UNIVERSIDADE NOVE DE JULHO - UNINOVE
PROGRAMA DE PÓS GRADUAÇÃO EM INFORMÁTICA E GESTÃO DO CONHECIMENTO - PPGIGC

EDSON MELO DE SOUZA

APLICAÇÃO DE CIÊNCIA DE DADOS NA ANÁLISE DO POSICIONAMENTO AUTORAL E CONTRIBUIÇÕES CIENTÍFICAS EM ARTIGOS

São Paulo
2021

EDSON MELO DE SOUZA

APLICAÇÃO DE CIÊNCIA DE DADOS NA ANÁLISE DO POSICIONAMENTO AUTORAL E CONTRIBUIÇÕES CIENTÍFICAS EM ARTIGOS

Tese apresentada ao Programa de Pós-Graduação em Informática e Gestão do Conhecimento (PPGIGC) da Universidade Nove de Julho - UNINOVE, como parte dos requisitos para a obtenção do título de Doutor em Informática e Gestão do Conhecimento.

Prof. Orientador: Dr. José Eduardo Storopoli
Prof. Coorientador: Dr. Wonder Alexandre Luz Alves

São Paulo
2021

Souza, Edson Melo de.
Aplicação de ciência de dados na análise do posicionamento autoral e contribuições científicas em artigos. / Edson Melo de Souza. 2021.
119 f.
Tese (Doutorado) - Universidade Nove de Julho - UNINOVE, São Paulo, 2021.
Orientador (a): Prof. Dr. José Eduardo Storopoli.
1. Ciência de dados. 2. Posição autoral. 3. Lista de contribuições. 4. *Byline*.
I. Storopoli, José Eduardo. II. Titulo.

CDU 004

PARECER – EXAME DE DEFESA

Parecer da Comissão Examinadora designada para o exame de defesa do Programa de Pós-Graduação em Informática e Gestão do Conhecimento, a qual se submeteu o aluno regularmente matriculado **Edson Melo de Souza**

Tendo examinado o trabalho apresentado para obtenção do título de "Doutor em Informática e Gestão do Conhecimento", com Tese intitulada **"**APLICAÇÃO DE CIÊNCIA DE DADOS NA ANÁLISE DO POSICIONAMENTO AUTORAL E CONTRIBUIÇÕES CIENTÍFICAS EM ARTIGOS", a Comissão Examinadora considerou o trabalho:

(X) Aprovado **( ) Aprovado condicionalmente**
( ) Reprovado com direito a novo exame **( ) Reprovado**

Parecer:
O candidato foi considerado aprovado.

EXAMINADORES

Prof. Dr. José Eduardo Storopoli (Orientador – PPGI/UNINOVE)

Prof. Dr. Diógenes de Souza Bido (Parcipante externo – Mackenzie)

Prof. Dr. Jesús Pascual Mena Chalco (Parcipante Externo – UFABC)

Prof. Dr. Leonardo Vils (Parcipante Externo – PPGCIS/UNINOVE)

Prof. Dr. Fellipe Silva Marns (Membro Interno – PPGI/UNINOVE)

Prof. Dr. Wonder Alexandre Luz Alves (Coorientador – PPGI/UNINOVE)

São Paulo, 28 de junho de 2021.

DEDICATÓRIA

Este trabalho é dedicado aos meus pais Irineu e Durvalina, à minha esposa, Andrea e aos meus filhos, Gustavo e Elisa, por serem meu porto seguro para o qual sempre pude retornar para recarregar as energias. Sem o companheirismo e apoio de vocês, esse sonho não seria possível de ser realizado.

AGRADECIMENTOS

Aos meus pais Irineu e Durvalina, por me darem a oportunidade de estar no mundo; apesar de toda a dificuldade, sempre me deram apoio e amor. Por isso, sou grato e sinto orgulho em tê-los como meus pais, além da admiração pelo esforço.

À minha esposa Andrea e meus filhos Gustavo e Elisa, que doaram incondicionalmente seu tempo e me apoiaram desde o início nesta jornada. Amo vocês.

Ao meu orientador e amigo, Professor Dr. José Eduardo Storopoli, pelos seus valiosos comentários, incentivo, disponibilidade, atenção, amizade e, principalmente, por acreditar em mim.

Ao meu amigo e coorientador, Professor Dr. Wonder Alexandre Luz Alves, pela amizade, confiança, ensinamentos, orientação, companheirismo e compreensão.

