[](http://crossmark.crossref.org/dialog/?doi=10.12688/f1000research.51117.2&domain=pdf&date_stamp=2023-10-09)

ĐÁNH GIÁ HỆ THỐNG

 [**Phương pháp trích xuất dữ liệu để đánh giá có hệ thống**](https://f1000research.com/articles/10-401/v2)

[**(bán tự động hóa): Cập nhật một đánh giá có hệ thống sống**](https://f1000research.com/articles/10-401/v2)

[Phiên bản 2; đánh giá ngang hàng: 3 đã được phê duyệt]

Trước đây có tiêu đề:Phương pháp trích xuất dữ liệu để đánh giá hệ thống (bán tự động hóa): Đánh giá có hệ thống sống

Lena Schmidt [](https://orcid.org/0000-0003-0709-8226)1-3, Ailbhe N. Finnerty Mutlu4, Rebecca Elmore [](https://orcid.org/0000-0003-1161-2064)2, Babatunde K. Olorisade3,5,6, James Thomas [](https://orcid.org/0000-0003-4805-4190)4, Julian P. T. Higgins3

1Đài quan sát đổi mới NIHR, Đại học Newcastle, Newcastle upon Tyne, NE4 5TG, Vương quốc Anh

2Sciome LLC, Công viên Tam giác Nghiên cứu, Bắc Carolina, 27713, Hoa Kỳ

3Trường Y Bristol, Đại học Bristol, Bristol, BS8 2PS, Vương quốc Anh

4Viện Nghiên cứu Xã hội UCL, Đại học College London, London, WC1H 0AL, Vương quốc Anh 5Trường Công nghệ Cardiff , Đại học Cardiff Metropolitan, Cardiff, CF5 2YB, Vương quốc Anh 6Evaluate Ltd, London, SE1 2RE, Vương quốc Anh

**v2 Xuất bản lần đầu:** 19 Tháng Năm 2021, **10**:401 <https://doi.org/10.12688/f1000research.51117.1>

**Xuất bản lần cuối:** 09 Oct 2023, **10**:401 <https://doi.org/10.12688/f1000research.51117.2>

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Đánh giá ngang hàng mở** | | | | |
|  | **Trạng thái phê duyệt** | | | |
|  | | | | |
|  |  | 1 | 2 | 3 |
|  | [**Phiên bản 2**](https://f1000research.com/articles/10-401/v2)  (cập nhật)  09 Tháng Mười 2023  [**Phiên bản 1**](https://f1000research.com/articles/10-401/v1)  19 Tháng Năm 2021 | [cảnh](https://f1000research.com/articles/10-401/v2#referee-response-85692) | [cảnh](https://f1000research.com/articles/10-401/v2#referee-response-89348) | [cảnh](https://f1000research.com/articles/10-401/v2#referee-response-89347) |
| 1. **Emma McFarlane** , Viện Quốc gia về Sức khỏe và Chăm sóc Xuất sắc, London, Vương quốc Anh 2. **Kathryn A. Kaiser** , Đại học Alabama tại Birmingham, Birmingham, Hoa Kỳ 3. **Carmen Amezcua-Prieto** , Đại học Granada, Granada, Tây Ban Nha   Bất kỳ báo cáo và phản hồi hoặc nhận xét nào về bài viết có thể được tìm thấy ở cuối bài viết. | | | | |



**Trừu tượng**

**Bối cảnh:** Tự động hóa (bán nửa) đáng tin cậy và có thể sử dụng được của trích xuất dữ liệu có thể hỗ trợ lĩnh vực đánh giá có hệ thống bằng cách giảm khối lượng công việc cần thiết để thu thập thông tin về tiến hành và kết quả của các nghiên cứu được đưa vào. Đánh giá có hệ thống sống này kiểm tra các phương pháp tiếp cận đã được công bố để trích xuất dữ liệu từ các báo cáo của các nghiên cứu lâm sàng.

**Phương pháp:** Chúng tôi tìm kiếm một cách có hệ thống và liên tục PubMed, ACL Anthology, arXiv, OpenAlex thông qua EPPI-Reviewer và  *thư mục khoa học máy tính dblp*. Sàng lọc toàn văn bản và trích xuất dữ liệu được thực hiện trong một ứng dụng đánh giá có hệ thống sống nguồn mở được tạo cho mục đích của đánh giá này. Bản cập nhật đánh giá sống này bao gồm các ấn phẩm cho đến tháng 12 năm 2022 và nội dung OpenAlex cho đến tháng 3 năm 2023.

**Kết quả:** 76 ấn phẩm được đưa vào đánh giá này. Trong số này, 64 (84%) các ấn phẩm đề cập đến việc trích xuất dữ liệu từ các bản tóm tắt, trong khi 19 (25%) sử dụng toàn bộ văn bản. Tổng cộng 71 (93%) ấn phẩm đã phát triển các bộ phân loại cho các thử nghiệm ngẫu nhiên có đối chứng. Hơn 30 thực thể đã được trích xuất, với PICO (dân số, can thiệp, so sánh, kết quả) là được trích xuất thường xuyên nhất. Dữ liệu có sẵn từ 25 (33%) và mã từ 30 (39%) ấn phẩm. Sáu (8%) đã triển khai các công cụ có sẵn công khai

**Kết luận:** Đánh giá có hệ thống sống này trình bày tổng quan về tài liệu trích xuất dữ liệu (bán tự động được quan tâm đối với các loại đánh giá tài liệu khác nhau. Chúng tôi đã xác định một cơ sở bằng chứng rộng rãi về

các ấn phẩm mô tả trích xuất dữ liệu cho các đánh giá can thiệp và một số lượng nhỏ các ấn phẩm trích xuất dữ liệu chính xác dịch tễ học hoặc chẩn đoán. Giữa các bản cập nhật đánh giá, xu hướng chia sẻ dữ liệu và mã tăng mạnh: trong đánh giá cơ sở, dữ liệu và mã có sẵn lần lượt là 13 và 19%, những con số này tăng lên 78 và 87% trong 23 ấn phẩm mới. So với đánh giá cơ sở, chúng tôi quan sát thấy một xu hướng nghiên cứu khác, tránh xa việc trích xuất dữ liệu đơn giản và hướng tới trích xuất thêm mối quan hệ giữa các thực thể hoặc tóm tắt văn bản tự động. Với đánh giá sống này, chúng tôi mong muốn xem xét tài liệu liên tục.

**Từ khoá**

Trích xuất dữ liệu, xử lý ngôn ngữ tự nhiên, khả năng tái tạo, đánh giá có hệ thống, khai thác văn bản

[](https://f1000research.com/collections/livingevidence)Bài viết này được bao gồm trong [bộ sưu tập Bằng](https://f1000research.com/collections/livingevidence)  chứng sống.

**Tác giả tương ứng:** Lena Schmidt ([lena.schmidt@bristol.ac.uk](mailto:lena.schmidt@bristol.ac.uk))

**Vai trò tác giả: Schmidt L**: Khái niệm, Điều tra, Phương pháp luận, Phần mềm, Trực quan, Viết - Chuẩn bị bản thảo gốc; **Finnerty Mutlu AN:** Quản lý dữ liệu, Điều tra, Viết – Đánh giá & Chỉnh sửa; **Elmore R**: Quản lý dữ liệu, Điều tra, Viết – Đánh giá & Chỉnh sửa; **Olorisade BK:** Khái niệm, Điều tra, Phương pháp, Phần mềm, Viết – Đánh giá & Chỉnh sửa; **Thomas J**: Khái niệm, Điều tra, Phương pháp luận, Viết – Đánh giá & Chỉnh sửa; **Higgins JPT**: Khái niệm, Thu thập tài trợ, Điều tra, Phương pháp luận, Viết – Đánh giá & Chỉnh sửa

**Lợi ích cạnh tranh:** Không có lợi ích cạnh tranh nào được tiết lộ.

**Thông tin tài trợ:** Chúng tôi ghi nhận tài trợ từ NIHR (LAM thông qua Học bổng Nghiên cứu Tiến sĩ NIHR (DRF-2018-11-ST2-048) và LS thông qua Học bổng Đánh giá Hệ thống NIHR (RM-SR-2017-09-028)). LAM là thành viên của Đơn vị Dịch tễ học Tích hợp MRC tại Đại học Bristol. Quan điểm được thể hiện trong bài viết này là của các tác giả và không nhất thiết đại diện cho quan điểm của NHS, NIHR, MRC hoặc Bộ Y tế và Chăm sóc Xã hội.

*Các nhà tài trợ không có vai trò trong việc thiết kế nghiên cứu, thu thập và phân tích dữ liệu, quyết định xuất bản hoặc chuẩn bị bản thảo.*

**Bản quyền:** © 2023 Schmidt L *và cộng sự*. Đây là một bài viết truy cập mở được phân phối theo các điều khoản của [Giấy phép](https://creativecommons.org/licenses/by/4.0/)  [Ghi công Creative Commons](https://creativecommons.org/licenses/by/4.0/), cho phép sử dụng, phân phối và sao chép không hạn chế trong bất kỳ phương tiện nào, miễn là tác phẩm gốc được trích dẫn đúng cách.

**Cách trích dẫn bài viết này:** Schmidt L, Finnerty Mutlu AN, Elmore R *và cộng sự.* **Phương pháp trích xuất dữ liệu để đánh giá hệ thống (bán tự động): Cập nhật đánh giá có hệ thống sống [phiên bản 2; bình duyệt: 3 đã được phê duyệt]** F1000Research 2023, **10**:401 <https://doi.org/10.12688/f1000research.51117.2>

**Xuất bản lần đầu:** 19 Tháng Năm 2021, **10**:401 <https://doi.org/10.12688/f1000research.51117.1>



*CẬP NHẬT* Các sửa đổi từ Phiên bản 1

Phiên bản này của LSR bao gồm 23 bài báo mới, một sự thay đổi trong tiêu đề cho thấy rằng phiên bản hiện tại là một bản cập nhật. Ailbhe Finnerty và Rebecca Elmore tham gia nhóm tác giả sau khi đóng góp vào việc sàng lọc và trích xuất dữ liệu; Luke A. McGuinness đã đóng góp vào đánh giá cơ sở nhưng không được liệt kê là tác giả trong bản cập nhật này. Bản tóm tắt và kết luận đã được cập nhật để phản ánh những thay đổi và xu hướng nghiên cứu mới như tăng tính sẵn có của bộ dữ liệu, mã nguồn, nhiều bài báo mô tả trích xuất và tóm tắt mối quan hệ hơn. Chúng tôi đã cập nhật các số liệu và bảng hiện có ngoại trừ Bảng 1 (kỹ thuật tiền xử lý), vì sự phụ thuộc vào tiền xử lý đã giảm trong những năm gần đây. Bảng 1 trong phụ lục đã được đổi tên thành 'Bảng A1' để tránh nhầm lẫn với Bảng 1 trong văn bản chính.

Trong đánh giá cơ sở, chúng tôi đánh giá các ấn phẩm bao gồm dựa trên danh sách 17 mục trong các lĩnh vực tái tạo (3.4.1), tính minh bạch (3.4.2), mô tả thử nghiệm (3.4.3), tính sẵn có của dữ liệu (3.4.4) và hiệu lực bên trong và bên ngoài (3.4.5). Danh sách các mục đã được giảm xuống còn sáu mục cho bản cập nhật, thông tin thêm về các mục đã bị xóa có thể được tìm thấy trong phần phương pháp của LSR này. Chúng tôi vẫn bao gồm các mục sau:

* 3.4.2.2 Có mô tả về bộ dữ liệu được sử dụng và các đặc điểm của nó không?
* 3.4.2.4 Mã nguồn có sẵn không?
* 3.4.3.2 Các chỉ số cơ bản có được báo cáo (dương tính và âm tính đúng/sai) không?
* 3.4.4.1 Chúng tôi có thể lấy phiên bản có thể chạy được của phần mềm dựa trên thông tin trong ấn phẩm không?
* 3.4.4.2 Tính bền vững: Dữ liệu có thể được truy xuất dựa trên thông tin được cung cấp trong ấn phẩm không?
* 3.4.5.1 Bộ dữ liệu hoặc thước đo đánh giá có cung cấp khả năng so sánh với các công cụ khác trong cùng một lĩnh vực không?

Ngoài ra, bảng tính với tất cả dữ liệu được trích xuất và số liệu cập nhật có sẵn dưới dạng Phụ lục D.

Bất kỳ câu trả lời nào khác từ những người phản biện có thể được tìm thấy ở cuối bài viết

1. Giới thiệu

Trong một đánh giá có hệ thống, trích xuất dữ liệu là quá trình nắm bắt các đặc điểm chính của các nghiên cứu ở dạng có cấu trúc và tiêu chuẩn hóa dựa trên thông tin trong các bài báo và báo cáo tạp chí. Nó là tiền đề cần thiết để đánh giá nguy cơ thiên vị trong các nghiên cứu riêng lẻ và tổng hợp những phát hiện của họ. Các đánh giá có hệ thống can thiệp, chẩn đoán hoặc tiên lượng thường trích xuất thông tin từ một tập hợp các lĩnh vực cụ thể có thể được xác định trước.[1](#_bookmark20) Các lĩnh vực phổ biến nhất để trích xuất trong các đánh giá can thiệp được xác định trong khuôn khổ PICO (dân số, can thiệp, so sánh, kết quả) và các khuôn khổ tương tự có sẵn cho các loại đánh giá khác. Nhiệm vụ trích xuất dữ liệu có thể tốn thời gian và lặp đi lặp lại khi được thực hiện bằng tay. Điều này tạo cơ hội hỗ trợ thông qua phần mềm thông minh, tự động xác định và trích xuất thông tin. Khi được áp dụng cho lĩnh vực nghiên cứu sức khỏe, tự động hóa (bán nửa) này nằm ở giao diện giữa y học dựa trên bằng chứng (EBM) và khoa học dữ liệu, và như được mô tả trong phần sau, sự quan tâm đến sự phát triển của nó đã tăng song song với sự quan tâm đến AI trong các lĩnh vực khác của khoa học máy tính.

* 1. Đánh giá và tổng quan hệ thống liên quan

Theo hiểu biết tốt nhất của chúng tôi, đánh giá này là đánh giá có hệ thống duy nhất (LSR) về các phương pháp trích xuất dữ liệu. Chúng tôi đã xác định bốn đánh giá trước đó về các công cụ và phương pháp trong lần lặp lại đầu tiên của đánh giá sống này (sau đây được gọi là đánh giá cơ sở),[2](#_bookmark22)–[5](#_bookmark23) và hai tài liệu cung cấp tổng quan và hướng dẫn liên quan đến chủ đề của chúng tôi.[3,6,7](#_bookmark25) Giữa đánh giá cơ sở và cập nhật này, chúng tôi đã xác định thêm sáu đánh giá tài liệu có liên quan (có hệ thống) sẽ được tóm tắt trong các đoạn sau.[8](#_bookmark29)–[13](#_bookmark40)

Các đánh giá liên quan đến năm 2014: Các đánh giá có hệ thống từ năm 2014 đến năm 2015 trình bày tổng quan về các phương pháp học máy cổ điển và xử lý ngôn ngữ tự nhiên (NLP) được áp dụng cho các nhiệm vụ như khai thác dữ liệu trong lĩnh vực y học dựa trên bằng chứng. Tại thời điểm xuất bản các tài liệu này, các phương pháp như mô hình hóa chủ đề (Phân bổ Dirichlet tiềm ẩn) và máy vectơ hỗ trợ (SVM) được coi là hiện đại cho các mô hình ngôn ngữ.

Năm 2014, Tsafnat *và cộng sự.* cung cấp một cái nhìn tổng quan rộng rãi về các công nghệ tự động hóa cho các giai đoạn khác nhau của việc tạo ra một đánh giá có hệ thống.[5](#_bookmark23) O'Mara-Eves *và cộng sự* đã xuất bản một đánh giá có hệ thống tập trung vào các phương pháp khai thác văn bản vào năm 2015.[4](#_bookmark21) Nó bao gồm một bản tóm tắt các phương pháp để đánh giá các hệ thống, chẳng hạn như khả năng thu hồi, độ chính xác và điểm F1 (trung bình hài hòa của khả năng nhớ lại và độ chính xác, một số liệu thường được sử dụng trong học máy). Các phản biện tập trung vào các nhiệm vụ liên quan đến phân loại PICO và hỗ trợ quá trình sàng lọc. Cùng năm đó, Jonnalagadda, Goyal và Huffman[3](#_bookmark24) đã mô tả các phương pháp trích xuất dữ liệu, tập trung vào PICO và các lĩnh vực liên quan. Tuổi của các ấn phẩm này có nghĩa là các phương pháp thần kinh và dựa trên nhúng tĩnh hoặc ngữ cảnh mới nhất không được bao gồm. Tuy nhiên, những phương pháp mới hơn này được sử dụng trong phần mềm tự động hóa đánh giá hệ thống đương đại sẽ được xem xét trong phạm vi của đánh giá sống này.

Đánh giá liên quan đến năm 2020: Đánh giá đến năm 2020 tập trung vào các cuộc thảo luận xung quanh việc phát triển và tích hợp công cụ trong thực tế, đồng thời đánh dấu ngày bắt đầu đưa các phương pháp tự động hóa dựa trên mạng nơ-ron. Beller *và cộng sự.*

Mô tả các nguyên tắc phát triển và tích hợp các công cụ để tự động hóa đánh giá có hệ thống.[6](#_bookmark25) Marshall và Wallace[7](#_bookmark28) trình bày một hướng dẫn về công nghệ tự động hóa, tập trung vào sự sẵn có của các công cụ và việc áp dụng vào thực tế. Họ kết luận rằng các công cụ tạo điều kiện sàng lọc có thể tiếp cận và sử dụng rộng rãi, trong khi các công cụ trích xuất dữ liệu vẫn đang ở giai đoạn thí điểm hoặc yêu cầu lượng đầu vào của con người cao hơn.

Một đánh giá có hệ thống về máy học để tự động hóa đánh giá có hệ thống, được xuất bản bằng tiếng Bồ Đào Nha vào năm 2020, bao gồm 35 ấn phẩm. Các tác giả đã kiểm tra các tạp chí trong đó các ấn phẩm về tự động hóa đánh giá có hệ thống được xuất bản và tiến hành phân tích tần suất thuật ngữ và trích dẫn. Họ phân loại các bài báo theo nhiệm vụ đánh giá có hệ thống và cung cấp một cái nhìn tổng quan ngắn gọn về các phương pháp trích xuất dữ liệu.[2](#_bookmark22)

Đánh giá liên quan sau năm 2020: Sáu bài đánh giá này bao gồm và thảo luận về các công cụ của người dùng cuối và bao gồm các nhiệm vụ khác nhau trong quy trình làm việc SR, bao gồm cả trích xuất dữ liệu. So với LSR này, các bài đánh giá này có phạm vi rộng hơn nhưng có ít tài liệu tham khảo hơn về tự động hóa trích xuất dữ liệu. Ruiz và Duffy[10](#_bookmark34) đã thực hiện một phân tích tài liệu và xu hướng cho thấy số lượng tài liệu tham khảo được công bố về tự động hóa SR đang tăng đều đặn. Sundaram và Berleant[11](#_bookmark36) phân tích 29 tài liệu tham khảo áp dụng khai thác văn bản cho các phần khác nhau của quy trình SR và lưu ý rằng 24 tài liệu tham khảo mô tả tự động hóa trong lựa chọn nghiên cứu trong khi lỗ hổng nghiên cứu nổi bật nhất đối với trích xuất, giám sát, đánh giá chất lượng và tổng hợp dữ liệu.[11](#_bookmark36) Khalil và cộng sự.[9 bao](#_bookmark32)  gồm 47 công cụ và mô tả các nghiên cứu xác nhận trong đánh giá phạm vi, trong đó 8 công cụ dành cho người dùng cuối có sẵn chủ yếu tập trung vào sàng lọc, nhưng cũng bao gồm trích xuất dữ liệu và đánh giá nguy cơ thiên vị. Họ thảo luận về những hạn chế của các công cụ như thiếu khả năng tổng quát, tích hợp, tài trợ và hiệu suất hoặc khả năng tiếp cận hạn chế.[9](#_bookmark32) Cierco Jimenez và cộng sự.[8 bao](#_bookmark29)  gồm 63 tài liệu tham khảo trong đánh giá bản đồ của máy học để hỗ trợ SR trong các bước quy trình làm việc khác nhau, trong đó 41 là các công cụ người dùng cuối có sẵn để sử dụng bởi các nhà nghiên cứu không có nền tảng tin học. Theo các đánh giá khác, họ mô tả sàng lọc là bước tự động thường xuyên nhất, trong khi các công cụ trích xuất dữ liệu tự động bị thiếu do sự phức tạp của nhiệm vụ. Zhang và cộng sự.[12 bao](#_bookmark38)  gồm 49 tài liệu tham khảo về tự động hóa các lĩnh vực trích xuất dữ liệu như bệnh, kết quả hoặc siêu dữ liệu. Họ tập trung vào việc trích xuất từ các văn bản y học cổ truyền Trung Quốc như các văn bản thử nghiệm lâm sàng đã xuất bản, hồ sơ sức khỏe hoặc văn học cổ.[12](#_bookmark38) Schmidt và cộng sự.[13 đã](#_bookmark40)  xuất bản một đánh giá tường thuật về các công cụ tập trung vào tự động hóa đánh giá có hệ thống sống. Họ thảo luận về các công cụ tự động hóa hoặc hỗ trợ truy xuất tài liệu liên tục là dấu hiệu của LSR, trong khi tự động hóa trích xuất dữ liệu tích hợp tốt (bán nửa) và tự động phổ biến hoặc trực quan hóa kết quả giữa các bản cập nhật đánh giá chính thức được hỗ trợ bởi một số người, nhưng ít phổ biến hơn.

* 1. Mục đích Chúng tôi mong muốn xem xét các phương pháp và công cụ đã xuất bản nhằm tự động hóa hoặc (bán phần) tự động hóa quá trình trích xuất dữ liệu trong bối cảnh đánh giá có hệ thống các nghiên cứu nghiên cứu y học. Chúng tôi sẽ làm điều này dưới dạng đánh giá có hệ thống sống, cập nhật thông tin và phù hợp với những thách thức mà những người đánh giá có hệ thống phải đối mặt bất cứ lúc nào.

Mục tiêu của chúng tôi trong việc xem xét tài liệu này là gấp đôi. Đầu tiên, chúng tôi muốn kiểm tra các phương pháp và công cụ từ góc độ khoa học dữ liệu, tìm cách giảm các nỗ lực trùng lặp, tóm tắt kiến thức hiện tại và khuyến khích khả năng so sánh các phương pháp đã xuất bản. Thứ hai, chúng tôi tìm cách làm nổi bật giá trị gia tăng của các phương pháp và công cụ từ quan điểm của những người đánh giá có hệ thống, những người muốn sử dụng tự động hóa (bán phần) để trích xuất dữ liệu, tức là mức độ tự động hóa là bao nhiêu? Nó có đáng tin cậy không? Chúng tôi giải quyết những vấn đề này bằng cách tóm tắt các cảnh báo quan trọng được thảo luận trong tài liệu, cũng như các yếu tố tạo điều kiện thuận lợi cho việc áp dụng các công cụ trong thực tế.

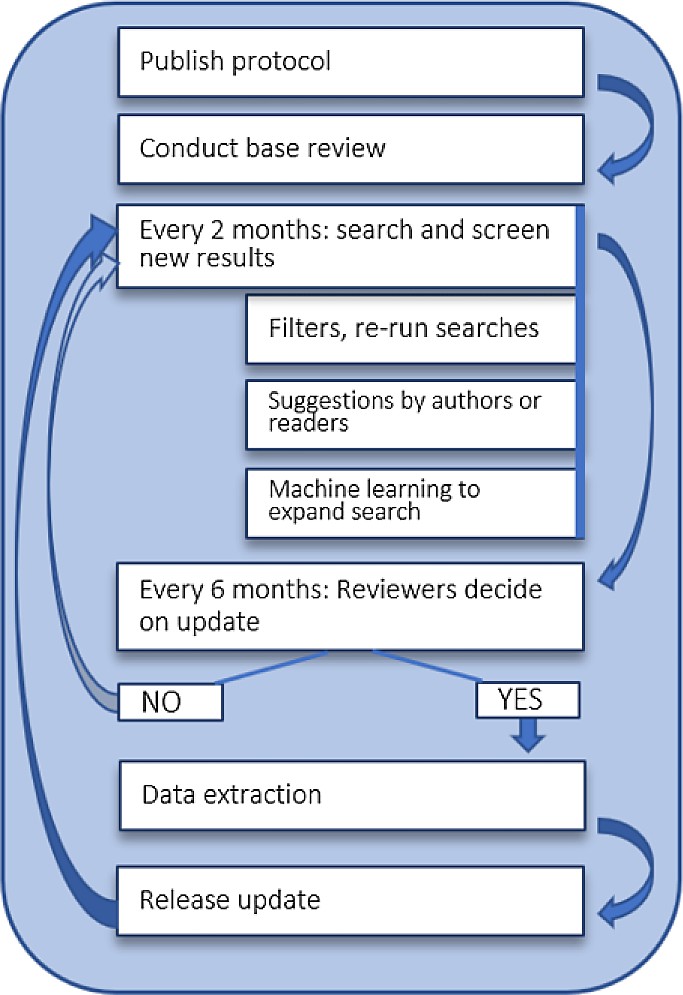
1. Phương pháp
   1. Đăng ký/giao thức

Đánh giá này được thực hiện theo một giao thức đã đăng ký trước và được công bố.[15](#_bookmark44) PROSPERO ban đầu được coi là nền tảng để đăng ký, nhưng nó chỉ giới hạn ở các đánh giá với kết quả liên quan đến sức khỏe. Bất kỳ sai lệch nào so với giao thức đã được mô tả dưới đây.

* 1. Phương pháp đánh giá cuộc sống

Chúng tôi đang tiến hành một đánh giá sống vì lĩnh vực tự động hóa đánh giá hệ thống (bán nửa) đang phát triển nhanh chóng cùng với những tiến bộ trong xử lý ngôn ngữ, học máy và học sâu.

Quá trình cập nhật bắt đầu như được mô tả trong giao thức[15](#_bookmark44) và được thể hiện trong [Hình 1](#_bookmark0). Tóm lại, chúng tôi sẽ liên tục cập nhật kết quả tìm kiếm tài liệu, sử dụng các chiến lược và phương pháp tìm kiếm được mô tả trong phần 'Tìm kiếm' bên dưới. Kết quả tìm kiếm PubMed và arXiv được cập nhật hàng ngày theo cách hoàn toàn tự động thông qua API. Các bài viết từ dblp, ACL và OpenAlex thông qua EPPI-Reviewer được thêm vào hai tháng một lần. Tất cả các kết quả tìm kiếm được tự động nhập vào ứng dụng web sàng lọc đánh giá sống và trích xuất dữ liệu của chúng tôi, được mô tả trong phần 'Thu thập và phân tích dữ liệu' bên dưới.



Hình 1. Cập nhật liên tục đánh giá cuộc sống. Hình ảnh này được sao chép theo các điều khoản của [giấy phép Creative Commons Attribution 4.0 International (CC-BY 4.0)](https://creativecommons.org/licenses/by/4.0/legalcode) từ Schmidt et al.[15](#_bookmark44)

Quyết định cập nhật đánh giá đầy đủ được đưa ra sáu tháng một lần dựa trên số lượng ấn phẩm mới được thêm vào đánh giá. Để biết thêm chi tiết về điều này, vui lòng tham khảo phác đồ hoặc [hướng dẫn đánh giá có hệ thống sống của Cochrane](https://community.cochrane.org/sites/default/files/uploads/inline-files/Transform/201912_LSR_Revised_Guidance.pdf). Giữa các bản cập nhật, quá trình sàng lọc và trạng thái hiện tại của việc trích xuất dữ liệu có thể hiển thị thông qua [trang web đánh giá sống](https://l-ena.github.io/living_review_data_extraction/).

* 1. Tiêu chí đủ điều kiện
* Chúng tôi bao gồm các ấn phẩm toàn văn mô tả phương pháp NLP ban đầu để trích xuất dữ liệu liên quan đến các nhiệm vụ đánh giá có hệ thống. Các trường dữ liệu quan tâm (ở đây được gọi là thực thể hoặc câu) được điều chỉnh từ Sổ tay Cochrane về Đánh giá Hệ thống về Can thiệp,1 và được xác định trong giao thức.[15](#_bookmark44) Chúng tôi bao gồm đầy đủ các phương pháp NLP (ví dụ: biểu thức chính quy, hệ thống dựa trên quy tắc, học máy và mạng nơ-ron sâu).
* Các ấn phẩm phải mô tả một chu kỳ đầy đủ của việc thực hiện và đánh giá một phương pháp. Ví dụ, họ phải báo cáo đào tạo và ít nhất một thước đo đánh giá hiệu suất của thuật toán trích xuất dữ liệu.
* Chúng tôi bao gồm các báo cáo được xuất bản từ năm 2005 cho đến ngày nay, tương tự như công trình trước đây.[3](#_bookmark24) Chúng tôi sẽ dịch các báo cáo không phải tiếng Anh, nếu chúng tôi tìm thấy bất kỳ báo cáo nào.
* Dữ liệu mà các ấn phẩm bao gồm sử dụng để khai thác phải là văn bản từ các thử nghiệm ngẫu nhiên có đối chứng, nghiên cứu thuần tập so sánh, nghiên cứu đối chứng ca bệnh hoặc nghiên cứu cắt ngang so sánh (ví dụ: để có độ chính xác của xét nghiệm chẩn đoán). Phạm vi của các phương pháp trích xuất dữ liệu có thể được áp dụng cho toàn văn hoặc cho các bản tóm tắt trong kho dữ liệu của mỗi ấn phẩm đủ điều kiện. Chúng tôi bao gồm các ấn phẩm trích xuất dữ liệu từ các loại nghiên cứu khác, miễn là ít nhất một trong các loại nghiên cứu quan tâm của chúng tôi có trong kho dữ liệu.

Chúng tôi đã loại trừ các ấn phẩm báo cáo:

* Các phương pháp và công cụ chỉ liên quan đến xử lý hình ảnh và nhập dữ liệu y sinh từ tệp PDF mà không cần bất kỳ phương pháp NLP nào, bao gồm trích xuất dữ liệu từ đồ thị.
* Bất kỳ nghiên cứu nào tập trung hoàn toàn vào việc chuẩn bị giao thức, tổng hợp dữ liệu đã được trích xuất, viết, chỉ xử lý trước văn bản hoặc phổ biến nó.
* Các phương pháp hoặc công cụ không cung cấp phương pháp xử lý ngôn ngữ tự nhiên và chỉ cung cấp giao diện tổ chức, quản lý tài liệu, cơ sở dữ liệu hoặc kiểm soát phiên bản
* Bất kỳ ấn phẩm nào liên quan đến báo cáo sức khỏe điện tử hoặc khai thác dữ liệu di truyền.
  1. Tìm kiếm

Đánh giá cơ sở: Chúng tôi đã tìm kiếm năm cơ sở dữ liệu điện tử, sử dụng các phương pháp tìm kiếm được mô tả trước đây trong phác đồ của chúng tôi.[15](#_bookmark44) Nói tóm lại, chúng tôi tìm kiếm MEDLINE thông qua Ovid, sử dụng chiến lược tìm kiếm được phát triển với sự giúp đỡ của một chuyên gia thông tin, và tìm kiếm Bộ sưu tập cốt lõi của Web of Science và IEEE bằng cách sử dụng các điều chỉnh của chiến lược này, được thực hiện bởi các tác giả đánh giá. Các tìm kiếm trên arXiv (khoa học máy tính) và dblp được thực hiện trên các kết xuất cơ sở dữ liệu đầy đủ bằng cách sử dụng chức năng tìm kiếm được mô tả bởi McGuinness và Schmidt.[16](#_bookmark46) Kết quả tìm kiếm đầy đủ và thông tin thêm về việc truy xuất tài liệu có sẵn trong *Dữ liệu cơ bản:* Phụ lục A và B.[127](#_bookmark142)

Ban đầu, chúng tôi dự định bao gồm một tìm kiếm tài liệu đầy đủ từ Bộ sưu tập cốt lõi của Web of Science. Do số lượng lớn các ấn phẩm được truy xuất thông qua tìm kiếm này (n = 7822), trước tiên chúng tôi quyết định sàng lọc các ấn phẩm từ tất cả các nguồn khác, để đào tạo một bộ phân loại tổng hợp học máy và chỉ thêm các ấn phẩm được dự đoán là có liên quan đến đánh giá sống của chúng tôi. Điều này làm giảm các ấn phẩm của Bộ sưu tập cốt lõi của Web of Science xuống còn 547 bản tóm tắt, được thêm vào các nghiên cứu trong bước sàng lọc ban đầu. Tập dữ liệu, mã và trọng số của các mô hình được đào tạo có sẵn trong *Dữ liệu cơ bản:* Phụ lục C.[Điều](#_bookmark142) này bao gồm các biểu đồ đánh giá của từng mô hình về diện tích dưới đường cong (AUC), độ chính xác, F1, khả năng thu hồi và phương sai của kết quả xác nhận chéo cho mọi số liệu.

Cập nhật: Theo kế hoạch, chúng tôi đã thay đổi sang API PubMed để tìm kiếm MEDLINE. Quyết định này được đưa ra để tạo điều kiện thuận lợi cho việc truy xuất tài liệu tham khảo liên tục. Chúng tôi chỉ tìm kiếm tài liệu in sẵn hoặc đã xuất bản và do đó không tìm kiếm các nguồn như GITHUB hoặc các kho lưu trữ mã nguồn khác.

Cập nhật: Chúng tôi đã tìm kiếm PubMed thông qua API, arXiv (khoa học máy tính), ACL-Anthology, dblp và sử dụng EPPI-Reviewer để thu thập các trích dẫn từ MicrosoftAcademic và sau đó là OpenAlex bằng phương pháp 'Bi-Citation AND Recommendations'.

* 1. Thu thập và phân tích dữ liệu
     1. *Lựa chọn nghiên cứu*

Sàng lọc ban đầu và trích xuất dữ liệu được tiến hành như đã nêu trong giao thức. Nói tóm lại, đối với đánh giá cơ sở, chúng tôi đã sàng lọc tất cả các ấn phẩm đã truy xuất bằng công cụ Abstrackr. Tất cả các bản tóm tắt được sàng lọc bởi hai người phản biện độc lập. Các phán đoán mâu thuẫn đã được giải quyết bởi các tác giả đã đưa ra quyết định sàng lọc ban đầu. Sàng lọc toàn văn được thực hiện theo cách tương tự như sàng lọc trừu tượng nhưng sử dụng ứng dụng web của chúng tôi cho LSR được mô tả trong phần sau.

Đối với đánh giá cập nhật, chúng tôi đã sử dụng ứng dụng web đánh giá sống của mình để truy xuất tất cả các ấn phẩm ngoại trừ các mục được truy xuất bởi EPPI-Reviewer (chúng được thêm vào tập dữ liệu riêng biệt). Chúng tôi tiếp tục sử dụng ứng dụng của mình để khử trùng lặp, sàng lọc và trích xuất dữ liệu tất cả các ấn phẩm.

