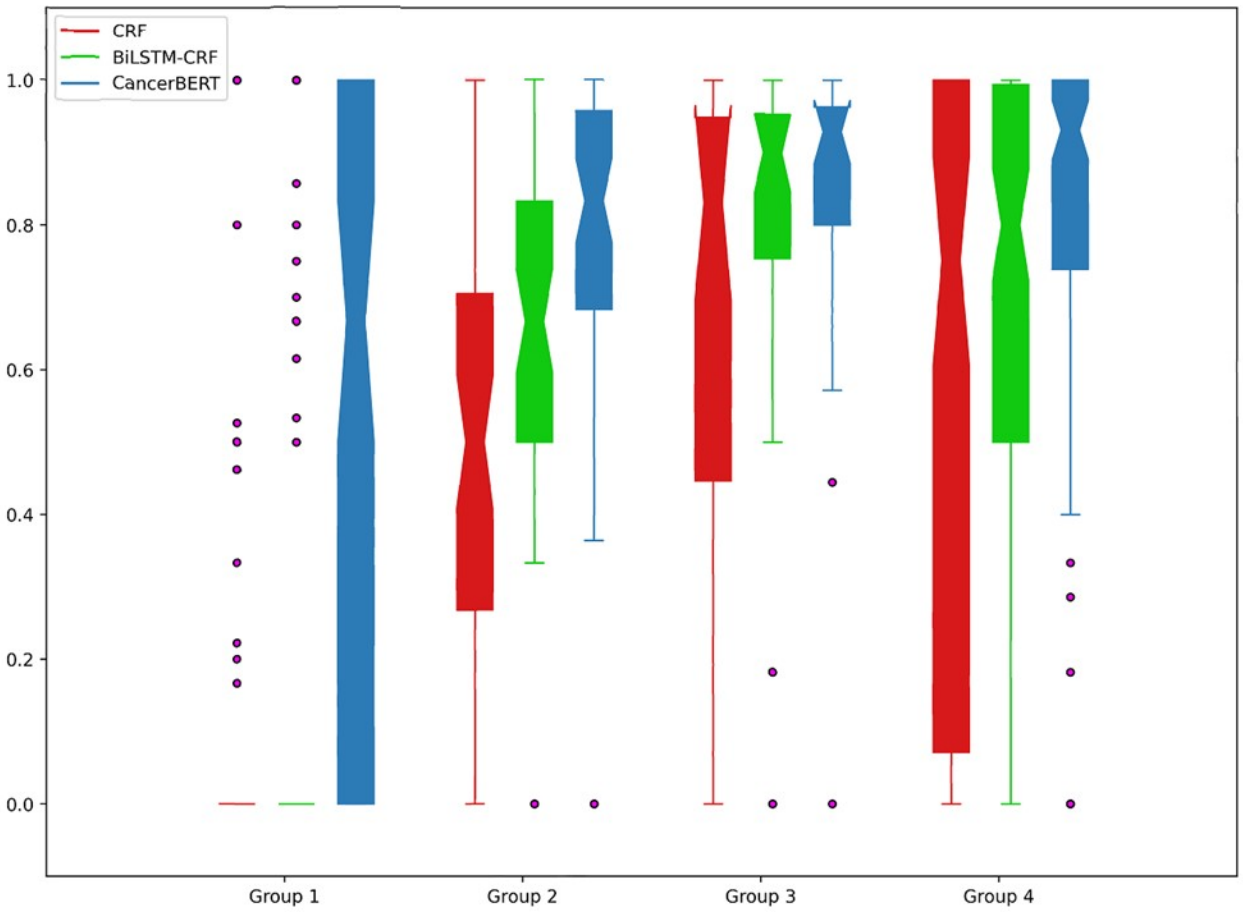
图2展示了各模型在不同测试集上的表现。原始测试集为 UMN 测试集，而可移植性测试集为 MC 测试集。所有模型均为仅使用 UMN 数据训练的 UMN 模型（条件随机场$_{UMN}$、双向长短期记忆网络-条件随机场$_{UMN}$ 和 CancerBERT$_{UMN_397}$）。CancerBERT$_{UMN_397}$ 模型在不同测试集上的表现最为一致。

图 2。CRF$_{UMN}$、双向长短期记忆网络-条件随机场$_{UMN}$ 和 CancerBERT$_{UMN_397}$ 模型在不同测试集上的表现。原始测试集为 UMN 测试集，可移植性测试集为 MC 测试集。所有模型均为仅使用 UMN 数据训练的 UMN 模型。



不同 ECR 组中各类模型对实体的表现见**图 3**：1) 0<= ECR <0.33，2) 0.33<=ECR <0.67，3) 0.67 <= ECR<1 和 4) ECR = 1 。三类模型在第 3 和第 4 组中的表现都较好。与另外两种模型相比，CancerBERT$_{UMN_397}$ 模型在第 1 和第 2 组中的表现显著更好。