Ao Professores Dr. André Felipe Henriques Librantz, Dr. Ivanir Costa, Dr. Marcos Antônio Gaspar, Dr. Sidnei Araújo, Dr. Fellipe Silva Martins e Dr. Leonardo Vils pelo incentivo e apoio e orientações.

Ao meu amigo, Dr. Charles Ferreira Gobber, por todo apoio e suporte nos momentos de dificuldades com o LaTeX e pelas conversas por horas a fio.

À Daniela Passarelo Moura da Fonseca por todo o apoio e suporte.

À Universidade Nove de Julho (UNINOVE), pela oportunidade da concessão da bolsa de estudos.

Por fim, a todos aqueles que, direta ou indiretamente, colaboraram para a realização deste sonho.

EPÍGRAFE

"Há pessoas que desejam saber só por saber, e isso é curiosidade; outras, para alcançarem fama, e isso é vaidade; outras, para enriquecerem com a sua ciência, e isso é um negócio torpe; outras, para serem edificadas, e isso é prudência; outras, para edificarem os outros, e isso é caridade".

Santo Agostinho (354-430)
Bispo católico, teólogo e filósofo.

RESUMO

Contexto: Nos últimos anos houve crescimento do número de autores listados em artigos e a questão do posicionamento autoral e das contribuições científicas ainda se encontra em aberto na literatura, uma vez que não há definição ou recomendação formal para o posicionamento dos autores. **Objetivo**: Neste trabalho foram estudadas as categorias de contribuição e a questão da posição autoral em relação às contribuições dos autores em publicações científicas na área de ciências biológicas e medicina. **Método**: O estudo foi realizado utilizando técnicas estatísticas, em especial Análise Fatorial, e de ciência de dados, em especial Regressão Linear sobre os dados de 2.024 artigos contendo 17.385 autores das fontes de dados: SCImago, Scopus e dos periódicos da área de ciências biológicas e medicina Annals of Internal Medicine (Anna. Intern. Med.), Journal of the American Medical Association (JAMA) e PLoS Medicine (PLoS Med). Para coleta e limpeza dos dados foram desenvolvidos robôs com técnicas de raspagem de dados para automatização dos processos. **Resultados**: O estudo mostrou a existência de dois grupos de contribuição (*"Theory"* e *"Methodology/Logistic"*), além da proposta de um modelo universal de contribuições com três categorias: teórica, metodológica e logística. Ademais, foram desenvolvidos algoritmos para automatização dos tratamentos de dados coletados pelos robôs. **Conclusão** As técnicas de ciência de dados permitiram a construção de robôs para automatização da coleta de dados em bases e publicações científicas em conjunto com os algoritmos desenvolvidos, reduzindo substancialmente o tempo de coleta e melhorando a acurácia dos processos. O agrupamento das contribuições científicas evidenciou que as maiores contribuições se encontram no grupo teórico, sinalizando que a experiência acadêmica dos autores é um fator preponderante, enquanto as contribuições metodológicas e logísticas representam contribuições substanciais. O modelo categórico proposto para o estudo da relação entre o posicionamento autoral e as contribuições científicas mostrou que autores com contribuições teóricas tendem a ser o primeiro autor. Já as contribuições logísticas tendem a posicionar um autor como último. As contribuições metodológicas não evidenciam impacto no posicionamento autoral. Por fim, o pequeno efeito registrado na correlação entre as variáveis de contribuição evidencia que o posicionamento autoral não é definido pelas contribuições dos autores.

Palavras-chave: Ciência de dados, Posição Autoral, Lista de Contribuições, *Byline*.

