arXiv:2303.05392v1 [cs.CL] 7 Mar 2023

自动总结临床试验证据：一个凸显当前挑战的原型

Sanjana Ramprasad东北大学
ramprasad.sa@northeastern.edu

DenisJered McInerney东北大学
mcinerney.de@northeastern.edu

Iain J. Marshall
伦敦国王学院
iainjmarshall@gmail.com

Byron C. Wallace
东北大学
b.wallace@northeastern.edu

摘要

我们介绍了 *TrialsSummarizer*，这是一个旨在自动总结与给定查询最相关的一组随机对照试验中所呈现证据的系统。在先前工作（Marshall 等人，2020）的基础上，该系统检索与指定条件、干预措施和结局组合的查询相匹配的试验发表文献，并根据样本量和估计的研究质量对其排序。排名前k 的研究将交由神经多文档摘要系统处理，从而生成这些试验的概要。我们考察了两种架构：一种基于 BART 的标准序列到序列模型（Lewis 等人，2019），以及一种旨在提高最终用户透明度的多头架构。两种模型都能针对查询生成流畅且相关的检索证据摘要，但它们倾向于引入无依据陈述，因此目前并不适用于这一领域。所提出的架构或可帮助用户核验输出，使用户能够追溯生成的词元对应的输入内容。演示视频见：https://vimeo.com/735605060 原型、源代码和模型权重见：https://sanjanaramprasad.github.io/trials-summarizer/。

1 引言

理想情况下，患者治疗决策应以所有可获取的相关证据为依据。然而，随着医学文献（本已浩如烟海）持续迅速扩增，实现循证医疗这一目标变得愈发困难（Bastian 等人，2010年）。如今，每天发表的随机对照试验报告超过100篇（Marshall 等人，2021年）。语言技术——尤其是自动摘要方法——有望为特定临床问题提供所有相关证据的简明概述，从而形成一种按需生成的系统综述（Wang 等人，2022年；DeYoung 等人，2021年；Wallace 等人，2021年）。

我们介绍一个演示系统 *TrialsSum-marizer*。该系统将临床试验文献检索与摘要模型相结合，为与临床问题相关的当前已发表证据提供叙述性概述。图 1 展示了在本系统中运行的一个示例查询及其输出结果。能够针对任何给定主题生成关于医学证据的准确 摘要的系统，可以大幅提升医护人员查阅完整证据基础以指导医疗护理的能力。

然而，目前的神经摘要系统容易在输出中引入错误信息（Kryscinski 等人，2020；Maynez 等人，2020；Pagnoni 等人，2021年；Ladhak 等人，2021年；Choubey 等人，2021年）。研究尤其表明，这一问题也存在于医学文献摘要领域（Wallace 等人，2021年；Otmakhova 等人，2022），而该领域对事实准确性的要求尤为严格。若没有高效的方法供用户评估输出的有效性，那么生成看似可信、却常常误导人的治疗效果比较摘要的系统便毫无用处。

为满足临床试验摘要对透明度的需求，我们实现了一种摘要架构和界面，旨在支持能够增强用户对输出信任度的交互。具体而言，模型将生成的摘要中的每个词元与从输入中提取的特定来源“方面”关联起来。这样一来，用户便能追溯输出文本的来源（即输入中的片段），从而进行某种形式的验证。该架构还提供了填充预定义模板摘要的功能，在模板所提供的控制力与抽象式摘要的灵活性之间取得平衡。我们在系统演示中实现了这一功能。

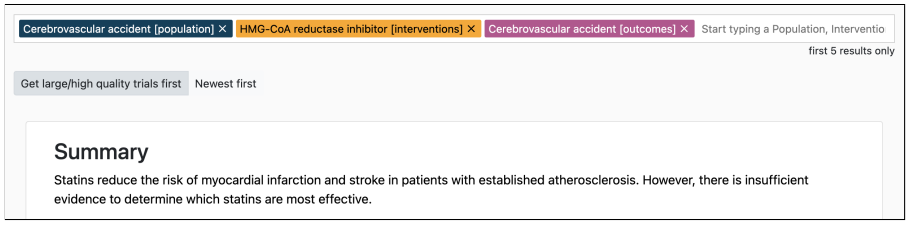


图1：系统提供的一个示例查询（关于使用他汀类药物 降低中风风险）及其摘要输出。在此示例中，摘要准确反映了证据，但并非总是如此。

2 相关工作

神经摘要系统的事实性（或其缺失）是一个活跃的研究领域（Chen 等人，2021年；Cao 等人，2020；Dong 等人，2020；Liu 等人，2020；Goyal 和 Durrett，2021年；Zhang 等人，2021年；Kryscinski 等，2020；Xie 等人，2021年）。本文的系统演示在特定领域和应用场景下探讨了这一问题。我们还探索了如何通过模板等方式实现可控性，以支持交互。这延续了此前关于混合式模板／神经摘要的研究（Hua 和 Wang，2020；Mishra 等人，2020；Wiseman 等人，2018）。

我们还注意到，本研究借鉴了有关摘要系统输出可视化（Vig 等人，2021年；Strobelt 等人，2018；Tenney 等人，2020）和生物医学文献摘要（Plaza 和 Carrillo-de Albornoz，2013；Demner- Fushman 和 Lin，2006；Mollá，2010；Sarker 等人，2017；Wallace 等人，2021年）的既有研究。据我们所知，这是首个能够按需生成（草拟）证据综述、且兼具可解释性和可编辑性的可运行原型。

3 系统概述

我们的界面基于 Trialstreamer（Marshall 等人，2020）构建。Trialstreamer 是一个自动化系统，用于识别新发表的人体随机对照试验（RCT）报告，并从中提取重要信息，将其存储在涵盖所有已发表试验信息的数据库中。我们的系统通过一种简单的检索技术（第3.1节）识别与给定查询相关的随机对照试验报告，然后将排名靠前的k 项输入多文档摘要模型（第3.2节）。对于后一个组件，我们考虑了标准的序列到序列方法，以及一种旨在提高透明度的方面结构化架构（第3.3节）。

3.1 检索文章

Trialstreamer（Marshall 等人，2020；Nye 等人，2020）监测研究数据库，具体包括：PubMed[1] 和世界卫生组织国际临床试验注册平台，以自动识别新近发表的关于人类随机对照试验的报告，使用此前验证过的分类器（Marshall 等人，2018）。

描述随机对照试验的文章随后会经过一系列机器学习模型的处理，这些模型会从试验报告中提取关键要素，包括：样本量；对试验人群、干预措施和结局的描述；关键结果；以及所报告证据的可靠性（通过近似偏倚风险评分表示；Higgins 等人 2019）。这些提取出的（半）结构化信息会存储在 Trialstreamer 关系数据库中。