图 3。CRF$_{UMN}$、双向长短期记忆网络-条件随机场$_{UMN}$ 和 CancerBERT$_{UMN_397}$ 模型对不同 ECR 组中实体的表现。第 1 组：0<= ECR <0.33；第 2 组：0.33<= ECR <0.67；第 3 组：0.67 <= ECR<1 ；第 4 组：ECR = 1。



讨论

电子健康记录中的临床文本不同于通用语言，包含许多专业短语、医学行话、首字母缩略词和缩写。不同研究机构的电子健康记录中的临床文本，其语义和上下文信息通常差异很大，在自然语言处理算法的开发过程中应予以充分考虑。从多个角度对 UMN 和梅奥诊所的语料库进行了比较。UMN 的标注数据中识别出的独特表型更多（673 对 592）。梅奥诊所语料库中，临床文本的乳腺癌表型密度更高（每 100 个词元中有 3.7 个表型，对比每 100 个词元中有 2.1 个表型）。乳腺癌表型之间的相似度显著高于两个语料库之间的相似度（0.9088 对 0.5411）；先前的一项研究也发现了类似现象 [1]。这表明，临床医生可能会使用相似的临床语言来描述临床文本中的特定医学概念，这为在不同临床机构之间迁移自然语言处理模型奠定了基础。

本研究主要探讨了基于BERT的模型（CancerBERT模型）以及另外两种经典机器学习模型，即条件随机场和双向长短期记忆网络-条件随机场的泛化能力。模型泛化能力的一项重要指标是模型的可移植性。我们通过癌症表型提取任务评估了模型的可移植性。结果显示，直接在 MC 测试数据上评估 UMN 模型时，只有CancerBERT模型取得了较为理想的表现，表明与双向长短期记忆网络-条件随机场和条件随机场模型相比，它们在可移植性方面更具优势。在基于

MC数据对模型进行优化后，它们的整体表现也最佳，而且更加稳定：CancerBERT模型的微平均F1分数平均下降了 0.067（$CancerBERT_{MC_refined}$ 与 $CancerBERT_{UMN}$相比）；条件随机场和双向长短期记忆网络-条件随机场的分数分别下降了 0.078（条件随机场$_{MC_refined}$ 与条件随机场$_{UMN}$相比）和 0.092（双向长短期记忆网络-条件随机场$_{MC_refined}$ 与双向长短期记忆网络-条件随机场$_{UMN}$相比）。尽管使用 MC数据从头开始训练的 $CancerBERT_{MC_397}$ 模型取得了最佳整体表现，但 $CancerBERT_{MC_refined}$ 模型的表现与 $CancerBERT_{MC_397}$ 模型相当（F1 分数仅低 0.007）。这表明，基于BERT的模型只需少量标注数据进行微调，无需使用大量语料库从头训练，就能以极低的成本迁移至另一家临床机构，同时保持相对较高的表现。从实际角度来看，在我们的数据集中，$CancerBERT_{MC_397}$ 模型平均每位患者可识别出 37.1 个表型，而 $CancerBERT_{MC_refined}$ 模型平均每位患者识别出 36.6 个表型（少 0.44 个）。相比之下，表现最佳的条件随机场$_{MC_refined}$ 和双向长短期记忆网络-条件随机场$_{MC_refined}$ 模型平均每位患者分别识别出 31.1 个和 34.0 个表型，分别少 6 个和 3.1 个。我们计算了两个语料库中癌症表型相似度与表现下降幅度之间的相关性（分别在 UMN 测试集上评估 UMN 模型，以及直接在 MC 测试集上评估 UMN 模型）。条件随机场$_{UMN}$、双向长短期记忆网络-条件随机场$_{UMN}$ 和 $CancerBERT_{UMN_397}$ 模型的皮尔逊相关系数分别为 -0.678、-0.345 和 -0.712，显示出从中度（双向长短期记忆网络-条件随机场$_{UMN}$）到强（条件随机场$_{UMN}$、$CancerBERT_{UMN_397}$）的负相关。这些结果表明，深度学习模型的可移植性与不同语料库中目标实体（例如本研究中的癌症表型）的相似程度有关。

在置换测试集上的评估反映了模型是在学习文本中的真实模式，还是仅仅记住训练集中出现的表型。在这项置换集评估中，与条件随机场和双向长短期记忆网络-条件随机场模型相比，CancerBERT 模型也展现出巨大优势。例如，**表4** 显示，$CancerBERT_{UMN_397}$ 模型的精确匹配（宽松匹配）宏平均F1分数仅下降了 0.270（0.169），而条件随机场$_{UMN}$ 和双向长短期记忆网络-条件随机场$_{UMN}$ 模型的相应宏平均F1分数分别下降了 0.440（0.355）和 0.524（0.438）。实际上，这些结果表明，在三种类型的模型中，稳健性最强的 CancerBERT 模型能够更好地学习实体的上下文信息，因此也能更好地处理此前未见过的新实体变体。