Một bản cập nhật phương pháp luận đối với quy trình sàng lọc bao gồm thay đổi đối với sàng lọc đơn để đánh giá tính đủ điều kiện ở cả cấp độ tóm tắt và toàn văn, giảm sàng lọc kép xuống còn 10% số ấn phẩm.

* + 1. *Trích xuất, đánh giá và quản lý dữ liệu*

Trước đây chúng tôi đã phát triển một ứng dụng web để tự động truy xuất tài liệu tham khảo cho các bản cập nhật đánh giá sống (xem Tính *khả dụng của phần mềm*[17](#_bookmark48)), để hỗ trợ sàng lọc cả bản tóm tắt và toàn văn để cập nhật đánh giá và quản lý quá trình trích xuất dữ liệu xuyên suốt.[17](#_bookmark48) Đối với các bản cập nhật trong tương lai của bài đánh giá sống này, chúng tôi sẽ sử dụng ứng dụng web, chứ không phải Abstrackr, để sàng lọc tài liệu tham khảo. Ứng dụng web này đã được sử dụng bởi một bài đánh giá sống khác.[18](#_bookmark49) Nó tự động truy xuất tài liệu tham khảo hàng ngày từ các nguồn đi kèm và có giao diện sàng lọc và trích xuất dữ liệu. Tất cả dữ liệu được trích xuất được lưu trữ trong cơ sở dữ liệu. Các số liệu và bảng có thể được xuất hàng ngày và tiến trình giữa các cập nhật đánh giá được chia sẻ trên trang web đánh giá sống của chúng tôi. Bảng tính đầy đủ của các mục được trích xuất từ mỗi tài liệu tham khảo được bao gồm có sẵn trong *Dữ liệu cơ bản.*[127](#_bookmark142) Như đã mô tả trước đây trong giao thức, chất lượng báo cáo và khả năng tái tạo ban đầu được đánh giá dựa trên danh sách kiểm tra đã được công bố trước đó về khả năng tái tạo trong khai thác văn bản, nhưng một số mục đã bị loại bỏ khỏi phạm vi cập nhật đánh giá này.[19](#_bookmark52)

Theo kế hoạch trong giao thức, một người đánh giá duy nhất đã tiến hành trích xuất dữ liệu và ngẫu nhiên 10% các ấn phẩm được bao gồm đã được kiểm tra bởi người đánh giá thứ hai.

* + 1. *Visualization*

Việc tạo tất cả các số liệu và biểu đồ tương tác trên trang web đánh giá sống và trong phần 'Kết quả' của đánh giá này được tự động hóa dựa trên nội dung có cấu trúc từ cơ sở dữ liệu đánh giá sống của chúng tôi (xem Phụ lục A và D, *Dữ liệu cơ bản*[127](#_bookmark142)). Chúng tôi tự động xuất báo cáo PDF cho mỗi ấn phẩm đi kèm. Việc tính toán tỷ lệ phần trăm, xuất văn bản được trích xuất và tạo số liệu cũng được tự động hóa.

* + 1. *Khả năng truy cập dữ liệu*

Tất cả dữ liệu và mã đều được truy cập miễn phí. Danh sách chi tiết các nguồn được đưa ra trong phần 'Tính khả dụng của dữ liệu' và 'Tính khả dụng của phần mềm'.

* 1. Thay đổi từ giao thức và giữa các bản cập nhật Trong giao thức, chúng tôi đã tuyên bố rằng dữ liệu sẽ có sẵn thông qua kho lưu trữ OSF. Thay vào đó, dữ liệu đánh giá đầy đủ có sẵn thông qua Harvard Dataverse, vì kho lưu trữ này cho phép chúng tôi giữ một DOI được chỉ định sau khi cập nhật kho lưu trữ với nội dung mới cho mỗi lần lặp lại của bài đánh giá sống này. Chúng tôi cũng tuyên bố rằng chúng tôi sẽ sàng lọc tất cả các ấn phẩm từ tìm kiếm của Web of Science. Thay vào đó, chúng tôi mô tả một cách tiếp cận đã thay đổi trong phần Phương pháp, trong phần 'Tìm kiếm'. Đối với các bản cập nhật đánh giá, Web of Science đã bị loại bỏ và thay thế bằng các tìm kiếm OpenAlex thông qua EPPI-Reviewer.

Chúng tôi đã thêm một mục trích xuất dữ liệu cho loại thông tin mà một ấn phẩm khai thác (ví dụ: P, IC, O) vào phần các mục chính quan tâm và chúng tôi chuyển loại định dạng đầu vào và đầu ra từ các mục chính sang các mục quan tâm phụ. Chúng tôi đã nhóm mục phụ quan tâm 'Các số liệu được báo cáo khác, chẳng hạn như tác động đến các quy trình đánh giá có hệ thống (ví dụ: thời gian tiết kiệm trong quá trình trích xuất dữ liệu)' với mục chính được quan tâm 'Các chỉ số hiệu suất được báo cáo được sử dụng để đánh giá'.

Mục 'Tính bền vững: tập dữ liệu có khả năng có sẵn để sử dụng trong tương lai không?' đã được thay đổi thành: 'Dữ liệu có thể được truy xuất dựa trên thông tin được cung cấp trong ấn phẩm không?'. Chúng tôi quyết định không suy đoán liệu một tập dữ liệu có khả năng có sẵn trong tương lai hay không và thay vào đó chọn ghi lại nếu tập dữ liệu có sẵn tại thời điểm chúng tôi cố gắng truy cập nó.

Mục 'Chúng ta có thể lấy một phiên bản có thể chạy được của phần mềm dựa trên thông tin trong ấn phẩm không?' đã được đổi thành

"Có ứng dụng nào có sẵn để khai thác dữ liệu, ví dụ như phiên bản ứng dụng web hoặc máy tính để bàn không?'.

Trong phiên bản đánh giá hiện tại này, chúng tôi vẫn chưa liên hệ với các tác giả của các ấn phẩm được đưa vào. Quyết định này được đưa ra do hạn chế về thời gian, tuy nhiên việc tiếp cận với các tác giả được lên kế hoạch như một phần của bản cập nhật đầu tiên cho bài đánh giá sống này.

Trong đánh giá cơ sở, chúng tôi đánh giá các ấn phẩm bao gồm dựa trên danh sách 17 mục trong các lĩnh vực tái tạo (3.4.1), tính minh bạch (3.4.2), mô tả thử nghiệm (3.4.3), tính sẵn có của dữ liệu (3.4.4) và hiệu lực bên trong và bên ngoài (3.4.5). Danh sách các vật phẩm đã được giảm xuống còn sáu mục cho bản cập nhật:

* 3.4.2.2 Có mô tả về bộ dữ liệu được sử dụng và các đặc điểm của nó không?
* 3.4.2.4 Mã nguồn có sẵn không?
* 3.4.3.2 Các chỉ số cơ bản có được báo cáo (dương tính và âm tính đúng/sai) không?
* 3.4.4.1 Chúng tôi có thể lấy phiên bản có thể chạy được của phần mềm dựa trên thông tin trong ấn phẩm không?
* 3.4.4.2 Tính bền vững: Dữ liệu có thể được truy xuất dựa trên thông tin được cung cấp trong ấn phẩm không?
* 3.4.5.1 Bộ dữ liệu hoặc thước đo đánh giá có cung cấp khả năng so sánh với các công cụ khác trong cùng một lĩnh vực không?

Các mục sau đây đã bị loại bỏ, mặc dù kết quả và thảo luận từ việc đánh giá các mục này trong đánh giá cơ sở vẫn nằm trong văn bản đánh giá:

* 3.4.1.1 Các nguồn dữ liệu đào tạo/thử nghiệm có được báo cáo không?
* 3.4.1.2 Nếu các kỹ thuật tiền xử lý được áp dụng cho dữ liệu, chúng có được mô tả không?
* 3.4.2.1 Có mô tả về các thuật toán được sử dụng không?
* 3.4.2.3 Có mô tả về phần cứng được sử dụng không?
* 3.4.3.1 Có biện minh / giải thích về đánh giá mô hình không?
* 3.4.3.3 Đánh giá có bao gồm bất kỳ thông tin nào về sự đánh đổi giữa khả năng nhớ lại hoặc độ chính xác (còn được gọi là độ nhạy và giá trị dự đoán dương tính) không?
* 3.4.4.3 Việc sử dụng các framework của bên thứ ba có được báo cáo không và chúng có thể truy cập được không?
* 3.4.5.2 Có phải giải thích cho ảnh hưởng của cả biến hiển thị và ẩn trong tập dữ liệu không?
* 3.4.5.3 Quy trình tránh quá khớp hoặc quá khớp có được mô tả không?
* 3.4.5.4 Quá trình tách đào tạo từ dữ liệu xác thực có được mô tả không?
* 3.4.5.5 Khả năng thích ứng của mô hình với các định dạng và / hoặc môi trường khác nhau ngoài dữ liệu đào tạo và thử nghiệm có được mô tả không?

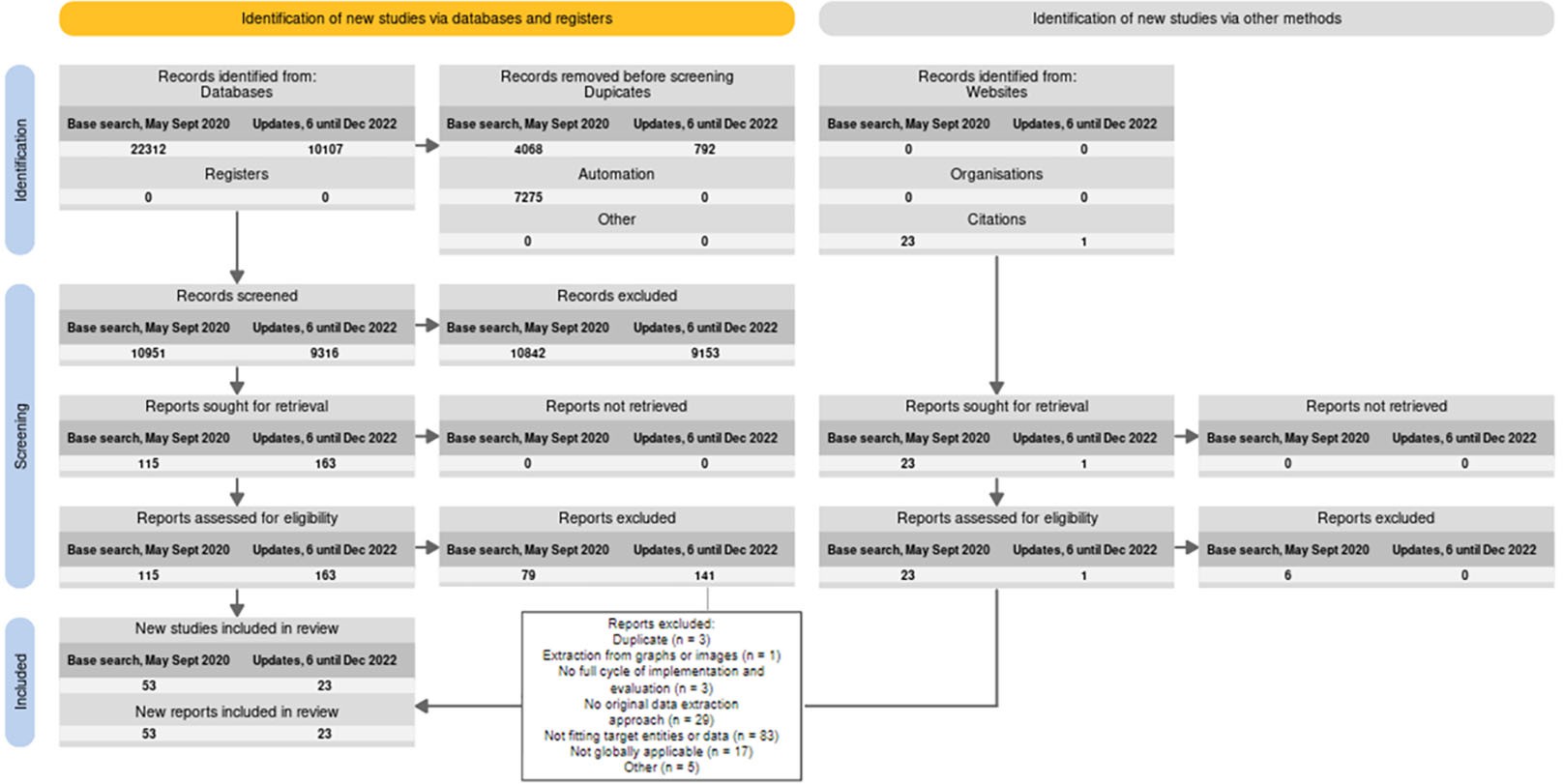
1. Kết quả
   1. Kết quả tìm kiếm

Các tìm kiếm cơ sở dữ liệu của chúng tôi đã xác định được 10.107 ấn phẩm sau khi các bản sao bị loại bỏ (xem [Hình 2](#_bookmark1)). Chúng tôi đã xác định thêm một ấn phẩm theo cách thủ công.

Lần lặp lại này của đánh giá sống bao gồm 76 ấn phẩm, được tóm tắt trong Bảng A1 trong *Dữ liệu cơ bản*[127](#_bookmark142)).

* + 1. *Các ấn phẩm bị loại trừ*

Trong quá trình đánh giá cơ sở và cập nhật, 216 ấn phẩm đã bị loại trừ ở giai đoạn sàng lọc toàn văn bản, với lý do phổ biến nhất để loại trừ là nó không phù hợp với các thực thể mục tiêu hoặc dữ liệu mục tiêu. Trong hầu hết các trường hợp, điều này là do các loại văn bản được khai thác trong các ấn phẩm. Hồ sơ sức khỏe điện tử và dữ liệu không thử nghiệm là phổ biến và chúng tôi đã tạo một danh sách các bộ dữ liệu sẽ bị loại trừ trong danh mục này (xem thêm thông tin trong *Dữ liệu cơ bản:* Phụ lục B[127](#_bookmark142)). Một số ấn phẩm đề cập đến loại văn bản phù hợp nhưng bị loại trừ vì không khai thác dữ liệu quan tâm đối với đánh giá này. Ví dụ, Norman, Leeflang và Névéol[23](#_bookmark60) đã thực hiện trích xuất dữ liệu để đánh giá độ chính xác của xét nghiệm chẩn đoán, nhưng tập trung vào việc trích xuất kết quả và dữ liệu để phân tích thống kê. Millard, Flach và Higgins[24 tuổi](#_bookmark62) và Marshall, Kuiper và Wallace[25](#_bookmark63) tuổi nhìn vào



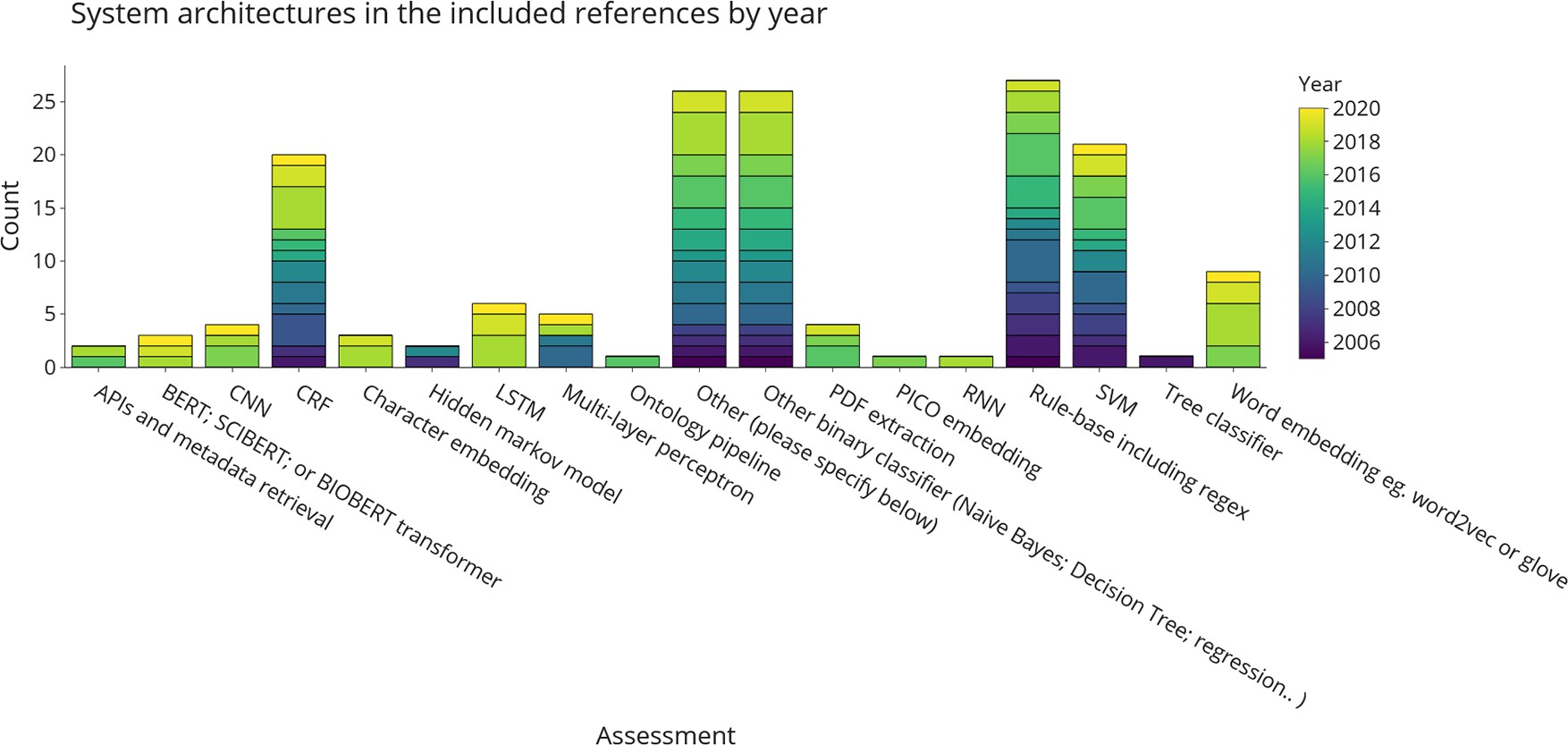
Hình 2. PRISMA2020 sơ đồ dòng chảy phù hợp với các bài đánh giá sống.[20](#_bookmark54)–[22](#_bookmark58)

phân loại nguy cơ thiên vị, nằm ngoài phạm vi của đánh giá này. Boudin, Nie và Dawes[26 đã](#_bookmark26)  phát triển một sơ đồ cân dựa trên phân tích vị trí nguyên tố PICO, để lại việc phát hiện các nguyên tố PICO đơn lẻ cho công việc trong tương lai. Luo *và cộng sự*.[27 trích](#_bookmark27)  xuất dữ liệu từ đăng ký thử nghiệm lâm sàng nhưng tập trung vào việc phân tích các tiêu chí đưa vào các thực thể sự kiện hoặc thời gian để hỗ trợ lựa chọn người tham gia cho các thử nghiệm ngẫu nhiên có đối chứng (RCT).

Lý do phổ biến thứ hai cho việc loại trừ nghiên cứu là họ 'không có phương pháp trích xuất dữ liệu gốc'. Ví dụ, Rathbone *và cộng sự*, [28](#_bookmark30) đã sử dụng các tìm kiếm Boolean thủ công cụ thể cho tiêu chí PICO của đánh giá có hệ thống để hỗ trợ quá trình sàng lọc đánh giá trong Endnote. Chúng tôi phân loại bài viết này là không có bất kỳ phương pháp trích xuất dữ liệu gốc nào vì nó không tạo ra bất kỳ đầu ra có cấu trúc nào cụ thể cho P, IC hoặc O. Malheiros *và cộng sự.*[29 người](#_bookmark31)  đã thực hiện khai thác văn bản trực quan, hỗ trợ các tác giả đánh giá có hệ thống bằng cách phân cụm tài liệu và đánh dấu văn bản. Tương tự, Fabbri *và cộng sự.*[30 đã](#_bookmark33)  triển khai một công cụ hỗ trợ toàn bộ quy trình đánh giá có hệ thống, từ giao thức đến trích xuất dữ liệu, thực hiện phân cụm và xác định các ấn phẩm tương tự. Các nhiệm vụ đánh giá có hệ thống khác có thể được hưởng lợi từ tự động hóa nhưng đã bị loại trừ khỏi đánh giá này được liệt kê trong *Dữ liệu cơ bản:* Phụ lục B.[127](#_bookmark142)

* 1. Kết quả trích xuất dữ liệu: Các mục quan tâm chính
     1. *Phương pháp tự động hóa được sử dụng*

[Hình 3](#_bookmark2) cho thấy các khía cạnh của kiến trúc hệ thống được triển khai trong các ấn phẩm đi kèm. Một bản tóm tắt ngắn gọn về những điều này cho mỗi ấn phẩm được cung cấp trong Bảng A1 trong *Dữ liệu cơ bản.*[127](#_bookmark142) Nếu có thể, chúng tôi đã cố gắng chia nhỏ các kiến trúc hệ thống lớn hơn thành các thành phần nhỏ hơn. Ví dụ, một kiến trúc kết hợp mạng nhúng từ + bộ nhớ ngắn hạn dài (LSTM) sẽ được chia thành hai thành phần phụ tương ứng. Chúng tôi nhóm các bộ phân loại nhị phân, chẳng hạn như Bayes ngây thơ và hồi quy logistic. Mặc dù SVM cũng là bộ phân loại nhị phân, nhưng nó được chỉ định là danh mục riêng biệt do mức độ phổ biến của nó. Các danh mục cuối cùng là sự kết hợp của tự động hóa không nghiêng về máy móc (giao diện lập trình ứng dụng (API) và truy xuất siêu dữ liệu, trích xuất PDF, cơ sở quy tắc), học máy cổ điển (Bayes ngây thơ, cây quyết định, SVM hoặc các bộ phân loại nhị phân khác) và các phương pháp tiếp cận thần kinh hoặc học sâu (mạng nơ-ron tích chập (CNN), LSTM, máy biến áp hoặc nhúng từ). Hình này cho thấy rằng không có sự lựa chọn rõ ràng về kiến trúc hệ thống cho nhiệm vụ này. Đối với bản cập nhật LSR, xu hướng mạnh nhất là việc áp dụng ngày càng tăng BERT (Biểu diễn bộ mã hóa hai chiều từ máy biến áp). BERT được xuất bản vào năm 2018 và các phiên bản kiến trúc giống hệt nhau khác của nó được điều chỉnh để sử dụng văn bản khoa học, chẳng hạn như SciBERT, được tóm tắt trong cùng một danh mục trong bài đánh giá này.[14,31](#_bookmark43) Trong đánh giá cơ sở, BERT đã được sử dụng ba lần, trong khi bây giờ xuất hiện 21 lần. Các kiến trúc dựa trên máy biến áp khác như phiên bản đào tạo trước sinh học của ELECTRA, cũng đang thu hút sự chú [ý,32,33](#_bookmark37) cũng như các mô hình dựa trên FLAIR.[34](#_bookmark41)–[36](#_bookmark45)



Hình 3. Kiến trúc hệ thống được sử dụng để tự động hóa trích xuất dữ liệu trong các ấn phẩm đi kèm. Kết quả được chia thành các loại khác nhau của phương pháp học máy và xử lý ngôn ngữ tự nhiên và được tô màu theo năm xuất bản. Có thể có nhiều hơn một thành phần kiến trúc cho mỗi ấn phẩm. Trong đó API, giao diện lập trình ứng dụng; BERT, biểu diễn bộ mã hóa hai chiều từ Transformers; CNN, mạng nơ-ron tích chập; CRF, các trường ngẫu nhiên có điều kiện; LSTM, trí nhớ ngắn hạn dài; PICO, dân số, can thiệp, so sánh, kết quả; RNN, mạng nơ-ron tuần hoàn; SVM, hỗ trợ máy vector.

Cơ sở quy tắc, bao gồm các phương pháp sử dụng phỏng đoán, danh sách từ và biểu thức chính quy, là một trong những kỹ thuật sớm nhất được sử dụng để trích xuất dữ liệu trong tài liệu EBM. Nó vẫn là cách tiếp cận được sử dụng thường xuyên nhất để tự động hóa. Chín ấn phẩm (12%) chỉ sử dụng cơ sở quy tắc, trong khi các ấn phẩm còn lại sử dụng chúng kết hợp với các bộ phân loại khác (dữ liệu được hiển thị trong *Dữ liệu cơ bản:* Phụ lục A và D[127](#_bookmark142)). Mặc dù được sử dụng thường xuyên hơn trong quá khứ, 11 ấn phẩm được xuất bản từ năm 2017 đến nay sử dụng cách tiếp cận này cùng với các kiến trúc khác như [BERT,33,37](#_bookmark39)–[39](#_bookmark51) trường ngẫu nhiên có điều kiện (CRF),[40](#_bookmark53) sử dụng nó với SVM[41](#_bookmark55) hoặc các bộ phân loại nhị phân khác.[42](#_bookmark56) Trong thực tế, các hệ thống này sử dụng các cơ sở quy tắc dưới dạng danh sách thủ công để xác định các cụm từ ứng cử viên cho các thực thể số tiền như cỡ mẫu[42,43](#_bookmark56) hoặc để tinh chỉnh kết quả thu được bởi bộ phân loại máy học ở cấp độ thực thể (ví dụ: các trường hợp trong đó một can thiệp hoặc kết quả cụ thể được trích xuất từ một câu).[40](#_bookmark53)

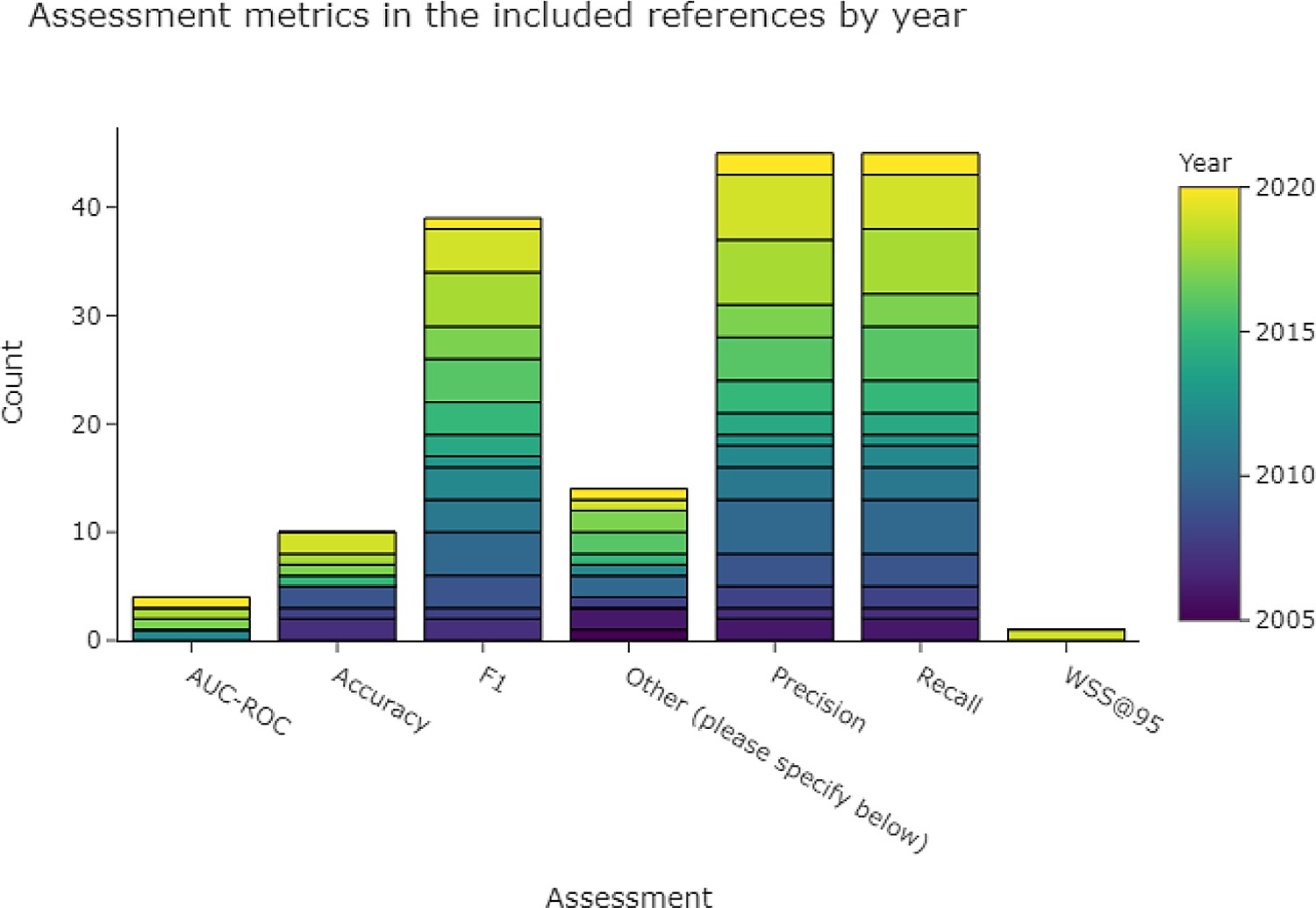
Bộ phân loại nhị phân, đáng chú ý nhất là Bayes và SVM ngây thơ, cũng là các thành phần hệ thống thường được sử dụng trong tài liệu trích xuất dữ liệu. Chúng thường được sử dụng trong các nghiên cứu được công bố từ năm 2005 đến nay nhưng việc sử dụng chúng bắt đầu giảm với sự ra đời của các mô hình thần kinh.

Nhúng và kiến trúc thần kinh đang ngày càng được sử dụng trong tài liệu trong bảy năm qua. Mạng nơ-ron tuần hoàn (RNN), mạng CNN và LSTM yêu cầu lượng dữ liệu đào tạo lớn hơn; bằng cách sử dụng nhúng dựa trên máy biến áp với các thuật toán đào tạo trước dựa trên dữ liệu không được gắn nhãn, chúng ngày càng trở nên thú vị hơn trong các lĩnh vực như trích xuất dữ liệu cho EBM- nơi dữ liệu đào tạo chất lượng cao rất khó và tốn kém.

Trong danh mục 'Khác', các công cụ được đề cập chủ yếu là các bộ phân loại khác như bộ phân loại entropy tối đa (n = 3), kLog, J48 và các thuật toán phân loại vị trí hoặc độ dài tài liệu khác nhau. Chúng tôi cũng thêm các phương pháp như giám sát từ xa có giám sát (n = 3, xem Tham khảo. [44](#_bookmark59)) và các phương pháp đào tạo mới đối với các kiến trúc thần kinh hiện có trong danh mục này.

* + 1. *Các chỉ số hiệu suất được báo cáo được sử dụng để đánh giá*

Độ chính xác (tức là giá trị dự đoán dương), khả năng nhớ lại (tức là độ nhạy) và điểm F1 (trung bình hài hòa của độ chính xác và khả năng thu hồi) là những số liệu được sử dụng rộng rãi nhất để đánh giá các bộ phân loại. Điều này được phản ánh trong [Hình 4](#_bookmark3), cho thấy ít nhất một trong các số liệu này đã được sử dụng trong phần lớn các ấn phẩm được đưa vào. Độ chính xác và diện tích dưới đường cong - đặc điểm của người vận hành máy thu (AUC-ROC) ít được sử dụng thường xuyên hơn.



Hình 4. Các chỉ số đánh giá phổ biến nhất được sử dụng trong các ấn phẩm bao gồm để đánh giá hiệu suất của hệ thống trích xuất dữ liệu. Có thể có nhiều hơn một số liệu cho mỗi ấn phẩm, có nghĩa là tổng số ấn phẩm được bao gồm (n = 76) thấp hơn tổng số lượng của các thanh trong con số này. AUC-ROC, khu vực dưới đường cong - đặc điểm của người vận hành máy thu; F1, trung bình sóng hài của độ chính xác và khả năng thu hồi.

Có một số cách tiếp cận và biện minh của việc sử dụng điểm chính xác, khả năng nhớ hoặc F1 trung bình vĩ mô hoặc vi mô trong các ấn phẩm bao gồm. Điểm vi mô hoặc vĩ mô được tính toán trong các trường hợp nhiều lớp và điểm cuối cùng có thể khác nhau bất cứ khi nào các lớp trong tập dữ liệu bị mất cân bằng (như trường hợp trong hầu hết các bộ dữ liệu được sử dụng để tự động trích xuất dữ liệu trong tự động hóa SR).

Cả điểm vi mô và vĩ mô đều được báo cáo bởi Singh et al. (2021), [45](#_bookmark61) Kilicoglu et al. (2021), [38](#_bookmark50) Kiritchenko et al. (2010), [46](#_bookmark64) Fiszman et al. (2007) [47](#_bookmark65) trong khi Karystianis et al. (2014, 2017) [48,49](#_bookmark67) báo cáo vi mô trên các tài liệu và vĩ mô trên các lớp.