描述研究人群、干预措施和结局（PICO 要素）的自由文本片段也会映射到医学主题词，[2] 这一过程通过重新实现 MetaMap Lite 完成（Demner-Fushman 等人，2017）。

为便于搜索，用户可以输入涉及部分人群、干预措施和结局的医学主题词，用于在数据库中的文章及其相应提取的关键数据中查找匹配项。随后，匹配的研究会根据样本量 s 和偏倚风险评分 rob：score $= s$/rob 的评分函数进行排序；也就是说，我们优先检索样本量大、质量高的试验报告。

该系统演示的创新之处在于加入了摘要组件，该组件接收排名靠前的k 检索到的试验（此处使用k=5），并以系统综述摘要的风格生成对这些证据的叙述性总结（Wallace 等人，2021年）。将该摘要模块与 Trialstreamer数据库结合后，我们便可实时总结所有符合给定查询条件的试验（图1）。

[1] https://pubmed.ncbi.nlm.nih.gov/ [2] MeSH 是医学主题词（Medical Subject Headings）的简称，是由美国国家医学图书馆（NLM）维护的受控词汇表。

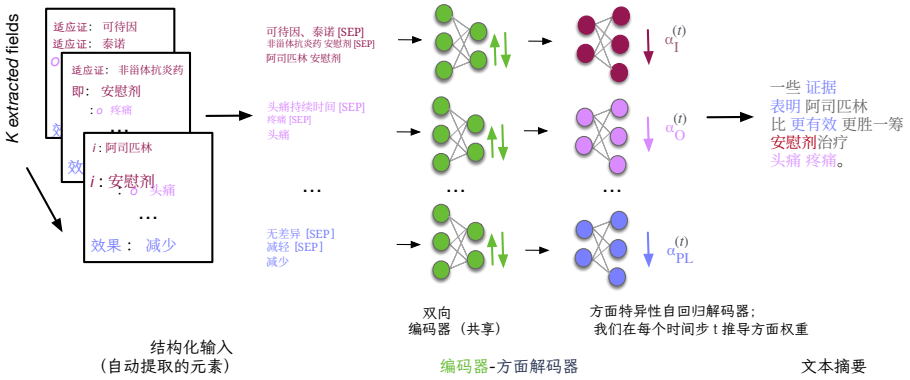


图 2：我们提出的结构化摘要方法旨在综合各个方面（在预处理步骤中自动抽取），并有条件地生成关于每个方面的文本。

3.2 试验摘要

我们考虑了摘要模块的两种实现方式。我们使用先前工作中介绍的数据集训练这两种模型，该数据集的输入是一组随机对照试验报告（从摘要中提取的 PICO要素），目标则是由 Cochrane协作网成员撰写的系统综述摘要中的“作者结论”部分（Wallace 等人，2021年）（见第4节）。

作为首个模型，我们采用 BART（Lewis 等人，2019），并使用 Longformer（Beltagy 等人，2020）作为编码器，以适应篇幅较长的多文档输入。作为模型输入，我们将从单项试验中抽取的、包含显著信息的文本片段拼接起来，其中包括研究人群、干预措施、结局以及“点睛句”。后者指抽取出的、似乎概括了主要结果或发现的片段，例如“死亡率显著上升.....”；更多详情见（Lehman 等人，2019）。我们用特殊标签括起这些片段。例如， <研究人群>参与者均为糖尿病患者.....</研究人群>。作为额外监督，我们对目标文本运行相同的抽取模型，并使用同一组标签对其进行标记。

用于此任务的标准序列到序列模型存在一个问题：它们没有自然的方法来评估输出中词元的来源，因此很难验证生成的摘要是否可信。接下来，我们将讨论一种旨在提升透明度和可控性的替代架构。

3.3 提升透明度的拟议方面结构化架构

我们采用一种类似于（Goyal 等人，2021年）的多头架构，显式生成与各个方面相对应的词元（图2）。我们假设输入被划分为文本片段，分别对应一组 K 字段或方面。这里，这些方面包括对试验人群、干预措施和结局的描述，以及报告主要研究发现的“妙语”片段。我们将每个 K 方面的输入记为 $\{x^{a_1}, ..., x^{a_K}\}$，其中 x^{a_k} 表示从输入 k 中提取的方面 x的文本。鉴于这是一个多文档设置（每个输入都包含多篇文章），x^{a_k} 由以下内容构成：将所有文档中的方面文本拼接起来，并使用特殊词元分隔各篇文章。

我们分别对各方面的文本进行编码，以获得方面专属嵌入 x^{a_k}编码器。随后，我们将这些表示分别输入方面专属解码器和共享的语言模型头，以获得词汇分布 $\hat{o}_t^{a_k}$。除最后两个解码器层由方面专属参数构成外，所有模型参数均为共享参数。重要的是，给定方面的表示仅基于与该方面相关的文本(x^{a_k})。

我们将最终输出建模为各方面分布的混合：$\hat{o}_t = \sum_{k=1}^{K} z_t^{a_k}(\hat{o}_t^{a_k})$。混合权重 $z_t = z_t^{a_1}, \ldots, z_t^{a_K}$表示在时间步 t 对各方面进行软选择，其计算方式为：将解码器的每个倒数第二层表示 $y_t^{a_k}$（在输入语言模型头之前）与一个可学习参数 $W_z \in R^D$进行点积。随后， K 逻辑值 $z_t^{a_k}$经 Softmax 归一化，再与方面专属词汇分布 $\hat{o}_t^{a_k}$相乘

追踪输出与输入的对应关系 这种架构允许查看生成的摘要中各个词元对应的混合权重，从而了解哪个方面（或哪些方面）对

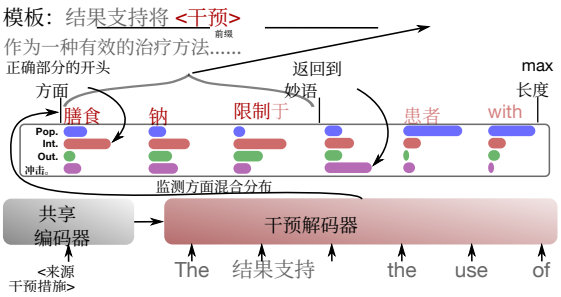


图 3：**模板生成**。为填充模板，我们强制模型从特定的头生成，并监测模型的混合分布，以确定何时停止。

输出的影响最大。进一步查看研究中与该方面对应的片段，可能有助于验证输出，和/或帮助发现错误及其可能的引入位置。

受控生成神经摘要模型往往难以恰当地综合相互矛盾的证据，从而对特定干预的总体效果得出正确判断。尽管摘要模型并不完美，但仍可能有用，因为它们可以快速起草证据综述，供人编辑。多头架构天然支持模板填充，因为用户可以明确地从对应感兴趣方面的头中提取词元。在我们的演示中，用户可以在不同模板之间切换，这些模板对应于对所讨论干预的总体效果所作的不同结论。（扩展此功能以允许用户指定自己的待填充模板也很简单。）