研究采用 ECR 分析可移植性评估结果。尽管 $CancerBERT_{UMN_397}$ 模型在所有组中的表现都显著更好，但其在第 0 组和第 1 组表型上的优势最大（ECR<0.67）。这些组包含的目标测试表型要么以不同标签出现在训练集中，要么未出现在训练集中，这表明这两组中的表型相对难以提取。结果表明，与双向长短期记忆网络-条件随机场$_{UMN}$ 和条件随机场$_{UMN}$ 模型相比，$CancerBERT_{UMN_397}$ 模型在学习目标表型及其上下文方面具有更强的学习能力。

本研究存在若干局限性。我们通过一项自然语言处理任务，评估了机器学习模型在两家医疗机构的临床文本中抽取乳腺癌表型的泛化能力。这是临床领域常见的命名实体识别任务；然而，自然语言处理任务还有许多其他类型，例如关系抽取和文本分类。其他临床自然语言处理任务值得进一步研究。我们重点关注自然语言处理模型在命名实体识别任务上的表现，以及自然语言处理表现的差异将如何进一步影响下游临床

应用，这仍有待未来探索。我们仅在两个临床语料库上评估了可移植性，进一步纳入更多研究机构的语料库开展探索，将有助于我们更好地了解自然语言处理模型的泛化能力。

结论

本研究通过一项信息抽取任务，评估了三种自然语言处理模型（条件随机场、双向长短期记忆网络-条件随机场和 CancerBERT）在临床领域的泛化能力。该任务旨在从 UMN 和梅奥诊所获取的临床文本中抽取乳腺癌表型。我们从不同角度评估了模型的泛化能力，发现 CancerBERT模型的泛化能力最佳，因为它们能够更好地学习目标表型的上下文信息，并处理临床文本中的文本变体。研究结果表明，在单一研究机构训练的 CancerBERT模型可以迁移至其他研究机构，并以较低成本取得相当的性能。此外，我们发现，在三种模型中，CancerBERT模型的鲁棒性最强。本研究首次分析了基于BERT的模型在临床领域的可移植性，我们的发现有助于指导不同临床机构之间自然语言处理模型的迁移与采用。

致谢

NA

数据可用性

数据　　本文所依据的数据因患者健康信息的隐私保护而无法公开共享　　n.

资助 声明

本研究部分由美国国家补充与综合健康中心（NCCIH）通过资助编号 R01AT009457（Zhang）提供支持；由美国国家老龄研究所通过资助编号 1R01AG078154-01（Zhang）提供支持；并由明尼苏达大学临床与转化科学研究所（CTSI）提供支持，该研究所获得美国国家转化科学促进中心通过资助编号 UL1TR002494 提供的资助。本文内容完全由作者负责，不一定代表美国国立卫生研究院的官方观点。

利益冲突声明

不存在利益冲突。

作者贡献声明

SZ 完成了主要实验。NW、SZ 和 LW 参与了数据收集与标注。HL 和 RZ 指导了研究设计、数据收集与分析。所有作者均参与了稿件撰写，并对稿件的重要学术内容进行了关键性修订。