Điểm vĩ mô đã được sử dụng trong một ấn phẩm.[37](#_bookmark47)

Điểm vi mô đã được sử dụng bởi Fiszman và cộng sự.[47](#_bookmark65) cho kết quả cấp lớp. Trong một ấn phẩm, trung bình hài hòa được sử dụng cho độ chính xác và thu hồi, trong khi tính điểm vi mô được sử dụng cho F1.[50](#_bookmark71) Điểm vi mô được sử dụng rộng rãi nhất, bao gồm Al-Hussaini et al. (2022), [32](#_bookmark37) Sanchez-Graillet et al. (2022), [51](#_bookmark73) Kim et al. (2011), [52](#_bookmark75) Verbeke et al. (2012), [53](#_bookmark77) và Jin and Szolovits (2020) [54](#_bookmark78) được sử dụng trong kịch bản đánh giá của Nye et al. (2018).[55](#_bookmark80)

Trong danh mục 'Khác', chúng tôi đã thêm một số trường hợp trong đó sự nới lỏng của một số liệu đã được giới thiệu, ví dụ: độ chính xác sử dụng các câu được phân loại hàng đầu [44,46,56](#_bookmark59)  hoặc độ chính xác trung bình trung bình và số liệu 'độ chính xác @rank 10' cho các bài tập xếp hạng câu.[57,58](#_bookmark85) [Một](#_bookmark87)  loại thư giãn khác cho các chỉ số tiêu chuẩn là thư giãn khoảng cách khi chuẩn hóa các thực thể thành các khái niệm trong tiêu đề chủ đề y tế (MesH) hoặc hệ thống ngôn ngữ y tế thống nhất (UMLS), để cho phép N bước nhảy giữa các khái niệm dự đoán và mục tiêu.[59](#_bookmark89)

Bản cập nhật LSR cho thấy xu hướng ngày càng tăng của thuật toán tóm tắt văn bản và trích xuất mối quan hệ. Sự tương đồng ROGUE, ΔEI hoặc Jaccard là các số liệu để tóm tắt.[60,61](#_bookmark91) Đối với trích xuất quan hệ F1, độ chính xác và khả năng nhớ lại vẫn là các chỉ số phổ biến nhất.[62,63](#_bookmark95)

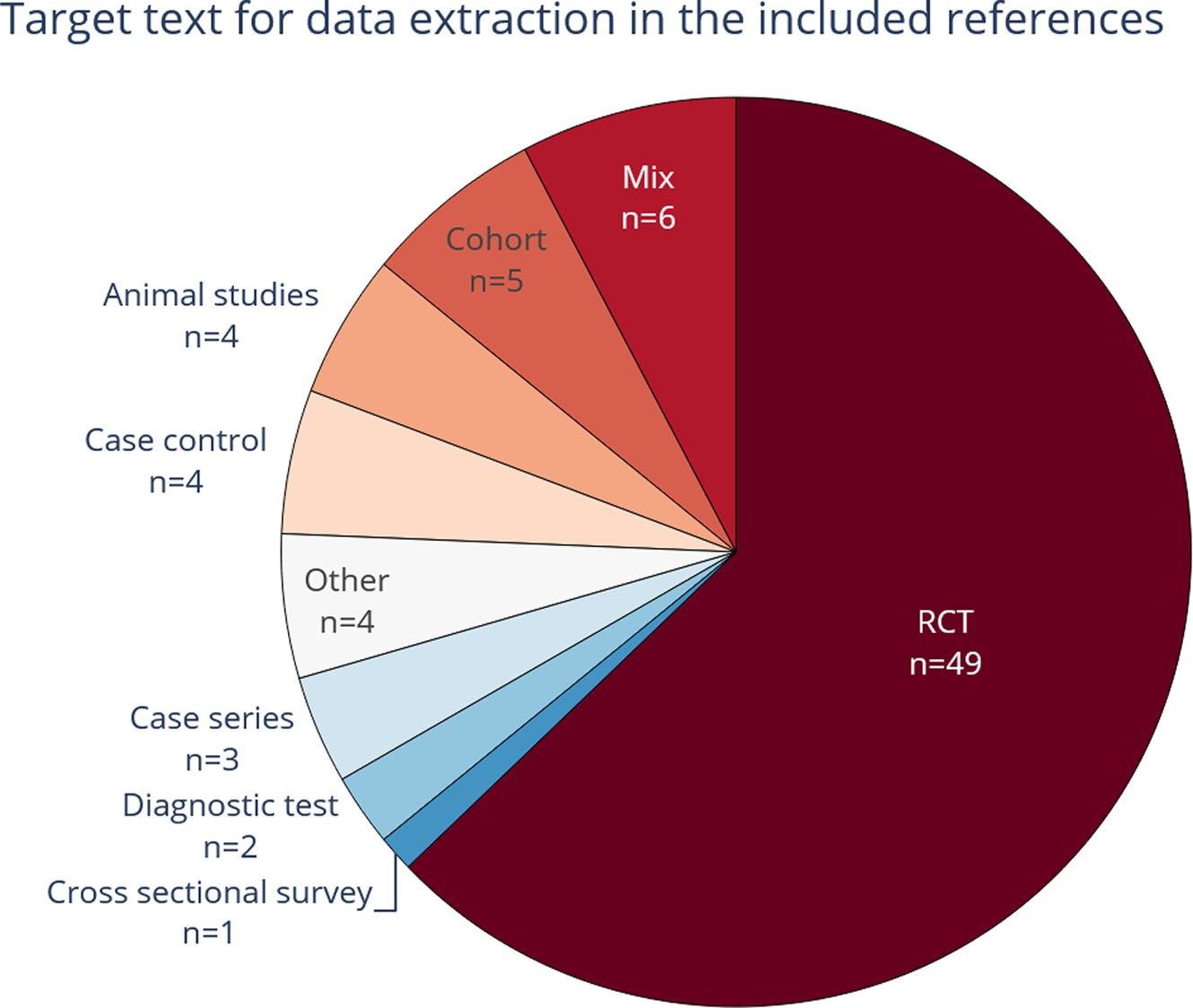
Các số liệu khác là kappa,58 xáo trộn ngẫu nhiên[64](#_bookmark99) hoặc kiểm tra tỷ lệ nhị thức[65](#_bookmark101) để kiểm tra ý nghĩa thống kê, được đưa ra với các khoảng tin cậy.[41](#_bookmark55) Các số liệu khác được bao gồm trong 'Khác' là tỷ lệ chênh lệch,66 mức tăng tích lũy chiết khấu được chuẩn hóa,44,67 'câu cần thiết để sàng lọc mỗi bài viết' để tìm ra một câu có liên quan,68 thử nghiệm McNemar,65 thống kê C (với 95% CI) và điểm Brier (với 95% CI).[69](#_bookmark66) Barnett (2022)[70](#_bookmark68) trích xuất cỡ mẫu và báo cáo sự khác biệt trung bình giữa số thật và số được trích xuất.

Các đánh giá thực tế, chẳng hạn như tỷ lệ phần trăm đầu ra cần chỉnh sửa của con người hoặc thời gian tiết kiệm cho mỗi bài báo, đã được báo cáo bởi hai ấn [phẩm,32,46](#_bookmark37)  và một đánh giá như một phần của hệ thống sàng lọc rộng hơn đã được thực hiện trong một ấn phẩm khác.[71](#_bookmark70)

* + 1. *Loại dữ liệu*
       1. Phạm vi và dữ liệu

Hầu hết việc trích xuất dữ liệu được thực hiện trên các bản tóm tắt (Xem Bảng A1 trong *Dữ liệu cơ* [bản,127](#_bookmark142) và bảng bổ sung cung cấp tổng quan về tất cả các ấn phẩm được bao gồm). Tóm tắt là lựa chọn thực tế nhất, do khả năng xuất chúng cùng với kết quả tìm kiếm tài liệu từ cơ sở dữ liệu như MEDLINE. Tổng cộng, 84% (N = 64) các ấn phẩm được đưa vào báo cáo trực tiếp bằng cách sử dụng bản tóm tắt. Trong số 19 tài liệu tham khảo (25%) báo cáo việc sử dụng toàn văn bản có đề cập cụ thể rằng điều này cũng bao gồm các bản tóm tắt nhưng không rõ liệu tất cả các văn bản có bao gồm văn bản trừu tượng hay không. Mô tả về lợi ích của việc sử dụng toàn bộ văn bản để trích xuất dữ liệu bao gồm có quyền truy cập vào một bộ dữ liệu hoàn chỉnh hơn, trong khi lợi ích của việc sử dụng tiêu đề (N = 4, 5%) bao gồm độ phức tạp thấp hơn cho nhiệm vụ trích xuất dữ liệu.[43](#_bookmark57) Xu et al. (2010) [72](#_bookmark72) chỉ sử dụng các tiêu đề, trong khi ba ấn phẩm còn lại đề cập cụ thể đến các tiêu đề cũng sử dụng tóm tắt trong bộ dữ liệu của họ. [43,73,74](#_bookmark74)

[Hình 5](#_bookmark4) cho thấy RCT là các văn bản thiết kế nghiên cứu phổ biến nhất được sử dụng để trích xuất dữ liệu trong các ấn phẩm bao gồm (xem thêm Bảng A1 mở rộng trong *Dữ liệu cơ bản*[127](#_bookmark142)). Điều này không có gì đáng ngạc nhiên, bởi vì các đánh giá có hệ thống về các can thiệp là loại đánh giá hệ thống phổ biến nhất và chúng thường tập trung vào bằng chứng từ RCT. Do đó, tài liệu về tự động hóa trích xuất dữ liệu tập trung vào RCT và các yếu tố PICO liên quan của chúng. Các đánh giá có hệ thống về độ chính xác của xét nghiệm chẩn đoán ít thường xuyên hơn và chỉ có một ấn phẩm bao gồm ấn phẩm đặc biệt tập trung vào văn bản và các thực thể liên quan đến các nghiên cứu [này,75](#_bookmark79) trong khi hai đề cập đến các quy trình chẩn đoán trong số các lĩnh vực quan tâm khác.35,76 Tám ấn phẩm tập trung vào việc trích xuất dữ liệu cụ thể từ nghiên cứu dịch tễ học, các nghiên cứu can thiệp không ngẫu nhiên hoặc bao gồm văn bản từ các nghiên cứu thuần tập cũng như văn bản RCT. [48,49,61,72–](#_bookmark69)[74,76,77](#_bookmark76)  Nhiều ấn phẩm khai thác dữ liệu từ các cuộc khảo sát, RCT động vật hoặc chuỗi trường hợp có thể được tìm thấy nếu tìm kiếm và xem xét của chúng tôi tập trung vào các loại văn bản này.



Hình 5. Các loại nghiên cứu mà dữ liệu được trích xuất. Thông thường, văn bản thử nghiệm ngẫu nhiên có đối chứng (RCT) là ít nhất một trong những loại văn bản mục tiêu được sử dụng trong các ấn phẩm được bao gồm.

* + - 1. Mục tiêu trích xuất dữ liệu

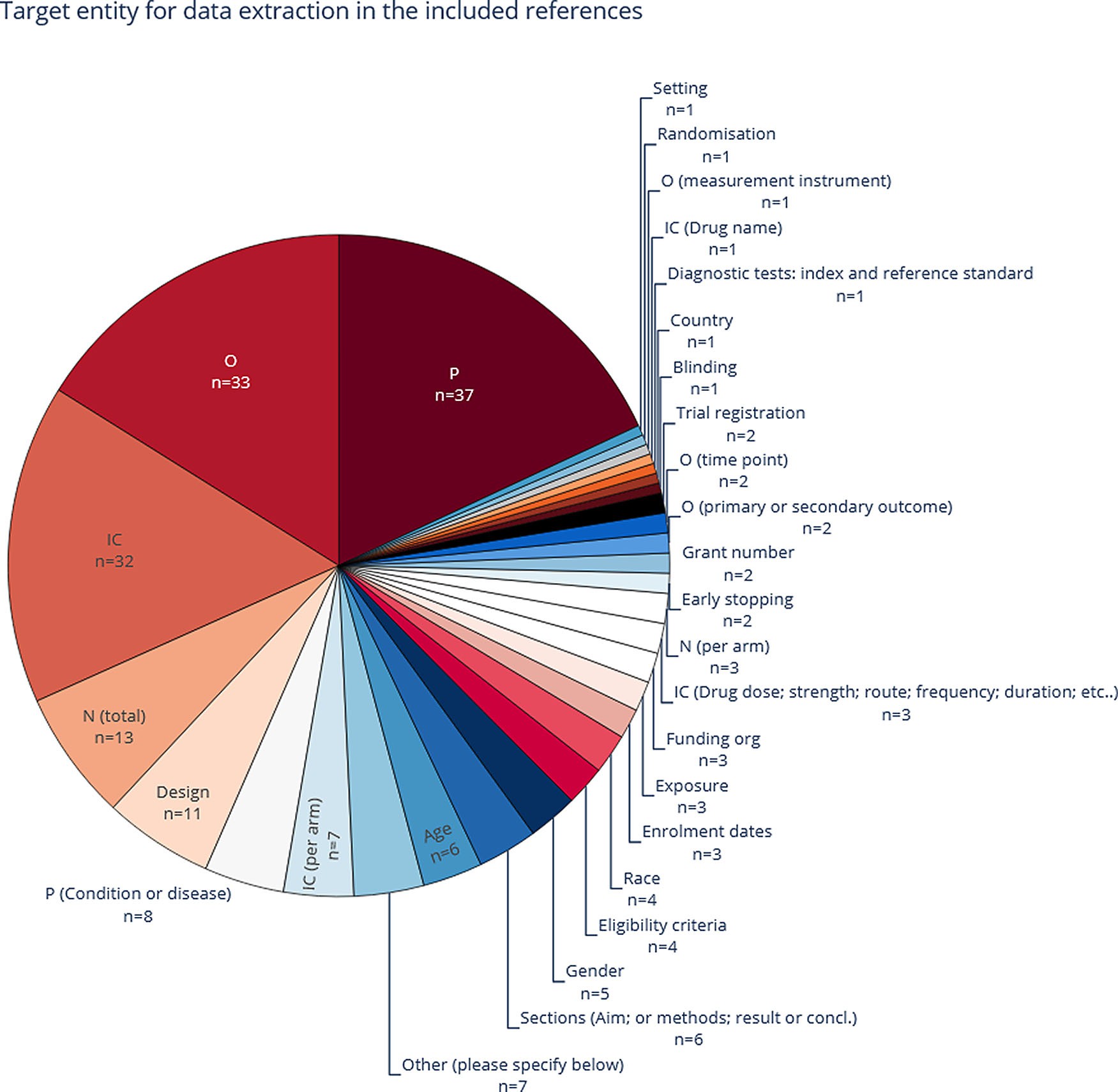
Khai thác các phần tử P, IC và O là nhiệm vụ phổ biến nhất được thực hiện trong tài liệu về tự động hóa đánh giá hệ thống (bán nửa) (xem Bảng A1 trong *Dữ liệu cơ bản*, [127](#_bookmark142) và [Hình 6](#_bookmark5)). Trong đánh giá cơ sở, P là thực thể phổ biến nhất. Sau khi cập nhật LSR, O (n = 52, 68%) đã trở thành phổ biến nhất, do xu hướng mới nổi của các mô hình trích xuất quan hệ tập trung vào mối quan hệ giữa các thực thể O và I và do đó có thể bỏ qua việc trích xuất tự động P. Một số mục tiêu trích xuất dữ liệu ít thường xuyên hơn trong tài liệu có thể được phân loại thành các lớp con của PICO,55 ví dụ, bằng cách chú thích theo thứ bậc nhiều loại thực thể như tình trạng sức khỏe, tuổi tác và giới tính theo lớp P. Loại thực thể 'P (Tình trạng và bệnh)', là thực thể phổ biến nhất có liên quan chặt chẽ đến lớp P, xuất hiện trong mười hai ấn phẩm bao gồm, trong đó bốn ấn phẩm được xuất bản vào năm 2021 hoặc muộn hơn. [35,36,51,55,63,71,75,76,78–](#_bookmark45)81

Đáng chú ý, mười một ấn phẩm đã chú thích hoặc làm việc với các bộ dữ liệu phân biệt giữa nhánh can thiệp và nhóm kiểm soát, bốn trong số này được xuất bản sau năm 2020 với xu hướng trích xuất mối quan hệ và các nhiệm vụ tóm tắt yêu cầu loại dữ liệu này. [46,47,51,56,62,63,66,82–](#_bookmark65)[84](#_bookmark96) Thông thường, I và C được hợp nhất (n = 47). Hầu hết các phương pháp trích xuất dữ liệu tập trung vào việc nhận ra các trường hợp của các lớp thực thể hoặc câu, và một số lượng nhỏ các ấn phẩm đã tiến thêm một bước nữa để chuẩn hóa các khái niệm thực tế và bao gồm các nguồn dữ liệu như UMLS (Hệ thống ngôn ngữ y tế thống nhất). [35,39,59,73,85](#_bookmark51)

Danh mục 'Khác' bao gồm một số chú thích thuốc chi tiết hơn[65](#_bookmark101) hoặc thông tin như yếu tố gây nhiễu[49](#_bookmark69) và các loại thực thể khác (xem bộ dữ liệu đầy đủ trong *Dữ liệu cơ bản:* Phụ lục A và D để biết thêm thông tin 127).

* 1. Kết quả trích xuất dữ liệu: Các mục phụ quan tâm
     1. *Độ chi tiết của trích xuất dữ liệu*

Tổng cộng có 54 ấn phẩm (71%) trích xuất ít nhất một loại thông tin ở cấp độ thực thể, trong khi 46 ấn phẩm (60%) sử dụng cấp câu (xem Bảng A1 phiên bản mở rộng trong *Dữ liệu cơ bản*[127](#_bookmark142)). Chúng tôi định nghĩa cấp độ thực thể là bất kỳ số lượng từ nào ngắn hơn toàn bộ câu, ví dụ: cụm danh từ hoặc văn bản phân đoạn khác. Các kiểu dữ liệu như P, IC hoặc O thường được trích xuất ở cả cấp độ thực thể và câu, trong khi 'N', số lượng người tham gia vào một nghiên cứu, thường chỉ được trích xuất ở cấp độ thực thể.



Hình 6. Các thực thể phổ biến nhất, như được trích xuất trong các ấn phẩm bao gồm. Có nhiều hơn một loại thực thể cho mỗi ấn phẩm là phổ biến, có nghĩa là tổng số ấn phẩm được bao gồm (n = 76) thấp hơn tổng số lượng trong con số này. P, dân số; I, can thiệp; C, so sánh; O, kết quả.

* + 1. *Loại đầu vào*

Phần lớn các ấn phẩm và kho dữ liệu chuẩn đã đề cập đến MEDLINE, thông qua PubMed, làm nguồn dữ liệu cho văn bản. Các tệp văn bản (n = 64), bên cạnh XML (n = 8) hoặc HTML (n = 3), là định dạng phổ biến nhất của dữ liệu được tải xuống từ các nguồn này. Do đó, hầu hết các hệ thống được mô tả bằng cách sử dụng hoặc được giả định sử dụng các tệp văn bản làm dữ liệu đầu vào. Tám ấn phẩm bao gồm được mô tả bằng cách sử dụng các tệp PDF làm đầu vào. 44,46,59,68,75,81,86,87

* + 1. *Loại đầu ra*

Một số lượng hạn chế các ấn phẩm mô tả các bản tóm tắt có cấu trúc là đầu ra của dữ liệu được trích xuất của họ (n = 14, xu hướng tăng giữa các bản cập nhật LSR). Các lựa chọn thay thế cho việc xuất bản tóm tắt có cấu trúc là JSON (n = 4), XML và HTML (n = 2 mỗi loại). Hai ấn phẩm đề cập đến đầu ra dữ liệu có cấu trúc dưới dạng bản thể luận.[51,88](#_bookmark73) Hầu hết các ấn phẩm chỉ đề cập đến điểm phân loại mà không chỉ định loại đầu ra. Trong những trường hợp này, chúng tôi giả định rằng đầu ra sẽ được lưu dưới dạng tệp văn bản, ví dụ như dưới dạng chú thích span thực thể hoặc danh sách câu (n = 55).

* 1. Đánh giá chất lượng báo cáo

Trong đánh giá cơ sở, chúng tôi đã sử dụng một danh sách gồm 17 mục để điều tra khả năng tái tạo, tính minh bạch, mô tả thử nghiệm, tính sẵn có của dữ liệu và tính hợp lệ bên trong và bên ngoài của các phương pháp tiếp cận trong mỗi ấn phẩm. Số lượng mục tối đa và tối thiểu được đánh giá tích cực lần lượt là 16 và 1, với mức trung bình là 10 (xem Bảng A1 trong *Dữ liệu cơ bản*[127](#_bookmark142)). Điểm số được cộng lại và tính toán dựa trên dữ liệu được cung cấp trong Phụ lục A và D (xem *Dữ liệu cơ bản*[127](#_bookmark142)), sử dụng các hàm tổng và trung bình được tích hợp trong Excel. Các ấn phẩm từ những năm gần đây đến năm 2021 cho thấy xu hướng báo cáo đầy đủ và rõ ràng hơn.

* + 1. *Khả năng tái tạo*
       1. Các nguồn dữ liệu đào tạo/thử nghiệm có được báo cáo không?

Trong số các ấn phẩm được đưa vào đánh giá cơ sở, 50 trong số 53 (94%) nêu rõ nguồn dữ liệu của họ được sử dụng để đào tạo và đánh giá. MEDLINE là nguồn dữ liệu phổ biến nhất, với các bản tóm tắt thường được mô tả là được truy xuất thông qua các tìm kiếm trên PubMed hoặc toàn bộ văn bản từ PubMed Central. Một số lượng nhỏ các ấn phẩm được mô tả bằng cách sử dụng văn bản từ các tạp chí cụ thể như PLoS Clinical Trials, New England Journal of Medicine, The Lancet hoặc BMJ.56,83 Văn bản và siêu dữ liệu từ Cochrane, được cung cấp đầy đủ hoặc được truy xuất thông qua PubMed, đã được sử dụng trong năm ấn phẩm. 57,59,68,75,86  [Corpora như tập dữ liệu ebm-nlp,55](#_bookmark80) [hoặc PubMed-PICO](#_bookmark78) 54 có sẵn để tải xuống trực tiếp. Các ấn phẩm được xuất bản trong những năm gần đây ngày càng báo cáo rằng họ đang sử dụng các bộ dữ liệu điểm chuẩn này thay vì tạo và chú thích kho dữ liệu của riêng họ (xem 4 để biết thêm chi tiết).

* + - 1. Nếu các kỹ thuật tiền xử lý được áp dụng cho dữ liệu, chúng có được mô tả không?

Trong số các ấn phẩm được đưa vào đánh giá cơ sở, 47 trong số 53 (89%) cho biết đã xử lý dữ liệu văn bản trước khi áp dụng / đào tạo các thuật toán để trích xuất dữ liệu. Các loại tiền xử lý khác nhau, với các ví dụ đại diện để sử dụng và triển khai, được liệt kê trong [Bảng 1](#_bookmark6) dưới đây.

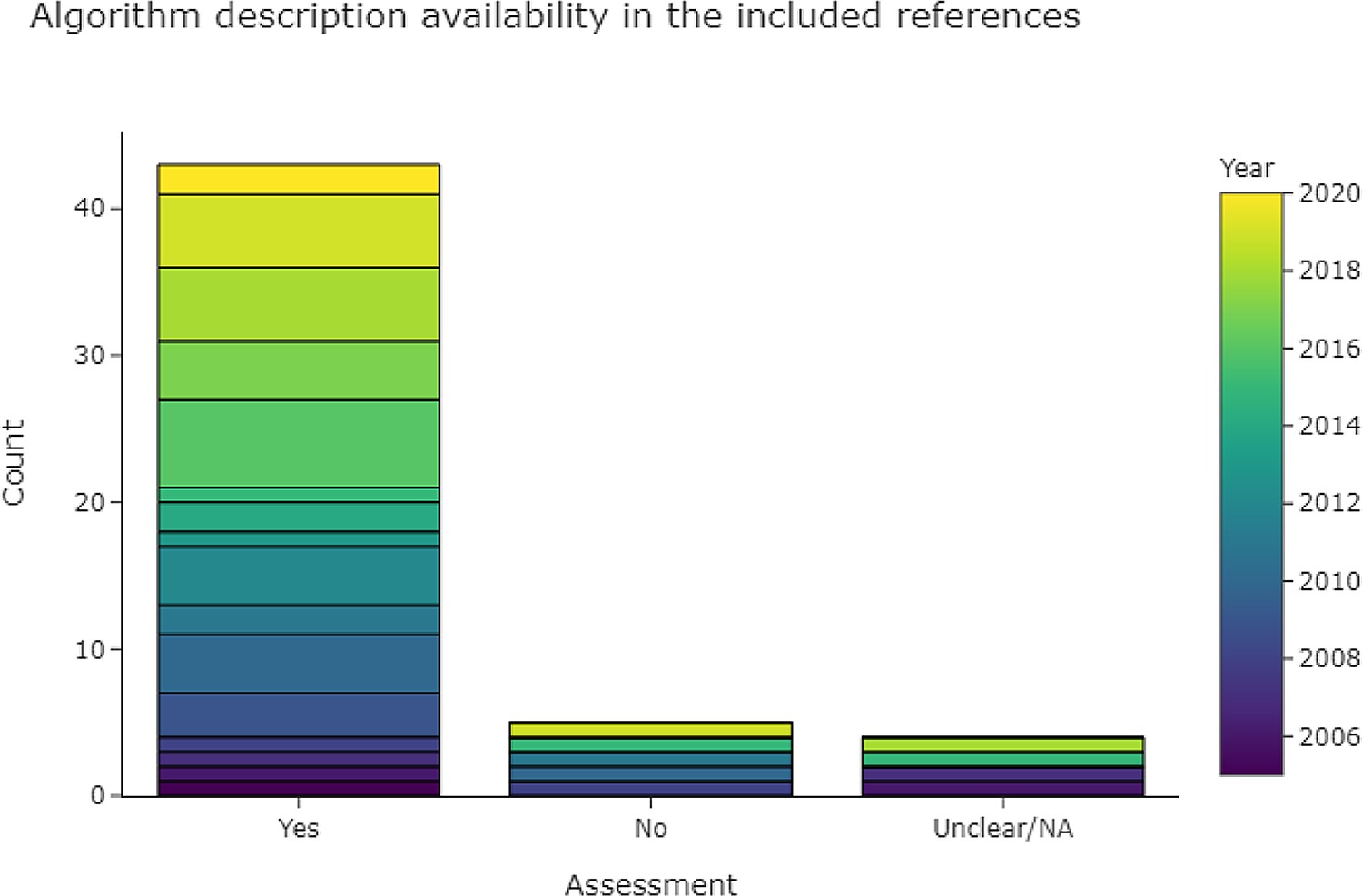
Sau khi xuất bản đánh giá cơ sở, các mô hình máy biến áp như BERT trở nên chiếm ưu thế trong tài liệu (xem [Hình 3](#_bookmark2)). Với vốn từ vựng từ ngữ, nhúng theo ngữ cảnh và đào tạo trước tự giám sát trên kho dữ liệu lớn không được dán nhãn, các mô hình này về cơ bản đã loại bỏ nhu cầu xử lý trước hầu hết ngoài chữ hoa thấp hơn được áp dụng tự động.[14,31](#_bookmark43) Do đó, chúng tôi sẽ không cập nhật bảng này trong điều này, hoặc bất kỳ lần lặp lại nào trong tương lai của LSR này. Chúng tôi để nó tham khảo cho các ấn phẩm có thể vẫn sử dụng các phương pháp này trong tương lai.

* + 1. *Tính minh bạch của các phương pháp*
       1. Có mô tả về các thuật toán được sử dụng không?

[Hình 7](#_bookmark7) cho thấy 43 trong số 53 ấn phẩm trong đánh giá cơ sở (81%) đã cung cấp mô tả về thuật toán trích xuất dữ liệu của họ. Trong trường hợp học máy và mạng nơ-ron, chúng tôi đã tìm kiếm mô tả về siêu tham số và tạo tính năng, cũng như các chi tiết về việc triển khai (ví dụ: khung học máy). Siêu tham số hiếm khi được mô tả đầy đủ, nhưng nếu khung (ví dụ: Scikit-learn, Mallet hoặc Weka) được đưa ra, ngoài mô tả về việc thực hiện và các tham số quan trọng cho từng bộ phân loại, thì chúng tôi đánh giá thuật toán như được mô tả đầy đủ.

Bảng 1. Kỹ thuật tiền xử lý, mô tả ngắn và ví dụ từ tài liệu.

|  |  |  |
| --- | --- | --- |
| Kỹ thuật | Chi tiết | Ví dụ trong văn học |
| Mã hóa | Tách văn bản theo cấp câu và từ | [56,83,88](#_bookmark94) |
| Chuẩn hóa | Thay thế các số nguyên, đơn vị, ngày, chữ thường | [65,89,90](#_bookmark106) |
| Từ điển hóa và bắt gốc | Giảm từ thành các dạng ngắn hơn hoặc phổ biến hơn | [53,91,92](#_bookmark110) |
| Loại bỏ từ dừng | Xóa các từ phổ biến, chẳng hạn như *'the'*, khỏi văn bản | [44,48,80](#_bookmark67) |
| Gắn thẻ một phần của bài phát biểu và phân tích cú pháp phần phụ thuộc | Gắn thẻ từ với vai trò ngữ pháp tương ứng của chúng | [41,78,88](#_bookmark84) |
| Chia nhỏ | Định nghĩa các phần câu, chẳng hạn như danh từ-cụm từ | [65,76,93](#_bookmark81) |
| Gắn thẻ khái niệm | Xử lý và gắn thẻ từ với các lớp hoặc khái niệm ngữ nghĩa, ví dụ: sử dụng danh sách từ hoặc MetaMap | [75,79,94](#_bookmark86) |



Hình 7. Biểu đồ thanh, hiển thị các mức mô tả thuật toán trong các ấn phẩm đi kèm.

Đối với các phương pháp dựa trên quy tắc, chúng tôi tìm kiếm mô tả về cách các quy tắc được rút ra và danh sách các quy tắc đầy đủ hoặc đại diện được đưa ra làm ví dụ. Khi mô tả nhiều phương pháp trích xuất dữ liệu, chúng tôi đã đánh giá tích cực nếu cách tiếp cận hiệu quả tốt nhất được mô tả.

* + - 1. Có mô tả về bộ dữ liệu được sử dụng và các đặc điểm của nó không?

Trong số các ấn phẩm được đưa vào bản cập nhật đánh giá, 73 trong số 76 (97%) đã cung cấp mô tả về bộ dữ liệu của họ và các đặc điểm của nó.

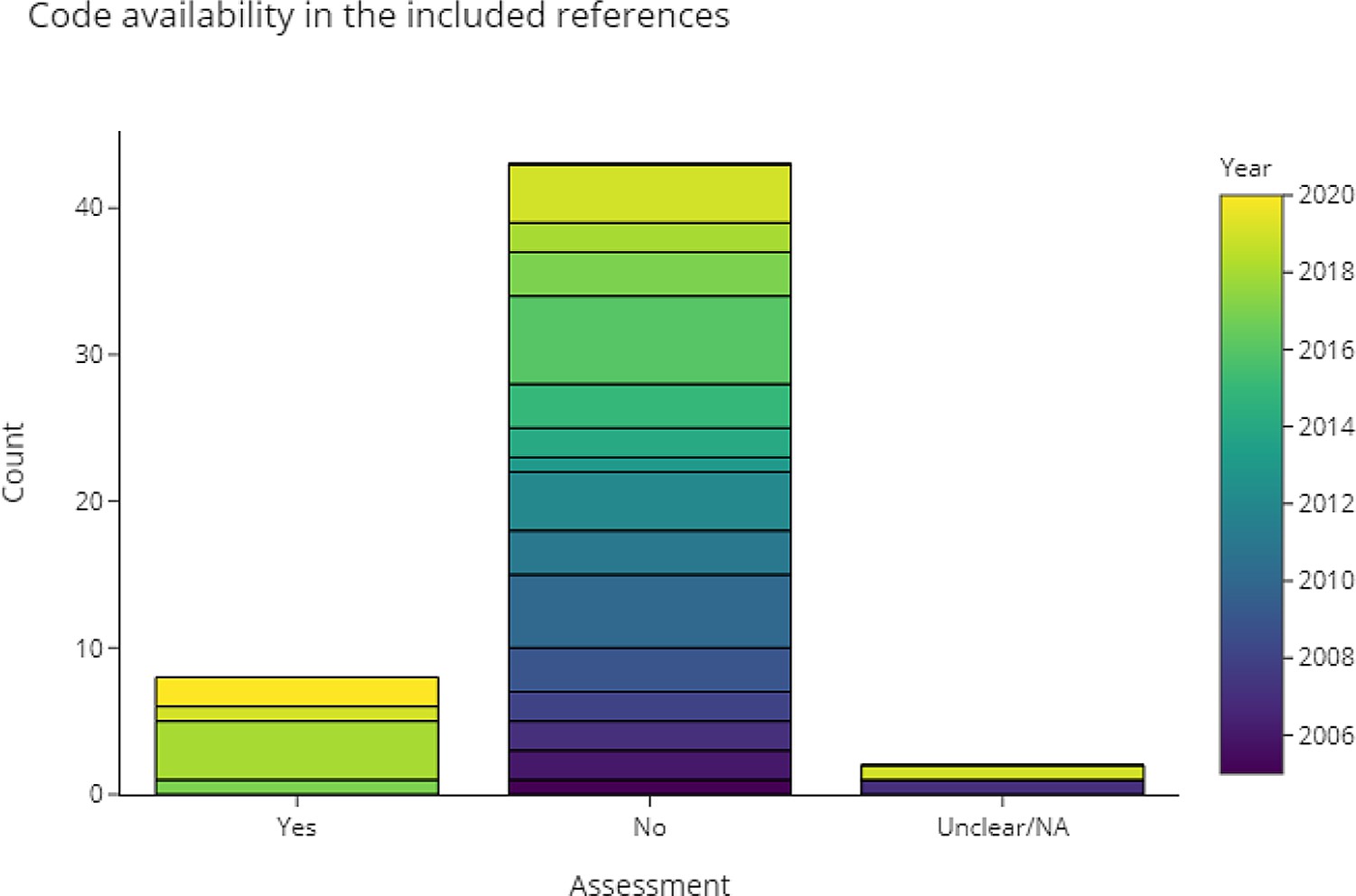
Hầu hết các ấn phẩm đều cung cấp mô tả về (các) bộ dữ liệu được sử dụng để đào tạo và đánh giá. Kích thước của mỗi bộ dữ liệu, cũng như tần suất của các lớp trong dữ liệu, là minh bạch và được mô tả cho hầu hết các ấn phẩm được bao gồm. Tất cả các trích dẫn tập dữ liệu, cùng với mô tả ngắn và tính khả dụng của dữ liệu, được thể hiện trong [Bảng 4](#_bookmark13).

* + - 1. Có mô tả về phần cứng được sử dụng không?

Hầu hết các ấn phẩm được đưa vào đánh giá cơ sở không báo cáo các thông số kỹ thuật phần cứng của họ, mặc dù năm ấn phẩm (9%) đã báo cáo. Ví dụ, một người đã áp dụng hệ thống của họ cho dữ liệu mới, không được dán nhãn và báo cáo rằng việc phân loại toàn bộ PubMed mất khoảng 20 giờ bằng cách sử dụng bộ xử lý đồ họa (GPU).[Trong](#_bookmark66) một ví dụ khác, các tác giả đã báo cáo sử dụng GPU Google Colab, cùng với ước tính thời gian tính toán cho các cài đặt đào tạo khác nhau.[95](#_bookmark118)

* + - 1. Mã nguồn có sẵn không?

[Hình 8](#_bookmark8) cho thấy hầu hết các ấn phẩm được bao gồm không cung cấp bất kỳ mã nguồn nào, mặc dù có một xu hướng rất mạnh mẽ hướng tới khả năng cung cấp mã tốt hơn trong các ấn phẩm từ bản cập nhật đánh giá (n = 19 mã đã xuất bản, 83% các ấn phẩm mới cung cấp mã). Các ấn phẩm đã cung cấp mã nguồn đã được xuất bản độc quyền hoặc cập nhật lần cuối trong bảy năm qua. GitHub là nền tảng phổ biến nhất để làm cho mã có thể truy cập được. Một số ấn phẩm cũng cung cấp liên kết đến sổ ghi chép trên Google Colab, một nền tảng dựa trên đám mây để phát triển và thực thi mã trực tuyến. Hai ấn phẩm cung cấp quyền truy cập vào các phần của mã, hoặc quyền truy cập bị hạn chế. Danh sách đầy đủ các kho lưu trữ mã từ các ấn phẩm đi kèm có sẵn trong [Bảng 2](#_bookmark9).



Hình 8. Biểu đồ này cho thấy mức độ mà các ấn phẩm bao gồm được cung cấp quyền truy cập vào mã nguồn của họ.