为填充模板，我们将空白处前面的模板文本用作上下文，然后从对应于指定方面的语言头生成文本。为动态确定片段长度，我们监测混合分布，并在其转向另一个方面时停止（图 3）。

3.4 用户界面

图 5 展示了我们构建的、集成了多头架构的界面。摘要中突出显示的方面有助于解读输出词元的来源，指出了哪些方面影响了这些词元的生成。用户还可以查看与这些方面相关的片段，这有助于识别生成的摘要中缺乏依据的内容。为此，用户点击某个词元时，我们会显示对其生成影响最大的那部分输入内容。

我们还会展示概览（即“关键结果”），以传达试验的主要发现，从而提供更多背景信息。由于标准的序列到序列模型无法将输出词元与输入方面关联起来，因此，对于该模型，我们会在摘要旁显示所有试验的全部方面（以及关键结果）。

利用模型上述的填空能力，我们还针对聚合发现的每种可能“方向”（显著效应与无效应）提供了预定义模板。第 5 节将结合示例介绍该界面。

4 数据集与训练详情

我们的目标是将描述针对同一临床问题的随机对照试验的标题和摘要集合，转化为能够综合其中证据的抽象式摘要。我们使用随机对照试验摘要数据集训练所有模型（Wallace 等人，2021年），并在预处理阶段借助现有模型从每条（非结构化）输入中提取临床上重要的要素，即我们的方面（Marshall 等人，2020）。

训练 我们使用 Huggingface Transformers 库（Wolf 等人，2020）实现这两个模型。两个模型均以 *bart-base* (Lewis 等人，2019) 初始化。我们以批次大小 2 对模型进行微调，共训练 3 个轮次，使用 Adam 优化器（Kingma 和 Ba，2014），学习率为 3e-5。

推理 我们使用束搜索，束宽设为 3。生成文本的最小长度和最大长度分别设为 10 和 300。

5 案例研究：验证与可控性

为展示该界面（及其支撑架构）的潜在用途，我们将逐一介绍两个案例研究。我们将重点说明所提出的方法能够提供何种用于验证的可解释性，同时展示其执行可控摘要的能力，以说明这种能力可能如何发挥作用。这些案例研究中使用的查询及相关调查由共同作者 IJM 完成；他是一位在循证医学领域经验丰富的医生。我们还比较了各模型，并在附录部分报告了 ROUGE 和事实性的自动评分A，结果发现两种模型的表现相当。

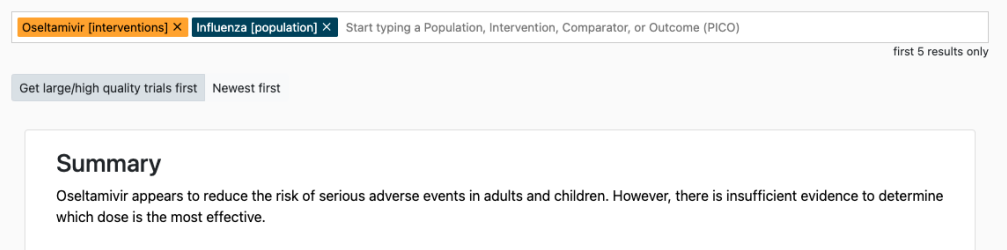


图 4：使用标准 BART（Lewis 等人，2019）模型生成的示例输出和界面。

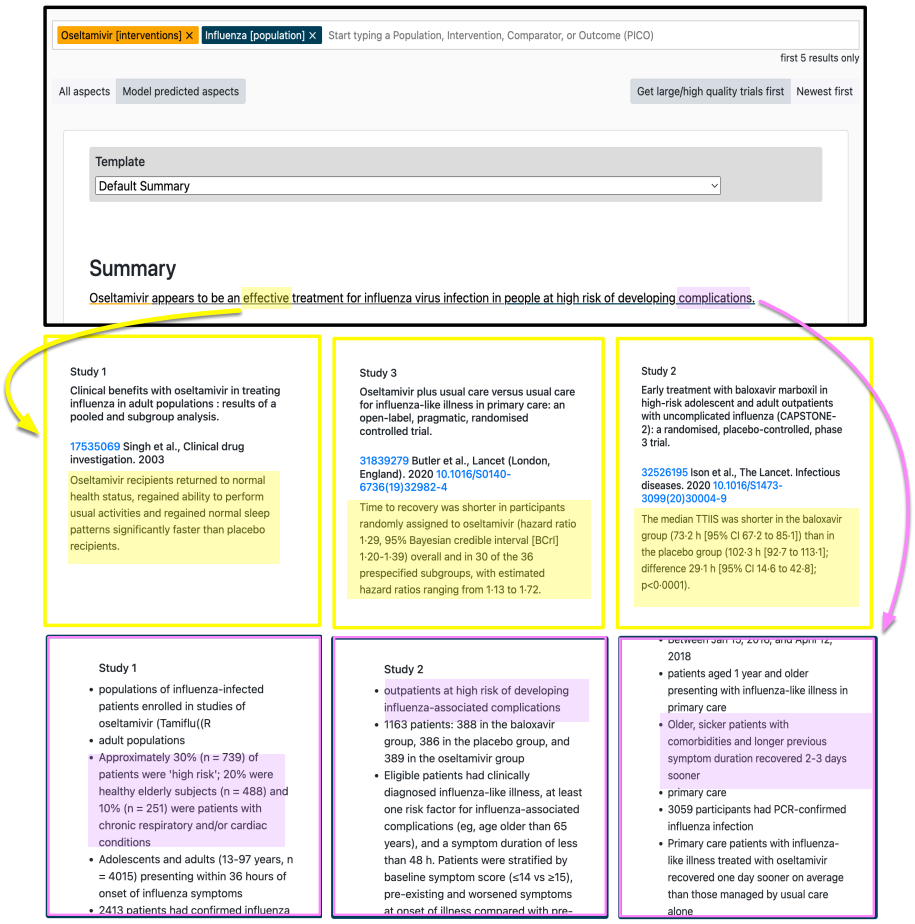


图 5：定性示例：结构化摘要模型（及其配套界面）支持对有关奥司他韦用于流感感染患者的摘要进行逐词元验证。该方法可轻松显示“有效”这一主张（上方；黄色）以及将人群描述为有“并发症”风险的个体（下方；紫色）所获得的支持。