ABSTRACT

In recent years, there has been an increase in the number of authors listed in articles and the issue of authorship and scientific contributions is still open in the literature since there is no formal definition or recommendation for the authors' position. **Objective**: In this dissertation, I've scrutinized the relationship between scientific contribution and authorship ethical issues. My main objective was to study contributions types and its impact in authorship position. **Methodology**: I've collected 2,024 articles containing 17,385 authors from the following data sources: SCImago, Scopus, and the following journals: Annals of Internal Medicine (Ann. Intern. Med.), Journal of the American Medical Association (JAMA), and PLoS Medicine (PLoS Med.). All of those being journals from the biological sciences and medicine areas. To analyze the data I've employed statistical techniques, specifically Factor Analysis; and also data science techniques, specifically Linear Regression. Furthermore, for data collection and cleaning automation, I've developed robots with data scraping techniques. **Results**: My study revealed the existence of mainly two scientific contribution categories (*"Theory"* and *"Methodology/Logistics"*). I've also proposed an universal categorical model of scientific contribution based on three main categories: theory, methodology, and logistic. Furthermore, I've also developed algorithms to automate data collection by robots/crawler. **Conclusion**: Data science techniques have allowed the construction of robots to automate data collection in scientific databases and publications together with developed algorithms, substantially reducing the collection time and improving the accuracy of the processes. The grouping of scientific contributions showed that the largest contributions are found in the theoretical group, indicating that the academic experience of the authors is a major factor, while the methodological and logistical contributions represent substantial contributions. The categorical model proposed for the study of the relationship between authorial positioning and scientific contributions shows that authors who contribute theoretically tend to be the first authors. Logistic contributions, on the other hand, tend to place an author last. Methodological contributions do not have an impact on author position. Finally, the small effect registered in the correlation between the contribution variables evidences that the author's position is not defined by the authors' contributions.

Keywords: Data Science, Author Position, Contribution List, Byline.

SUMÁRIO

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE QUADROS

LISTA DE ALGORITMOS

1 INTRODUÇÃO

1.1 Contextualização do Tema

Com o avanço da ciência nos últimos 30 anos, a colaboração científica entre diferentes áreas do conhecimento mostrou um forte crescimento no número de autores listados nos artigos (ROSENZWEIG et al., 2008; PATIENCE et al., 2019). De forma geral, as principais contribuições acadêmicas ocorrem pelos autores posicionados na primeira e na última posição na ordem de autoria, onde a primeira é geralmente ocupada por um estudante de doutorado ou pós-doutorado e a última posição pelo diretor do laboratório ou responsável pela pesquisa, enquanto os autores do meio colaboram com contribuições menos expressivas para artigos acadêmicos (MONGEON et al., 2017). Entretanto, podem ocorrer variações de acordo com a área de pesquisa (LIU; FANG, 2014a) como na economia, matemática e na física de altas energias que listam os autores em ordem alfabética (MARQUES, 2011). O processo para determinar este posicionamento não está definido de forma clara na literatura, além de existirem desafios para determinar o grau de colaboração e responsabilidade de cada um, gerando assimetrias nas contribuições (CHANG, 2019).

No decorrer dos últimos anos, esse assunto vem despertando interesse na comunidade científica para compreender o cenário (PEIDU, 2019), pois o posicionamento autoral é uma exigência em todas as publicações científicas, de modo a identificar os responsáveis através de suas contribuições. Entretanto, não existe consenso sobre a declaração das contribuições dos autores em um artigo (HILÁRIO et al., 2018), apesar da criação da taxonomia CRediT em 2014, que orienta sobre a denominação de 14 categorias de contribuição (LARIVIÈRE; PONTILLE; SUGIMOTO, 2021). Segundo Allen, O'Connell e Kiermer (2019), em 2018 era usado em mais de 120 periódicos, com perspectivas de crescimento para os próximos. Na área das ciências da saúde, classificadas entre as 27 áreas temáticas do *SCImago Journal & Country Rank* (SCIMAGO, 2020), a descrição das contribuições de cada autor seguem as diretrizes recomendadas nos *Guidelines* do *International Committee of Medical Journal Editors* (ICMJE) (SAUERMANN; HAEUSSLER, 2017), devendo ser informadas dentro de cada uma das seguintes categorias: *"Conception & design or acquisition of data or analysis and interpretation"*, *"Drafting or revising"* e *"Final approval"* (DECULLIER; MAISONNEUVE, 2019).

Diante do exposto, este trabalho visa compreender a questão da posição autoral em relação às contribuições dos autores em publicações científicas, usando como amostragem a área de ciências biológicas e medicina. Tal compreensão pode determinar qual a relação entre posição autoral e colaboração em cada uma das categorias de contribuição. Para endereçar essa lacuna, a aplicação de ciência de dados pode ser uma maneira eficiente, ágil, gerando *insights* e conhecimento (PRIESTLEY; MCGRATH, 2019), permitindo avanços

no campo da ciência (SCHOENHERR; SPEIER-PERO, 2015; LOWNDES et al., 2017), e em especial a compreensão da natureza das contribuições científicas.