Bảng 2. Kho lưu trữ chứa mã nguồn cho các ấn phẩm được bao gồm.

|  |  |  |
| --- | --- | --- |
| Công bố | Mã | LSR |
| [81](#_bookmark90) | Có sẵn theo: <https://github.com/ijmarshall/robotreviewer>, phiên bản cũ hơn: [https://](https://figshare.com/articles/Spa/997707)  [figshare.com/articles/Spa/997707](https://figshare.com/articles/Spa/997707) | Căn cứ |
| [96](#_bookmark120) | Có sẵn: https://github.com/jind11/LSTM-PICO-Detection | Căn cứ |
| [55](#_bookmark80) | Có sẵn theo: <https://github.com/bepnye/EBM-NLP>[https://colab.research.google.](https://colab.research.google.com/drive/1Ir52OmkJ2C_Iy9V_eS-_KFVLircJ4MXp) [com/drive/1Ir52OmkJ2C\_Iy9V\_eS-\_KFVLircJ4MXp](https://colab.research.google.com/drive/1Ir52OmkJ2C_Iy9V_eS-_KFVLircJ4MXp) <https://colab.research.google.com/drive/1YbbQojM147Ybt1nEcyoXTqlvefmwMg-q> | Căn cứ |
| [54](#_bookmark78) | Có sẵn: <https://github.com/jind11/Deep-PICO-Detection> | Căn cứ |
| [97](#_bookmark122) | Có sẵn theo: <https://ii.nlm.nih.gov/DataSets/index.shtml> | Căn cứ |
| [85](#_bookmark98) | Có sẵn: <https://github.com/Tian312/PICO_Parser> | Căn cứ |
| [95](#_bookmark118) | Có sẵn theo: <https://github.com/L-ENA/HealthINF2020> <https://www.kaggle.com/lenaschmidt0493/qa-integrated-biomedical-ner-classifier-for-pico> | Căn cứ |
| [69](#_bookmark66) | Có sẵn theo: <https://github.com/ijmarshall/trialstreamer> | Căn cứ |
| [47](#_bookmark65) | Không rõ liệu mã Java có thể truy cập được hay không, đang chờ truy cập của người dùng: [https://semrep.nlm.nih.gov/](https://semrep.nlm.nih.gov/SemRep.v1.8_Installation.html#Download)  [SemRep.v1.8\_Installation.html#Download](https://semrep.nlm.nih.gov/SemRep.v1.8_Installation.html#Download) | Căn cứ |
| [75](#_bookmark79) | Đã sử dụng triển khai công khai của Google về transformers + [https://zenodo.org/ record](https://zenodo.org/record/1303259#.X4wSoaySk2w) [/1303259#. X4wSoaySk2w](https://zenodo.org/record/1303259#.X4wSoaySk2w) | Căn cứ |
| [60](#_bookmark91) | Có sẵn theo: [https://github.com/smileslab/Brain\_Aneurysm\_Research/tree/ master](https://github.com/smileslab/Brain_Aneurysm_Research/tree/master/BioMed_Summarizer)  [/ BioMed\_Summarizer](https://github.com/smileslab/Brain_Aneurysm_Research/tree/master/BioMed_Summarizer) | Cập nhật |
| [74](#_bookmark76) | Có sẵn theo: <https://github.com/nstylia/pico_entities/> | Cập nhật |
| [98](#_bookmark123) | Có sẵn theo: <https://github.com/wds-seu/Aceso> | Cập nhật |
| [62](#_bookmark95) | Có sẵn theo: <https://github.com/jayded/evidence-inference> | Cập nhật |
| [61](#_bookmark93) | Có sẵn: <https://github.com/allenai/ms2> | Cập nhật |
| [99](#_bookmark125) | Có sẵn theo: <https://github.com/Tian312/MD-Attention> | Cập nhật |
| [38](#_bookmark50) | Có sẵn theo: <https://github.com/kilicogluh/CONSORT-TM> | Cập nhật |
| [35](#_bookmark42) | Có sẵn theo: <https://github.com/lcampillos/Medical-NER> | Cập nhật |

|  |  |  |
| --- | --- | --- |
| Bảng 2. *Tiếp tục* | | |
| Công bố | Mã | LSR |
| [36](#_bookmark45) | Có sẵn theo: <https://gitlab.com/tomaye/ecai2020-transformer_based_am> | Cập nhật |
| [50](#_bookmark71) | Có sẵn theo: <https://github.com/jetsunwhitton/RCT-ART> | Cập nhật |
| [34](#_bookmark41) | Có sẵn theo: <https://github.com/LivNLP/ODP-tagger> | Cập nhật |
| [33](#_bookmark39) | Có sẵn theo: <https://data.mendeley.com/datasets/ccfnn3jb2x/1> | Cập nhật |
| [82](#_bookmark92) | Có sẵn theo: <https://osf.io/2dqcg/> | Cập nhật |
| [51](#_bookmark73) | Có sẵn theo: <https://zenodo.org/record/6365890> | Cập nhật |
| [45](#_bookmark61) | Có sẵn theo: <https://github.com/gauravsc/pico-tagging> | Cập nhật |
| [67](#_bookmark105) | Có sẵn trong: [Mô](https://github.com/MichealAbaho/Label-Context-Aware-Attention-Model)  [hình https://github.com/MichealAbaho/Label-Context-Aware-Attention](https://github.com/MichealAbaho/Label-Context-Aware-Attention-Model) | Cập nhật |
| [100](#_bookmark127) | Có sẵn: <https://github.com/evidence-surveillance/sent2span> | Cập nhật |
| [70](#_bookmark68) | Có sẵn theo: <https://zenodo.org/record/6647853#.ZBnpLXbP2Uk> | Cập nhật |
| [37](#_bookmark47) | Có sẵn theo: <https://github.com/anjani-dhrangadhariya/distant-PICO> | Cập nhật |

* + 1. *Thử nghiệm*
       1. Có biện minh / giải thích cho đánh giá mô hình không?

Trong số các ấn phẩm được đưa vào đánh giá cơ sở, 47 trong số 53 (89%) đưa ra đánh giá chi tiết về các thuật toán trích xuất dữ liệu của họ. Chúng tôi đánh giá mục này là tiêu cực nếu chỉ đưa ra điểm hiệu suất, tức là nếu không có phân tích lỗi nào được thực hiện và không có giải thích hoặc ví dụ nào được đưa ra để minh họa hiệu suất của mô hình. Trong hầu hết các ấn phẩm, một phân tích lỗi ngắn gọn là phổ biến, ví dụ như thảo luận về các ví dụ đại diện cho âm tính giả và dương tính [giả,47](#_bookmark65) nguồn lỗi chính[90](#_bookmark108) hoặc làm nổi bật các lỗi liên quan đến mọi lớp thực thể.[76](#_bookmark81) Cả hai tham khảo. [52](#_bookmark75), [53](#_bookmark77) đã sử dụng các bản tóm tắt có cấu trúc và phi cấu trúc, và do đó thảo luận về ý nghĩa của dữ liệu văn bản phi cấu trúc đối với điểm phân loại.

Một số lượng nhỏ các ấn phẩm đã thực hiện đánh giá thực tế, trong đó thuật toán trích xuất dữ liệu được áp dụng cho các bộ dữ liệu khác nhau, không được dán nhãn và thường lớn hơn nhiều hoặc được thử nghiệm trong khi tiến hành đánh giá có hệ thống thực tế. [46,58,63,69,48,95,101,102](#_bookmark87)

* + - 1. Các chỉ số cơ bản có được báo cáo (dương tính và âm tính đúng/sai) không?

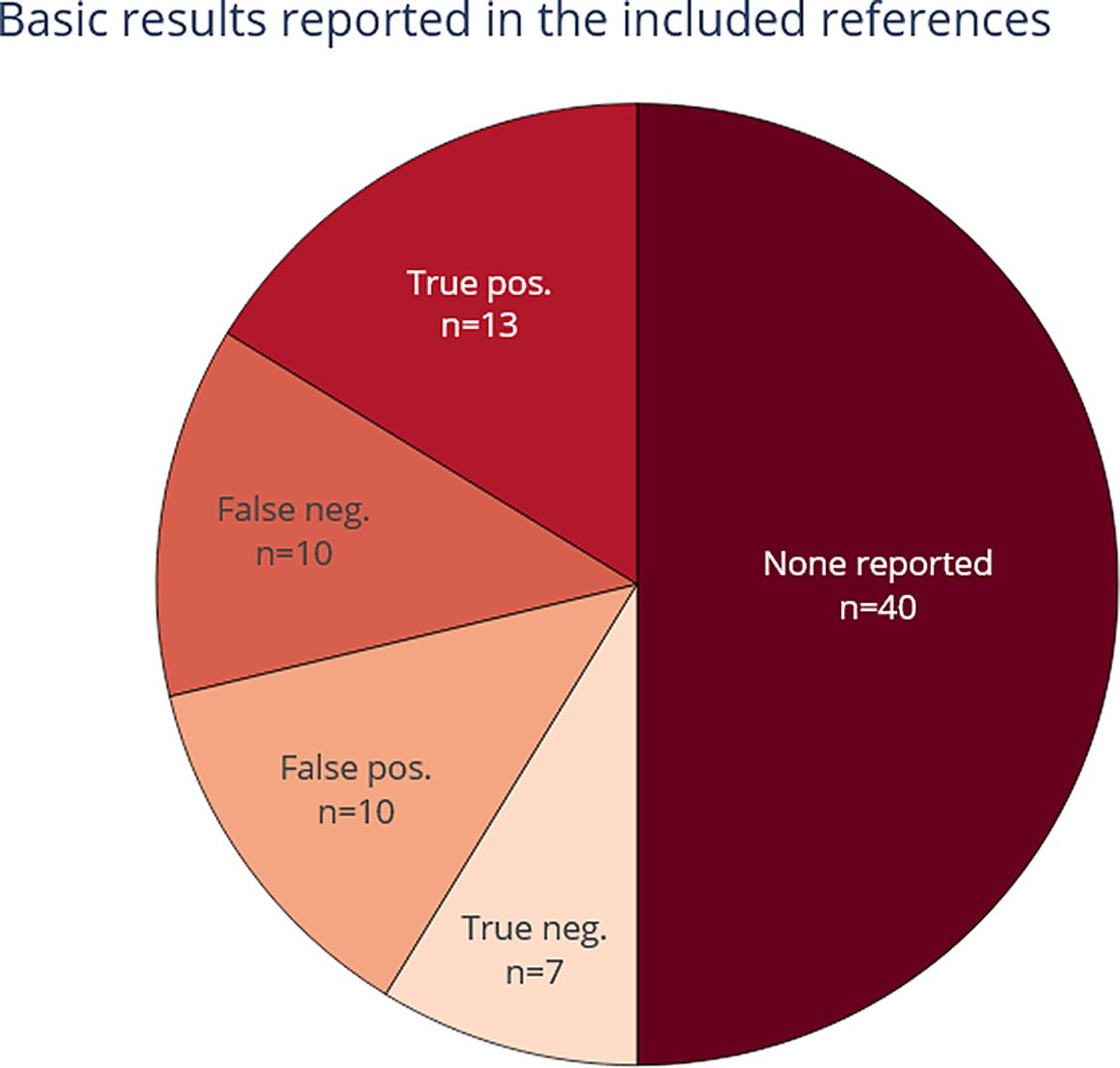
[Hình 9](#_bookmark10) cho thấy mức độ mà tất cả các chỉ số cơ bản thô, chẳng hạn như dương tính thực, đã được báo cáo trong các ấn phẩm bao gồm trong bản cập nhật LSR. Trong hầu hết các ấn phẩm (n = 62) các số liệu cơ bản này không được báo cáo và có xu hướng giữa đánh giá cơ sở và cập nhật này theo hướng không báo cáo những số liệu này. Tuy nhiên, các số liệu cơ bản có thể thu được vì phần lớn các ấn phẩm mới được bao gồm đã cung cấp mã nguồn và sử dụng các bộ dữ liệu có sẵn công khai. Khi xử lý trích xuất dữ liệu cấp thực thể, có thể là một thách thức để xác định số lượng thực thể tiêu cực thực sự. Điều này đúng đặc biệt nếu các thực thể được gắn nhãn và trích xuất dựa trên các đoạn văn bản, bởi vì có thể có nhiều sự kết hợp của các cụm từ và mã thông báo tạo thành một thực thể.[47](#_bookmark65) Vấn đề này đã được giải quyết trong các ấn phẩm gần đây hơn bằng cách tiến hành đánh giá dựa trên mã thông báo tính toán điểm trên mọi mã thông báo, do đó đạt được khả năng ghi điểm khớp một phần cho các thực thể nhiều từ.[55](#_bookmark80)

* + - 1. Đánh giá có bao gồm bất kỳ thông tin nào về sự đánh đổi giữa khả năng nhớ lại hoặc độ chính xác (còn được gọi là độ nhạy và giá trị dự đoán tích cực) không?

Trong số các ấn phẩm được đưa vào đánh giá cơ sở, 17 trong số 53 (32%) mô tả sự đánh đổi hoặc cung cấp các biểu đồ hoặc bảng cho thấy sự phát triển của điểm đánh giá nếu một số thông số nhất định bị thay đổi hoặc nới lỏng. Nhớ lại (tức là độ nhạy) thường được mô tả là thước đo quan trọng nhất cho các nhiệm vụ tự động hóa đánh giá có hệ thống, vì đó là một yêu cầu phương pháp luận rằng các đánh giá có hệ thống không loại trừ bất kỳ dữ liệu đủ điều kiện nào.

Tài liệu tham khảo [56](#_bookmark83) và [76](#_bookmark81) cho thấy quyết định trích xuất hai hoặc N dự đoán hàng đầu ảnh hưởng như thế nào đến điểm đánh giá, ví dụ như độ chính xác hoặc khả năng nhớ lại. Tài liệu tham khảo [102](#_bookmark132) cho thấy các biểu đồ thu hồi độ chính xác cho các ngưỡng phân loại khác nhau. Tham khảo [72](#_bookmark72) cho thấy bốn điểm cắt, trong khi Tham khảo. [95](#_bookmark118) cho thấy các ngưỡng xác suất khác nhau cho bộ phân loại của chúng và mô tả tác động của điều này đối với độ chính xác, khả năng thu hồi và đường cong F1.

Một số kiến trúc máy học cần chuyển đổi văn bản thành tính năng trước khi thực hiện phân loại. Ví dụ, một đặc điểm có thể là số lần một từ nhất định xuất hiện hoặc độ dài của một bản tóm tắt. Số lượng tính năng được sử dụng,



Hình 9. Báo cáo các chỉ số cơ bản (dương tính thực, dương tính giả, âm tính thực và âm tính giả). Đối với mỗi tờ giấy đi kèm. Có thể có nhiều hơn một lựa chọn, có nghĩa là tổng số ấn phẩm được đưa vào ( n = 76) thấp hơn tổng số lượng trong con số này.

ví dụ: đối với các thuật toán CRF, được đưa ra trong nhiều ấn phẩm,92 cùng với một cuộc thảo luận về các bộ phân loại nên được sử dụng trong khả năng nhớ cao là cần thiết.[42.103](#_bookmark56) cho thấy các đường cong ROC định lượng lượng dữ liệu đào tạo và tác động của nó đối với điểm số.

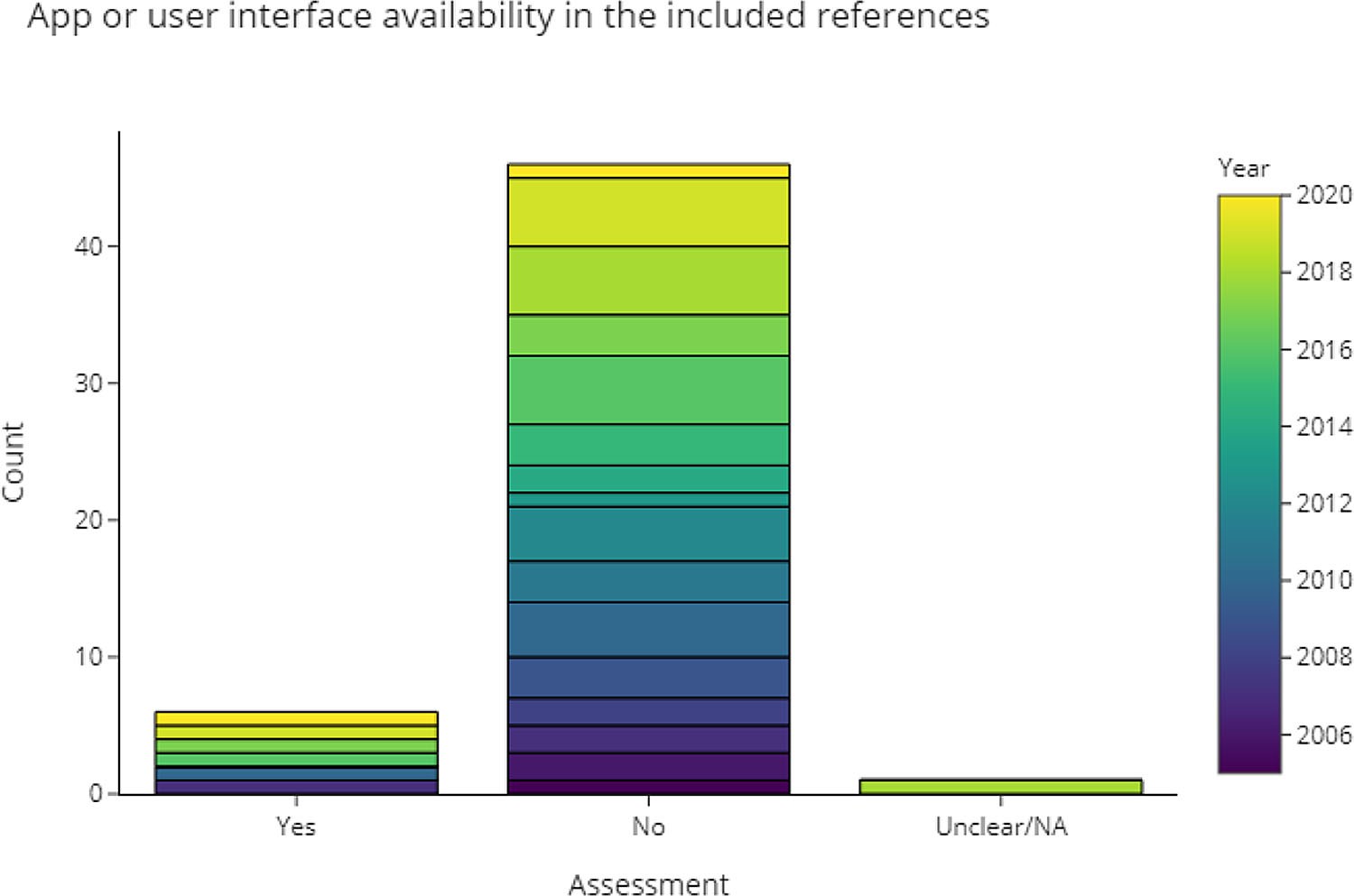
* + 1. *Tính khả dụng của mô hình hoặc công cụ cuối cùng*
       1. Chúng tôi có thể lấy một phiên bản có thể chạy được của phần mềm dựa trên thông tin trong ấn phẩm không?

Việc biên soạn và kiểm tra mã từ mọi ấn phẩm nằm ngoài phạm vi của bài đánh giá này. Thay vào đó, trong [Hình 10](#_bookmark11) và [Bảng 3](#_bookmark12) , chúng tôi đã ghi lại các ấn phẩm có sẵn giao diện (web) hoặc ứng dụng hoàn chỉnh. Tính RobotReviewer và Trialstreamer là các dự án riêng biệt, 12% các ấn phẩm được bao gồm có ứng dụng liên quan đến nó, nhưng chỉ có 5 (6%) có sẵn và có thể sử dụng trực tiếp qua ứng dụng web. Các ứng dụng có sẵn dưới dạng mã nguồn mở, hoàn toàn miễn phí hoặc phiên bản cơ bản miễn phí với các tính năng tùy chọn có thể được mua hoặc đăng ký.

* + - 1. Tính bền bỉ: Dữ liệu có thể được truy xuất dựa trên thông tin được cung cấp trong ấn phẩm không?

Chúng tôi quan sát thấy xu hướng ngày càng tăng về tính sẵn có của bộ dữ liệu và các ấn phẩm sử dụng lại kho dữ liệu điểm chuẩn trong bản cập nhật LSR. Chỉ có bảy trong số các ấn phẩm được đưa vào đánh giá cơ sở (13%) công khai bộ dữ liệu của họ, trong số 36 kho dữ liệu duy nhất được tìm thấy sau đó.

Sau khi cập nhật LSR, chúng tôi đã tích lũy được 55 ấn phẩm mô tả kho dữ liệu mới độc đáo. Trong số này, 23 kho dữ liệu có sẵn trực tuyến và tổng cộng 40 ấn phẩm được đề cập bằng cách sử dụng một trong những bộ điểm chuẩn công khai này. [Bảng 4](#_bookmark13) cho thấy tóm tắt về kho dữ liệu, kích thước, lớp, liên kết đến bộ dữ liệu và tham chiếu chéo đến các ấn phẩm đã biết sử dụng lại từng tập dữ liệu. Đối với đánh giá cơ sở, chúng tôi đã thu thập kho dữ liệu, cung cấp một liên kết trung tâm đến tất cả các bộ dữ liệu và sẽ thêm các bộ dữ liệu khi chúng có sẵn trong suốt vòng đời của đánh giá sống này (xem *Dữ liệu cơ bản*[127,128](#_bookmark142) bên dưới). Do số lượng kho có sẵn tăng lên, chúng tôi đã ngừng tải xuống dữ liệu và thay vào đó cung cấp các liên kết. Khi một bộ dữ liệu được cung cấp miễn phí mà không có rào cản (tức là tải xuống trực tiếp văn bản và nhãn), thì bất kỳ nhà nghiên cứu nào cũng có thể sử dụng lại dữ liệu và công bố kết quả từ các mô hình khác nhau, có thể so sánh với nhau. Các vấn đề bản quyền xung quanh việc chia sẻ dữ liệu là



Hình 10. Ấn phẩm cung cấp cho các ứng dụng giao diện người dùng.

Bảng 3. Các ấn phẩm cung cấp giao diện người dùng cho hệ thống trích xuất dữ liệu cuối cùng của họ.

|  |  |
| --- | --- |
| Giấy | Truy cập |
| [42](#_bookmark56) | Không rõ ràng: Một liên kết đã được đưa ra, nhưng công cụ chưa trực tuyến: <https://ihealth.uemc.es/> |
| [43](#_bookmark57) | <https://www.tripdatabase.com/#pico> |
| [44](#_bookmark59), [81](#_bookmark90) | <https://www.robotreviewer.net/> |
| [46](#_bookmark64) | <https://exact.cluster.gctools.nrc.ca/ExactDemo/> |
| [47](#_bookmark65) | <https://semrep.nlm.nih.gov/SemRep.v1.8_Installation.html>, SemMed là một ứng dụng dựa trên web được xuất bản sau khi ấn phẩm này được phát hành: [https://skr3.nlm.nih.gov/](https://skr3.nlm.nih.gov/SemMed/semmed.html)  [SemMed/semmed.html](https://skr3.nlm.nih.gov/SemMed/semmed.html) |
| [69](#_bookmark66) | Cơ sở dữ liệu với tất cả dữ liệu được trích xuất có sẵn trực tuyến: [https://trialstreamer.robotreviewer.](https://trialstreamer.robotreviewer.net/) [mạng/](https://trialstreamer.robotreviewer.net/) |
| [58](#_bookmark87) | Đang chờ xử lý: bài viết đề cập rằng một ứng dụng đang được triển khai. |
| [36](#_bookmark45) | <http://ns.inria.fr/acta/> |
| [82](#_bookmark92) | Mã ứng dụng để triển khai riêng có sẵn tại đây: <https://osf.io/2dqcg/> |

ghi chú bởi Ref. [75](#_bookmark79), do đó, họ đã chia sẻ các chú thích tiêu chuẩn vàng được sử dụng làm dữ liệu đào tạo hoặc đánh giá và thông tin về cách lấy các văn bản.

* + - 1. Việc sử dụng các khuôn khổ của bên thứ ba có được báo cáo không và chúng có thể truy cập được không?

Trong số các ấn phẩm được đưa vào đánh giá cơ sở, 47 trong số 53 (88%) mô tả việc sử dụng ít nhất một khuôn khổ của bên thứ ba cho hệ thống trích xuất dữ liệu của họ. Danh sách sau đây có thể không đầy đủ, do mã không có sẵn và báo cáo không đầy đủ trong các ấn phẩm đi kèm. Thông thường nhất, có mô tả về bộ công cụ học máy (Mallet, N = 12; Weka, N = 6; lưu lượng tensor, N = 5; scikit-learn, N = 3). Các bộ công cụ xử lý ngôn ngữ tự nhiên như Stanford parser / CoreNLP (N = 12) hoặc NLTK (N = 3), cũng thường được báo cáo cho các bước xử lý trước và phân tích cú pháp phụ thuộc trong các ấn phẩm. Công cụ MetaMap đã được sử dụng trong chín ấn phẩm và trình gắn thẻ GENIA trong bốn ấn phẩm. Để biết danh sách đầy đủ các framework, vui lòng xem Phụ lục A và D trong *Dữ liệu cơ bản.*[127](#_bookmark142)

* + 1. *Hiệu lực bên trong và bên ngoài của mô hình*
       1. Bộ dữ liệu hoặc thước đo đánh giá có cung cấp khả năng so sánh với các công cụ khác trong cùng một lĩnh vực không?

Với mục này, chúng tôi nhằm mục đích đánh giá các ấn phẩm để xem liệu kết quả đánh giá từ các mô hình có so sánh với kết quả từ các mô hình khác hay không. Lý tưởng nhất, một ấn phẩm sẽ báo cáo kết quả của một mô hình phân loại khác trên cùng một bộ dữ liệu, bằng cách tự triển khai lại mô hình[96](#_bookmark120) hoặc bằng cách mô tả kết quả của các mô hình khác khi sử dụng bộ dữ liệu điểm chuẩn.[Điều](#_bookmark99) này hiếm khi xảy ra đối với các ấn phẩm trong đánh giá cơ sở, vì hầu hết các bộ dữ liệu chỉ được tuyển chọn và sử dụng trong các ấn phẩm đơn lẻ. Tuy nhiên, việc tái sử dụng kho dữ liệu điểm chuẩn đã tăng lên với các ấn phẩm trong bản cập nhật LSR, nơi chúng tôi tìm thấy 40 ấn phẩm báo cáo kết quả trên một trong những bộ dữ liệu điểm chuẩn đã xuất bản trước đó (xem [Bảng 4](#_bookmark13)).

Ngoài ra, trong đánh giá cơ sở, trong 40 ấn phẩm (75%) dữ liệu đã được mô tả rõ ràng và họ sử dụng các thực thể thường được sử dụng và các chỉ số đánh giá phổ biến, chẳng hạn như độ chính xác, khả năng nhớ và điểm F1, dẫn đến khả năng so sánh kết quả hạn chế. Trong những trường hợp này, khả năng so sánh bị hạn chế vì các ấn phẩm đó sử dụng các bộ dữ liệu khác nhau, điều này có thể ảnh hưởng đến độ khó của nhiệm vụ trích xuất dữ liệu và dẫn đến kết quả tốt hơn trong ví dụ như các bộ dữ liệu có cấu trúc hoặc bộ dữ liệu theo chủ đề cụ thể.

* + - 1. Giải thích cho ảnh hưởng của cả biến hiển thị và ẩn trong tập dữ liệu có được đưa ra không?

Mục này chỉ liên quan đến các ấn phẩm sử dụng máy học hoặc mạng nơ-ron. Các hệ thống phân loại dựa trên quy tắc (N = 8, 15% báo cáo cơ sở quy tắc như cách tiếp cận duy nhất) không áp dụng cho mục này, bởi vì các quy tắc dẫn đến quyết định được lựa chọn có chủ ý bởi những người tạo ra hệ thống và do đó luôn hiển thị.

Mười ấn phẩm trong đánh giá cơ sở (19%) thảo luận về các biến ẩn.[83](#_bookmark94) thảo luận rằng việc xác định thực thể nhóm điều trị mang lại kết quả tốt nhất. Tuy nhiên, khi cả hai từ 'nhóm' và 'cánh tay' đều không có trong văn bản thì hệ thống đã gặp vấn đề với việc xác định thực thể. 'Trigger tokens'[104](#_bookmark136) và ảnh hưởng của các cụm từ phổ biến cũng được mô tả bởi Ref. [68](#_bookmark107), sau này cho thấy rằng hệ thống của họ có thể mang lại một số phân loại tích cực trong trường hợp không có các cụm từ chung.[103](#_bookmark134) đã tiến thêm một bước nữa và cung cấp một bảng với các từ có tác động nhiều nhất đến dự đoán của mỗi lớp.[57](#_bookmark85) mô tả việc loại bỏ các tiêu đề câu trong các bản tóm tắt có cấu trúc để tránh tạo ra một hệ thống thiên về các thuật ngữ chung, trong khi Ref. [90](#_bookmark108) thảo luận về chữ viết tắt và ngữ pháp là các yếu tố ảnh hưởng đến kết quả. Độ dài của văn bản đầu vào[59](#_bookmark89) và vị trí của một câu trong một đoạn văn hoặc tóm tắt, ví dụ: điểm phân loại thấp hơn tới 10% đối với một số kết hợp câu nhất định trong các bản tóm tắt không có cấu trúc, đã được thể hiện trong một số ấn phẩm. [46,66,102](#_bookmark103)

* + - 1. Quá trình tránh quá khớp hoặc thiếu khớp có được mô tả không?

'Overfitted' là một thuật ngữ được sử dụng để mô tả một hệ thống cho thấy kết quả đánh giá đặc biệt tốt trên một bộ dữ liệu cụ thể vì nó đã học cách phân loại nhiễu và các biến thể nội tại khác trong dữ liệu như một phần của mô hình của nó.[105](#_bookmark138)

Trong số các ấn phẩm được đưa vào đánh giá cơ sở, 33 trong số 53 (62%) báo cáo rằng họ đã sử dụng các phương pháp để tránh quá khớp. Tám (15%) trong số tất cả các ấn phẩm báo cáo phân loại dựa trên quy tắc là cách tiếp cận duy nhất của họ, cho phép họ không dễ bị quá khớp bởi máy học.

Hơn nữa, 28 ấn phẩm đã báo cáo xác nhận chéo để tránh quá khớp. Hầu hết các bộ phân loại này nằm trong lĩnh vực học máy, ví dụ như SVM. Thông thường nhất, 10 nếp gấp được sử dụng (N = 15), nhưng tùy thuộc vào kích thước của kho đánh giá, 3, 6, 5 hoặc 15 nếp gấp cũng được mô tả. Hai ấn phẩm[55,85](#_bookmark80) cảnh báo rằng xác nhận chéo với số lượng gấp cao (ví dụ: 10) gây ra phương sai cao trong kết quả đánh giá khi sử dụng các bộ dữ liệu nhỏ như NICTA-PIBOSO. Một ấn phẩm[104 phân](#_bookmark136)  tầng các nếp gấp theo lớp để tránh sự khác biệt này trong đánh giá dẫn đến một nếp gấp được gây ra bởi sự thưa thớt của các trường hợp tích cực.