模型可解释性 为展示所提出的架构和界面在支持验证方面的潜力，我们考虑一个有关奥司他韦作为流感感染患者干预措施效果的查询。标准架构会生成一份摘要，概述与该查询最相关的随机对照试验，如图4所示。摘要包含两项主张：（1）研究表明，该干预措施能够降低成人和儿童发生不良事件的风险；（2）对于最有效的剂量，尚未达成共识。我们可以检查输入以尝试验证这些说法。这样做后，我们发现，报告的结果往往表明不良事件的风险有所降低，而且部分研究纳入了青少年和成人，这说明第一项说法是准确的。第二项说法则较难通过检查来验证；输入中并未明确传达剂量方面存在此类不确定性。使用标准的序列到序列架构验证这些说法十分费力，因为这类模型具有抽象式特性，难以将输出中的部分内容追溯到输入。

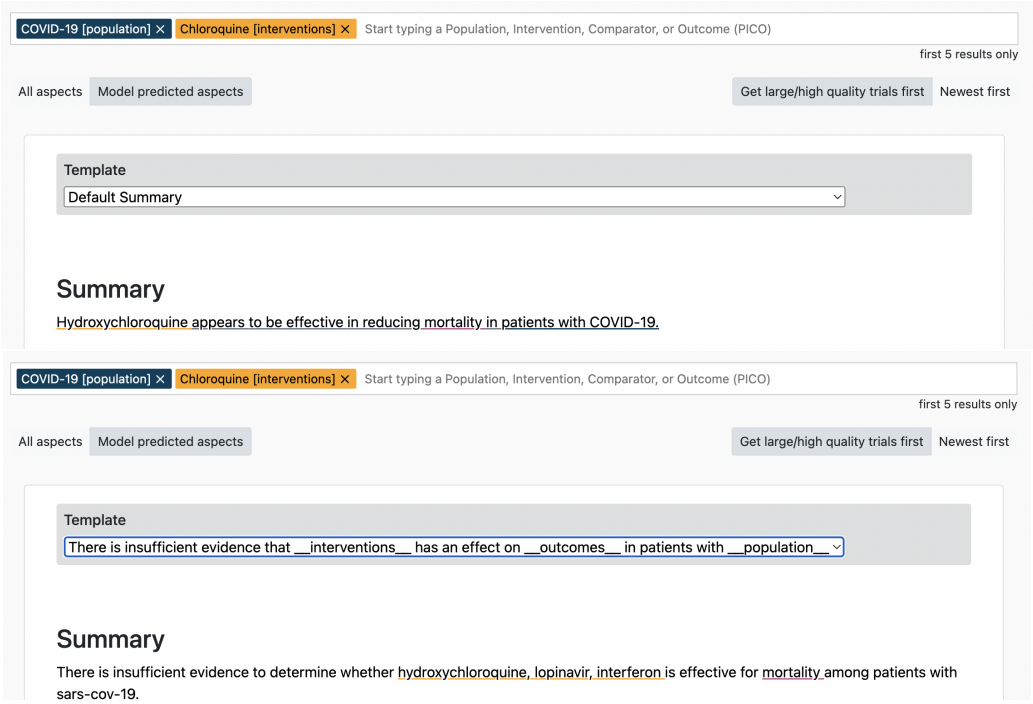


图 6：结构化模型生成的关于氯喹对新冠肺炎患者影响的不准确摘要（上）。使用结构化模型生成的模板控制摘要（下）。

因此，验证需要通读整个输入，才能核实不同方面的信息。

多头架构使我们能够提供一种交互式界面，旨在让验证变得更加容易。具体而言，将每个输出词元与特定方面关联起来，为检查输入中可能支持生成文本的片段提供了一种自然的机制。图5展示了上述奥司他韦与流感示例。这里，我们展示了如何点击输出中的“有效”词元，从而查看影响其生成的方面（图2）；在本例中，这可以追溯到提取出的、传达主要研究发现的“点睛之句”。由此可以很容易地看出，该说法有证据支持。同样，我们可以通过追溯到作为该输出依据的人群片段，验证有关人群是并发症高风险个体的内容。

可控性 如上所述，所提出架构的另一项潜在优势，是能够通过“填充”模板，为神经生成模型赋予可控性。具体而言，鉴于总体（汇总）治疗效果在此情境中至关重要，我们预定义了用于表达效应方向的模板。其思路是，如果验证后发现模型得出的总体效应方向有误，就可以使用与正确方向相对应的预定义模板，按需生成更准确的摘要。

我们在图的上半部分展示了一个由结构化模型生成的摘要示例6。利用上文讨论的可解释性验证功能，我们发现模型错误地表示干预氯喹对治疗新冠肺炎有效。不过，借助交互式界面，我们能够立即生成一份新的摘要，其中呈现了经更正的综合结果（方向），如图 6下半部分所示，无需手动撰写。

我们在附录 B 节提供了更多案例研究B。

6 结论

我们介绍了 TrialsSummarizer，这是一种原型系统，可根据给定的查询自动总结随机对照试验。神经摘要模型生成的摘要可读且（大多）相关，但它们往往会在输出中引入缺乏依据或不准确的信息，因此目前尚不适合用于这一领域。

我们实现了一种多头架构，旨在提高透明度。我们提供了定性示例，以展示

其在加快验证和实现可控生成方面的潜力。未来需要通过用户试验来检验这一功能的实用性，并为新的架构提供参考，以进一步提高用于总结生物医学证据的模型的准确性和透明度。

局限性与伦理问题

局限性这项工作存在若干局限性。首先，如上所述，尽管自动总结生物医学证据的前景令人振奋，但现有模型容易引入事实错误，因此尚不适合执行这项任务。我们的原型系统在一定程度上旨在凸显这一问题，并推动相关工作解决可靠性和可信度方面的问题。

在这篇演示论文中，我们还尝试通过提出的结构化摘要模型及配套界面，在缓解此类问题方面取得一些进展，并提供了定性示例来展示其潜力。不过，确实应开展正式的用户研究，以评估其效用。任务本身的难度使评估变得复杂：要评估自动摘要的事实准确性，需要深厚的领域专业知识，并且需要花费大量时间阅读构成摘要的输入内容，以判断生成的摘要是否属实。

这项工作的另一个局限性是，我们在当前的原型系统中做出了一些临时性的设计决策。例如，目前我们会将针对给定查询检索到的文章中排名前 5 的文章（依据试验样本量和估计的可靠性）传入摘要系统。未来的工作可以通过探索依据更充分的方法来选择应纳入哪些研究以及纳入多少项研究，从而解决这一问题。

伦理 准确总结生物医学证据有望通过支持循证医学实践，最终改善患者护理。然而，目前这类模型仍存在固有风险。尤其是，人们可能会不加判断地信任模型输出；鉴于当前摘要技术的局限性，这种做法并不可取。