参考文献

[1] Sohn S, Wang Y, Wi CI, Krusemark EA, Ryu E, Ali MH, Juhn YJ, Liu H. 临床文档差异与自然语言处理系统的可移植性：一项针对不同机构哮喘出生队列的案例研究。《美国医学信息学会杂志》。2018年3月；25(3):353-9。[2] Xie F, Lee J, Munoz-Plaza CE, Hahn EE, Chen W. 在一体化医疗机构中应用文本信息抽取系统实时识别癌症病例。《病理信息学杂志》。2017年1月1日；8(1):48。[3] Carchiolo V, Longheu A, Reitano G, Zagarella L. 医疗处方分类：一种基于自然语言处理的方法。载于《2019年计算机科学与信息系统联合会议（FedCSIS）》，2019年9月1日（第605-609页）。IEEE。[4] Vijayakrishnan R, Steinhubl SR, Ng K, Sun J, Byrd RJ, Daar Z, Williams BA, Defilippi C, Ebadollahi S, Stewart WF. 通过对电子健康记录进行文本与数据挖掘，识别大型初级保健人群中心力衰竭体征和症状的患病率。《心力衰竭杂志》。2014年7月1日；20(7):459-64。[5] Wen A, Fu S, Moon S, El Wazir M, Rosenbaum A, Kaggal VC, Liu S, Sohn S, Liu H, Fan J. 加速医疗人工智能发展的自然语言处理服务交付愿景，以及梅奥诊所的自然语言处理即服务实践。《NPJ数字医学》。2019年12月17日；2(1):1-7。[6] Wang Y, Wang L, Rastegar-Mojarad M, Moon S, Shen F, Afzal N, Liu S, Zeng Y, Mehrabi S, Sohn S, Liu H. 临床信息抽取应用：文献综述。《生物医学信息学杂志》。2018年1月1日；77:34-49。[7] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. 注意力机制即所需的一切。《神经信息处理系统进展》。2017；30。[8] Li L, Chen X, Ye H, Bi Z, Deng S, Zhang N, Chen H. 基于BERT的关系抽取的鲁棒性与偏差分析。载于《中国知识图谱与语义计算会议》，2021年11月4日（第43-59页）。Springer，新加坡。[9] Fan JW, Prasad R, Yabut RM, Loomis RM, Zisook DS, Mattison JE, Huang Y. 临床文本的词性标注：机构之间的壁垒还是桥梁? 载于《AMIA年度研讨会论文集》，2011年（第2011卷，第382页）。美国医学信息学会。[10] Friedman C, Kra P, Rzhetsky A. 两种生物医学子语言：基于Zellig Harris理论的描述。《生物医学信息学杂志》。2002年8月1日；35(4):222-35。[11] Mehrabi S, Krishnan A, Roch AM, Schmidt H, Li D, Kesterson J, Beesley C, Dexter P, Schmidt M, Palakal M, Liu H. 识别有胰腺癌家族史的患者——对自然语言处理系统可移植性的调查。《健康技术与信息学研究》。2015；216:604。[12] Liu M, Shah A, Jiang M, Peterson NB, Dai Q, Aldrich MC, Chen Q, Bowton EA, Liu H, Denny JC, Xu H. 现有吸烟状态检测模块跨机构可移植性研究。载于《AMIA年度研讨会论文集》，2012年（第2012卷，第577页）。美国医学信息学会。[13] Zhou S, Wang N, Wang L, Liu H, Zhang R. CancerBERT：一种癌症领域专用语言模型，用于从电子健康记录中提取乳腺癌表型。《美国医学信息学会杂志》。2022年3月25日。

[14] Peng Y, Chen Q, Lu Z. 基于 BERT 的多任务学习在生物医学文本挖掘中的实证研究。arXiv 预印本 arXiv:2005.02799。2020年5月6日。[15] Liu Z, Tang B, Wang X, Chen Q. 通过循环神经网络和条件随机场对临床笔记进行去标识化。《生物医学信息学杂志》。2017年11月1日；75：S34-42。[16] Chapman AB, Peterson KS, Alba PR, DuVall SL, Patterson OV. 使用快速训练的分类模型检测药物不良事件。《药物安全》。2019年1月；42(1)：147-56。[17] Unanue IJ, Borzeshi EZ, Piccardi M. 使用专用词嵌入的循环神经网络进行健康领域命名实体识别。《生物医学信息学杂志》。2017年12月1日；76：102-9。[18] Pennington J, Socher R, Manning C. GloVe：词表示全局向量。载于《2014年自然语言处理实证方法会议论文集》（自然语言处理实证方法会议）。2014年；第1532-1543页。[19] Mikolov T, Sutskever I, Chen K, et al. 词和短语的分布式表示及其组合性。载于《神经信息处理系统进展》。2013年；第3111-3119页。[20] Yang X, Bian J, Hogan WR, Wu Y. 使用变换器进行临床概念抽取。《美国医学信息学会杂志》。2020年12月9日；27(12)：1935-42。[21] Schutte D, Vasilakes J, Bompelli A, Zhou Y, Fiszman M, Xu H, Kilicoglu H, Bishop JR, Adam T, Zhang R. 利用从生物医学文献生成的 SuppKG 发现新型药物与补充剂相互作用。《生物医学信息学杂志》。2022年7月1日；131：104120。[22] Fu J, Liu P, Zhang Q. 重新思考神经模型的泛化：命名实体识别案例研究。载于《AAAI人工智能会议论文集》2020年4月3日（第34卷，第05期，第7732-7739页）。[23] Devlin J, Chang MW, Lee K, Toutanova K. BERT：用于语言理解的深度双向变换器预训练。arXiv 预印本 arXiv:1810.04805。2018年10月11日。

缩写：

EHR：电子健康记录 NLP：自然语言处理 UMN：明尼苏达大学 BERT：基于变换器的双向编码器表示 MC：梅奥诊所 CRF：条件随机场 BiLSTM-CRF：双向长短期记忆网络-条件随机场 IRB：机构审查委员会 ECR：实体覆盖率 IAA：标注者间一致性 TF-IDF：词频-逆文档频率 NER：命名实体抽取 GloVe：词表示全局向量