De origem multidisciplinar, faz amplo uso de "matemática aplicada (probabilidade)" (EGGHE; LIANG; ROUSSEAU, 2003), estatística (HARDIN et al., 2015), ciência da computação (HE et al., 2013) e metaciência (IOANNIDIS, 2005; FORTUNATO et al., 2018). A resolução de problemas com o uso da ciência de dados ocorre em etapas, mais conhecidas como *pipeline*, onde cada etapa é interdependente e possui sua complexidade individual. As técnicas disponíveis usam diferentes algoritmos, destacando-se os de aprendizagem de máquina (MITCHELL et al., 1997; BISHOP, 2006).

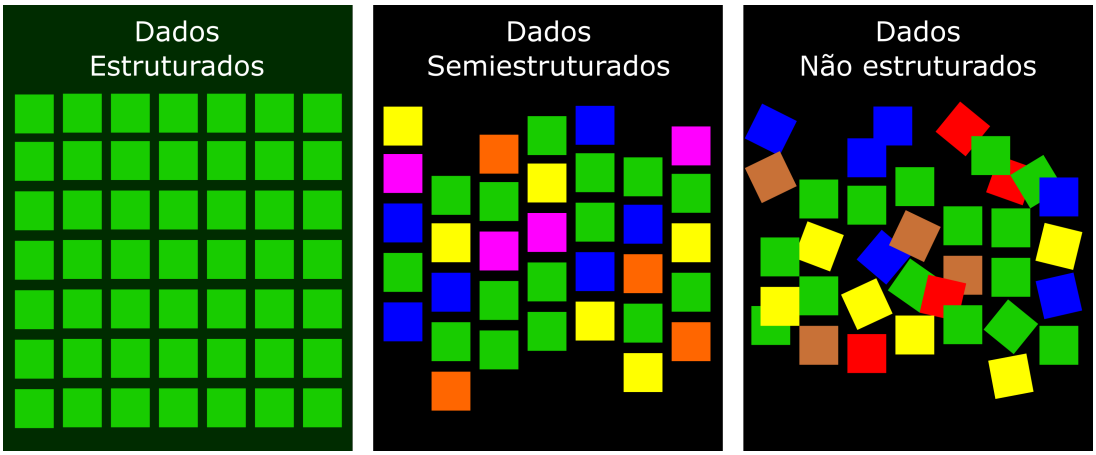
Para que a ciência de dados possa auxiliar de maneira assertiva no processo de geração de conhecimento, é importante que os dados sejam identificados e coletados em suas fontes primárias. Posteriormente devem passar por uma fase de pré-processamento para seu tratamento (de maneira transparente e replicável), de modo a identificar parâmetros ausentes ou irrelevantes, detecção de valores duplicados, fora de padrão, anômalos, entre outros (BAJPAI; METKEWAR, 2016). A ausência ou realização incorreta desses procedimentos pode influenciar na geração de modelos de aprendizado de máquina tendenciosos que não fornecerem condições de realizar inferências ou detecção correta de padrões (VELLIDO; MARTÍN-GUERRERO; LISBOA, 2012). Segundo Porto e Ziviani (2014), a ciência de dados vem sendo "[...] aplicada na área das ciências por biólogos, astrônomos, físicos, e demais pesquisadores, visando enfrentarem problemas computacionais na chamada *e*-ciência, que se tornam barreiras para as suas descobertas.". Nesse sentido, Cole (2020) utilizou a ciência de dados na área da química para o desenvolvimento de um método para a descoberta acelerada de novos produtos químicos para aplicações funcionais. Na área de ciências biológicas e medicina, Corey et al. (2018) desenvolveram modelos de aprendizagem de máquina para identificar pacientes com alto risco cirúrgico. Já Kavakiotis et al. (2017) aplicaram métodos de aprendizado de máquina e mineração de dados em pesquisas sobre diabetes. Esses estudos mostram a viabilidade da utilização da ciência de dados no campo da pesquisa científica, não só para a realização de pesquisas científicas, mas também para estudar a própria ciência (metaciência): o foco desta tese; **o estudo da relação entre o posicionamento autoral e as contribuições dos autores**.

1.2 Problema e Justificativa de Pesquisa

Devido à existência de aproximadamente 11.000 artigos na área da medicina (consulta realizada em nov. 2019), o volume de dados a ser coletado para o desenvolvimento desta pesquisa é expressivo. Esses artigos são originados de bases científicas indexadas e dos diversos periódicos. Dessa maneira, a manipulação de um grande volume de dados com técnicas tradicionais, dificulta o processo de coleta, processamento e análise de forma eficaz (KATAL; WAZID; GOUDAR, 2013; CAI; ZHU, 2015).