我们的原型演示系统旨在强调这一领域当前面临的挑战，这些挑战必须在任何模型真正投入使用之前得到解决（除此之外，我们还强调验证输出的必要性，这也是本项工作的重点）。在界面中，我们通过一则醒目提示明确指出，该系统仅供研究使用，这些摘要并不可靠，不可轻信。

致谢

本研究部分得到美国国立卫生研究院（NIH）资助（项目编号 R01LM012086），并得到美国国家科学基金会（NSF）资助（项目编号 1901117 和 2211954）。本文内容完全由作者负责，不一定代表 NIH 或 NSF 的官方观点。

参考文献

H Bastian、P Glasziou 和 I Chalmers。2010年。每天七十五项试验和十一篇系统综述：我们究竟如何才能跟上？《*PLoS Medicine*》，7(9)。Iz Beltagy、Matthew E Peters 和 Arman Cohan。2020年。Longformer：长文档 Transformer。*arXiv* 预印本 *arXiv:2004.05150*。Meng Cao、Yue Dong、Jiapeng Wu 和 Jackie Chi Kit Cheung。2020年。抽象式摘要模型的事实错误纠正。*arXiv* 预印本 *arXiv:2010.08712*。Sihao Chen、Fan Zhang、Kazoo Sone 和 Dan Roth。2021年。通过对比候选生成与选择提升抽象式摘要的忠实性。*arXiv* 预印本 *arXiv:2104.09061*。Prafulla Kumar Choubey、Jesse Vig、Wenhao Liu 和 Nazneen Fatema Rajani。2021年。Mofe：用于控制抽象式摘要中幻觉的事实专家混合模型。*arXiv* 预印本 *arXiv:2110.07166*。Dina Demner-Fushman 和 Jimmy Lin。2006年。用于临床问答的答案抽取、语义聚类和抽取式摘要。载于第*21*届计算语言学国际会议暨计算语言学协会第*44*届年会论文集，第841–848页。Dina Demner-Fushman、Willie J Rogers 和 Alan R Aronson。2017年。MetaMap Lite：对 MetaMap 的一种新 Java 实现的评估。《美国医学信息学会杂志》，24(4):841–844。Jay DeYoung、Iz Beltagy、Madeleine van Zuylen、Bailey Kuehl 和 Lucy Lu Wang。2021年。MS2：医学研究的多文档摘要。*arXiv* 预印本 *arXiv:2104.06486*。Yue Dong、Shuohang Wang、Zhe Gan、Yu Cheng、Jackie Chi Kit Cheung 和 Jingjing Liu。2020年。

抽象式文本摘要中的多事实纠正。*arXiv*预印本 *arXiv:2010.02443*。

Tanya Goyal 和 Greg Durrett。2021年。为摘要中的细粒度事实性进行标注和建模。*arXiv* 预印本 *arXiv:2104.04302*。

Tanya Goyal、Nazneen Fatema Rajani、Wenhao Liu 和 Wojciech Kry´sci´nski。2021年。Hydrasum：使用多解码器模型解耦文本摘要中的风格特征。*arXiv* 预印本 *arXiv:2110.04400*。

Julian PT Higgins、Jelena Savovi´c、Matthew J Page、Roy G Elbers 和 Jonathan AC Sterne。2019年。评估随机试验中的偏倚风险。《*Cochrane*干预措施系统综述手册》，第205–228页。

Xinyu Hua 和 Lu Wang。2020年。Pair：预训练变换器中的规划与迭代优化，用于长文本生成。*arXiv* 预印本 *arXiv:2010.02301*。

Diederik P Kingma 和 Jimmy Ba。2014年。Adam：一种随机优化方法。*arXiv* 预印本 *arXiv:1412.6980*。

Wojciech Kryscinski、Bryan McCann、Caiming Xiong 和 Richard Socher。2020年。评估抽象式文本摘要的事实一致性。载于*2020*年自然语言处理经验方法会议论文集（*EMNLP*），第9332–9346页，线上。计算语言学协会。

Faisal Ladhak、Esin Durmus、He He、Claire Cardie 和 Kathleen McKeown。2021年。忠实还是抽取式？论如何缓解抽象式摘要中忠实性与抽象程度之间的权衡。*arXiv* 预印本 *arXiv:2108.13684*。

Eric Lehman、Jay DeYoung、Regina Barzilay 和 Byron C Wallace。2019年。根据临床试验报告推断哪些医学治疗有效。*arXiv* 预印本 *arXiv:1904.01606*。

Mike Lewis、Yinhan Liu、Naman Goyal、Marjan Ghazvininejad、Abdelrahman Mohamed、Omer Levy、Ves Stoyanov 和 Luke Zettlemoyer。2019。BART：面向自然语言生成、翻译和理解的去噪序列到序列预训练。*arXiv* 预印本 *arXiv:1910.13461*。

Zhenghao Liu、Chenyan Xiong、Zhuyun Dai、Si Sun、Maosong Sun 和 Zhiyuan Liu。2020。通过领域内语言建模，将开放领域事实抽取和验证适配到 COVID-Fact。在计算语言学协会论文集：*2020*年*EMNLP*，第2395–2400页。

Iain J Marshall、Anna Noel-Storr、Joël Kuiper、James Thomas 和 Byron C Wallace。2018年。用于识别随机对照试验的机器学习：评估与实践者指南。研究综述方法，9(4):602–614。

Iain J Marshall、Benjamin Nye、Joël Kuiper、Anna Noel-Storr、Rachel Marshall、Rory Maclean、Frank Soboczenski、Ani Nenkova、James Thomas 和 Byron C Wallace。2020年。Trialstreamer：一个持续更新的临床试验报告自动数据库。美国医学信息学会杂志，27(12):1903–1912。Iain James Marshall、Veline L' Esperance、Rachel Marshall、James Thomas、Anna Noel-Storr、Frank Soboczenski、Benjamin Nye、Ani Nenkova 和 Byron C Wallace。2021年。证据现状：临床试验全球差异调查。*BMJ*全球健康，6(1):e004145。

Joshua Maynez、Shashi Narayan、Bernd Bohnet 和 Ryan McDonald。2020。关于抽象式摘要中的忠实性和事实准确性。*arXiv* 预印本 *arXiv:2005.00661*。

Abhijit Mishra、Md Faisal Mahbub Chowdhury、Sagar Manohar、Dan Gutfreund 和 Karthik Sankaranarayanan。2020。模板可控的关键词到文本生成。*arXiv* 预印本 *arXiv:2011.03722*。

Diego Mollá。2010。一个用于循证医学摘要的语料库。载于澳大利亚语言技术协会研讨会论文集，第 76–80 页。

Benjamin E Nye、Ani Nenkova、Iain J Marshall 和 Byron C Wallace。2020。Trialstreamer：实时映射和浏览医学证据。载于计算语言学协会北美分会会议论文集，第 2020 卷，第 63 页。美国国立卫生研究院公共获取。