Các ấn phẩm trong lĩnh vực thần kinh và học sâu mô tả các phương pháp tiếp cận như dừng sớm, bỏ học, chính quy hóa L2 hoặc suy giảm cân nặng. [59,96,106](#_bookmark89)  Một số ấn phẩm không thảo luận cụ thể về quá khớp trong văn bản, nhưng mã nguồn mở của họ chỉ ra rằng các kỹ thuật sau đã được sử dụng.[55,75](#_bookmark80)

F1000Research 2023, 10:401 Cập nhật lần cuối: 14 NOV 2024

Trang 21 của 46

Bảng 4. Corpora được sử dụng trong các ấn phẩm đi kèm. RCT, thử nghiệm ngẫu nhiên có đối chứng; IR, truy xuất thông tin; PICO, dân số, can thiệp, so sánh, kết quả; UMLS, hệ thống ngôn ngữ y tế thống nhất.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Công bố | Cũng được sử dụng bởi | Sự miêu tả | Lớp học | Kích thước / loại | Sẵn sàng | Ghi |
| [96](#_bookmark120) | [39,54,87,95,98](#_bookmark51) Bộ dữ liệu  Chuyển thể: [60](#_bookmark91) | Nhãn câu được dán nhãn tự động từ các bản tóm tắt có cấu trúc cho đến tháng 8 năm17 | P, IC, O, Phương pháp | 24,668  Tóm tắt | Có, [https://github.com/jind11/](https://github.com/jind11/PubMed-PICO-Detection)  [PubMed-PICO-Detection](https://github.com/jind11/PubMed-PICO-Detection) |  |
| [55](#_bookmark80) | [32,33,36,61,74,85,95,98,100,106](#_bookmark39)  Thích ứng bộ dữ liệu:  [34,37,50,67](#_bookmark47) | Thực thể | P, IC, O + tuổi, giới tính và các thực thể khác | 5,000  Tóm tắt | Có, <https://github.com/bepnye/EBM-NLP> |  |
| [97](#_bookmark122) |  | Thực thể | I và liều lượng liên quan | 694 Tóm tắt/ Toàn văn | Có, [https://ii.nlm.nih.gov/DataSets/](https://ii.nlm.nih.gov/DataSets/index.shtml)  [index.shtml](https://ii.nlm.nih.gov/DataSets/index.shtml) | Can thiệp dựa trên thuốc miền |
| [48](#_bookmark67) |  | Thực thể | P, O, Thiết kế, Phơi sáng | 60 + 30  Tóm tắt | Vâng, [http://gnteam.cs.manchester.ac.](http://gnteam.cs.manchester.ac.uk/old/epidemiology/data.html) [uk/old/epidemiology/data.html](http://gnteam.cs.manchester.ac.uk/old/epidemiology/data.html) | Béo phì miền |
| [75](#_bookmark79) |  | Mức câu 90.000 chú thích giám sát từ xa, 1000 hướng dẫn sử dụng. | Điều kiện mục tiêu, kiểm tra chỉ số và tiêu chuẩn tham chiếu | 90,000 + 1000  Câu | Có (nhãn, không phải văn bản), [https://zenodo.](https://zenodo.org/record/1303259) [org/record/1303259](https://zenodo.org/record/1303259) | Xét nghiệm chẩn đoán miền |
| [52](#_bookmark75) | [64](#_bookmark99)  (bao gồm các bộ phân loại từ), [40,53,54,102,107–](#_bookmark53)110 | Tóm tắt có cấu trúc và không có cấu trúc, đa nhãn trên câu. | P, IC, O, Thiết kế | 1000  Tóm tắt | Có, [https://drive.google.com/file/](https://drive.google.com/file/d/1M9QCgrRjERZnD9LM2FeK-3jjvXJbjRTl/view?usp=sharing)  [d/1M9QCgrRjERZnD9LM2FeK-](https://drive.google.com/file/d/1M9QCgrRjERZnD9LM2FeK-3jjvXJbjRTl/view?usp=sharing)  [3jjvXJbjRTl/view?usp=chia sẻ](https://drive.google.com/file/d/1M9QCgrRjERZnD9LM2FeK-3jjvXJbjRTl/view?usp=sharing) | Câu đa nhãn |
| [47](#_bookmark65) |  | Câu | Can thiệp thuốc và tuyên bố so sánh cho từng nhánh | 300 (500 trong  dữ liệu có sẵn) câu | Có, [https://dataverse.harvard.edu/](https://dataverse.harvard.edu/file.xhtml?fileId=4171005&version=1.0)  [file.xhtml?fileId=4171005&version](https://dataverse.harvard.edu/file.xhtml?fileId=4171005&version=1.0) [=1.0](https://dataverse.harvard.edu/file.xhtml?fileId=4171005&version=1.0) | Can thiệp dựa trên thuốc miền |
| [98](#_bookmark123) |  | Câu | P, IC, O | 5099  câu từ các tài liệu tham khảo có trong SR, được dán nhãn bằng cách sử dụng học tập tích cực | Có, [https://github.com/wds-seu/ Aceso](https://github.com/wds-seu/Aceso/tree/master/datasets) [/tree/master/datasets](https://github.com/wds-seu/Aceso/tree/master/datasets) | Bệnh tim miền |

F1000Research 2023, 10:401 Cập nhật lần cuối: 14 NOV 2024

Trang 22 của 46

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Bảng 4. *Tiếp tục* | | | | | | |
| Công bố | Cũng được sử dụng bởi | Sự miêu tả | Lớp học | Kích thước / loại | Sẵn sàng | Ghi |
| [62](#_bookmark95)dựa trên  trên[111](#_bookmark111) | [32,61,99](#_bookmark93) | Câu | P, I, O | Toàn văn: 12,616  nhắc nhở bắt nguồn từ 3.346 bài báo; Chỉ tóm tắt: 6375 lời nhắc | Vâng, <http://evidence-inference.ebm-nlp.com/download/> | Bộ ba để trích xuất quan hệ |
| [61](#_bookmark93) |  | Câu, Thực thể | P, IC, O | 470 nghiên cứu từ 20k đánh giá, nhãn thực thể ban đầu được chỉ định thông qua mô hình được đào tạo trên EBM-NLP | Vâng, <https://github.com/allenai/ms2> | Trích xuất mối quan hệ với hướng của nhãn hiệu ứng |
| [35](#_bookmark42) |  | Thực thể | P, IC, xét nghiệm chẩn đoán | 500 bản tóm tắt  và 700 hồ sơ xét xử | Vâng, [http://www.lllf.uam.es/ESP/](http://www.lllf.uam.es/ESP/nlpmedterm_en.html)  [nlpmedterm\_en.html](http://www.lllf.uam.es/ESP/nlpmedterm_en.html) | Bộ dữ liệu tiếng Tây Ban Nha, chuẩn hóa UMLS |
| [36](#_bookmark45) |  | Thực thể | P, O | 660 RCT  Tóm tắt | Vâng, [https://gitlab.com/tomaye/](https://gitlab.com/tomaye/abstrct)  [bất khả xâm phạm](https://gitlab.com/tomaye/abstrct) | Chiết xuất quan hệ, ung thư miền, tăng nhãn áp, viêm gan, tiểu đường, tăng huyết áp |
| [112](#_bookmark113) | [50](#_bookmark71) | Thực thể | P, IC, O, Thiết kế | 99 RCT  Tóm tắt | Có, [https://github.com/](https://github.com/jetsunwhitton/RCT-ART)  [jetsunwhitton/RCT-ART](https://github.com/jetsunwhitton/RCT-ART) | Loại trừ vì chỉ chứa các xét nghiệm tăng nhãn áp |
| [34](#_bookmark41) | [67](#_bookmark105) | Thực thể | O | 300 bản tóm tắt | Vâng, <https://github.com/LivNLP/ODP-tagger> | Dữ liệu riêng + sự thích ứng của EBM-NLP với chuẩn hóa thành 38 miền  và 5 lĩnh vực kết quả |

F1000Research 2023, 10:401 Cập nhật lần cuối: 14 NOV 2024

Trang 23 của 46

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Bảng 4. *Tiếp tục* | | | | | | |
| Công bố | Cũng được sử dụng bởi | Sự miêu tả | Lớp học | Kích thước / loại | Sẵn sàng | Ghi |
| [33](#_bookmark39) |  | Thực thể | Tôi | 1807  tóm tắt, được dán nhãn tự động bằng cách khớp các chuỗi can thiệp từ đăng ký thử nghiệm lâm sàng | Có, [https://data.mendeley.com/](https://data.mendeley.com/datasets/ccfnn3jb2x/1)  [datasets/ccfnn3jb2x/1](https://data.mendeley.com/datasets/ccfnn3jb2x/1) |  |
| [60](#_bookmark91) |  | Câu | P, IC, O | 42000  Câu | Có, [https://github.com/](https://github.com/smileslab/Brain_Aneurysm_Research/tree/master/BioMed_Summarizer)  [smileslab/Brain\_Aneurysm\_Research/tree](https://github.com/smileslab/Brain_Aneurysm_Research/tree/master/BioMed_Summarizer) [/master/BioMed\_Summarizer](https://github.com/smileslab/Brain_Aneurysm_Research/tree/master/BioMed_Summarizer) | Dữ liệu riêng về phình động mạch não + bộ dữ liệu hiện có từ Jin và Szolovits[96](#_bookmark120) |
| [74](#_bookmark76) |  | Câu, Thực thể | P, IC, O | 130 bản tóm tắt từ MEDLINE's  Giao diện PICO trực tuyến PubMed | Có, [https://github.com/nstylia/](https://github.com/nstylia/pico_entities/)  [pico\_entities/](https://github.com/nstylia/pico_entities/) |  |
| [99](#_bookmark125) |  | Thực thể | TÔI, C, O | 10 RCT  Tóm tắt | Có, [https://www.ncbi.nlm.nih.gov/](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8135980/bin/ocab077_supplementary_data.pdf)  [pmc/articles/PMC8135980/](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8135980/bin/ocab077_supplementary_data.pdf) [bin/ocab077\_supplementary\_data.](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8135980/bin/ocab077_supplementary_data.pdf) [Pdf](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8135980/bin/ocab077_supplementary_data.pdf) | Trích xuất quan hệ, miền COVID-19 |
| [38](#_bookmark50) |  | Câu | P, IC, O, N + CONSORT  Mục | 50 RCT toàn văn | Có, [https://github.com/kilicogluh/](https://github.com/kilicogluh/CONSORT-TM)  [CONSORT-TM](https://github.com/kilicogluh/CONSORT-TM) |  |
| [82](#_bookmark92) |  | Thực thể, câu | I, C, O + thực thể động vật | 400 RCT  Tóm tắt trong kho dữ liệu đầu tiên, 10k tóm tắt trong kho dữ liệu bổ sung từ dữ liệu khai thác | Vâng, <https://osf.io/2dqcg/> | RCT động vật miền |
| [51](#_bookmark73) |  | Thực thể | P, TÔI, C, O | 211 RCT  Tóm tắt và 20 văn bản đầy đủ | Có, [https://zenodo.org/ record](https://zenodo.org/record/6365890) [/6365890](https://zenodo.org/record/6365890) |  |

F1000Research 2023, 10:401 Cập nhật lần cuối: 14 NOV 2024

Trang 24 của 46

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Bảng 4. *Tiếp tục* | | | | | | |
| Công bố | Cũng được sử dụng bởi | Sự miêu tả | Lớp học | Kích thước / loại | Sẵn sàng | Ghi |
| [70](#_bookmark68) |  | Thực thể | N | 200 RCT  toàn văn từ PMC,  chú thích N từ bảng cơ sở | Có, [https://zenodo.org/ record](https://zenodo.org/record/6647853#.ZCa9dXbMJPY) [/6647853#. ZCa9dXbMJPY](https://zenodo.org/record/6647853#.ZCa9dXbMJPY) |  |
| [63](#_bookmark97)dựa trên  trên[111](#_bookmark111) |  | Thực thể | Tôi, C, O | Kho dữ liệu đầu tiên 160 tóm tắt, kho thứ hai 20 | Có, [https://github.com/bepnye/](https://github.com/bepnye/evidence_extraction/blob/master/data/exhaustive_ico_fixed.csv)  [evidence\_extraction/blob/master/](https://github.com/bepnye/evidence_extraction/blob/master/data/exhaustive_ico_fixed.csv) [data/exhaustive\_ico\_fixed.csv](https://github.com/bepnye/evidence_extraction/blob/master/data/exhaustive_ico_fixed.csv) | Tiểu thể thứ hai là ung thư miền |
| [39](#_bookmark51) |  | Câu, Thực thể | P, IC, O | 500 bản tóm tắt được gắn nhãn cho câu và 100 cho các thực thể P, O | Không |  |
| [73](#_bookmark74) |  | Thực thể | O | 1300  Tóm tắt với 3100  Tuyên bố kết quả | Không | Ung thư miền |
| [63.111](#_bookmark97) |  |  |  |  |  |  |
| [45](#_bookmark61) |  | Thực thể | P, IC, O | Bộ dữ liệu do Cochrane cung cấp với 10137  Tóm tắt | Không |  |
| [61](#_bookmark93) | [113](#_bookmark115) | Câu và thực thể | P, N, các phần | 3657  Tóm tắt có cấu trúc với thẻ câu, 204 bản tóm tắt với N (tổng số) thực thể | Không |  |
| [57](#_bookmark85) |  | Tóm tắt RCT có cấu trúc, được dán nhãn tự động với thẻ câu và 378 tài liệu với thẻ truy vấn IR cấp thực thể | P, IC, O | 15,000  Tóm tắt + 378  tài liệu có thẻ IR | Không |  |

F1000Research 2023, 10:401 Cập nhật lần cuối: 14 NOV 2024

Trang 25 của 46

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Bảng 4. *Tiếp tục* | | | | | | |
| Công bố | Cũng được sử dụng bởi | Sự miêu tả | Lớp học | Kích thước / loại | Sẵn sàng | Ghi |
| [84](#_bookmark96) | [83](#_bookmark94)  (không rõ ràng) | Câu và thực thể | IC, O, N (tổng + mỗi cánh tay) | 263 bản tóm tắt | Không |  |
| [76](#_bookmark81) | [53](#_bookmark77), [58](#_bookmark87) | 100 bản tóm tắt với P, Điều kiện, IC, có thể ở cấp độ thực thể. Đối với O, 633 bản tóm tắt được chú thích ở cấp câu. | P, Tình trạng, IC, 0 | 633 bản tóm tắt cho O, 100 cho các lớp khác | Không |  |
| [77](#_bookmark82) |  | Thực thể | Tuổi, Thiết kế, Bối cảnh (Quốc gia), IC, N, ngày học và các tổ chức liên kết | 185 Toàn văn  (ít nhất 93 được dán nhãn) | Không |  |
| [79](#_bookmark86) |  | Câu và thực thể | P, IC, Tuổi, Giới tính, Thiết kế, Tình trạng, Chủng tộc | 2000  câu từ tóm tắt | Không |  |
| [93](#_bookmark114) |  | 200 bản tóm tắt, 140 chứa nhãn câu và thực thể | P, IC | 200 bản tóm tắt | Không |  |
| [114](#_bookmark117) |  | Tóm tắt có cấu trúc được dán nhãn tự động, cấp câu. | P, IC, O | 14200+  Tóm tắt | Không |  |
| [94](#_bookmark116) |  | Thực thể | P, tuổi, giới tính, chủng tộc | 50 bản tóm tắt | Không |  |
| [115](#_bookmark119) |  | Câu (và thực thể?) | P, IC, O | 3000  Tóm tắt | Không |  |
| [42](#_bookmark56) |  | Thực thể | N (tổng) | 648 bản tóm tắt | Không |  |
| [90](#_bookmark108) |  | Thực thể | IC | 330 bản tóm tắt | Không |  |
| [66](#_bookmark103) |  | Văn bản tiếng Indonesia có chú thích câu | P, I, C, O | 200 bản tóm tắt | Không |  |
| [68](#_bookmark107) |  | Câu từ 69  (tim) +24 RCT (ngẫu nhiên) được đưa vào đánh giá Cochrane | Tiêu chí đưa vào | 69 + 24 toàn văn | Không | Tim mạch miền |
| [80](#_bookmark88) |  | Câu và thực thể | P, IC, Tuổi, Giới tính, P (Tình trạng hoặc bệnh tật) | 200 bản tóm tắt | Không |  |

F1000Research 2023, 10:401 Cập nhật lần cuối: 14 NOV 2024

Trang 26 của 46

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Bảng 4. *Tiếp tục* | | | | | | |
| Công bố | Cũng được sử dụng bởi | Sự miêu tả | Lớp học | Kích thước / loại | Sẵn sàng | Ghi |
| [71](#_bookmark70) |  | 4.824 câu từ  18 Cập nhật  tài liệu và 714 câu từ các trích dẫn MEDLINE cho  P. Đối với I: CLEF 2013  nhiệm vụ chung và 852 trích dẫn MEDLINE | P, IC, P (Tình trạng hoặc bệnh) | tóm tắt, toàn văn | Không | Chủ đề chung và lĩnh vực tim mạch |
| [41](#_bookmark55) | [102](#_bookmark132) | Chú thích thực thể dưới dạng cụm danh từ | O, IC | 100 + 132  Câu từ toàn văn | Không | Tạp chí bệnh tiểu đường và nội tiết làm nguồn |
| [92](#_bookmark112) | [103](#_bookmark134) | Câu trừu tượng RCT có cấu trúc tự động được gắn nhãn.[92](#_bookmark112) có 19.854 câu, giả định cùng ngữ liệu với tác giả và kỹ thuật là giống nhau. | P, IC, O | 23,472  Tóm tắt | Không |  |
| [46](#_bookmark64) |  | RCTs tóm tắt và toàn văn văn bản: 132 + 50 bài báo | IC (mỗi nhánh), IC (thực thể thuốc.), O (thời điểm), O (kết quả chính hoặc phụ), N (tổng số), Tiêu chí đủ điều kiện, Ngày đăng ký, Tổ chức tài trợ, Số tài trợ, Dừng sớm, Đăng ký thử nghiệm, Siêu dữ liệu | 132 + 50  Tóm tắt và toàn văn | Không |  |
| [86](#_bookmark100) |  | Câu và thực thể | P, IC, O, N (mỗi cánh tay + tổng) | 48 Toàn văn | Không |  |
| [49](#_bookmark69) |  | Các nghiên cứu từ 5 đánh giá có hệ thống về phơi nhiễm sức khỏe môi trường, thực thể | P, O, Quốc gia, Phơi nhiễm | Nghiên cứu từ 5 đánh giá hệ thống | Không | Các nghiên cứu quan sát về tiếp xúc với sức khỏe môi trường ở người |

F1000Research 2023, 10:401 Cập nhật lần cuối: 14 NOV 2024

Trang 27 của 46

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Bảng 4. *Tiếp tục* | | | | | | |
| Công bố | Cũng được sử dụng bởi | Sự miêu tả | Lớp học | Kích thước / loại | Sẵn sàng | Ghi |
| [44](#_bookmark59) |  | Được dán nhãn thông qua giám sát từ xa có giám sát. Toàn văn (~12500 mỗi lớp), 50  + 133 chú thích thủ công để đánh giá. | P, IC, O | 12700+ toàn văn | Không |  |
| [89](#_bookmark106) |  | Nhãn câu, tóm tắt có cấu trúc và không có cấu trúc. Chú thích thủ công: 344 IC,  341 O, và 144 P và hơn thế nữa có được bằng cách dán nhãn tự động. | P, IC, O | 344+  Tóm tắt | Không |  |
| [88](#_bookmark104) |  | Thực thể | P, IC, O, O dưới dạng  "Dụng cụ" hoặc "Biến nghiên cứu" | 20 Toàn văn / Tóm tắt | Không |  |
| [85](#_bookmark98) |  | Thực thể (định dạng Brat, IOB) | P, IC, O | 170 bản tóm tắt | Không |  |
| [59](#_bookmark89) |  | Các thực thể được gán cho các khái niệm UMLS (có thể là kho dữ liệu Cochrane, kích thước không rõ ràng). '88 trường hợp, được chú thích tổng cộng với 76, 87 và 139 [P, IC, O tương ứng]' | P, IC, O | Không rõ ràng, ít nhất 88 tài liệu | Không |  |
| [43](#_bookmark57) |  | Câu và thực thể | P, IC (mỗi cánh tay), N (tổng) | 1750 tiêu đề hoặc tóm tắt | Không |  |
| [116](#_bookmark121) |  | Loại trừ giấy, không có hệ thống trích xuất dữ liệu. Kho dữ liệu của bệnh nhân, dân số, vấn đề, phơi nhiễm, can thiệp, so sánh, kết quả, thời lượng và kết quả trong bản tóm tắt. |  |  | Không | Bị loại khỏi đánh giá, nhưng mô tả kho dữ liệu có liên quan |
| [56](#_bookmark83) |  | Câu và thực thể | P, IC (mỗi cánh tay), O, nhiều hơn | 88 Toàn văn | Không |  |

* + - 1. Quá trình tách đào tạo từ dữ liệu xác thực có được mô tả không?

Phân bổ ngẫu nhiên cho các nhóm điều trị là một mục quan trọng khi đánh giá sai lệch trong RCT, vì phân bổ có chọn lọc có thể dẫn đến sự khác biệt ban đầu.[1](#_bookmark20) Tương tự như vậy, quá trình tách một bộ dữ liệu một cách ngẫu nhiên, hoặc theo cách phân tầng, thành dữ liệu đào tạo (hoặc tạo quy tắc) và dữ liệu thử nghiệm là rất quan trọng khi xây dựng các bộ phân loại và hệ thống thông minh.[117](#_bookmark124)

Tất cả các ấn phẩm bao gồm trong đánh giá cơ sở đều đưa ra các dấu hiệu về cách thu được các bộ dữ liệu đào tạo và đánh giá khác nhau. Thông thường nhất là có một bộ dữ liệu và tỷ lệ phân tách chỉ ra rằng việc phân tách là ngẫu nhiên. Thông tin này được cung cấp trong 36 ấn phẩm (68%).

Đối với các ấn phẩm đề cập đến xác nhận chéo (N = 28, 53%), chúng tôi giả định rằng việc phân tách là ngẫu nhiên. Tỷ lệ phân tách (ví dụ: 80:20 đối với dữ liệu đào tạo và thử nghiệm) là rõ ràng trong các trường hợp xác nhận chéo và được mô tả trong phần còn lại của các ấn phẩm.

Các ấn phẩm cũng thường sử dụng các bộ dữ liệu hoàn toàn khác nhau hoặc lặp lại nhiều lần tách, đào tạo và thử nghiệm (N = 13, 24%). Ví dụ: Ref. [56 đã](#_bookmark83)  sử dụng xác nhận chéo để đào tạo và đánh giá mô hình của họ, và sau đó sử dụng một kho dữ liệu bổ sung sau quá trình xác nhận chéo. Tương tự tham khảo. [59](#_bookmark89), sử dụng phân tách đào tạo/thử nghiệm 60:40, nhưng sau đó tạo ra một kho dữ liệu bổ sung gồm 88 tài liệu để xác nhận thêm hiệu suất của mô hình trên dữ liệu chưa từng thấy trước đây.

* + - 1. Khả năng thích ứng của mô hình với các định dạng và/hoặc môi trường khác nhau ngoài dữ liệu đào tạo và thử nghiệm có được mô tả không?

Đối với mục này, chúng tôi nhằm mục đích tìm hiểu xem có bao nhiêu ấn phẩm được bao gồm trong đánh giá cơ sở đã thử nghiệm các thuật toán trích xuất dữ liệu của họ trên các bộ dữ liệu khác nhau. Một hạn chế thường được ghi nhận trong tài liệu là các trình chú thích tiêu chuẩn vàng có phong cách và sở thích khác nhau, và bộ dữ liệu nhỏ và giới hạn trong một tìm kiếm tài liệu cụ thể. Đánh giá một mô hình trên nhiều bộ dữ liệu độc lập cung cấp khả năng định lượng dữ liệu có thể được trích xuất tốt như thế nào trên các miền và mức độ linh hoạt của mô hình trong ứng dụng thực tế với các bộ dữ liệu hoàn toàn mới. Trong số các ấn phẩm được đưa vào, 19 (36%) thảo luận về cách mô hình của họ hoạt động trên các bộ dữ liệu có các đặc điểm khác với các đặc điểm được sử dụng để đào tạo và thử nghiệm. Tuy nhiên, trong một số trường hợp, đánh giá này là định tính khi các mô hình được áp dụng cho các bộ dữ liệu lớn không được gắn nhãn, ngoài đời thực. [46,58,69,48,95,101,102](#_bookmark87)

* + 1. *Khác*
       1. Lưu ý

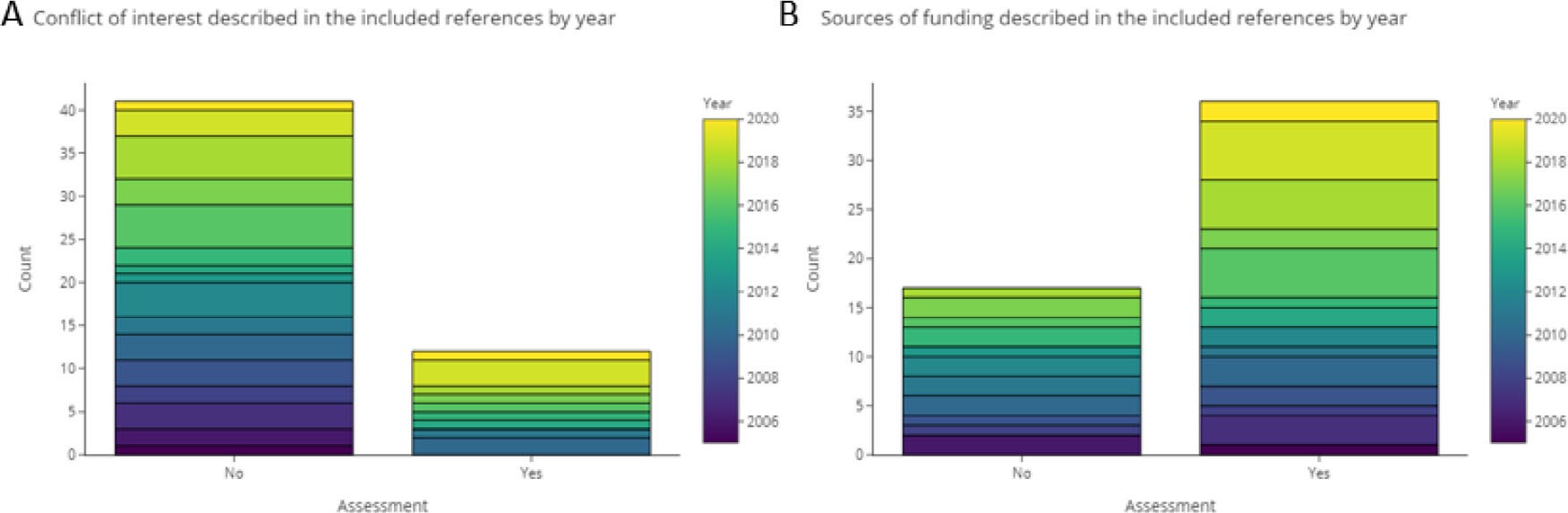
Các cảnh báo được trích xuất dưới dạng văn bản tự do. Các ấn phẩm bao gồm (N = 64, 86%) báo cáo nhiều cảnh báo. Sau khi trích xuất, chúng tôi cấu trúc chúng thành sáu lĩnh vực khác nhau:

* + - * 1. Sự bất đồng về chất lượng nhãn và giữa các người chú thích
        2. Các biến thể trong văn bản
        3. Khả năng thích ứng và so sánh miền
        4. Ý nghĩa của tính toán hoặc kiến trúc hệ thống
        5. Thiếu thông tin trong văn bản hoặc cơ sở kiến thức
        6. Ý nghĩa thực tế

Những điều này được thảo luận thêm trong phần 'Thảo luận' của bài đánh giá sống này.

* + - 1. Nguồn tài trợ và xung đột lợi ích

[Hình 11](#_bookmark14) cho thấy hầu hết các ấn phẩm được đưa vào đánh giá cơ sở không tuyên bố bất kỳ xung đột lợi ích nào. Điều này đúng với hầu hết các ấn phẩm được xuất bản trước năm 2010 và khoảng 50% tài liệu được xuất bản trong những năm gần đây. Tuy nhiên, các nguồn tài trợ được công bố phổ biến hơn, với 69% tất cả các ấn phẩm bao gồm các tuyên bố cho mục này. Điều này phản ánh xu hướng báo cáo đầy đủ hơn trong những năm gần đây.



Hình 11. Tuyên bố các nguồn tài trợ và xung đột lợi ích trong các nghiên cứu bao gồm.

1. Sự thảo luận
   1. Tóm tắt các phát hiện chính
      1. *Kiến trúc hệ thống*

Các hệ thống được mô tả trong các ấn phẩm đi kèm đang thay đổi theo thời gian. Trích xuất dữ liệu không học máy thông qua cơ sở quy tắc và API là một trong những cách tiếp cận sớm nhất và được sử dụng thường xuyên nhất. Các bộ phân loại máy học cổ điển khác nhau như Bayes và SVM ngây thơ rất phổ biến trong các tài liệu được xuất bản từ năm 2005-2018. Cho đến năm 2020, đã có xu hướng nhúng từ và mạng nơ-ron như LSTM. Từ năm 2020 đến năm 2022, chúng tôi quan sát thấy xu hướng hướng tới máy biến áp, đặc biệt là kiến trúc BERT, RoBERTa và ELECTRA được đào tạo trước về văn bản y sinh hoặc khoa học.

* + 1. *Đánh giá*

Chúng tôi nhận thấy rằng độ chính xác, khả năng nhớ và F1 được sử dụng làm số liệu đánh giá trong hầu hết các ấn phẩm, mặc dù đôi khi các chỉ số này được điều chỉnh hoặc nới lỏng để tính đến các trận đấu một phần hoặc tương tự.

* + 1. *Phạm vi*

Hầu hết các ấn phẩm bao gồm tập trung vào việc trích xuất dữ liệu từ các bản tóm tắt. Lý do cho điều này bao gồm sự sẵn có của dữ liệu và dễ dàng truy cập, cũng như độ bao phủ thông tin cao và sự sẵn có của các bản tóm tắt có cấu trúc có thể tự động lấy dữ liệu đào tạo được gắn nhãn. Một số lượng nhỏ hơn nhiều trong số các ấn phẩm được bao gồm (n = 19, 25%) trích xuất dữ liệu từ toàn văn bản. Một nửa số hệ thống trích xuất dữ liệu từ toàn văn đã được xuất bản trong vòng bảy năm qua. Trong thực tiễn đánh giá có hệ thống, trích xuất thủ công dữ liệu từ các bản tóm tắt nhanh hơn và dễ dàng hơn so với trích xuất thủ công dữ liệu từ toàn văn bản. Do đó, khả năng tiết kiệm thời gian và tiện ích của việc trích xuất dữ liệu toàn văn cao hơn nhiều vì tự động hóa có thể tiết kiệm nhiều thời gian hơn và nó cung cấp tự động hóa phản ánh chặt chẽ hơn công việc được thực hiện bởi những người đánh giá có hệ thống trong thực tế. Tuy nhiên, tài liệu trích xuất dữ liệu về toàn văn vẫn còn thưa thớt và trích xuất từ các bản tóm tắt có thể có giá trị hạn chế đối với người phản biện trong thực tế vì nó có nguy cơ thiếu thông tin. Bất cứ khi nào một ấn phẩm báo cáo trích xuất toàn văn bản, chúng tôi cố gắng tìm hiểu xem điều này có bao gồm văn bản trừu tượng hay không, trong trường hợp đó chúng tôi sẽ tính ấn phẩm trong cả hai loại. Tuy nhiên, thông tin này không phải lúc nào cũng được báo cáo rõ ràng.

* + 1. *Văn bản mục tiêu*

Báo cáo về các thử nghiệm ngẫu nhiên có đối chứng là các văn bản phổ biến nhất được sử dụng để trích xuất dữ liệu. Bằng chứng liên quan đến việc trích xuất dữ liệu từ các loại nghiên cứu khác rất hiếm và được thảo luận thêm trong các phần sau.

* 1. Đánh giá chất lượng báo cáo

Chúng tôi chỉ đánh giá chất lượng báo cáo đầy đủ trong đánh giá cơ sở và đánh giá các mục được chọn trong quá trình cập nhật đánh giá. Chất lượng báo cáo trong các nghiên cứu được đưa vào đánh giá cơ sở đang được cải thiện theo thời gian Chúng tôi đánh giá các ấn phẩm được đưa vào dựa trên danh sách 17 mục trong các lĩnh vực tái tạo, tính minh bạch, mô tả thử nghiệm, tính sẵn có của dữ liệu và tính hợp lệ bên trong và bên ngoài.

Đánh giá cơ sở: Khả năng tái tạo cao xuyên suốt, với thông tin về các nguồn dữ liệu đào tạo và đánh giá được báo cáo trong 94% tất cả các ấn phẩm và tiền xử lý được mô tả trong 89%.

Đánh giá cơ sở: Về tính minh bạch, 81% các ấn phẩm cung cấp mô tả rõ ràng về thuật toán của họ, 94% mô tả các đặc điểm của bộ dữ liệu của họ, nhưng chỉ 9% đề cập đến các thông số kỹ thuật phần cứng hoặc tính khả thi của việc sử dụng thuật toán của họ trên các bộ dữ liệu lớn trong thế giới thực như PubMed.

Cập nhật: Tính khả dụng của mã nguồn cao trong các ấn phẩm được thêm vào trong bản cập nhật LSR (N = 19, 83%). Trước khi cập nhật, 15% tất cả các ấn phẩm được đưa vào đã cung cấp mã của họ. Nhìn chung, 39% (N = 30) hiện có sẵn mã của họ và tất cả các liên kết đến kho lưu trữ mã được hiển thị trong [Bảng 2](#_bookmark9).

Đánh giá cơ sở: Thử nghiệm các hệ thống thường được mô tả, 89% đưa ra đánh giá chi tiết về các thuật toán của chúng. Sự đánh đổi giữa độ chính xác và khả năng thu hồi đã được thảo luận trong 32%.

Cập nhật: Các chỉ số cơ bản chỉ được báo cáo trong 19% (N = 14) các ấn phẩm được đưa vào, đây là xu hướng giảm từ 24% trong đánh giá cơ sở. Tuy nhiên, báo cáo đầy đủ hơn về mã nguồn và bộ dữ liệu công khai vẫn dẫn đến tăng tính minh bạch và khả năng so sánh.

Cập nhật: Tính khả dụng của các mô hình cuối cùng làm công cụ người dùng cuối rất kém. Chỉ có 12% trong số các ấn phẩm được bao gồm có một ứng dụng liên quan đến nó, nhưng chỉ có 5 (6%) có sẵn và có thể sử dụng trực tiếp qua các ứng dụng web (xem [Bảng 3](#_bookmark12) để biết các liên kết). Hơn nữa, không rõ có bao nhiêu công cụ khác được mô tả trong tài liệu được sử dụng trong thực tế, ngay cả khi chỉ được sử dụng nội bộ trong các nhóm nghiên cứu tác giả của họ. Có một xu hướng mạnh mẽ đáng ngạc nhiên đối với việc chia sẻ và sử dụng lại kho dữ liệu đã được xuất bản trong bản cập nhật LSR. Trước đó, dữ liệu đào tạo và đánh giá được dán nhãn có sẵn từ 13% các ấn phẩm và chỉ có 32% trong số tất cả các ấn phẩm được báo cáo bằng cách sử dụng một trong những bộ dữ liệu có sẵn này. Trong bản cập nhật LSR, 22 kho dữ liệu đã có sẵn trực tuyến và ít nhất 40 tài liệu khác bao gồm đề cập đến ấn phẩm sử dụng chúng. [Bảng 4](#_bookmark13) cung cấp các nguồn của tất cả các kho dữ liệu và các ấn phẩm sử dụng chúng. Đối với nhận dạng thực thể có tên, EBM-NLP[55](#_bookmark80) là bộ dữ liệu phổ biến nhất, được sử dụng bởi ít nhất 10 ấn phẩm khác và được điều chỉnh và sử dụng bởi bốn ấn phẩm khác. Để phân loại câu, tiêu chuẩn vàng NICTA[52](#_bookmark75) được sử dụng bởi tám người khác, và kho dữ liệu được dán nhãn tự động bởi Jin và Szolovits[96](#_bookmark120) được sử dụng bởi năm người khác và đã được điều chỉnh một lần. Để trích xuất mối quan hệ, kho dữ liệu EvidenceInference 2.0 đang thu hút sự chú ý, được sử dụng trong ít nhất ba ấn phẩm khác.

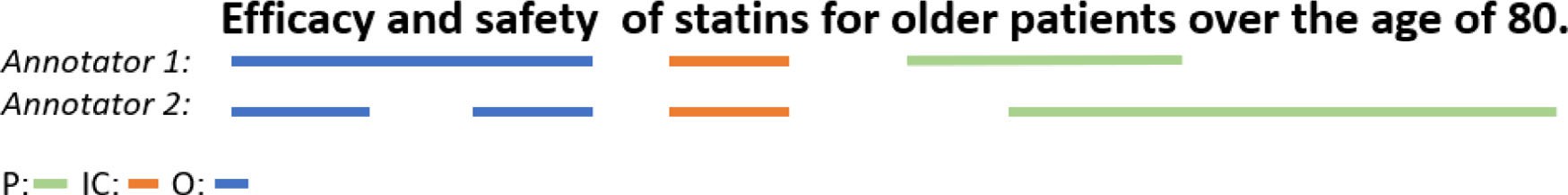
Đánh giá cơ sở: Tổng cộng 88% các ấn phẩm được mô tả sử dụng ít nhất một khuôn khổ của bên thứ ba có thể truy cập cho hệ thống trích xuất dữ liệu của họ. Tính hợp lệ bên trong và bên ngoài của mỗi mô hình được đánh giá dựa trên khả năng so sánh của nó với các công cụ khác (75%), đánh giá các biến hiển thị và ẩn trong dữ liệu (19%), tránh quá phù hợp (62%, không áp dụng cho các hệ thống không phải máy học), mô tả về việc tách đào tạo từ dữ liệu xác nhận (100%), và khả năng thích ứng và thử nghiệm bên ngoài trên các bộ dữ liệu có các đặc điểm khác nhau (36%). Những mục này, cùng với các cảnh báo và hạn chế được lưu ý trong các ấn phẩm đi kèm được thảo luận trong phần sau.

* 1. Cảnh báo và thách thức đối với đánh giá hệ thống (bán tự động hóa)

Trong phần sau, chúng tôi thảo luận về các cảnh báo và thách thức được nhấn mạnh bởi các tác giả của các ấn phẩm đi kèm. Chúng tôi tìm thấy nhiều chủ đề được thảo luận trong các ấn phẩm này và tóm tắt chúng theo bảy lĩnh vực khác nhau. Do xu hướng ngày càng tăng của các mô hình trích xuất quan hệ và tóm tắt văn bản, chúng tôi hiện tóm tắt bất kỳ thách thức hoặc cảnh báo nào liên quan đến những điều này trong văn bản cập nhật ở cuối mỗi miền áp dụng.

* + 1. *Sự bất đồng về chất lượng nhãn và giữa các người chú thích*

Chất lượng của nhãn trong bộ dữ liệu có chú thích đã được xác định là một vấn đề bởi một số tác giả. Độ dài của thực thể được chú thích, ví dụ như các thực thể O hoặc P, thường gây ra bất đồng giữa những người chú thích. [46,48,58,69,95,101,102](#_bookmark64)  Chúng tôi đã tạo ra một ví dụ trong [Hình 12](#_bookmark15), cho thấy hai chú thích có khả năng đúng, nhưng vẫn khác nhau trên cùng một câu.