Em ciência de dados, os processos mais trabalhosos são os de extração e tratamento dos dados. Para isso, são necessários diversos passos, porque geralmente existem múltiplas fontes de dados com diferentes formatos (estruturados, semiestruturados e não estruturados) (LI et al., 2008), representados na Figura 1.1.

Figura 1.1 – *Representação visual entre as diferentes estruturas de dados, onde os dados "estruturados" são organizados de forma homogênea, no formato tabular. Já os dados "semiestruturados" são heterogêneos, possuindo estrutura irregular. Por fim, os dados "não estruturados" não possuem padrão definido, podendo ser obtidos em diferentes formatos.*



Fonte: Autor

Diante desse contexto, a coleta de dados de artigos e de informações dos autores deve ser realizada a partir de múltiplas fontes, englobando múltiplos formatos, que representa grande esforço manual e elevado consumo de tempo. Para mitigar o esforço humano em processos repetitivos e exaustivos, são necessárias ferramentas automáticas que suportem todas as etapas do processo (LOWNDES et al., 2017). Na literatura é possível encontrar trabalhos que utilizaram ferramentas baseadas em técnicas de ciência de dados para análise de dados sobre publicações científicas (ROSENZWEIG et al., 2008), ambiguidade autoral (KIM; KIM; OWEN-SMITH, 2019) e coautoria (ABBASI; ALTMANN; HOSSAIN, 2011). Entretanto, nos trabalhos pesquisados não foram apontadas ferramentas que realizem, de forma automática, as tarefas de coleta e tratamento de dados, surgindo a primeira lacuna de pesquisa.

Lacuna de Pesquisa 1. Identificar padrões para os processos de coleta, limpeza e preparação dos dados de bases científicas indexadas.

A partir da lacuna de pesquisa 1, surgem algumas perguntas:

***Pergunta* 1.1.** Como automatizar o processo de coleta de dados em bases científicas indexadas através de robôs usando técnicas de raspagem de dados (*web scraping*)?

***Pergunta* 1.2.** Como realizar o tratamento dos dados coletados para eliminação de anomalias, identificação de erros, dados ausentes e duplicados de forma automática?

A revisão da literatura mostra a existência de diversos trabalhos sobre a questão do posicionamento autoral utilizando parâmetros como: estudo sobre indicadores de produção (WALTERS, 2016), posicionamento autoral baseado em ordenação alfabética (LIU; FANG, 2014b), idade dos autores (COSTAS; BORDONS, 2011) e senioridade em pesquisa (ABRAMO; D'ANGELO; MURGIA, 2016).

Especificamente na área de ciências biológicas e medicina, pela necessidade de seguir as diretrizes sobre contribuições do ICMJE (*Conception & design or acquisition of data or analysis and interpretation*, *Drafting or revising* e *Final approval* (MATHESON, 2011)), há ampla literatura que se debruça sobre a relação entre as contribuições declaradas e a posição autoral (CLEMENT, 2014; JIA et al., 2016; SIVERTSEN; ROUSSEAU; ZHANG, 2019). Porém, ainda não há avanços na classificação da posição autoral em relação à lista de contribuições. Assim, emerge a segunda lacuna de pesquisa.

Lacuna de Pesquisa 2. Identificar padrões na relação entre posição autoral e contribuições na autoria científica.

Uma vez explorada a lacuna de pesquisa 2, as seguintes perguntas surgem a respeito da relação entre posição autoral e contribuições dos autores:

***Pergunta* 2.1.** Quais categorias emergem ao mapear a contribuição científica utilizando estatística e ciência de dados?

***Pergunta* 2.2.** Qual a relação entre a posição autoral e categorias de contribuição?

1.3 Objetivos

Para responder às perguntas de pesquisa, o objetivo geral desse trabalho é **aplicar estatística e ciência de dados para investigar a relação entre a posição autoral e as categorias de contribuições nas produções científicas**. Especificamente, pretende-se atingir os seguintes objetivos específicos:

1. Desenvolver robôs para coleta de dados utilizando técnicas de raspagem de dados para automatizar o processo de extração de:
    - Dados de publicações de artigos em bases científicas indexadas.
    - Dados sobre contribuição autoral.
2. Criar procedimentos automatizados para realizar o pré-processamento dos dados.
3. Modelar e processar os dados com algoritmos de aprendizagem de máquina supervisionada para estudar a relação entre o posicionamento autoral e a lista de contribuições.
4. Propor um método de categorização de contribuições, visando mostrar quais são mais relevantes e que refletem o posicionamento autoral.