Yulia Otmakhova、Karin Verspoor、Timothy Baldwin 和 Jey Han Lau。2022。患者死多生少：探索生物医学文献多文档摘要的现状。载于计算语言学协会第 *60* 届年会论文集（第 *1* 卷：长篇论文），第 5098–5111 页，爱尔兰都柏林。计算语言学协会。

Artidoro Pagnoni、Vidhisha Balachandran 和 Yulia Tsvetkov。2021年。理解抽象式摘要中的事实性：用于事实性指标的基准。*arXiv*预印本 *arXiv:2104.13346*。

Laura Plaza 和 Jorge Carrillo-de Albornoz。2013。评估生物医学文献摘要中用于句子选择的不同位置策略。*BMC*生物信息学，14(1)：1–11。

Abeed Sarker、Diego Molla 和 Cecile Paris。2017。自动文本摘要与循证医学：对两个领域的综述。

Hendrik Strobelt、Sebastian Gehrmann、Michael Behrisch、Adam Perer、Hanspeter Pfister 和 Alexander M Rush。2018。S eq 2s eq-v is：一种用于序列到序列模型的可视化调试工具。*IEEE Transactions on Visualization and Computer Graphics*，25(1):353–363。

Ian Tenney、James Wexler、Jasmijn Bastings、Tolga Bolukbasi、Andy Coenen、Sebastian Gehrmann、Ellen Jiang、Mahima Pushkarna、Carey Radebaugh、Emily Reif 等。2020。语言可解释性工具：面向自然语言处理模型的可扩展交互式可视化与分析。*arXiv* 预印本 *arXiv:2008.05122*。

Jesse Vig、Wojciech Kry´sci´nski、Karan Goel 和 Nazneen Fatema Rajani。2021 年。SummVis：用于文本摘要的模型、数据和评估的交互式可视化分析。*arXiv* 预印本 *arXiv:2104.07605*。

Byron C Wallace、Sayantan Saha、Frank Soboczenski 和 Iain J Marshall。2021年。生成（事实性？）随机对照试验的叙事摘要：神经网络多文档摘要实验。载于*AMIA*年度研讨会论文集，第2021卷，第605页。美国医学信息学会。

Lucy Lu Wang、Jay DeYoung 和 Byron Wallace。2022年。MSLR2022概述：面向文献综述的多文档摘要共享任务。载于第三届学术文档处理研讨会论文集，第175–180页，韩国庆州。计算语言学协会。

Sam Wiseman、Stuart M Shieber 和 Alexander M Rush。2018。学习用于文本生成的神经模板。*arXiv* 预印本 *arXiv:1808.10122*。

Thomas Wolf、Lysandre Debut、Victor Sanh、Julien Chaumond、Clement Delangue、Anthony Moi、Pierric Cistac、Tim Rault、Rémi Louf、Morgan Funtowicz 等。2020。Transformers：最先进的自然语言处理。在*2020*年自然语言处理实证方法会议论文集：系统演示，第 38–45 页。

Yuexiang Xie、Fei Sun、Yang Deng、Yaliang Li 和 Bolin Ding。2021年。通过反事实估计评估文本摘要的事实一致性。*arXiv* 预印本 *arXiv:2108.13134*。

Sen Zhang、Jianwei Niu 和 Chuyuan Wei。2021年。抽象式摘要模型的细粒度事实一致性评估。在*2021*年自然语言处理实证方法会议，第107–116 页。

附录

自动评估

我们报告了相对于目标（人工撰写的）Cochrane摘要的 ROUGE 分数，涵盖开发集和测试集。我们报告了标准 BART模型以及我们提出的旨在提高可验证性和可控性的多头模型的分数。这些模型在这一指标上的表现大致相当，如表1所示。

然而，ROUGE 指标基于（精确的）n元语法重叠，无法衡量生成文本的事实性。衡量事实性通常仍是一个未解决的问题，而评估生物医学综述的事实准确性尤其复杂，因为该领域和文本本身都很复杂。尽管如此，先前的研究已针对这一特定任务提出了自动化指标（Wallace 等人，2021年；DeYoung 等人，2021年）。这些指标基于模型，用于推断报告结果的方向性，例如，摘要是否表明所述治疗有效。更具体地说，我们对生成摘要和参考摘要是否报告了显著结果进行二元预测，然后计算前者相对于后者的 F1 分数。

模型	ROUGE-L（开发集）	ROUGE-L（测试集）
BART	20.4	19.7
多头	19.9	19.3

表1：标准 BART 模型和我们提出的多头架构在开发集和测试集上取得的 ROUGE 分数。

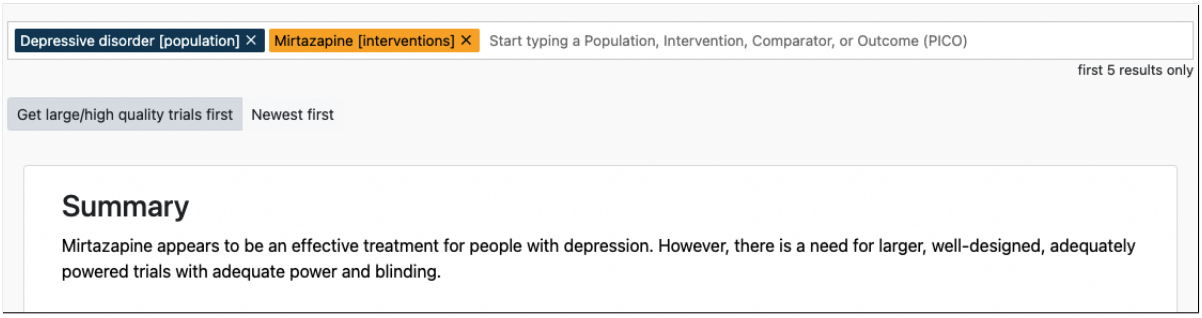
模型	直接（开发集）	直接（测试集）
BART	49.6	51.8
多头	49.3	52.7