Hình 12. Ví dụ về sự bất đồng giữa các người chú thích. P, dân số; I, can thiệp; C, so sánh; O, kết quả.

Những bất đồng tương tự,65,85,104 cùng với các chú thích bị bỏ lỡ,72 tốn nhiều thời gian để dung hòa 97 và làm cho điểm số kém tin cậy hơn.[95](#_bookmark118) Ví dụ về điều này, hai ấn phẩm quan sát thấy rằng hệ thống của họ hoạt động kém hơn đối với các lớp có sự bất đồng cao.[75.104](#_bookmark79) Có những lời giải thích khác nhau cho hiệu suất kém hơn trong những trường hợp này. Các mô hình có thể khó học hỏi từ dữ liệu được gắn nhãn với sự khác biệt có hệ thống bên trong. Một lý do khác là mô hình học các dự đoán dựa trên một kiểu chú thích và do đó các lỗi nhân tạo được tạo ra khi được đánh giá dựa trên dữ liệu được dán nhãn khác nhau, hoặc bản thân nhiệm vụ chú thích tự nhiên khó hơn trong trường hợp có sự bất đồng giữa các chú thích cao, và do đó hiệu suất thấp hơn từ các mô hình có thể được giải thích được. Tổng quan về các ấn phẩm bao gồm thảo luận về điều này, cùng với điểm số bất đồng giữa các người chú thích của họ, được đưa ra trong [Bảng 5](#_bookmark16).

Để giảm thiểu những vấn đề này, cần được đào tạo cẩn thận và hướng dẫn cho các chuyên gia chú thích.[58,77](#_bookmark87) Ví dụ, thông tin nên được cung cấp về việc có nhiều thực thể cơ bản hay một chú thích thực thể dài hơn hay không.[85 Các](#_bookmark98)  chú thích có nguồn gốc từ cộng đồng có thể chứa thông tin ồn ào hoặc không chính xác và có độ tin cậy của người đánh giá thấp. Tuy nhiên, chúng có thể được tổng hợp để cải thiện chất lượng.[55](#_bookmark80) Trong các ấn phẩm gần đây, các kết quả trùng khớp thực thể một phần (tức là đánh giá theo mã thông báo) thường được ưu tiên hơn phát hiện hoàn toàn, điều này giúp giảm thiểu tác động của vấn đề này đối với điểm đánh giá cuối cùng.[55,83](#_bookmark80)

Bảng 5. Ví dụ về các báo cáo về những bất đồng giữa các người chú thích trong các ấn phẩm được bao gồm. Vui lòng xem từng ấn phẩm đi kèm để biết thêm chi tiết về chất lượng kho dữ liệu.

|  |  |  |
| --- | --- | --- |
| Công bố | Kiểu | Điểm số hoặc phạm vi từ hạng kém nhất đến tốt nhất |
| [43](#_bookmark57) | Độ chính xác trung bình giữa các trình chú thích | Phạm vi: 0,62 đến 0,70 |
| [48](#_bookmark67) | Tỷ lệ thỏa thuận | 80% |
| [65](#_bookmark101) | *Kappa của* Cohen | 0,84 tổng thể, giảm xuống 0,59 cho hạng tồi tệ nhất |
| [104](#_bookmark136) | *Kappa của* Cohen | Phạm vi: 0,41 đến 0,71 |
| [75](#_bookmark79) | Nhớ lại liên chú thích | Phạm vi: 0,38 đến 0,86 |
| [55](#_bookmark80) | Kappa của Cohen giữa các chuyên gia | Phạm vi: 0,5 đến 0,59 |
| [55](#_bookmark80) | Độ chính xác của công nhân trung bình vĩ mô so với tổng hợp, khả năng thu hồi, F1 (xem ấn phẩm để biết điểm đầy đủ) | Phạm vi: 0,39 đến 0,70 |
| [116](#_bookmark121) (chỉ mô tả việc tạo kho dữ liệu PECODR, không được xem xét) | Thỏa thuận ban đầu giữa các người chú thích | Phạm vi: 85-87% |
| [52](#_bookmark75) | Mức trung bình và phạm vi thỏa thuận | 62%, Phạm vi: 41-71 |
| [58](#_bookmark87) | Câu trung bình được dán nhãn bởi chuyên gia so với sinh viên cho mỗi bản tóm tắt | 1.9 so với 4.2 |
| [58](#_bookmark87) | Chuyên *gia Kappa* của Cohen so với sinh viên | 0.42 |
| [61](#_bookmark93) | Hiệp định; Kappa của Cohen | 86%; 0.76 |
| [38](#_bookmark50) | Biện pháp MASI (Thỏa thuận đo lường về các hạng mục có giá trị đặt) cho mặt hàng/ mức độ lựa chọn; Alpha của Krippendorff cho cấp lớp | MASI 0,6 phạm vi 0,5-0,89; Krippendorf  0,53 cho I, 0,57 cho O, dao động từ 0,06-  0,96 giữa tất cả các lớp |
| [35](#_bookmark42) | F1 nghiêm ngặt so với thoải mái, ở đầu và cuối giai đoạn chú thích | 85,6% so với 93,9% vào cuối; Điểm thư giãn tăng từ 86% khi bắt đầu giai đoạn chú thích lên 93,9% khi kết thúc |
| [36](#_bookmark45) | Kappa của Fleisstrên 47 bản tóm tắt cho kết quả và trên 30 cho trích xuất mối quan hệ | Kết quả 0,81; Quan hệ 0,62-0,72 |
| [63](#_bookmark97) | Điểm B3, MUC, Đo lường F liên kết thực thể bị hạn chế (CEAFe) | B3 0,40; MUC 0,46; và CEAFe 0.42 |
| [51](#_bookmark73) | Kappa cho các thực thể và F1 cho các thực thể phức tạp có các lớp con hoặc quan hệ | Phạm vi Kappa 0,74-0,68; thực thể phức tạp 0,81 |
| [37](#_bookmark47) | Kappa của Cohen về sự thích ứng EBM-NLP của họ so với bộ dữ liệu ban đầu | Từ 0,53 đối với P-0,69 đối với O |

Đối với dữ liệu được dán nhãn tự động hoặc được giám sát từ xa, chất lượng nhãn thường thấp hơn. Điều này chủ yếu là do chú thích không đầy đủ do thiếu tiêu đề, hoặc do sự mơ hồ trong dữ liệu câu, được thảo luận như một phần của miền tiếp theo. [44,57,103](#_bookmark85)

* + 1. *Sự mơ hồ*

Nguồn phổ biến nhất của sự mơ hồ trong các nhãn được mô tả trong các ấn phẩm bao gồm được liên kết với dữ liệu cấp câu được dán nhãn tự động. Ví dụ về điều này là các câu có thể thuộc nhiều loại, ví dụ: những câu nên có cả nhãn 'P' và 'I', hoặc các câu được gán cho lớp 'khác' trong khi chứa thông tin PICO (Tham khảo. [54](#_bookmark78), [95](#_bookmark118), [96](#_bookmark120), trong số những người khác). Sự mơ hồ cũng đã được thảo luận liên quan đến các thuật ngữ can thiệp[76](#_bookmark81) hoặc khi phân biệt giữa các nhánh 'kiểm soát' và 'can thiệp'.[46](#_bookmark64) Khi sử dụng, hoặc ánh xạ đến các khái niệm UMLS, sự mơ hồ đã được thảo luận trong Refs. [41](#_bookmark55), [52](#_bookmark75), [72](#_bookmark72).

Ở cấp độ văn bản, sự mơ hồ xung quanh ý nghĩa của các từ ngữ cụ thể đã được thảo luận như một thách thức, ví dụ, từ 'tập trung' có thể là một thước đo định lượng hoặc một khái niệm tinh thần.[41](#_bookmark55) Các con số cũng được mô tả là khó khăn do sự mơ hồ, bởi vì chúng có thể đề cập đến tổng số người tham gia, số lượng trên mỗi nhánh của một thử nghiệm hoặc chỉ có thể đề cập đến một con số liên quan đến kết quả.[84.113](#_bookmark96) Khi phân loại những người tham gia, thực thể hoặc câu P thường bị quá tải vì nó bao gồm quá nhiều thông tin về các thực thể khác nhau, nhỏ hơn trong đó, chẳng hạn như tuổi, giới tính hoặc chẩn đoán.[89](#_bookmark106)

Sự mơ hồ trong trích xuất mối quan hệ có thể bao gồm các trường hợp trong đó các can thiệp và chất so sánh được phân loại riêng biệt trong một thử nghiệm với nhiều hơn hai nhóm, do đó dẫn đến sự phức tạp ngày càng tăng trong việc nhóm và trích xuất dữ liệu chính xác cho mỗi so sánh riêng biệt.

* + 1. *Các biến thể trong văn bản*

Các biến thể trong ngôn ngữ tự nhiên, từ ngữ hoặc ngữ pháp được xác định là thách thức trong nhiều tài liệu tham khảo xem xét kỹ hơn các văn bản trong kho dữ liệu của chúng. Sự thay đổi như vậy có thể phát sinh khi mô tả các thực thể hoặc câu (ví dụ: Refs. [48](#_bookmark67), [79](#_bookmark86), [97](#_bookmark122)) hoặc có thể phản ánh các đặc điểm cụ thể cho một nguồn dữ liệu, ví dụ: vị trí của các thực thể trong một tạp chí cụ thể.[46](#_bookmark64) Đặc biệt, các phong cách hoặc cách diễn đạt khác nhau đã được ghi nhận như những cảnh báo trong các hệ thống dựa trên quy tắc. [42,48,80](#_bookmark67)

Có sự khác biệt đáng kể trong cách báo cáo một thực thể, ví dụ giữa các loại can thiệp (thuốc, liệu pháp, lộ trình áp dụng)[56](#_bookmark83) hoặc trong các biện pháp đo kết quả.[46](#_bookmark64) Đặc biệt, sự khác biệt về phong cách giữa các bản tóm tắt có cấu trúc và phi cấu trúc[65,78](#_bookmark101) và độ dài mô tả và chi tiết59,79 có thể gây ra kết quả không nhất quán trong việc trích xuất dữ liệu, ví dụ như không phát hiện thông tin chính xác hoặc trích xuất thông tin không mong muốn. Cấu trúc câu phức tạp được đề cập như một cảnh báo đặc biệt đối với các hệ thống dựa trên quy tắc.[80](#_bookmark88) Một ví dụ về một cấu trúc phức tạp là khi nhiều hơn một thực thể được mô tả (ví dụ: Refs. [93](#_bookmark114), [102](#_bookmark132)) hoặc khi các thực thể như 'I' và 'O' được đề cập gần nhau.[57](#_bookmark85) Cuối cùng, các tên khác nhau cho cùng một thực thể trong một bản tóm tắt là một nguồn gốc tiềm ẩn của các vấn đề.[84](#_bookmark96) Khi sử dụng các văn bản không phải tiếng Anh, chẳng hạn như các bài viết tiếng Tây Ban Nha, người ta lưu ý rằng việc dịch bắt buộc các tiêu đề có thể dẫn đến lỗi chính tả và lỗi dịch.[35](#_bookmark42)

Một biến thể phổ biến khác trong văn bản là thông tin ngụ ý. Ví dụ, thay vì nêu liều lượng cụ thể, một văn bản thử nghiệm có thể báo cáo liều lượng '10 hoặc 20 mg', trong đó đơn vị 'mg' được ngụ ý cho số 10, làm cho nó trở thành một thực thể 'liều lượng'. [46,48,90](#_bookmark67)

Thông tin ngụ ý cũng được đề cập là vấn đề trong lĩnh vực trích xuất mối quan hệ, với Nye et al. (2021) [63](#_bookmark97) thảo luận về tầm quan trọng của việc khớp chính xác và giải quyết tên cánh tay can thiệp chỉ ngụ ý can thiệp nào đã được sử dụng. Ví dụ như sử dụng 'Nhóm 1' thay vì đề cập đến tên can thiệp thực tế hoặc ngụ ý tác động trên một nhóm kết quả, chẳng hạn như tất cả các tác dụng phụ.[63](#_bookmark97)

* + 1. *Khả năng thích ứng và so sánh miền*

Do sự khác biệt rộng rãi giữa các lĩnh vực y tế, không có gì đảm bảo rằng hệ thống trích xuất dữ liệu được phát triển trên một bộ dữ liệu sẽ tự động thích ứng để tạo ra kết quả đáng tin cậy trên các bộ dữ liệu khác nhau liên quan đến các lĩnh vực khác. Cấu hình siêu tham số hoặc cơ sở quy tắc được sử dụng để hình thành một hệ thống có thể không truy xuất các kết quả tương đương trong một lĩnh vực y tế khác.[40,68](#_bookmark53) Do đó, điểm số có thể không giống nhau giữa các bộ dữ liệu khác nhau, đặc biệt là đối với các bộ phân loại dựa trên quy tắc,80 khi bộ dữ liệu [nhỏ,35,49](#_bookmark42) khi cấu trúc và phân bố của lớp quan tâm khác nhau,40 hoặc khi các hướng dẫn chú thích khác nhau.[85 Ví](#_bookmark98)  dụ, một mô hình phát hiện kết quả có thể học cách thiên vị về các kết quả thường xuất hiện trong một lĩnh vực nhất định, chẳng hạn như kết quả liên quan đến hóa trị liệu cho tài liệu ung thư hoặc nó có thể ưu tiên phát hiện kết quả thường xuyên hơn trong các văn bản thử nghiệm cũ hơn nếu dữ liệu đào tạo cơ bản cũ hơn hoặc lỗi thời.[73](#_bookmark74) Một cảnh báo khác

được đề cập bởi Refs. [59](#_bookmark89), [85](#_bookmark98) là kích thước của không gian nhãn phải được xem xét khi so sánh điểm số, vì các mô hình chuẩn hóa theo các khái niệm cụ thể hơn là phát hiện các thực thể có xu hướng có độ chính xác, khả năng nhớ và điểm F1 thấp hơn.

Khả năng so sánh giữa các mô hình có thể giảm hơn nữa bằng cách so sánh kết quả giữa các ấn phẩm sử dụng phương pháp đánh giá thoải mái so với nghiêm ngặt để đánh giá dựa trên mã thông [báo,34](#_bookmark41) hoặc các ấn phẩm sử dụng cùng một bộ dữ liệu nhưng có các hạt giống ngẫu nhiên khác nhau để chia dữ liệu đào tạo và thử nghiệm.[33.118](#_bookmark39)

Do đó, một số ấn phẩm thảo luận rằng một số lượng lớn các bộ dữ liệu điểm chuẩn với sự phân tách tiêu chuẩn cho các bộ dữ liệu đào tạo, phát triển và đánh giá và các tập lệnh đánh giá được tiêu chuẩn hóa có thể làm tăng khả năng so sánh giữa các hệ thống đã xuất bản. [46.92.114](#_bookmark112)

* + 1. *Ý nghĩa của tính toán hoặc kiến trúc hệ thống*

Chi phí tính toán và khả năng mở rộng đã được mô tả trong hai ấn phẩm.[53.114](#_bookmark77) Các vấn đề trong hệ thống, ví dụ: mã hóa[97](#_bookmark122) hoặc lỗi trích xuất PDF[75](#_bookmark79) dẫn đến các vấn đề ở hạ nguồn và cuối cùng dẫn đến sự thiên vị, ủng hộ các bài viết từ các nhà xuất bản lớn có dữ liệu được định dạng tốt hơn.[75](#_bookmark79) Tương tự, ngữ pháp và phân tích cú pháp các phần của bài phát biểu và/hoặc lỗi phân đoạn (Tham khảo. [76](#_bookmark81), [80](#_bookmark88), [90](#_bookmark108), trong số những người khác) hoặc các cây phân tích cú pháp bị lỗi[78](#_bookmark84) có thể làm giảm hiệu suất của hệ thống nếu nó dựa vào quyền truy cập vào cấu trúc ngữ pháp chính xác. Về mặt đánh giá hệ thống, xác nhận chéo 10 lần gây ra phương sai cao trong kết quả khi sử dụng các bộ dữ liệu nhỏ như NICTA-PIBOSO,54,85,104 mô tả rằng vấn đề tương tự cần được giải quyết thông qua phân tầng các trường hợp tích cực của mỗi lớp trong các nếp gấp.

* + 1. *Thiếu thông tin trong văn bản hoặc cơ sở kiến thức*

Thông tin trong văn bản có thể không đầy đủ.[Ví](#_bookmark117) dụ, số lượng bệnh nhân trong một nghiên cứu có thể không được báo cáo rõ ràng,76 hoặc các bản tóm tắt thiếu thông tin về thiết kế và phương pháp nghiên cứu có thể xuất hiện, đặc biệt là trong các bản tóm tắt không có cấu trúc và các văn bản thử nghiệm cũ hơn.[91,96](#_bookmark110) Trong một số trường hợp, các bản tóm tắt có thể bị thiếu hoàn toàn. Những vấn đề này đôi khi có thể được giải quyết bằng cách xem xét việc sử dụng toàn bộ văn bản làm đầu vào.[71,87](#_bookmark70)

Khi một mô hình dựa trên các tính năng, ví dụ: MetaMap, thì việc thiếu phạm vi bảo hiểm UMLS sẽ gây ra lỗi.[Điều](#_bookmark72) này cũng áp dụng cho các mô hình như CNN gán các khái niệm cụ thể, trong đó các thực thể vô hình không được xác định trong không gian nhãn đầu ra.[59](#_bookmark89)

Về mặt tóm tắt tự động và trích xuất mối quan hệ, người ta cũng cảnh báo rằng việc dựa vào các bản tóm tắt sẽ dẫn đến độ nhạy thấp của thông tin được truy xuất, vì không phải tất cả các thông tin quan tâm đều có thể được báo cáo đủ chi tiết để cho phép thực hiện các bản tóm tắt hoặc tuyên bố toàn diện về mối quan hệ giữa các can thiệp và kết quả.[60,63](#_bookmark91)

* + 1. *Thực tế và các tác động khác*

Trái ngược với vấn đề thiếu thông tin, quá nhiều thông tin cũng có thể có ý nghĩa thực tế. Ví dụ, thường có nhiều câu với mỗi nhãn, trong đó một câu là 'chìa khóa', ví dụ: mô tả về tiêu chí bao gồm và loại trừ thường kéo dài nhiều câu và đối với một hệ thống trích xuất dữ liệu, có thể khó tìm ra câu nào là câu chính. Vấn đề tương tự cũng áp dụng cho các phương pháp chọn và xếp hạng n câu hàng đầu cho mỗi mục tiêu trích xuất dữ liệu, trong đó một hệ thống có nguy cơ bao gồm quá nhiều hoặc không đủ kết quả tùy thuộc vào số lượng câu được lưu giữ.[46](#_bookmark64)

Khả năng nhớ thấp là một hàm ý thực tế quan trọng,53 đặc biệt là trong các thực thể xuất hiện không thường xuyên trong dữ liệu đào tạo, và do đó không được thể hiện tốt trong quá trình đào tạo của hệ thống phân loại.[48](#_bookmark67) Nói cách khác, một thực thể như 'Chủng tộc' có thể không được dán nhãn thường xuyên là một kho dữ liệu huấn luyện, và bị bỏ sót một cách có hệ thống hoặc phân loại sai khi hệ thống trích xuất dữ liệu được sử dụng trên các văn bản mới. Do đó, sự tham gia của con người là cần thiết,86 và điểm số cần được cải thiện.[41](#_bookmark55) Thật khó để tìm ra tập hợp siêu tham số tốt nhất[106](#_bookmark140) và điều chỉnh độ chính xác và đánh đổi thu hồi để tối đa hóa tiện ích của hệ thống trong khi minh bạch về số lượng điểm dữ liệu có thể bị bỏ sót khi tăng độ chính xác của hệ thống để tiết kiệm công việc cho người đánh giá. [69,95,101](#_bookmark118)

Đối với các nhiệm vụ trích xuất hoặc chuẩn hóa mối quan hệ, lan truyền lỗi được ghi nhận là một vấn đề thực tế trong các mô hình chung.63,67 Để trích xuất các mối quan hệ, trước tiên cần một mô hình để xác định các thực thể, và sau đó một mô hình khác để phân loại các mối quan hệ được áp dụng trong một đường ống. Cả con người và máy móc đều không thể thực hiện trích xuất hoặc ghi nhãn dữ liệu hoàn hảo ngay lập tức,37 và do đó các lỗi được thực hiện trong các bước phân loại trước đó có thể được thực hiện và tích lũy.

Để trích xuất và tóm tắt mối quan hệ, tầm quan trọng của đánh giá định tính trong thế giới thực đã được thảo luận. Điều này là do thiếu sự rõ ràng về mức độ liên quan của các số liệu tóm tắt với tính hữu ích hoặc tính đầy đủ thực tế của một bản tóm tắt và bởi vì những thách thức như mâu thuẫn hoặc phủ định trong và giữa các văn bản thử nghiệm cần được đánh giá trong bối cảnh đánh giá chứ không chỉ là bản thân một thử nghiệm.[61,63](#_bookmark93)

Một cảnh báo thực tế riêng biệt với các mô hình trích xuất quan hệ là sự phụ thuộc dài hơn, tức là khoảng cách lớn hơn giữa các mẩu thông tin nổi bật trong văn bản dẫn đến kết luận. Điều này dẫn đến tăng độ phức tạp của nhiệm vụ và do đó giảm hiệu suất.[99](#_bookmark125)

Trong tuyên bố của họ về các mối quan tâm về đạo đức, DeYoung et al. (2021) [61](#_bookmark93) đề cập rằng các mô hình tóm tắt và mối quan hệ phức tạp này có thể tạo ra các tuyên bố đúng nhưng không chính xác thực tế và có nguy cơ được áp dụng trong thực tế mà không cần thận trọng thêm.

* 1. Khả năng giải thích và giải thích của hệ thống trích xuất dữ liệu

Các mạng nơ-ron hoặc mô hình học máy từ các ấn phẩm có trong bài đánh giá này học cách phân loại và trích xuất dữ liệu bằng cách điều chỉnh trọng số và bằng cách áp dụng các hàm toán học cho các tập hợp trọng số này. Do đó, quá trình ra quyết định đằng sau việc phân loại một câu hoặc một thực thể có thể so sánh với một hộp đen, bởi vì rất khó để hiểu làm thế nào hoặc tại sao mô hình đưa ra dự đoán của nó. Một bình luận gần đây được công bố trên tạp chí Nature đã kêu gọi phân tích và giải thích chuyên sâu hơn về quá trình ra quyết định trong mạng nơ-ron.[Cuối](#_bookmark124) cùng, các xu hướng ẩn trong dữ liệu đào tạo có thể ảnh hưởng đến quá trình ra quyết định của mô hình trích xuất dữ liệu một cách không minh bạch. Nhiều ví dụ được thảo luận trong bình luận có liên quan đến chăm sóc sức khỏe, nhưng trên thực tế, có sự hiểu biết rất hạn chế về những thành kiến cố hữu của chúng mặc dù có ứng dụng rộng rãi của máy học và mạng nơ-ron.[117](#_bookmark124)

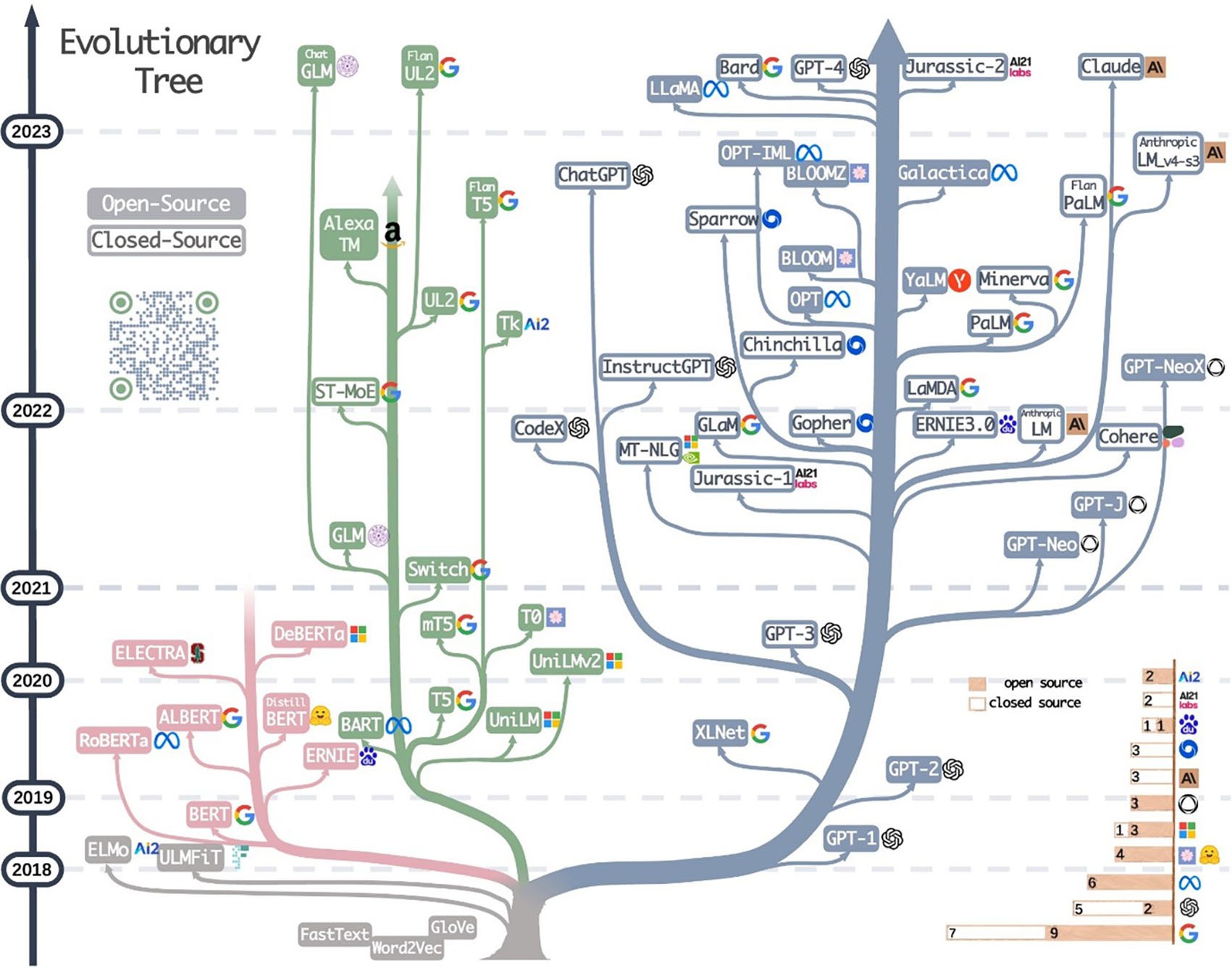
Hiểu sâu hơn về những gì xảy ra giữa việc nhập dữ liệu và điểm dự đoán có thể mang lại lợi ích cho hiệu suất chung của hệ thống, vì nó phát hiện ra những thiếu sót trong quá trình đào tạo. Những thiếu sót này có thể liên quan đến thành phần của dữ liệu đào tạo (ví dụ: đại diện quá mức hoặc đại diện thiếu của các nhóm), kiến trúc hệ thống chung hoặc các xu hướng ngoài ý muốn khác trong dự đoán của hệ thống.[119](#_bookmark128) Một số lượng nhỏ các ấn phẩm được đưa vào đánh giá cơ sở (N = 10) đã thảo luận về các vấn đề liên quan đến các biến ẩn như một phần của phân tích lỗi mở rộng (xem phần 3.5.2). Thành phần của dữ liệu đào tạo và kiểm tra đã được mô tả trong hầu hết các ấn phẩm, nhưng không có ấn phẩm nào đề cập cụ thể đến các vấn đề về khả năng giải thích hoặc khả năng giải thích được tìm thấy.

* 1. Tính khả dụng của kho và các vấn đề về bản quyền Có một số kho dữ liệu được mô tả trong tài liệu, nhiều có nhãn tiêu chuẩn vàng thủ công (xem [Bảng 4](#_bookmark13)). Vẫn còn những ấn phẩm với bộ dữ liệu tùy chỉnh, không được chia sẻ. Lý do có thể cho điều này là lo ngại về bản quyền hoặc liên kết tải xuống trục trặc từ các trang web được đề cập trong các ấn phẩm cũ hơn. Lý tưởng nhất là các thuật toán trích xuất dữ liệu nên được đánh giá trên các bộ dữ liệu khác nhau để phát hiện sự phù hợp quá mức, để kiểm tra cách các hệ thống phản ứng với dữ liệu từ các miền khác nhau và các trình chú thích khác nhau, đồng thời cho phép so sánh các hệ thống một cách đáng tin cậy. Để bổ sung cho bản thảo này, chúng tôi đã thu thập các liên kết đến các bộ dữ liệu trong [Bảng 4](#_bookmark13) và khuyến khích các nhà nghiên cứu chia sẻ các nhãn và văn bản được chú thích tự động hoặc thủ công của họ để các nhà nghiên cứu khác có thể sử dụng chúng để phát triển và đánh giá các hệ thống trích xuất dữ liệu mới.
  2. Những phát triển mới nhất và nghiên cứu sắp tới

Đây là bản cập nhật LSR kịp thời, vì nó có một ngưỡng ngay trước khi xuất hiện một thế hệ công cụ mới: 'Mô hình ngôn ngữ lớn' (LLM) tổng quát, chẳng hạn như ChatGPT từ OpenAI, dựa trên mô hình GPT-3.5 [[1](#_bookmark17)].[Như](#_bookmark129) vậy, nó có thể đánh dấu tình trạng hiện tại của lĩnh vực này vào cuối giai đoạn nghiên cứu đầy thách thức, nơi những hạn chế của các phương pháp học máy gần đây đã được thể hiện rõ ràng và việc tự động hóa trích xuất dữ liệu khá hạn chế.

Sự xuất hiện của các phương pháp dựa trên máy biến áp vào năm 2018 đánh dấu sự thay đổi lớn cuối cùng trong lĩnh vực này, như được ghi nhận bởi LSR này. Các phương pháp của các bài báo bao gồm của chúng tôi hiếm khi tiến bộ vượt quá kiến trúc BERT ban đầu,14 thay đổi chủ yếu chỉ về các bộ dữ liệu được sử dụng trong đào tạo trước. Rất ít người sử dụng các mô hình chỉ khác một chút so với BERT, chẳng hạn như RoBERTa với chiến lược đào tạo trước đã được thay đổi.[121](#_bookmark131) Tuy nhiên, [Hình 13](#_bookmark18) (sao chép từ Yang et al. (2023) [122](#_bookmark133)) cho thấy đã có một lượng lớn nghiên cứu NLP và toàn bộ các phương pháp mới chưa được thử nghiệm để thúc đẩy nhiệm vụ mục tiêu của chúng tôi là trích xuất dữ liệu. Ví dụ, trong báo cáo kỹ thuật GPT-4 mới, OpenAI mô tả hiệu suất tăng lên, khả năng dự đoán và tuân thủ chặt chẽ hơn hành vi dự kiến của mô hình của [họ,123](#_bookmark135) và một số LLM (mã nguồn mở) khác được hiển thị trong [Hình 13](#_bookmark18) có thể có tiềm năng tương tự.

1<https://openai.com/blog/chatgpt> (truy cập lần cuối ngày 22/05/2023).



Hình 13. Cây tiến hóa của các mô hình ngôn ngữ, được sao chép từ Yang et al.[122](#_bookmark133) như được công bố trong bài báo của họ 'Khai thác sức mạnh của LLM trong thực tế: Khảo sát về ChatGPT và hơn thế nữa'.

Các đánh giá ban đầu về LLM cho thấy rằng các mô hình này có thể tạo ra một bước thay đổi về cả độ chính xác và hiệu quả của việc trích xuất thông tin tự động, đồng thời giảm nhu cầu về dữ liệu đào tạo được dán nhãn đắt tiền: một bản in trước của Shaib và cộng sự.[124](#_bookmark137) mô tả một bộ dữ liệu mới [[2](#_bookmark19)] và đánh giá tóm tắt RCT do GPT-3 tạo ra;[124](#_bookmark137) Wadhwa, DeYoung, và cộng sự.[125](#_bookmark139) sử dụng bộ dữ liệu Suy luận bằng chứng và chú thích của bộ ba kết quả can thiệp-so sánh-kết quả RCT để đào tạo và đánh giá BRAN, DyGIE++, ELI, BART, T5-base và một số mô hình FLAN trong bản in trước;[125](#_bookmark139) và trong một bản in trước riêng biệt Wadhwa, Amir, et al.[126 người](#_bookmark141)  đã sử dụng các mô hình Flan-T5 và GPT-3 để trích xuất và dự đoán mối quan hệ giữa thuốc và các tác dụng phụ.[126](#_bookmark141) Trong tương lai gần, chúng tôi hy vọng số lượng nghiên cứu trong đánh giá này sẽ tăng lên, khi nhiều đánh giá về LLM chuyển sang tài liệu in sẵn hoặc đã xuất bản.

* + 1. *Hạn chế của bài đánh giá sống này*

Đánh giá này tập trung vào việc trích xuất dữ liệu từ các báo cáo về thử nghiệm lâm sàng và nghiên cứu dịch tễ học. Điều này chủ yếu bao gồm trích xuất dữ liệu từ các báo cáo của các thử nghiệm ngẫu nhiên có đối chứng, trong đó can thiệp và các chất so sánh thường được trích xuất cùng nhau và chỉ một phần rất nhỏ của bằng chứng đề cập đến các loại nghiên cứu quan trọng khác (ví dụ: nghiên cứu về độ chính xác chẩn đoán). Trong quá trình sàng lọc, chúng tôi đã loại trừ tất cả các ấn phẩm liên quan đến dữ liệu lâm sàng (chẳng hạn như hồ sơ sức khỏe điện tử) và các ấn phẩm trích xuất dữ liệu bệnh tật, dân số hoặc can thiệp từ nghiên cứu di truyền và sinh học. Có rất nhiều bằng chứng và dữ liệu đào tạo và đánh giá tiềm năng trong các ấn phẩm này, nhưng không khả thi để đưa chúng vào đánh giá sống.

5. Kết luận

LSR này trình bày tổng quan về tài liệu trích xuất dữ liệu được quan tâm đối với các loại đánh giá có hệ thống khác nhau. Chúng tôi bao gồm một cơ sở bằng chứng rộng rãi của các ấn phẩm mô tả trích xuất dữ liệu cho các đánh giá hệ thống can thiệp (tập trung vào các nhóm P, IC và O và dữ liệu RCT), và một số lượng rất nhỏ các ấn phẩm trích xuất dịch tễ học và chẩn đoán

2<https://github.com/cshaib/summarizing-medical-evidence> (truy cập lần cuối ngày 22/05/2022).

dữ liệu chính xác. Trong bản cập nhật LSR, chúng tôi đã xác định các xu hướng nghiên cứu như sự xuất hiện của các phương pháp trích xuất mối quan hệ, sự thống trị hiện tại của mạng nơ-ron biến áp hoặc tăng tính khả dụng của mã và bộ dữ liệu trong giai đoạn 2020-2022. Tuy nhiên, số lượng các công cụ có thể truy cập có thể giúp người đánh giá có hệ thống trích xuất dữ liệu vẫn còn rất thấp. Hiện tại, chỉ có khoảng một trong mười ấn phẩm được liên kết với một công cụ có thể sử dụng hoặc mô tả một triển khai đang diễn ra.