1.4 Limitações da Pesquisa

Embora esta pesquisa tenha como objetivo aplicar estatística e ciência de dados para o entendimento do posicionamento autoral e a contribuição em publicações científicas. O objetivo e amplo e não será esgotado nesta tese. Com isso, temos algumas limitações de pesquisa que geram oportunidades para trabalhos futuros:

a. O estudo se restringe à investigação do posicionamento autoral e as contribuições na área de ciências biológicas e medicina, visto que há forte recomendação da declaração das contribuições nestas áreas pelo ICMJE, que estão entre as 27 áreas categorizadas pelo SCImago.

b. Os dados analisados limitam-se a três principais periódicos da área de ciências biológicas e medicina *Annals of Internal Medicine* (Ann. Intern. Med.), *Journal of the American Medical Association* (JAMA) e *PLoS Medicine* (PLoS Med.).

c. O índice H (H-*index*) de cada autor, utilizado como uma medida aproximada de experiência em pesquisa, limita-se ao ano de 2019, visto que foi o único ano disponível para o índice na base de dados do Scopus. Isso implica em uma limitação dessa métrica, pois a amostra foi coletada entre 2000 e 2019. Apesar disso, índice H é considerado uma *proxy* adequada para a mensuração de experiência em pesquisa.

1.5 Organização do Trabalho

O presente trabalho apresenta mais sete capítulos, que expõem e tratam dos seguintes tópicos:

- Capítulo 2 (Base Teórica). Aborda os procedimentos para realização do levantamento bibliográfico sobre os conceitos teóricos presentes nesta tese no que diz respeito à ciência de dados, posicionamento autoral e contribuições em publicações científicas, bem como trabalhos correlatos.

- Capítulo 3 (Fontes de Dados). Descreve os métodos para identificação das fontes de dados utilizadas neste trabalho. Também são descritas as etapas e procedimentos para a fase preparação de coleta dos dados, descrita no Capítulo 4.

- Capítulo 4 (Coleta de Dados dos Artigos, Autores, Citações e Contribuições). Apresenta em detalhes o processo de coleta dos dados dos artigos e dos autores.

- Capítulo 5 (Análise de Contribuições). Apresenta um estudo sobre as categorias de contribuição utilizadas nos periódicos Ann. Intern. Med., JAMA e PLoS.

- Capítulo 6 (Posição Autoral e Contribuições) Apresenta os procedimentos para a modelagem dos dados das contribuições autorais com a técnica de análise fatorial e as definições do modelo de classificação de contribuições utilizado para estudar a relação entre o posicionamento autoral e as contribuições em publicações científicas.
- Capítulo 7 (Conclusões). Apresenta as conclusões sobre os estudos apresentados neste trabalho.
- Capítulo 8 (Trabalhos Futuros). Apresenta sugestões para trabalhos futuros.

2 BASE TEÓRICA

Resumo do capítulo

Este capítulo apresenta os procedimentos para realização do levantamento bibliográfico sobre os conceitos teóricos presentes nesta tese, abordando ciência de dados, posicionamento autoral e contribuições em publicações científicas.

2.1 Levantamento Bibliográfico

O levantamento bibliográfico foi realizado através de buscas nas seguintes bases científicas indexadas: Scopus, Web of Science e Scielo para os seguintes termos em lingua inglesa: *author byline*, *authorship byline*, *author order*, *authorship order*, *author contribution*, *authorship contribution*, *author responsibilities*, *authorship responsibilities*, *author position*, *authorship position*, *author impact*, *authorship impact*, *author prestigious*, *authorship prestigious*, *byline contribution*, *byline order*, *byline position*, *byline impact*, *coauthorship*, *data science*, *data science pipeline*, *data gathering*, *data cleaning*, *data scraping* e *data science pipeline* no período compreendido entre os anos de 2013 e 2019. A base de dados do PubMed não foi utilizada, uma vez que o processo de busca não apresenta resultados sobre métricas de publicação dos autores.

Os termos foram pesquisados nas palavras-chave, título e resumo. Em alguns casos, o filtro foi limitado apenas ao título e/ou palavras-chaves devido ao elevado número de artigos encontrados que, em alguns casos, apresentou valores próximos de 50.000.