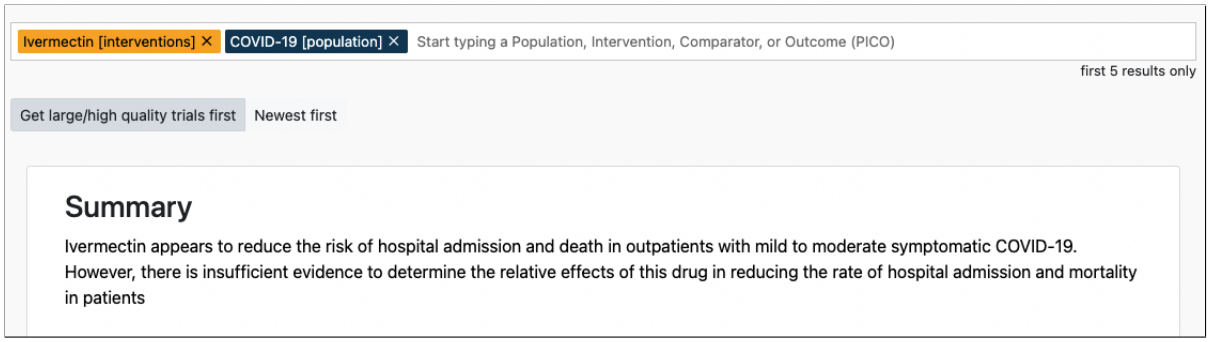
表 2：基础 BART 模型和我们提出的多头架构在开发集和测试集上的方向性得分。

B 附加案例研究

本节将重点介绍更多用例，以说明可解释性和可控性的必要性。

可解释性 我们首先列举一些示例，说明如果没有专门设计来提供可解释性功能的界面，验证模型生成的摘要十分困难。在图 7（a）中，我们展示了一个示例：模型生成的摘要准确概括了米氮平对抑郁症患者的影响。不过，摘要中还包含一项陈述，指出需要开展充分且设计良好的试验。由于这项陈述比较笼统





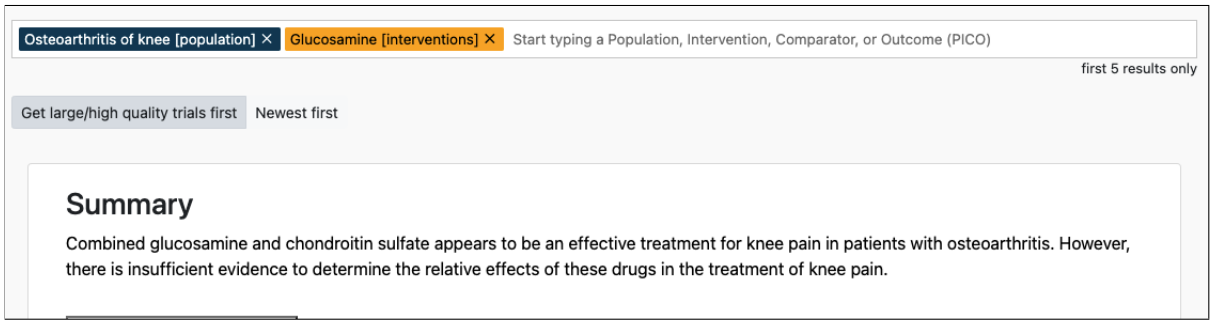


图7：a) 查询使用米氮平 治疗抑郁症 时，BART 生成的摘要；b) 查询使用伊维菌素治疗新冠肺炎时，BART 生成的摘要）

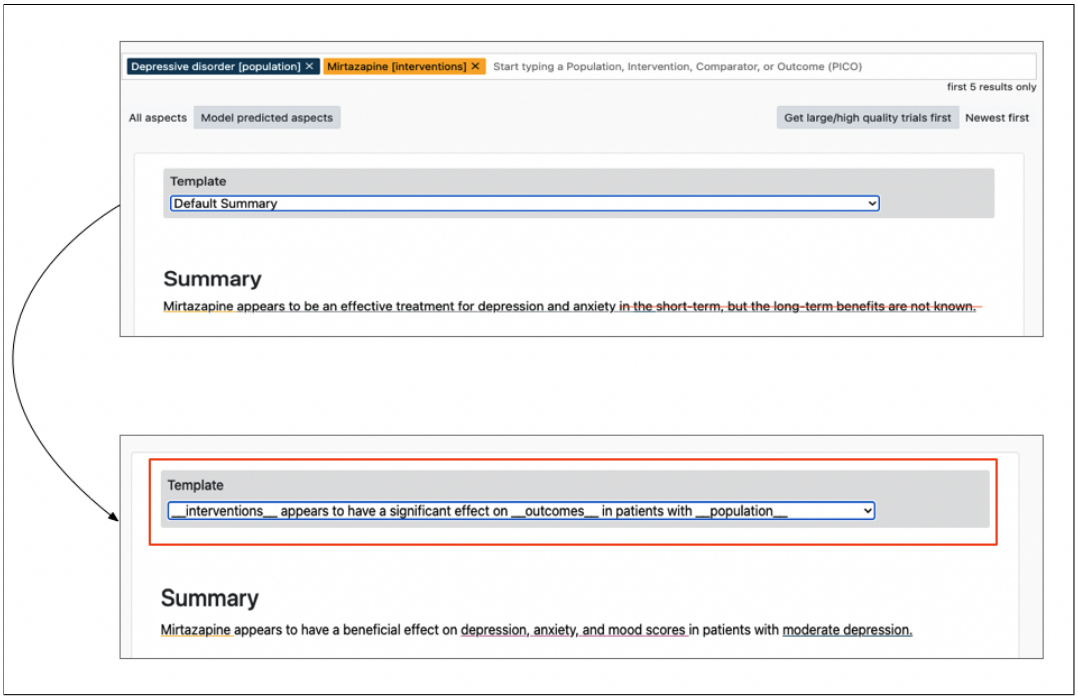


图8：上方摘要显示了多头模型在查询米氮平对抑郁症的影响时生成的默认摘要。下方摘要则显示了使用预定义模板生成的受控摘要。

而且没有指出要讨论任何 PICO 要素，因此无法确定是哪个要素导致生成了该陈述。因此，用户需要审阅所有（原始）输入文本。

在图7 (b) 中，模型生成的摘要包含两个相互矛盾的句子。第一句指出，使用伊维菌素可减少新冠肺炎患者的住院和死亡人数；第二句则称，关于这一点的证据不足。然而，如果不具备可解释性，就无法排查并验证是否是同一组要素导致了相互矛盾的陈述。

图7(c) 中的示例展示了这样一种情况：模型首先准确综合了研究中关于氨基葡萄糖与硫酸软骨素联合使用对膝痛影响的发现。然而，接下来的陈述讨论的是两者的相对效果。同样，在这种情况下，究竟是哪个要素导致生成了该陈述并不直观；要进行验证，就必须仔细审阅所有文本及其对各个要素的含义。

可控性 接下来，我们将展示一些示例，说明如何利用我们的模型所具备的模板填充能力，有效控制摘要的生成，避免生成原本会出错的内容。尽管可解释性特征有助于高效验证，但模型仍难以持续生成事实准确的摘要。我们将展示一些实例，说明如何借助模型实现的可控性（模板填充），快速得到更准确的摘要。