Các thuật toán trích xuất dữ liệu và đặc điểm của dữ liệu mà chúng được đào tạo và đánh giá đã được báo cáo tốt. Khoảng ba trong số mười ấn phẩm đã cung cấp bộ dữ liệu của họ cho công chúng và hơn một nửa số ấn phẩm bao gồm báo cáo đào tạo hoặc đánh giá trên các bộ dữ liệu này. Thật không may, việc sử dụng các tập lệnh đánh giá khác nhau, các phương pháp khác nhau để tính trung bình kết quả hoặc điều chỉnh tùy chỉnh cho bộ dữ liệu vẫn gây khó khăn cho việc đưa ra kết luận về hệ thống nào hoạt động tốt nhất. Ngoài ra, trích xuất dữ liệu là một nhiệm vụ rất khó khăn. Nó thường yêu cầu giải quyết xung đột giữa các chuyên gia đánh giá có hệ thống khi được thực hiện thủ công và do đó tạo ra các vấn đề khi tạo ra các tiêu chuẩn vàng được sử dụng để đào tạo và đánh giá các thuật toán trong bài đánh giá này.

Chúng tôi đã liệt kê nhiều thách thức đang diễn ra trong lĩnh vực trích xuất dữ liệu để tự động hóa đánh giá hệ thống (bán nửa), bao gồm sự mơ hồ trong văn bản thử nghiệm lâm sàng, dữ liệu không đầy đủ và dữ liệu chưa từng thấy trước đây. Với bài đánh giá sống này, chúng tôi mong muốn xem xét các tài liệu liên tục khi nó có sẵn. Do đó, phiên bản đánh giá mới nhất, cùng với số lượng bản tóm tắt được sàng lọc và đưa vào sau khi xuất bản lần đánh giá này, có sẵn trên trang web của chúng tôi.

Tính khả dụng của dữ liệu

Dữ liệu cơ bản

Harvard Dataverse: Phụ lục để xem xét cơ sở. <https://doi.org/10.7910/DVN/LNGCOQ>.[127](#_bookmark142) Dự án này chứa các dữ liệu cơ bản sau:

* Appendix\_A.zip (cơ sở dữ liệu đầy đủ với tất cả trích xuất dữ liệu và các trường khác cho dữ liệu đánh giá cơ sở)
* Phụ lục B.docx (thông tin thêm về các ấn phẩm bị loại trừ)
* Appendix\_C.zip (mã, trọng số, dữ liệu, điểm số của bộ phân loại trừu tượng cho nội dung Web of Science)
* Appendix\_D.zip (cơ sở dữ liệu đầy đủ với tất cả trích xuất dữ liệu và các trường khác để cập nhật LSR)
* Supplementary\_key\_items.docx (tổng quan về các mục được trích xuất cho mỗi nghiên cứu bao gồm)
* Bảng 1. csv và bảng 1\_long.csv (Bảng A1 ở định dạng csv, phiên bản dài bao gồm thêm dữ liệu)
* bảng 1\_long\_updated.csv (cập nhật LSR cho Bảng A1 ở định dạng csv, phiên bản dài bao gồm dữ liệu bổ sung)
* included.ris và background.ris (tài liệu tham khảo từ đánh giá cơ sở)

Harvard Dataverse: Bộ dữ liệu có sẵn để tự động hóa SR. <https://doi.org/10.7910/DVN/0XTV25>.[128](#_bookmark144) Dự án này chứa các dữ liệu cơ bản sau:

* Bộ dữ liệu được chia sẻ bởi các tác giả của các ấn phẩm bao gồm

Dữ liệu có sẵn theo các điều khoản của từ bỏ dữ liệu Creative Commons Zero "Không được bảo lưu" (CC0 1.0 Public domain dedication).

Dữ liệu mở rộng

Khung khoa học mở: Phương pháp trích xuất dữ liệu để đánh giá có hệ thống (bán tự động hóa: Giao thức đánh giá sống). <https://doi.org/10.17605/OSF.IO/ECB3T>.[15](#_bookmark44)

Dự án này chứa các dữ liệu mở rộng sau:

* Đánh giá giao thức
* Additional\_Fields.docx (tổng quan về các trường dữ liệu quan tâm để khai thác văn bản trong thử nghiệm lâm sàng)
* Search.docx (thông tin bổ sung về các tìm kiếm, bao gồm cả chiến lược tìm kiếm đầy đủ)
* Danh sách kiểm tra PRISMA P cho 'Phương pháp trích xuất dữ liệu để đánh giá hệ thống (bán tự động hóa): Một giao thức đánh giá sống.'

Dữ liệu có sẵn theo các điều khoản của [giấy phép quốc tế Creative Commons Attribution 4.0](https://creativecommons.org/licenses/by/4.0/)  (CC-BY 4.0).

Nguyên tắc báo cáo

Harvard Dataverse: Danh sách kiểm tra PRISMA cho 'Phương pháp trích xuất dữ liệu để đánh giá hệ thống (bán tự động hóa: Đánh giá có hệ thống sống) <https://doi.org/10.7910/DVN/LNGCOQ>.[127](#_bookmark142)

Dữ liệu có sẵn theo các điều khoản của [Creative Commons Zero "Không được bảo lưu" miễn trừ dữ liệu](http://creativecommons.org/publicdomain/zero/1.0/) (CC0 1.0 Hiến tặng phạm vi công cộng).

Tính khả dụng của phần mềm

Phiên bản phát triển của phần mềm để tìm kiếm tự động có sẵn từ Github: [https://github.com/](https://github.com/mcguinlu/COVID_suicide_living)  [mcguinlu/COVID\_suicide\_living](https://github.com/mcguinlu/COVID_suicide_living).

Mã nguồn lưu trữ tại thời điểm xuất bản: <http://doi.org/10.5281/zenodo.3871366>.[17](#_bookmark48) Giấy phép: MIT

Đóng góp của tác giả

LS: Khái niệm, Điều tra, Phương pháp luận, Phần mềm, Trực quan, Viết - Chuẩn bị bản thảo gốc ANFM: Quản lý dữ liệu, Điều tra, Viết - Đánh giá & Chỉnh sửa

RE: Quản lý dữ liệu, điều tra, viết - xem xét và chỉnh sửa

BKO: Khái niệm, Điều tra, Phương pháp luận, Phần mềm, Viết – Đánh giá & Chỉnh sửa JT: Khái niệm hóa, Điều tra, Phương pháp luận, Viết – Đánh giá & Chỉnh sửa

JPTH: Khái niệm, Thu thập tài trợ, Điều tra, Phương pháp, Viết – Đánh giá & Chỉnh sửa

Lời cảm ơn

Chúng tôi cảm ơn Luke McGuinness vì những đóng góp của anh ấy cho đánh giá cơ sở, đặc biệt là lập trình ứng dụng web LSR, sàng lọc, giải quyết xung đột và phản hồi của anh ấy đối với bản thảo đánh giá cơ sở.

Chúng tôi cảm ơn Patrick O'Driscoll vì sự giúp đỡ của anh ấy trong việc kiểm tra dữ liệu, số lượng và từ ngữ trong bản thảo và phụ lục.

Chúng tôi cảm ơn Sarah Dawson đã phát triển và đánh giá chiến lược tìm kiếm, và cung cấp lời khuyên về cơ sở dữ liệu để tìm kiếm cho đánh giá này. Cũng rất cảm ơn Alexandra McAleenan và Vincent Cheng vì đã cung cấp phản hồi có giá trị về đánh giá này và giao thức của nó.

Tham khảo

1. Higgins J, *và cộng sự*: Sổ tay Cochrane về Đánh giá Hệ thống về Can thiệp phiên bản 6.1 (cập nhật Tháng Chín 2020). 2020: Cochrane.
2. Fukumi Tsunoda D, Conceição Moreira P, Ribeiro Guimarães A: Học máy và đánh giá có hệ thống tài liệu tự động: đánh giá có hệ thống. *Tạp chí Công nghệ và Xã hội.* năm 2020; 16(45).
3. Jonnalagadda SR, Goyal P, Huffman MD: Tự động trích xuất dữ liệu trong các đánh giá có hệ thống: một đánh giá có hệ thống.

*Đánh giá có hệ thống.* năm 2015; 4 (1): 78.

[Tóm tắt PubMed](http://www.ncbi.nlm.nih.gov/pubmed/26073888) [|Nhà xuất bản Toàn văn](https://doi.org/10.1186/s13643-015-0066-7) [|Toàn văn miễn phí](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4514954)

1. O*'*Mara-Eves A, *và cộng sự*: Sử dụng khai thác văn bản để xác định nghiên cứu trong các đánh giá có hệ thống: đánh giá có hệ thống về các phương pháp tiếp cận hiện tại. *Syst Rev.* 2015; 4 (1): 5.

[Tóm tắt PubMed](http://www.ncbi.nlm.nih.gov/pubmed/25588314) [|Nhà xuất bản Toàn văn](https://doi.org/10.1186/2046-4053-4-5) [|Toàn văn miễn phí](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4320539)

1. Tsafnat G, *và cộng sự*: Công nghệ tự động hóa đánh giá có hệ thống.

*Syst Rev.* 2014; 3 (1): 74.

[Nhà xuất bản toàn văn](https://doi.org/10.1186/2046-4053-3-74)

1. Beller E, *và cộng sự*: Tiến bộ với việc tự động hóa các đánh giá có hệ thống: các nguyên tắc của Hợp tác Quốc tế về Tự động hóa Đánh giá Hệ thống (ICASR). *Syst. Rev.* 2018; 7 (1): 77.
2. Marshall IJ, Wallace BC: Hướng tới tự động hóa đánh giá có hệ thống: hướng dẫn thực tế để sử dụng các công cụ học máy trong tổng hợp nghiên cứu. *Syst Rev.* 2019; 8 (1): 163.
3. Cierco Jimenez R, Lee T, Rosillo N, *và cộng sự*: Các công cụ tính toán học máy để hỗ trợ thực hiện các đánh giá có hệ thống: Đánh giá bản đồ. *BMC Med Res Methodol.* năm 2022; 22 (1): 322.

[Tóm tắt PubMed](http://www.ncbi.nlm.nih.gov/pubmed/36522637)[|Nhà xuất bản đầy đủ Nhắn tin](https://doi.org/10.1186/s12874-022-01805-4)[|Miễn phí đầy đủ Nhắn tin](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC9756658)

1. Khalil H, Ameen D, Zarnegar A: Các công cụ hỗ trợ tự động hóa các đánh giá có hệ thống: đánh giá phạm vi. *J Clin Epidemiol.* năm 2022; 144: 22*–*42.

[Tóm tắt PubMed](http://www.ncbi.nlm.nih.gov/pubmed/34896236) [|Nhà xuất bản toàn văn](https://doi.org/10.1016/j.jclinepi.2021.12.005)

1. Ruiz RL, Duffy VG: Tự động hóa trong Đánh giá hệ thống chăm sóc sức khỏe. Trong Stephanidis C, Duffy VG, Krömker H, *và cộng sự*. *HCI International 2021 - Bài báo đột phá muộn: Ứng dụng HCI trong Y tế, Giao thông vận tải và Công nghiệp.* Chăm. Năm 2021.
2. Sundaram G, Berleant D: Tự động hóa Đánh giá Tài liệu Có Hệ thống với Xử lý Ngôn ngữ Tự nhiên và Khai thác Văn bản: Đánh giá Tài liệu Có Hệ thống. *bản in trước arXiv arXiv.: 2211.15397.* Năm 2022.
3. Zhang T, Huang Z, Wang Y, *và cộng sự*: Trích xuất thông tin từ dữ liệu văn bản về y học cổ truyền Trung Quốc: Đánh giá về nhiệm vụ, thách thức và phương pháp từ năm 2010 đến năm 2021. *Evid dựa trên bổ sung Alternat Med.* năm 2022; 2022: 1679589.

[Tóm tắt PubMed](http://www.ncbi.nlm.nih.gov/pubmed/35600940) [|Nhà xuất bản Toàn văn](https://doi.org/10.1155/2022/1679589) [|Toàn văn miễn phí](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC9122692)

1. Schmidt L, Sinyor M, Webb RT, *và cộng sự*.: Đánh giá tường thuật về các công cụ và đổi mới gần đây hướng tới tự động hóa các đánh giá có hệ thống sống và tổng hợp bằng chứng. *Tạp chí Bằng chứng, Giáo dục Thường xuyên và chất lượng trong chăm sóc sức khỏe.* 2023; S1865-9217(23)00140-X. [Nhà xuất bản toàn văn](https://doi.org/10.1016/j.zefq.2023.06.007)
2. Devlin J, Chang MW, Lee K, và *cộng sự*: Bert: Đào tạo trước các máy biến áp hai chiều sâu để hiểu ngôn ngữ. *bản in trước arXiv arXiv.* 2018; 1810: 04805.
3. Schmidt L, *và cộng sự*: Phương pháp trích xuất dữ liệu để đánh giá hệ thống (bán tự động hóa): Một giao thức đánh giá sống. *F1000Res.* năm 2020; 9 (210).

[Tóm tắt PubMed](http://www.ncbi.nlm.nih.gov/pubmed/32724560) [|Nhà xuất bản Toàn văn](https://doi.org/10.12688/f1000research.22781.2) [|Toàn văn miễn phí](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC7338918)

1. McGuinness LA, Schmidt L: medrxivr: Truy cập và tìm kiếm dữ liệu medRxiv và bioRxivpreprint trong R. *JOSS.* Năm 2020.

[Nhà xuất bản toàn văn](https://doi.org/10.21105/joss.02651)

1. McGuinness LA, Schmidt L: mcguinlu / COVID\_suicide\_living: Bản phát hành ban đầu (Phiên bản v1.0.0). *Zenodo.* 2020, ngày 1 tháng 6. [Nhà xuất bản toàn văn](https://doi.org/10.5281/zenodo.3871366)
2. John A, *và cộng sự*: Tác động của đại dịch COVID-19 đối với hành vi tự làm hại bản thân và tự tử: giao thức cho một đánh giá có hệ thống sống [phiên bản 1; đánh giá ngang hàng:

1 đã được phê duyệt, 1 được chấp thuận với đặt trước]. *F1000Res.* năm 2020;

9 (644).

[Tóm tắt PubMed](http://www.ncbi.nlm.nih.gov/pubmed/33604025) [|Nhà xuất bản Toàn văn](https://doi.org/10.12688/f1000research.25522.1) [|Toàn văn miễn phí](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC7871358)

1. Olorisade BK, Brereton P, Andras P: Khả năng tái tạo của các nghiên cứu về khai thác văn bản để sàng lọc trích dẫn trong các đánh giá có hệ thống : Đánh giá và danh sách kiểm tra. *J Biomed Thông báo.* năm 2017; 73: 1*–*13.

[Tóm tắt PubMed](http://www.ncbi.nlm.nih.gov/pubmed/28711679) [|Nhà xuất bản toàn văn](https://doi.org/10.1016/j.jbi.2017.07.010)

1. Haddaway NR: livingPRISMA\_flow *: Gói R và ShinyApp để tạo sơ đồ dòng chảy kiểu PRISMA cho các đánh giá có hệ thống sống (Phiên bản 0.0.1).* Trong Zenodo. xxx 2021.

[Toàn văn miễn phí](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC9828146)

1. Kahale LA, Elkhoury R, El Mikati I, *và cộng sự*: Sơ đồ dòng chảy PRISMA 2020 phù hợp cho các đánh giá có hệ thống sống: một cuộc khảo sát phương pháp luận và một đề xuất. *F1000Res.* năm 2021; 10: 192.

[Tóm tắt PubMed](http://www.ncbi.nlm.nih.gov/pubmed/35136567)[|Nhà xuất bản đầy đủ Nhắn tin](https://doi.org/10.12688/f1000research.51723.3)[|Miễn phí đầy đủ Nhắn tin](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC8804909)

1. Page MJ, McKenzie JE, Bossuyt PM, và *cộng sự*: Tuyên bố PRISMA 2020: hướng dẫn cập nhật để báo cáo các đánh giá có hệ thống. *BMJ.* năm 2021; 372, n71.

[Tóm tắt PubMed](http://www.ncbi.nlm.nih.gov/pubmed/33782057)[|Nhà xuất bản đầy đủ Nhắn tin](https://doi.org/10.1136/bmj.n71)[|Miễn phí đầy đủ Nhắn tin](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC8005924)

1. Norman C, Leeflang M, Névéol A: Trích xuất và tổng hợp dữ liệu trong các đánh giá có hệ thống về độ chính xác của xét nghiệm chẩn đoán: Một kho dữ liệu để tự động hóa và đánh giá quy trình. *AMIA Annu Symp Proc.* 2018; 2018: 817*–*826.

[Tóm tắt PubMed](http://www.ncbi.nlm.nih.gov/pubmed/30815124) [|Toàn văn miễn phí](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC6371350)

1. Millard LA, Flach PA, Higgins JP: Học máy để hỗ trợ đánh giá rủi ro thiên vị trong các đánh giá có hệ thống. *Int J Epidemiol.* 2016; 45 (1): 266*–*277.

[Tóm tắt PubMed](http://www.ncbi.nlm.nih.gov/pubmed/26659355) [|Nhà xuất bản Toàn văn](https://doi.org/10.1093/ije/dyv306) [|Toàn văn miễn phí](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4795562)

1. Marshall IJ, Kuiper J, Wallace B: RobotReviewer: đánh giá một hệ thống để tự động đánh giá sự thiên vị trong các thử nghiệm lâm sàng. *J Am Med Thông báo Assoc.* 2016; 23 (1): 193*–*201.

[Tóm tắt PubMed](http://www.ncbi.nlm.nih.gov/pubmed/26104742) [|Nhà xuất bản Toàn văn](https://doi.org/10.1093/jamia/ocv044) [|Toàn văn miễn phí](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4713900)

1. Boudin F, Nie JY, Dawes M: Truy xuất thông tin lâm sàng bằng cách sử dụng tài liệu và cấu trúc PICO. *Assoc. Compu. Nhà ngôn ngữ học.* 2010: 822*–*830.
2. Luo Z, *và cộng sự*: Trích xuất các ràng buộc thời gian từ các tiêu chí đủ điều kiện nghiên cứu lâm sàng bằng cách sử dụng các trường ngẫu nhiên có điều kiện. *AMIA Annu Symp Proc.* 2011; 2011: 843*–*852.

[Tóm tắt PubMed](http://www.ncbi.nlm.nih.gov/pubmed/22195142) [|Toàn văn miễn phí](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3243135)

1. Rathbone J, *et al*.: Sàng lọc trích dẫn nhanh bằng cách sử dụng sàng lọc chỉ tiêu đề dựa trên PICo để xác định các nghiên cứu trong tìm kiếm phạm vi và đánh giá nhanh. *Syst Rev.* 2017; 6 (1): 233.

[Tóm tắt PubMed](http://www.ncbi.nlm.nih.gov/pubmed/29178925) [|Nhà xuất bản Toàn văn](https://doi.org/10.1186/s13643-017-0629-x) [|Toàn văn miễn phí](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5702220)

1. Malheiros V, *và cộng sự*: Phương pháp khai thác văn bản trực quan cho các đánh giá có hệ thống. trong: *Hội nghị chuyên đề quốc tế đầu tiên về kỹ thuật và đo lường phần mềm thực nghiệm (ESEM 2007).* Năm 2007.
2. Fabbri S, *và cộng sự*: Sử dụng trực quan hóa thông tin và khai thác văn bản để tạo điều kiện thuận lợi cho việc tiến hành các đánh giá tài liệu có hệ thống. trong Hệ *thống thông tin doanh nghiệp.* 2013.Berlin, Heidelberg: Springer Berlin Heidelberg.
3. Beltagy I, Lo K, Cohan A: SciBERT: Một mô hình ngôn ngữ được đào tạo trước cho văn bản khoa học. *bản in trước arXiv arXiv: 1903.10676.* Năm 2019.
4. Al-Hussaini I, An DN, Lee AJ, *và cộng sự*: CCS Explorer: Dự đoán mức độ liên quan, Tóm tắt trích xuất và Nhận dạng thực thể được đặt tên từ các nghiên cứu thuần tập lâm sàng. *Hội nghị Quốc tế IEEE 2022 về Dữ liệu lớn (Big Data).* 2022, 17-20 Tháng Mười Hai 2022.

[Toàn văn miễn phí](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC7647812)

1. Tsubota T, Bollegala D, Zhao Y, *và cộng sự*: Cải thiện phát hiện thông tin can thiệp để sàng lọc tài liệu lâm sàng tự động trong quá trình xem xét có hệ thống. *J Biomed Thông báo.* năm 2022; 134: 104185.

[Tóm tắt PubMed](http://www.ncbi.nlm.nih.gov/pubmed/36038066) |Nhà xuất bản toàn văn

1. Abaho M, Bollegala D, Williamson PR, *và cộng sự*: Đánh giá các biểu diễn theo ngữ cảnh trong việc phát hiện các cụm từ kết quả trong các thử nghiệm lâm sàng. *bản in trước arXiv arXiv: 2203.03547.* Năm 2022.
2. Campillos-Llanos L, Valverde-Mateos A, Capllonch-Carrión A, *và cộng sự*: Một kho thử nghiệm lâm sàng được chú thích với các thực thể UMLS để tăng cường khả năng tiếp cận với y học dựa trên bằng chứng. *BMC Med thông báo Decis Mak.* năm 2021; 21 (1): 69.

[Tóm tắt PubMed](http://www.ncbi.nlm.nih.gov/pubmed/33618727) [|Nhà xuất bản Toàn văn](https://doi.org/10.1186/s12911-021-01395-z) [|Toàn văn miễn phí](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC7898014)

1. Mayer T, Marro S, Cabrio E, *và cộng sự*: Tăng cường y học dựa trên bằng chứng với phân tích tranh luận ngôn ngữ tự nhiên của các thử nghiệm lâm sàng. *Artif Intell Med.* năm 2021; 118: 102098.

[Tóm tắt PubMed](http://www.ncbi.nlm.nih.gov/pubmed/34412851) |Nhà xuất bản toàn văn

1. Dhrangadhariya A, Müller H: PICO không quá yếu: tận dụng sự giám sát yếu cho những người tham gia, can thiệp và nhận biết kết quả để tự động hóa đánh giá có hệ thống. *JAMIA Mở cửa.* 2023; 6 (1): OOAC107.

[Tóm tắt PubMed](http://www.ncbi.nlm.nih.gov/pubmed/36632329) [|Nhà xuất bản Toàn văn](https://doi.org/10.1093/jamiaopen/ooac107) [|Toàn văn miễn phí](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC9828146)

1. Kilicoglu H, Rosemblat G, Hoàng L, *và cộng sự*: Hướng tới đánh giá các ấn phẩm thử nghiệm lâm sàng để báo cáo minh bạch. *J Biomed Thông báo.* năm 2021; 116, 103717.

[Tóm tắt PubMed](http://www.ncbi.nlm.nih.gov/pubmed/33647518) |Nhà xuất bản Toàn văn [|Toàn văn miễn phí](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC8112250)

1. Zhang T, Yu Y, Mei J, và *cộng sự*: Mở khóa sức mạnh của chiết xuất pico sâu: Nhận dạng người y tế từng bước. *bản in trước arXiv arXiv.:* 2005.06601*.* Năm 2020.
2. Chabou S, Iglewski M: Sự kết hợp của trường ngẫu nhiên có điều kiện với phương pháp dựa trên quy tắc trong việc trích xuất các phần tử PICO. *BMC Med thông báo Decis Mak.* 2018; 18: 14.

[Tóm tắt PubMed](http://www.ncbi.nlm.nih.gov/pubmed/30509272) [|Nhà xuất bản Toàn văn](https://doi.org/10.1186/s12911-018-0699-2) [|Toàn văn miễn phí](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC6278016)

1. Lucic A, Blake CL: Cải thiện phát hiện điểm cuối để hỗ trợ đánh giá có hệ thống tự động. *AMIA Annu Symp Proc.* 2016; Năm 2016:

trang 1900*–*1909.

[Tóm tắt PubMed](http://www.ncbi.nlm.nih.gov/pubmed/28269949) [|Toàn văn miễn phí](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5333237)

1. Baladron C, *và cộng sự*: Công cụ lọc kết quả tìm kiếm PubMed theo kích thước mẫu. *J Am Med Thông báo Assoc.* 2018; 25 (7): 774*–*779. [Tóm tắt PubMed](http://www.ncbi.nlm.nih.gov/pubmed/29409012) [|Nhà xuất bản Toàn văn](https://doi.org/10.1093/jamia/ocx155) [|Toàn văn miễn phí](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC7647020)
2. Brassey J, Price C, Edwards J, *và cộng sự*: Phát triển một công cụ tổng hợp bằng chứng hoàn toàn tự động để xác định, đánh giá và đối chiếu bằng chứng. *BMJ Evid dựa trên Med.* năm 2021; 26 (1): 24*–*27.

[Tóm tắt PubMed](http://www.ncbi.nlm.nih.gov/pubmed/31467247) [|Nhà xuất bản toàn văn](https://doi.org/10.1136/bmjebm-2018-111126)

1. Wallace BC, *và cộng sự*: Trích xuất các câu PICO từ các báo cáo thử nghiệm lâm sàng bằng cách sử dụng Giám sát từ xa có giám sát. *J Mach Học Res.* 2016; 17.

[Tóm tắt PubMed](http://www.ncbi.nlm.nih.gov/pubmed/27746703) [|Toàn văn miễn phí](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5065023)

1. Singh G, Sabet Z, Shawe-Taylor J, và *cộng sự*: Xây dựng dữ liệu nhân tạo để tinh chỉnh cho việc gắn thẻ văn bản y sinh có tài nguyên thấp với các ứng dụng trong chú thích PICO. Trong Shaban-Nejad A, Michalowski M, Buckeridge DL, biên tập viên. *AI có thể giải thích trong chăm sóc sức khỏe và y học: Xây dựng văn hóa minh bạch và trách nhiệm.* Nhà xuất bản Quốc tế Springer; trang 131*–*145. Năm 2021.

[Nhà xuất bản toàn văn](https://doi.org/10.1007/978-3-030-53352-6_12)

1. Kiritchenko S, *et al*.: ExaCT: tự động trích xuất các đặc điểm thử nghiệm lâm sàng từ các ấn phẩm tạp chí. *BMC Med thông báo Decis Mak.* 2010; 10: 17. BMC Med thông báo Decis Mak.
2. Fiszman M, *và cộng sự*: Giải thích các cấu trúc so sánh trong văn bản y sinh. 2007: 137*–*144.
3. Karystianis G, Buchan I, Nenadic G: Khai thác đặc điểm của các nghiên cứu dịch tễ học từ Medline: một nghiên cứu điển hình về béo phì. *J Ngữ nghĩa sinh học.* 2014; 5: 11.

[Tóm tắt PubMed](http://www.ncbi.nlm.nih.gov/pubmed/24949194) [|Nhà xuất bản Toàn văn](https://doi.org/10.1186/2041-1480-5-22) [|Toàn văn miễn phí](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4062908)

1. Karystianis G, *và cộng sự*: Đánh giá phương pháp dựa trên quy tắc để phân loại tài liệu dịch tễ học hướng tới tự động hóa các đánh giá có hệ thống. *J Biomed Thông báo.* năm 2017; 70: 27*–*34.

[Tóm tắt PubMed](http://www.ncbi.nlm.nih.gov/pubmed/28455150) [|Nhà xuất bản toàn văn](https://doi.org/10.1016/j.jbi.2017.04.004)

1. Whitton J, Hunter A: Lập bảng tự động kết quả thử nghiệm lâm sàng: Một thực thể chung và phương pháp trích xuất mối quan hệ với các biểu diễn ngôn ngữ dựa trên máy biến áp. *bản in trước arXiv arXiv.:* 2112.05596*.* Năm 2021.
2. Sanchez-Graillet O, Witte C, Grimm F, *và cộng sự*: Một kho dữ liệu có chú thích của các ấn phẩm thử nghiệm lâm sàng hỗ trợ trích xuất thông tin quan hệ dựa trên lược đồ.*J. Sinh học. Ngữ nghĩa học.* năm 2022; 13 (1): 14. [Tóm tắt PubMed](http://www.ncbi.nlm.nih.gov/pubmed/35606797) [|Nhà xuất bản Toàn văn](https://doi.org/10.1186/s13326-022-00271-7) [|Toàn văn miễn phí](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC9128209)
3. Kim S, *và cộng sự*: Tự động phân loại các câu để hỗ trợ Y học dựa trên bằng chứng. *Thông tin sinh học BMC.* 2011; 12 (S-2): S5. [Tóm tắt PubMed](http://www.ncbi.nlm.nih.gov/pubmed/21489224) [|Nhà xuất bản Toàn văn](https://doi.org/10.1186/1471-2105-12-S2-S5) [|Toàn văn miễn phí](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3073185)
4. Verbeke M, *và cộng sự*: Phương pháp tiếp cận học tập quan hệ thống kê để xác định các danh mục y học dựa trên bằng chứng. Năm 2012.

trang 579*–*589.

1. Jin D, Szolovits P: Nâng cao phát hiện phần tử PICO trong văn bản y sinh thông qua mạng nơ-ron sâu. Thông *tin sinh học.* năm 2020; 36 (12): 3856*–*3862.

[Tóm tắt PubMed](http://www.ncbi.nlm.nih.gov/pubmed/32311009) [|Nhà xuất bản toàn văn](https://doi.org/10.1093/bioinformatics/btaa256)

1. Nye B, *và cộng sự*: Một kho dữ liệu với chú thích đa cấp về bệnh nhân, can thiệp và kết quả để hỗ trợ xử lý ngôn ngữ cho tài liệu y khoa. *Proc Conf Assoc Comput Linguist Meet.* 2018; 2018: 197*–*207.

[Tóm tắt PubMed](http://www.ncbi.nlm.nih.gov/pubmed/30305770) [|Toàn văn miễn phí](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC6174533)

1. de Bruijn B, *và cộng sự*: Trích xuất thông tin tự động về các yếu tố thiết kế thử nghiệm chính từ các ấn phẩm thử nghiệm lâm sàng. *AMIA Annu Symp Proc.* 2008; tr. 141*–*5.

[Tóm tắt PubMed](http://www.ncbi.nlm.nih.gov/pubmed/18999067) [|Toàn văn miễn phí](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2655966)

1. Boudin F, Shi L, Nie JY: Cải thiện truy xuất thông tin y tế với phát hiện phần tử PICO. 2010. trang 50*–*61.

[Nhà xuất bản toàn văn](https://doi.org/10.1007/978-3-642-12275-0_8)

1. Demner-Fushman D, *và cộng sự*: Bài nghiên cứu: Tự động xác định thông tin kết quả sức khỏe trong hồ sơ MEDLINE.

*J. Am. Hiệp hội Tin học Y tế.* 2006; 13 (1): 52*–*60.

[Tóm tắt PubMed](http://www.ncbi.nlm.nih.gov/pubmed/16221937) [|Nhà xuất bản Toàn văn](https://doi.org/10.1197/jamia.M1911) [|Toàn văn miễn phí](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1380197)

1. Singh G, *và cộng sự*: Kiến trúc bộ chọn ứng cử viên thần kinh cho chú thích văn bản lâm sàng có cấu trúc tự động. *Proc ACM Int conf inf knowl manag.* năm 2017; 2017: 1519*–*1528.

[Tóm tắt PubMed](http://www.ncbi.nlm.nih.gov/pubmed/29308293) [|Nhà xuất bản Toàn văn](https://doi.org/10.1145/3132847.3132989) [|Toàn văn miễn phí](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5752318)

1. Afzal M, Alam F, Malik KM, *và cộng sự*: Tóm tắt văn bản y sinh nhận biết bối cảnh lâm sàng bằng mạng nơ-ron sâu: Phát triển và xác nhận mô hình. *J Med Internet Res.* năm 2020; 22 (10): e19810.

[Tóm tắt PubMed](http://www.ncbi.nlm.nih.gov/pubmed/33095174) [|Nhà xuất bản Toàn văn](https://doi.org/10.2196/19810) [|Toàn văn miễn phí](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC7647812)

1. DeYoung J, Beltagy I, van Zuylen M, và *cộng sự*: Ms2: Tóm tắt nhiều tài liệu về các nghiên cứu y học. *bản in trước arXiv arXiv: 2104.06486.* Năm 2021.
2. DeYoung J, Lehman E, Nye B, và *cộng sự*: Suy luận bằng chứng 2.0: Nhiều dữ liệu hơn, mô hình tốt hơn. *bản in trước arXiv arXiv: 2005.04177.* Năm 2020.
3. Nye BE, DeYoung J, Lehman E, *và cộng sự*: Hiểu các báo cáo thử nghiệm lâm sàng: Trích xuất các thực thể y tế và mối quan hệ của chúng. *Hội nghị thượng đỉnh AMIA Jt Transl Sci Proc.* năm 2021; 2021: 485*–*494.

[Tóm tắt PubMed](http://www.ncbi.nlm.nih.gov/pubmed/34457164)

1. Amini I, Martínez D, Aliod DM: Tổng quan về ALTA. *Nhiệm vụ được chia sẻ.*

năm 2012; 2012: 124*–*129.

1. Guo J, Blake C, Guan Y: Đánh giá chiết xuất thực thể tự động đối với các chiến lược điều trị bằng thuốc và không dùng thuốc.

*J Biomed Thông báo.* năm 2019; 94: 103177.

[Tóm tắt PubMed](http://www.ncbi.nlm.nih.gov/pubmed/30986506) [|Nhà xuất bản toàn văn](https://doi.org/10.1016/j.jbi.2019.103177)

1. Suwarningsih W, Purwarianti A, Supriana I: Phân loại câu hỏi y tế Indonesia với khớp mẫu. trong: *Hội nghị quốc tế 2015 về Tự động hóa, Khoa học Nhận thức, Quang học, Hệ thống Cơ điện vi mô và Công nghệ Thông tin (ICACOMIT).* Năm 2015.
2. Abaho M, Bollegala D, Williamson P, *và cộng sự*: Phát hiện và phân loại - Phát hiện và phân loại nhịp chung cho kết quả sức khỏe. *Kỷ yếu của Hội nghị năm 2021 về các phương pháp thực nghiệm trong xử lý ngôn ngữ tự nhiên.* Trực tuyến và Punta Cana, Cộng hòa Dominica; 2021, Tháng Mười Một.
3. Basu T, *và cộng sự*: Một khuôn khổ mới để đẩy nhanh đánh giá có hệ thống bằng cách tự động xây dựng kho đào tạo khai thác thông tin. *Corr.* 2016. ABS / 1606.06424.
4. Marshall IJ, *và cộng sự*: Trialstreamer: Một cơ sở dữ liệu sống, được cập nhật tự động về các báo cáo thử nghiệm lâm sàng. *J Am Med Thông báo Assoc.* năm 2020; 27(12): 1903*–*1912.

[Tóm tắt PubMed](http://www.ncbi.nlm.nih.gov/pubmed/32940710) [|Nhà xuất bản Toàn văn](https://doi.org/10.1093/jamia/ocaa163) [|Toàn văn miễn phí](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC7727361)

1. Barnett A: Tự động phát hiện sự phân tán quá mức và dưới mức trong các bảng cơ sở trong các thử nghiệm ngẫu nhiên có đối chứng. *F1000Nghiên cứu.* năm 2022; 11(783).