As palavras-chave foram selecionadas a partir da análise da representatividade de cada uma sobre os temas: ciência de dados, posicionamento autoral e contribuições em publicações científicas, além do estudo dos trabalhos de (PERNEGER et al., 2017; YANG; WOLFRAM; WANG, 2017), considerados importantes para esta pesquisa.

Os *scripts* utilizados para a realização das buscas nas bases de dados do Scopus, Web of Science e Scielo se encontram nos Apêndices A, B e C, respectivamente. As buscas resultaram em um total de 2.084 registros com a seguinte divisão:

1. Autoria, posicionamento autoral e contribuições.
   (a) 661 no Scopus.
   (b) 721 na Web of Science.
   (c) 13 no Scielo.
2. Ciência de Dados
   (a) 511 no Scopus.
   (b) 148 na Web of Science.

(c) 30 no Scielo.

Os 2.084 registros dos artigos foram exportados de cada uma das bases, agregados e processados para filtragem, utilizando os seguintes critérios:

1. Registros que possuíam alguma restrição explícita de citação (n=6).
2. Não continham palavras-chave (n=429).
3. Registros duplicados entre as bases (n=238).
4. Artigos que não possuíam citações (n=244).
5. Termos que estavam presentes apenas no resumo (n=129).
6. Termos que estavam presentes apenas nas palavras-chave (n=235).
7. Termos presentes no resumo e que não estavam no título e nas palavras-chave (n=346).

O resultado da filtragem contabilizou 457 artigos, os quais foram recuperados de forma digital e catalogados com o programa gerenciador de bibliografias Mendeley (YAMAKAWA et al., 2014). Foram incluídas 21 publicações originadas de outras fontes de pesquisa, totalizando 478 artigos.

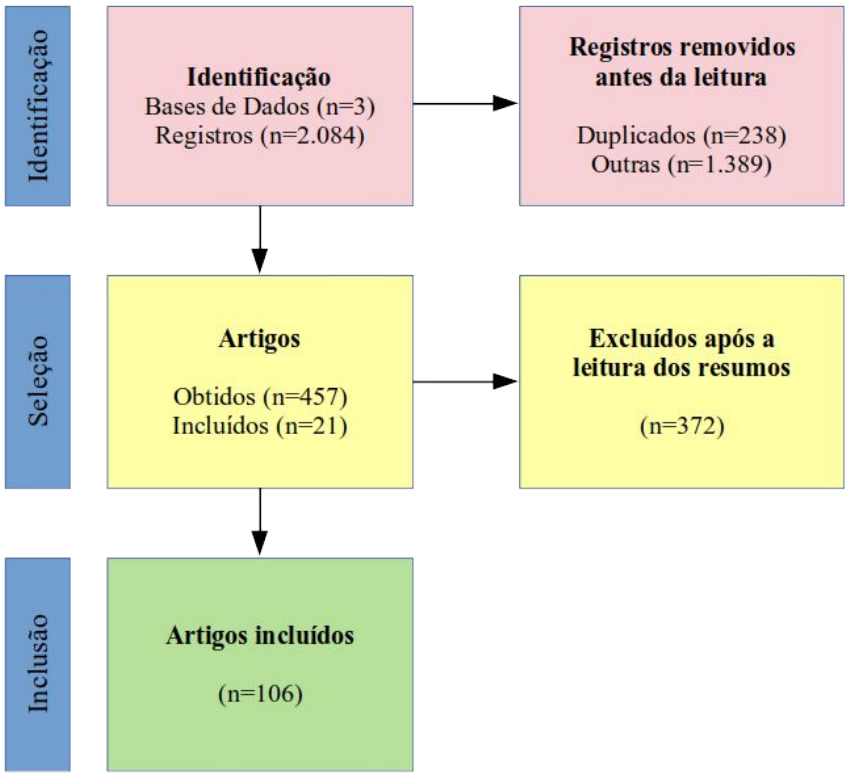
Para leitura dos resumos foram consideradas a relevância do tema, resultados obtidos, contribuição para esta pesquisa, recência da publicação, número de autores no artigo e quantidade de citações.

Após a leitura de cada um dos resumos, foram excluídos 372 artigos. Desta forma, o referencial teórico desta tese é composto por 106 referências.