在图8 所示的示例中，默认摘要准确地综合了效果。然而，模型生成的摘要讨论了根据研究关键结果得出的短期和长期获益。阅读提取出的“关键结果”后，我们发现研究指出了停药后出现的问题，但未必提供有关长期用药的信息。填充模板会限制输出，但仍能利用生成式方法所带来的灵活性，从而生成更准确的摘要。例如，在本例中，我们可以看到，借助模板编辑后的摘要更加准确。

同样，在图9 中，当我们询问多头模型氨基葡萄糖对膝骨关节炎的影响时，可以看到模型自行生成的摘要传达了一个不正确的研究总体效应。我们可以验证

这一点，方法是检查负责生成摘要的各个元素，如上文所述。随后，我们使用所示模板得出了更准确的摘要。

图中的示例10 展示了模型犯下的一个有趣错误。由于结局可以用相同的信息以正向或负向的方向呈现（例如，体重减轻与体重增加），模型必须准确推断所有研究的影响。在这个例子中，模型生成的摘要虽然正确描述了影响，却将体重减轻视为不良影响。这里我们再次选择一个模板，让模型快速填入内容，从而得到更准确的摘要。

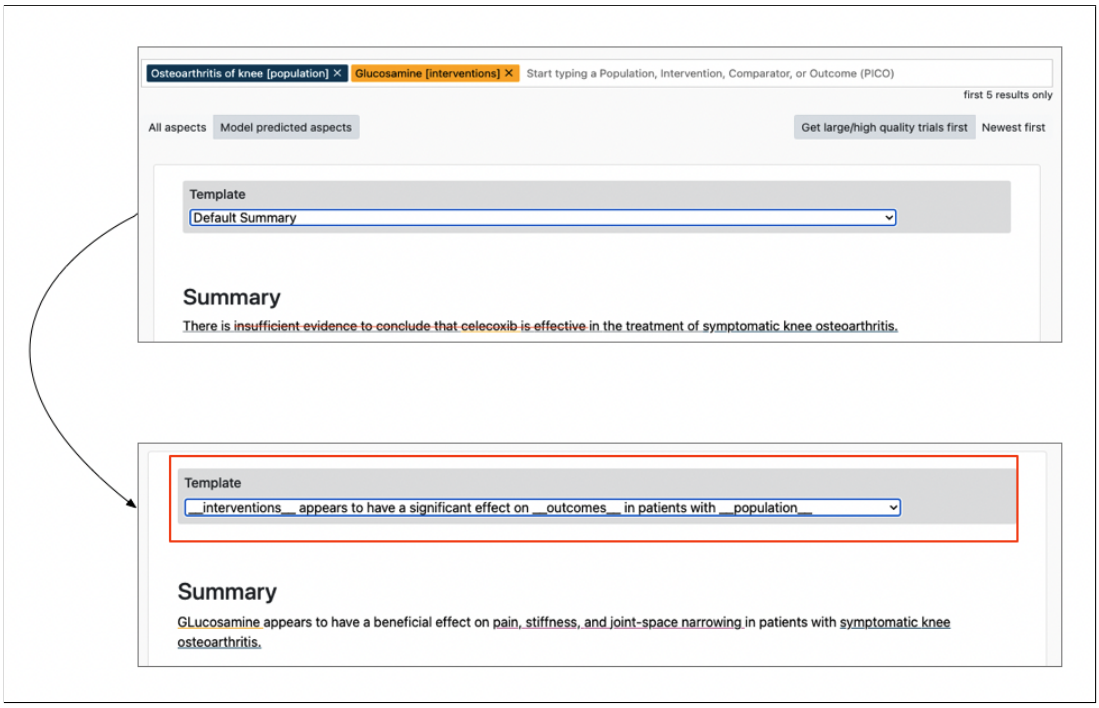


图9：上方的摘要展示了查询氨基葡萄糖 对膝骨关节炎的影响时，多头模型生成的默认摘要。下方的摘要展示了使用预定义模板编辑后的摘要

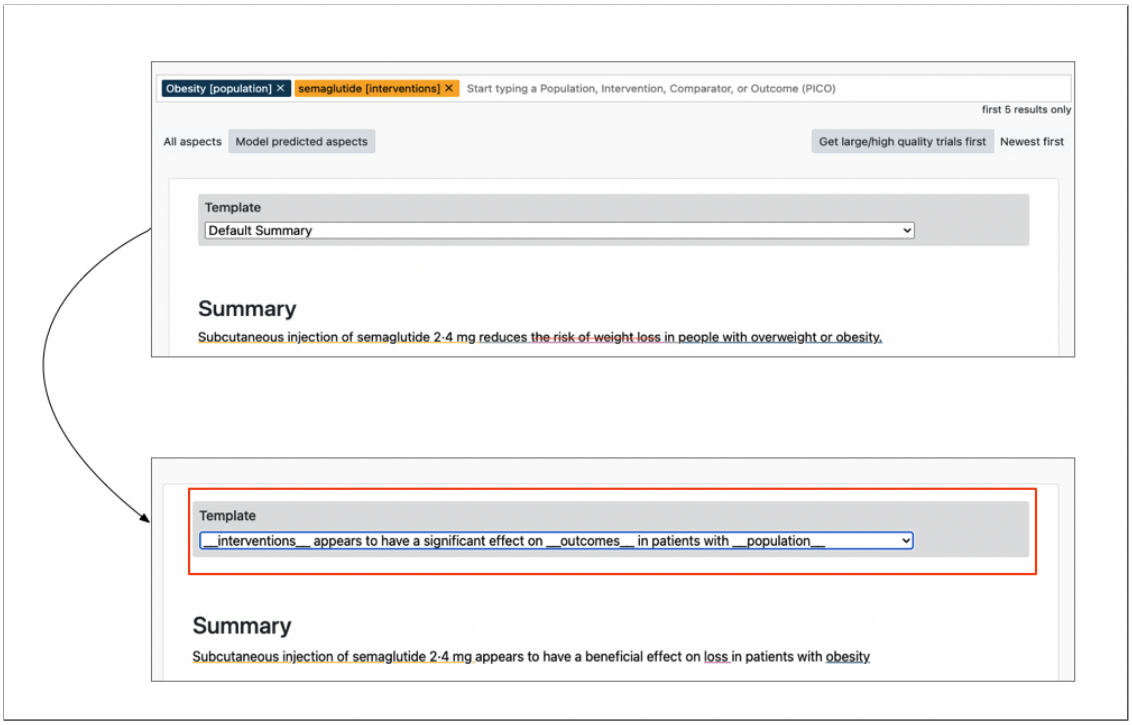


图 10：上方的摘要展示了查询司美格鲁肽 对肥胖 患者的影响时，多头模型生成的默认摘要。下方的摘要展示了使用预定义模板编辑后的摘要