[Nhà xuất bản toàn văn](https://doi.org/10.12688/f1000research.123002.1)

1. Raja K, *và cộng sự*: Thuật toán truy xuất trích dẫn kết hợp để tóm tắt kiến thức lâm sàng dựa trên bằng chứng: Kết hợp trích xuất khái niệm, tương tự vectơ và mở rộng truy vấn để có độ chính xác cao. *Corr.* 2016. ABS / 1609.01597.
2. Xu H, *và cộng sự*: Khai thác tài liệu y sinh cho các thuật ngữ liên quan đến phơi nhiễm dịch tễ học. *AMIA Annu Symp Proc.* 2010; 2010: 897*–*901.

[Tóm tắt PubMed](http://www.ncbi.nlm.nih.gov/pubmed/21347108) [|Toàn văn miễn phí](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3041399)

1. Saiz FS, Sanders C, Stevens R, *và cộng sự*: Công cụ bằng chứng lâm sàng trí tuệ nhân tạo để tự động xác định, ưu tiên và trích xuất nghiên cứu ung thư lâm sàng có liên quan. *Thông báo về ung thư JCO Clin.* năm 2021; 5: 102*–*111.

[Tóm tắt PubMed](http://www.ncbi.nlm.nih.gov/pubmed/33439724) [|Nhà xuất bản Toàn văn](https://doi.org/10.1200/cci.20.00087) [|Toàn văn miễn phí](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC8140792)

1. Stylianou N, Razis G, Goulis DG, *và cộng sự*: EBM +: Nâng cao Y học dựa trên bằng chứng thông qua xác định tự động hai cấp độ của Dân số, Can thiệp, Kết quả trong tài liệu y khoa. *Artif Intell Med.* năm 2020; 108, 101949.

[Tóm tắt PubMed](http://www.ncbi.nlm.nih.gov/pubmed/32972669) [|Nhà xuất bản toàn văn](https://doi.org/10.1016/j.artmed.2020.101949)

1. Norman CR, Leeflang M, Spijker R, và *cộng sự*: Một bộ dữ liệu được giám sát từ xa để trích xuất dữ liệu tự động từ các nghiên cứu chẩn đoán. *Kỷ yếu của Hội thảo BioNLP lần thứ 18 và nhiệm vụ chung.* Florence, Ý. 2019. trang 105*–*114.

[Nhà xuất bản toàn văn](https://doi.org/10.18653/v1/W19-5012)

1. Demner-Fushman D, Lin J: *Trích xuất kiến thức để trả lời câu hỏi lâm sàng: Kết quả sơ bộ.* Năm 2005.
2. Lin S, *và cộng sự*: Trích xuất siêu dữ liệu các bài báo nghiên cứu lâm sàng công thức và văn bản miễn phí bằng cách sử dụng các trường ngẫu nhiên có điều kiện. Năm 2010.

trang 90*–*95.

1. Xu R, và *cộng sự*: Trích xuất thông tin nhân khẩu học đối tượng từ các bản tóm tắt của các báo cáo thử nghiệm lâm sàng ngẫu nhiên. Năm 2007.

trang 550*–*554.

[Tóm tắt PubMed](http://www.ncbi.nlm.nih.gov/pubmed/17911777)

1. Zhao J, Bysani P, Kan MY: Khai thác mối tương quan phân loại để trích xuất thông tin thực hành dựa trên bằng chứng. 2012. [Tóm tắt PubMed](http://www.ncbi.nlm.nih.gov/pubmed/23304383) [|Toàn văn miễn phí](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3540431)
2. Raja K, *và cộng sự*: Hướng tới Y học chính xác dựa trên bằng chứng: Trích xuất thông tin dân số từ văn bản y sinh bằng cách sử dụng các bộ phân loại nhị phân và các mẫu cú pháp. *Hội nghị thượng đỉnh AMIA Jt Transl Sci Proc.* 2016; 2016: 203*–*212.

[Tóm tắt PubMed](http://www.ncbi.nlm.nih.gov/pubmed/27570671) [|Toàn văn miễn phí](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5001749)

1. Marshall IJ, *và cộng sự*: Tự động hóa tổng hợp bằng chứng y sinh: RobotReviewer. Trong: *Kỷ yếu của Hội nghị thường niên lần thứ 55 của Hiệp hội Ngôn ngữ học Tính toán.* biên tập Bansal M, Ji H. 2017 Stroudsburg: Assoc Computational Linguistics-Acl. 7*–*12.
2. Wang Q, Liao J, Lapata M, *và cộng sự*: Chiết xuất thực thể PICO cho tài liệu động vật tiền lâm sàng. *Syst Rev.* 2022; 11 (1): 209. [Tóm tắt PubMed](http://www.ncbi.nlm.nih.gov/pubmed/36180888) [|Nhà xuất bản Toàn văn](https://doi.org/10.1186/s13643-022-02074-4) [|Toàn văn miễn phí](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC9524079)
3. Summerscales RL, A S, Hupert J, *và cộng sự*: Xác định các phương pháp điều trị, nhóm và kết quả trong các bản tóm tắt y tế. Năm 2009.
4. Summerscales RL, *và cộng sự*: Tóm tắt tự động kết quả từ các thử nghiệm lâm sàng. trong: *Hội nghị quốc tế IEEE 2011 về Tin sinh học và Y sinh.* Năm 2011.
5. Kang T, Zou S, Weng C: Đào tạo trước để nhận ra các yếu tố PICO từ tài liệu thử nghiệm ngẫu nhiên có đối chứng. *Thông báo về công nghệ sức khỏe Stud.* năm 2019; 264: 188*–*192.

[Tóm tắt PubMed](http://www.ncbi.nlm.nih.gov/pubmed/31437911) [|Nhà xuất bản Toàn văn](https://doi.org/10.3233/SHTI190209) [|Toàn văn miễn phí](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC6852618)

1. Bui DDA, *et al*.: Hệ thống tóm tắt văn bản trích xuất để hỗ trợ trích xuất dữ liệu từ toàn văn trong phát triển đánh giá có hệ thống.

*J Biomed Thông báo.* 2016; 64: 265*–*272.

[Tóm tắt PubMed](http://www.ncbi.nlm.nih.gov/pubmed/27989816) [|Nhà xuất bản Toàn văn](https://doi.org/10.1016/j.jbi.2016.10.014) [|Toàn văn miễn phí](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5362293)

1. Xia Y, *và cộng sự*: Trích xuất các phần tử PICO từ các bản tóm tắt RCT bằng cách sử dụng phân tích 1-2gram và phân loại đa nhiệm. *Corr.* 2019. ABS / 901.08351.

[Nhà xuất bản toàn văn](https://doi.org/10.1145/3340037.3340043)

1. Valdez J, Rueschman M, Kim M, và *cộng sự*: Một đường ống xử lý ngôn ngữ tự nhiên hỗ trợ bản thể để trích xuất siêu dữ liệu xuất xứ từ văn bản y sinh. trong: *Đang di chuyển sang các hệ thống internet có ý nghĩa: otm.* Hội nghị năm 2016, Debruyne C, *và cộng sự*, Biên tập viên. 2016; Nhà xuất bản Springer Int Ag: Cham. trang 699*–*708.
2. Chung GY: Truy xuất câu cho các bản tóm tắt của các thử nghiệm ngẫu nhiên có đối chứng. *BMC Med thông báo Decis Mak.* 2009; 9: 13. [Tóm tắt PubMed](http://www.ncbi.nlm.nih.gov/pubmed/19208256) [|Nhà xuất bản Toàn văn](https://doi.org/10.1186/1472-6947-9-10) [|Toàn văn miễn phí](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2657779)
3. Chung GYC: Hướng tới xác định các nhánh can thiệp trong các thử nghiệm ngẫu nhiên có đối chứng: Trích xuất phối hợp

Cấu trúc. *J Biomed Thông báo.* 2009; 42 (5): 790*–*800.

[Tóm tắt PubMed](http://www.ncbi.nlm.nih.gov/pubmed/19166975) [|Nhà xuất bản toàn văn](https://doi.org/10.1016/j.jbi.2008.12.011)

1. Chung G, Coiera EW: Nghiên cứu về các bản tóm tắt lâm sàng có cấu trúc và phân loại ngữ nghĩa của câu. 2007. trang 121*–*128.
2. Huang K, *và cộng sự*: Phân loại các phần tử PICO theo các đặc điểm văn bản được trích xuất một cách có hệ thống từ các bản tóm tắt của PubMed. *Hội nghị Quốc tế IEEE 2011 về Điện toán chi tiết.* Năm 2011.
3. Hara K, Matsumoto Y: Trích xuất thông tin thiết kế thử nghiệm lâm sàng từ bản tóm tắt MEDLINE. *Gener mới. Tính toán.* 2007; 25 (3): 263*–*275.

[Nhà xuất bản toàn văn](https://doi.org/10.1007/s00354-007-0017-5)

1. Zhu H, *và cộng sự*: Tự động trích xuất các thuộc tính liên quan đến bệnh nhân: bệnh tật, tuổi tác, giới tính và chủng tộc. *Thông báo về công nghệ sức khỏe Stud.* năm 2012; 180: 589*–*593.

[Tóm tắt PubMed](http://www.ncbi.nlm.nih.gov/pubmed/22874259)

1. Schmidt L, Weeds J, Higgins JPT: Khai thác dữ liệu trong văn bản thử nghiệm lâm sàng: Máy biến áp cho các nhiệm vụ phân loại và trả lời câu hỏi. 2020. trang 83*–*94.
2. Jin D, Szolovits P: Phát hiện phần tử PICO trong văn bản y tế thông qua mạng nơ-ron trí nhớ ngắn hạn dài. *Kỷ yếu của Hội thảo BioNLP 2018.* Melbourne, Úc. 2018. tr. 67*–*75. [Nhà xuất bản toàn văn](https://doi.org/10.18653/v1/W18-2308)
3. Demner-Fushman D, *và cộng sự*: Tìm liều lượng thuốc trong liteature. *AMIA Annu Symp Proc.* 2018; 2018: tr. 368*–*376. [Tóm tắt PubMed](http://www.ncbi.nlm.nih.gov/pubmed/30815076) [|Toàn văn miễn phí](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC6371291)
4. Zhang X, Geng P, Zhang T, và *cộng sự*: Aceso: Tóm tắt bằng chứng có hướng dẫn PICO về tài liệu y khoa. *Thông tin sức khỏe IEEE J Biomed.* năm 2020; 24 (9): 2663*–*2670.

[Tóm tắt PubMed](http://www.ncbi.nlm.nih.gov/pubmed/32275627) [|Nhà xuất bản toàn văn](https://doi.org/10.1109/JBHI.2020.2984704)

1. Kang T, Turfah A, Kim J, *và cộng sự*: Một phương pháp biểu tượng thần kinh để hiểu bằng chứng y tế văn bản tự do. *J Am Med Thông báo Assoc.* năm 2021; 28 (8): 1703*–*1711.

[Tóm tắt PubMed](http://www.ncbi.nlm.nih.gov/pubmed/33956981) [|Nhà xuất bản Toàn văn](https://doi.org/10.1093/jamia/ocab077) [|Toàn văn miễn phí](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC8135980)

1. Liu S, Sun Y, Li B, và *cộng sự*: Sent2Span: phát hiện nhịp để chiết xuất PICO trong văn bản y sinh mà không có chú thích nhịp. *bản in trước arXiv arXiv: 2109.02254.* Năm 2021.

[Tóm tắt PubMed](http://www.ncbi.nlm.nih.gov/pubmed/21489224)

1. Nye BE, *và cộng sự*: Trialstreamer: Lập bản đồ và duyệt bằng chứng y tế trong thời gian thực. *Corr.* 2020. ABS / 2005.10865.
2. Blake C, Lucic A: Tự động phát hiện điểm cuối để hỗ trợ quá trình đánh giá có hệ thống. *J Biomed Thông báo.* năm 2015; 56: 42*–*56. [Tóm tắt PubMed](http://www.ncbi.nlm.nih.gov/pubmed/26003938) [|Nhà xuất bản toàn văn](https://doi.org/10.1016/j.jbi.2015.05.004)
3. Huang KC, *et al*.: Phát hiện phần tử PICO trong văn bản y tế mà không có siêu dữ liệu: câu đầu tiên có đủ không? *J Biomed Thông báo.* năm 2013; 46 (5): 940*–*946.

[Tóm tắt PubMed](http://www.ncbi.nlm.nih.gov/pubmed/23899909) [|Nhà xuất bản toàn văn](https://doi.org/10.1016/j.jbi.2013.07.009)

1. Hassanzadeh H, Groza T, Hunter J: Xác định các đồ tạo tác khoa học trong tài liệu y sinh: Trường hợp sử dụng Y học dựa trên bằng chứng. *J Biomed Thông báo.* 2014; 49: 159*–*170.

[Tóm tắt PubMed](http://www.ncbi.nlm.nih.gov/pubmed/24530879) [|Nhà xuất bản toàn văn](https://doi.org/10.1016/j.jbi.2014.02.006)

1. Burnham KP, Anderson DR: Lựa chọn mô hình và suy luận đa mô hình (ấn bản thứ 2). 2002; Springer-Verlag.
2. Brockmeier AJ, *và cộng sự*: Cải thiện ưu tiên tham chiếu với nhận dạng PICO. *BMC Med thông báo Decis Mak.* năm 2019; 19 (1): 14. [Tóm tắt PubMed](http://www.ncbi.nlm.nih.gov/pubmed/31805934) [|Nhà xuất bản Toàn văn](https://doi.org/10.1186/s12911-019-0992-8) [|Toàn văn miễn phí](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC6896258)
3. Gella S, Long DT: Bộ phân loại câu tự động sử dụng các tính năng sắp xếp câu cho Y học dựa trên sự kiện: Mô tả hệ thống tác vụ được chia sẻ. 2012. trang 130*–*133.
4. Lui M: Tính năng xếp chồng để phân loại câu trong bằng chứng-

Y học dựa trên. 2012: 134*–*138.

1. Mollá D: Thí nghiệm với các tính năng dựa trên phân cụm để phân loại câu trong các ấn phẩm y tế: Sự tham gia của Macquarie Test trong nhiệm vụ chung ALTA 2012. 2012: 139*–*142.
2. Sarker A, *và cộng sự*: *Một cách tiếp cận để phân loại nhiều nhãn tự động của các câu y tế.* NICTA: Eveleigh NSW; Năm 2013.
3. Lehman E, DeYoung J, Barzilay R, và *cộng sự*: Suy ra phương pháp điều trị y tế nào hoạt động từ các báo cáo thử nghiệm lâm sàng. *bản in trước arXiv arXiv.: 1904.01606.* Năm 2019.
4. Trenta A, Hunter A, Riedel S: Trích xuất các bảng bằng chứng từ các bản tóm tắt của các thử nghiệm lâm sàng ngẫu nhiên sử dụng bộ phân loại entropy tối đa và các ràng buộc toàn cầu. *CoRR,* *abs./1509.05209.* Năm 2015.

[Nguồn tham khảo](http://arxiv.org/abs/1509.05209)

1. Hansen MJ, Rasmussen G, Fau - Chung NØ, và *cộng sự*: Một phương pháp trích xuất số lượng người tham gia thử nghiệm từ

tóm tắt mô tả các thử nghiệm ngẫu nhiên có đối chứng. *(1758-1109 (Điện tử)).*

1. Boudin F, *và cộng sự*: Kết hợp các bộ phân loại để phát hiện phần tử PICO mạnh mẽ. *BMC Med thông báo Decis Mak.* 2010; 10: 29.

[Tóm tắt PubMed](http://www.ncbi.nlm.nih.gov/pubmed/20470429) [|Nhà xuất bản Toàn văn](https://doi.org/10.1186/1472-6947-10-29) [|Toàn văn miễn phí](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2891622)

1. Chabou S, Iglewski MIeee: Chiết xuất PICO bằng

Kết hợp tính mạnh mẽ của các phương pháp học máy với các phương pháp dựa trên quy tắc. *Đại hội Thế giới về Công nghệ Thông tin và Ứng dụng Máy tính năm 2015.*2015. New York: Ieee.

1. Dawes M, *và cộng sự*: Việc xác định các yếu tố quan trọng về mặt lâm sàng trong các bản tóm tắt tạp chí y khoa: Bệnh nhân- Vấn đề dân số, Phơi nhiễm-Can thiệp, So sánh, Kết quả, Thời gian và Kết quả (PECODR). *Thông báo cho Prim Care.* 2007; 15 (1): 9*–*16.

[Tóm tắt PubMed](http://www.ncbi.nlm.nih.gov/pubmed/17612476)

1. Riley P: Ba cạm bẫy cần tránh trong học máy. *Tính.* năm 2019;

572 (7767).

[Tóm tắt PubMed](http://www.ncbi.nlm.nih.gov/pubmed/31363197) [|Nhà xuất bản toàn văn](https://doi.org/10.1038/d41586-019-02307-y)

1. Amir S, van de Meent JW, Wallace BC: Về tác động của hạt giống ngẫu nhiên đối với tính công bằng của các bộ phân loại lâm sàng. *bản in trước arXiv arXiv: 2104.06338.* Năm 2021.
2. Mehrabi N, *và cộng sự*: Một cuộc khảo sát về sự thiên vị và công bằng trong học máy. *arXiv.* Năm 2019.
3. Brown T, Mann B, Ryder N, *và cộng sự*: *Mô hình ngôn ngữ là những người học ít bắn.* Năm 2020.

[Nguồn tham khảo](https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf)

1. Liu Y, Ott M, Goyal N, và *cộng sự*: Roberta: Một phương pháp đào tạo trước bert được tối ưu hóa mạnh mẽ. *bản in trước arXiv arXiv: 1907.11692.* Năm 2019.
2. Yang J, Jin H, Tang R, *và cộng sự*: Khai thác sức mạnh của LLM trong thực tế: Khảo sát về ChatGPT và hơn thế nữa. *bản in trước arXiv arXiv.: 2304.13712.* Năm 2023.

[Tóm tắt PubMed](http://www.ncbi.nlm.nih.gov/pubmed/37396605)

1. OpenAI.: Báo cáo kỹ thuật GPT-4. *ArXiv.* 2023; ABS/2303.08774.
2. Shaib C, Li ML, Joseph S, *và cộng sự*: Tóm tắt, đơn giản hóa và tổng hợp bằng chứng y tế bằng GPT-3 (với thành công khác nhau). *bản in trước arXiv arXiv: 2305.06299.* Năm 2023.

[Toàn văn miễn phí](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC9128209)

1. Wadhwa S, DeYoung J, Nye B, *và cộng sự*: Cùng trích xuất các can thiệp, kết quả và phát hiện từ các báo cáo RCT với LLM. *bản in trước arXiv arXiv.: 2305.03642.* Năm 2023.
2. Wadhwa S, Amir S, Wallace BC: Xem lại việc trích xuất mối quan hệ trong kỷ nguyên của các mô hình ngôn ngữ lớn. *bản in sẵn arXiv arXiv.: 2305.05003.* Năm 2023.
3. Schmidt L; Phụ lục để xem xét cơ sở. *Harvard Dataverse, V4, UNF:*

*6:0z0ZlKmB1VglRVObRackrw== [fileUNF].* Năm 2020.

[Nhà xuất bản toàn văn](https://doi.org/10.7910/DVN/LNGCOQ)

1. Schmidt L; Bộ dữ liệu có sẵn để tự động hóa SR. *Harvard Dataverse, V1.* Năm 2021.

[Nhà xuất bản toàn văn](https://doi.org/10.7910/DVN/0XTV25)

**Đánh giá ngang hàng mở**

**Trạng thái đánh giá ngang hàng hiện tại: **



**Phiên bản 1**

Báo cáo phản biện 26 Tháng Tám 2021

<https://doi.org/10.5256/f1000research.54235.r89347>

**© 2021 Amezcua-Prieto C.** Đây là một báo cáo đánh giá ngang hàng truy cập mở được phân phối theo các điều khoản của [Giấy phép ghi công](https://creativecommons.org/licenses/by/4.0/)  [Creative](https://creativecommons.org/licenses/by/4.0/) Commons, cho phép sử dụng, phân phối và sao chép không hạn chế trong bất kỳ phương tiện nào, miễn là tác phẩm gốc được trích dẫn đúng cách.

# Carmen Amezcua-Prieto

1 Khoa Y tế Dự phòng và Y tế Công cộng, Đại học Granada, Granada, Tây Ban Nha

2 Khoa Y tế Dự phòng và Y tế Công cộng, Đại học Granada, Granada, Tây Ban Nha

Trích xuất dữ liệu trong một đánh giá có hệ thống là một nhiệm vụ khó khăn và tốn thời gian. Tự động hóa (nửa) trích xuất dữ liệu trong các đánh giá có hệ thống là một lợi thế cho các nhà nghiên cứu và cuối cùng là cho thực hành lâm sàng dựa trên bằng chứng. Đánh giá có hệ thống trực tiếp này kiểm tra các phương pháp tiếp cận đã được công bố để trích xuất dữ liệu từ các báo cáo về các nghiên cứu lâm sàng được công bố cho đến ngày giới hạn là ngày 22 tháng 4 năm 2020. Các tác giả đã bao gồm hơn 50 ấn phẩm trong phiên bản đánh giá này của họ đề cập đến việc trích xuất dữ liệu từ các bản tóm tắt, trong khi ít hơn (26%) sử dụng toàn bộ văn bản. Họ xác định nhiều ấn phẩm mô tả trích xuất dữ liệu cho các đánh giá can thiệp. Các ấn phẩm trích xuất dữ liệu chính xác dịch tễ học hoặc chẩn đoán bị hạn chế.

Các vấn đề quan trọng chính đã được giải quyết trong đánh giá hệ thống:

* Đánh giá có hệ thống sống động này đã được chứng minh. Lĩnh vực tự động hóa đánh giá hệ thống (bán phần) đang phát triển nhanh chóng cùng với những tiến bộ trong xử lý ngôn ngữ, học máy và học sâu.
* Lịch trình tìm kiếm và cập nhật đã được xác định rõ ràng, được thể hiện trong Hình 1.
* Có đầy đủ chi tiết về các phương pháp và phân tích được cung cấp để cho phép sao chép.
* Các kết luận được rút ra đầy đủ được hỗ trợ bởi các kết quả được trình bày trong đánh giá.

Một cân nhắc nhỏ được đề xuất:

* Một câu không đầy đủ trong Phương pháp: 'Chúng tôi bao gồm các báo cáo được xuất bản từ năm 2005 cho đến ngày nay, tương tự với'.

# Phương pháp sống có hợp lý không?

Có

# Lịch trình tìm kiếm và cập nhật đã được xác định rõ ràng và hợp lý chưa?

Có

# Cơ sở lý luận và mục tiêu của Đánh giá Hệ thống có được nêu rõ ràng không?

Có

# Có đủ chi tiết về các phương pháp và phân tích được cung cấp để cho phép người khác sao chép không?

Có

# Phân tích thống kê và giải thích nó có phù hợp không?

Không áp dụng

# Các kết luận rút ra có được hỗ trợ đầy đủ bởi các kết quả được trình bày trong đánh giá không?

Có

# Nếu đây là Đánh giá có hệ thống sống, phương pháp 'sống' có phù hợp không và lịch trình tìm kiếm có được xác định rõ ràng và hợp lý không? ('Đánh giá có hệ thống sống' hoặc một biến thể của thuật ngữ này nên được đưa vào tiêu đề.)

Có

***Lợi ích cạnh tranh:*** Không có lợi ích cạnh tranh nào được tiết lộ.

# Tôi xác nhận rằng tôi đã đọc bài đệ trình này và tin rằng tôi có trình độ chuyên môn thích hợp để xác nhận rằng nó có tiêu chuẩn khoa học có thể chấp nhận được.

Báo cáo phản biện 12 Tháng Tám 2021

<https://doi.org/10.5256/f1000research.54235.r89348>

**© 2021 Kaiser K.** Đây là một báo cáo đánh giá ngang hàng truy cập mở được phân phối theo các điều khoản của [Giấy phép ghi công](https://creativecommons.org/licenses/by/4.0/)  [Creative Commons](https://creativecommons.org/licenses/by/4.0/), cho phép sử dụng, phân phối và sao chép không hạn chế trong bất kỳ phương tiện nào, miễn là tác phẩm gốc được trích dẫn đúng cách.

# Kathryn A. Kaiser

1 Khoa Hành vi Sức khỏe, Trường Y tế Công cộng, Đại học Alabama tại Birmingham, Birmingham, AL, Hoa Kỳ

2 Khoa Hành vi Sức khỏe, Trường Y tế Công cộng, Đại học Alabama tại Birmingham,

Birmingham, AL, Hoa Kỳ

Các tác giả đã thực hiện và ghi lại các bước được thực hiện để theo dõi một lĩnh vực phương pháp nghiên cứu quan trọng đối với nhiều người trên thế giới bằng cách sử dụng "đánh giá có hệ thống sống". Trọng tâm cụ thể là trích xuất dữ liệu tự động hoặc bán tự động xung quanh cấu trúc PICO thường được sử dụng trong y sinh, cho dù đó là tóm tắt một phần tài liệu tường thuật hay sử dụng các kỹ thuật phân tích tổng hợp. Một điều trớ trêu đáng kể về phần thân của các bài báo có trong bài đánh giá này là có một lượng lớn sự thiếu sót liên quan đến việc thực hiện các phương pháp như vậy. Những người tiến hành đánh giá có hệ thống biết rõ mức độ thiếu thông tin được tìm cách tóm tắt một nhóm nghiên cứu.

Những độc giả quan tâm nhất đến công việc đang diễn ra này có thể theo dõi tiến trình của các tác giả trong việc xác định các hoạt động trong không gian này. Tuy nhiên, không rõ nguồn tài trợ sẽ hỗ trợ nỗ lực này trong bao lâu hoặc các tác giả sẽ tiếp tục tham gia vào việc thúc đẩy dự án này trong bao lâu. Dữ liệu được trình bày trong bài báo này không mang lại cho độc giả niềm tin rằng cộng đồng đang tiếp cận các phương pháp có thể chấp nhận được vượt trội hơn so với các phương pháp khác, ít tự động hơn (phương pháp sau này không được thảo luận kỹ lưỡng).

Một số khía cạnh của bài báo sẽ được hưởng lợi từ chi tiết bổ sung (không theo thứ tự quan trọng cụ thể):

1. Trò chơi cuối cùng cho việc theo dõi lĩnh vực văn học này không được mô tả rõ ràng trong bản tóm tắt, cũng như không được thảo luận ở mức độ lớn ở cuối bài báo. Phần lớn các kết quả được trình bày không vẽ ra một tương lai tươi sáng cho lĩnh vực nghiên cứu này như các điều kiện hiện tại. Mặc dù mục tiêu được trình bày tốt trong phần 1.2, nhưng một lượng lớn dữ liệu hiệu suất bị thiếu (được báo cáo là 87%) không thể giải quyết câu hỏi "Nó có đáng tin cậy không?". Người ta có thể nghi ngờ rằng nếu hiệu suất đặc biệt xuất sắc được chứng minh bởi một dự án, những dữ liệu đó sẽ được quảng cáo nổi bật. Do đó, bước liên hệ với các tác giả chưa được thực hiện sẽ khai sáng nếu có thể thu thập dữ liệu hiệu suất hoặc nếu các tác giả vẫn im lặng về yêu cầu đó. Nhiệm vụ tiếp theo này sẽ là một điểm quan tâm chính đối với nhiều người sẽ theo dõi các bản cập nhật của bài báo này. Có khả năng bối cảnh nghiên cứu cụ thể (ví dụ: xem Phạm và *cộng sự*, 20211) sẽ có ảnh hưởng lớn đến các chỉ số hiệu suất nếu chúng có thể được xác định.
2. Mô tả về cách xác định 17 "Mục quan tâm chính" và liệu có kế hoạch đưa chúng làm hướng dẫn phương pháp luận hoặc danh sách kiểm tra báo cáo hay không sẽ hữu ích. Một trong hai điều này sẽ giúp thúc đẩy lĩnh vực này hơn nữa.
3. Trên Trang 5, các loại trừ được liệt kê có việc sử dụng tiền xử lý văn bản, nhưng kết quả thảo luận về nhiều bài báo dường như đã sử dụng điều đó trong phương pháp của họ. Có lẽ đây là một sai lệch so với giao thức ban đầu sau khi đánh giá bắt đầu (một quyết định dễ hiểu)?
4. Trong phần 2.4 về tìm kiếm Pubmed, các tác giả có thể làm rõ rằng API hoặc GUI của Pubmed 2.0 sẽ được sử dụng để truy cập tài liệu của ứng cử viên không?
5. Cũng liên quan đến phần 2.4 về tìm kiếm, vì GITHUB rất phổ biến, đây có thể là một nơi hiệu quả để tìm kiếm thường xuyên?
6. Làm rõ khả năng có được các gói phần mềm được trích dẫn (cho dù miễn phí hoặc với một số chi phí) sẽ hữu ích.
7. Hình 3 giải thích PICO là một lỗi chính tả – "PCIO".
8. Bảng 5 được hiển thị trước Bảng 1. Vui lòng kiểm tra và sửa luồng và tham chiếu đến số bảng (5,1,4,2,3 là dòng bây giờ).
9. Một trong những hạn chế chính cần lưu ý là vấn đề đáng tiếc là thiếu dữ liệu cụ thể trong các bản tóm tắt về các can thiệp và so sánh.

# Tham khảo

1. Phạm B, Jovanovic J, Bagheri E, Antony J, và cộng sự: Khai thác văn bản để hỗ trợ sàng lọc trừu tượng cho tổng hợp kiến thức: quy trình làm việc bán tự động. *Đánh giá có hệ thống*. 2021; **10** (1). Nhà [xuất bản toàn văn](https://doi.org/10.1186/s13643-021-01700-x)

# Phương pháp sống có hợp lý không?

Có

# Lịch trình tìm kiếm và cập nhật đã được xác định rõ ràng và hợp lý chưa?

Có

# Cơ sở lý luận và mục tiêu của Đánh giá Hệ thống có được nêu rõ ràng không?

Có

# Có đủ chi tiết về các phương pháp và phân tích được cung cấp để cho phép người khác sao chép không?

Có

# Phân tích thống kê và giải thích nó có phù hợp không?

Có

# Các kết luận rút ra có được hỗ trợ đầy đủ bởi các kết quả được trình bày trong đánh giá không?

Có

# Nếu đây là Đánh giá có hệ thống sống, phương pháp 'sống' có phù hợp không và lịch trình tìm kiếm có được xác định rõ ràng và hợp lý không? ('Đánh giá có hệ thống sống' hoặc một biến thể của thuật ngữ này nên được đưa vào tiêu đề.)

Có

***Lợi ích cạnh tranh:*** Không có lợi ích cạnh tranh nào được tiết lộ.

***Chuyên môn của người đánh giá:*** Đánh giá có hệ thống về các chủ đề y sinh, các vấn đề về thời gian và công sức cần thiết để hoàn thành đánh giá bằng các công cụ có sẵn chung.

# Tôi xác nhận rằng tôi đã đọc bài đệ trình này và tin rằng tôi có trình độ chuyên môn thích hợp để xác nhận rằng nó có tiêu chuẩn khoa học có thể chấp nhận được.

Báo cáo phản biện 08 Tháng Sáu 2021

<https://doi.org/10.5256/f1000research.54235.r85692>

**© 2021 McFarlane E.** Đây là một báo cáo đánh giá ngang hàng truy cập mở được phân phối theo các điều khoản của [Giấy phép ghi công](https://creativecommons.org/licenses/by/4.0/)  [Creative](https://creativecommons.org/licenses/by/4.0/) Commons, cho phép sử dụng, phân phối và sao chép không hạn chế trong bất kỳ phương tiện nào, miễn là tác phẩm gốc được trích dẫn đúng cách.



# Emma McFarlane

1 Trung tâm Hướng dẫn, Viện Quốc gia về Sức khỏe và Chăm sóc Xuất sắc, London, Vương quốc Anh

2 Trung tâm Hướng dẫn, Viện Quốc gia về Sức khỏe và Chăm sóc Xuất sắc, London, Vương quốc Anh

Đây là một đánh giá có hệ thống sống động về các phương pháp và công cụ đã xuất bản nhằm tự động hóa hoặc bán tự động hóa quá trình trích xuất dữ liệu trong bối cảnh đánh giá có hệ thống. Tự động hóa trích xuất dữ liệu là một lĩnh vực được quan tâm trong y học dựa trên bằng chứng.

Các phương pháp được mô tả đầy đủ để được sao chép, nhưng các chi tiết phân tích để xác định các mục quan tâm sẽ hữu ích để liên kết với kết quả. Ngoài ra, các tác giả có thể muốn xem xét bình luận về các lĩnh vực chủ đề được đề cập trong các nghiên cứu bao gồm và liệu điều đó có tác động đến bất kỳ số liệu nào được đo lường hay không.

Trong phần thảo luận, điều thú vị là ít nghiên cứu trích xuất dữ liệu từ toàn văn hơn. Các tác giả có thể bình luận về ý nghĩa của việc sử dụng các công cụ trong đánh giá trực tiếp vì không phổ biến chỉ trích xuất dữ liệu từ một bản tóm tắt.

# Phương pháp sống có hợp lý không?

Có

# Lịch trình tìm kiếm và cập nhật đã được xác định rõ ràng và hợp lý chưa?

Có

# Cơ sở lý luận và mục tiêu của Đánh giá Hệ thống có được nêu rõ ràng không?

Có

# Có đủ chi tiết về các phương pháp và phân tích được cung cấp để cho phép người khác sao chép không?

Một phần

# Phân tích thống kê và giải thích nó có phù hợp không?

Không áp dụng

# Các kết luận rút ra có được hỗ trợ đầy đủ bởi các kết quả được trình bày trong đánh giá không?

Có

# Nếu đây là Đánh giá có hệ thống sống, phương pháp 'sống' có phù hợp không và lịch trình tìm kiếm có được xác định rõ ràng và hợp lý không? ('Đánh giá có hệ thống sống' hoặc một biến thể của thuật ngữ này nên được đưa vào tiêu đề.)

Có

***Lợi ích cạnh tranh:*** Không có lợi ích cạnh tranh nào được tiết lộ.

***Chuyên môn của người đánh giá:*** Y học dựa trên bằng chứng, đánh giá có hệ thống, kỹ thuật tự động hóa.

**Tôi xác nhận rằng tôi đã đọc bài đệ trình này và tin rằng tôi có trình độ chuyên môn thích hợp để xác nhận rằng nó có tiêu chuẩn khoa học có thể chấp nhận được.**



Lợi ích của việc xuất bản với F1000Research:

* Bài viết của bạn được xuất bản trong vòng vài ngày, không thiên vị biên tập
* Bạn có thể xuất bản các bài báo truyền thống, kết quả null/âm tính, báo cáo ca bệnh, ghi chú dữ liệu và hơn thế nữa
* Quá trình đánh giá ngang hàng minh bạch và hợp tác
* Bài viết của bạn được lập chỉ mục trong PubMed sau khi vượt qua đánh giá ngang hàng
* Hỗ trợ khách hàng tận tâm ở mọi giai đoạn

Đối với các thắc mắc trước khi gửi, hãy liên hệ với [research@f1000.com](mailto:research@f1000.com)