Em complemento, foram pesquisados artigos no site dos periódicos: *Annals of Internal Medicine* (Ann. Intern. Med.), *Journal of the American Medical Association* (JAMA) e no *PLoS Medicine* (PLoS Med.), visto que, além de manterem publicações da área das Ciências da Saúde, os artigos foram estudados para a compreensão do formato de apresentação/disponibilização. Esta etapa foi importante para estruturar os métodos de extração dos dados neles contidos acerca dos autores e das informações de contribuição.

Na Figura 2.1 é mostrado o PRISMA sobre o levantamento bibliográfico realizado.

Figura 2.1 – *Diagrama de fluxo PRISMA.*



Fonte: Autor

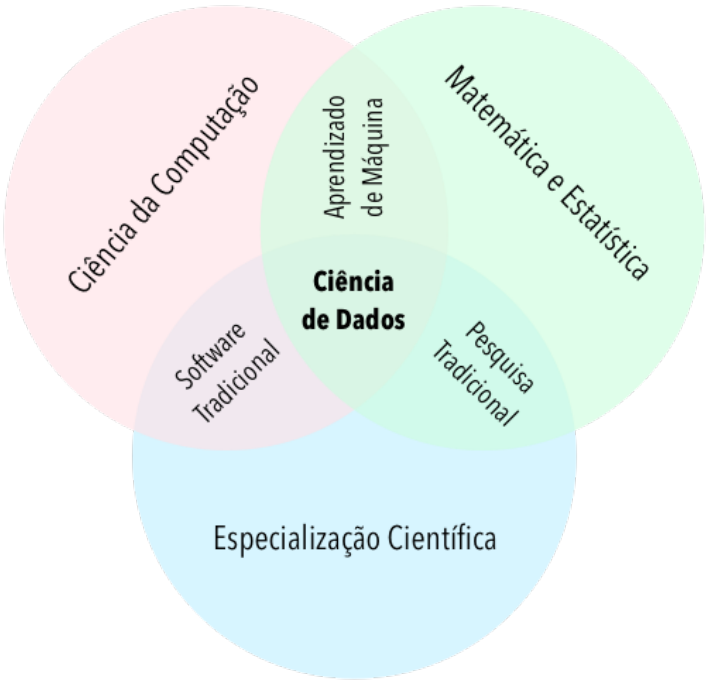
2.2 Ciência de Dados

2.2.1 Visão Geral

O termo Ciência de Dados surgiu na década de 1960 com Peter Naur (NAUR, 1992), sendo descrita como uma filosofia emergente multidisciplinar, que impacta uma mudança cultural nas organizações e no meio acadêmico (AZAM, 2016; HARDIN et al., 2015; CAO, 2017). De outra forma, pode-se dizer que a Ciência de Dados combina estruturas de armazenamento, programação, estatística e técnicas de visualização para tirar conclusões em meio a grandes quantidades de dados (BAUDISCH, 2016). Na Figura 2.2 pode ser visualizada a interdisciplinaridade da ciência de dados.

No campo das ciências, surgiu como uma nova maneira de investigação (WRIGHT, 2014), despertando atenção em físicos, astrônomos, biólogos, entre outros pesquisadores das mais diversas áreas do conhecimento para o desenvolvimento de suas pesquisas (PORTO; ZIVIANI, 2014). Já na indústria, é utilizada nos processos de descoberta do conhecimento para predição de falhas e controle de produção, alinhando-se com o conceito de Indústria 4.0 (WILSON; LADEIRA, 2017). Na área da saúde tem sido aplicada no desenvolvimento de modelos para identificação de riscos cirúrgicos (COREY et al.,

Figura 2.2 – *Diagrama interdisciplinar da ciência de dados (Data Science Venn Diagram).*



Fonte: Adaptado de (DREW, 2020)

2018), estudo e controle do diabetes (KAVAKIOTIS et al., 2017), entre outros. No âmbito governamental, é aplicada, por exemplo, para desenvolver estratégias no setor público, visando melhorar a relação de atendimento com o cidadão (AMADEU, 2017); e detecção de fraudes em energia elétrica (ARAUJO; ALMEIDA; MELLO, 2019), entre outras áreas.

A explosão da internet na década de 1990 e o surgimento da tecnologia de computação em nuvem nos anos seguintes, criaram um ambiente *on-line* favorável para produção e compartilhamento de dados através de tecnologias da informação (TI). Estes ambientes proporcionaram o crescimento acelerado do volume de dados gerado pelas aplicações utilizadas em diversas áreas (MIKALEF et al., 2020; CAMPÊLO et al., 2020).

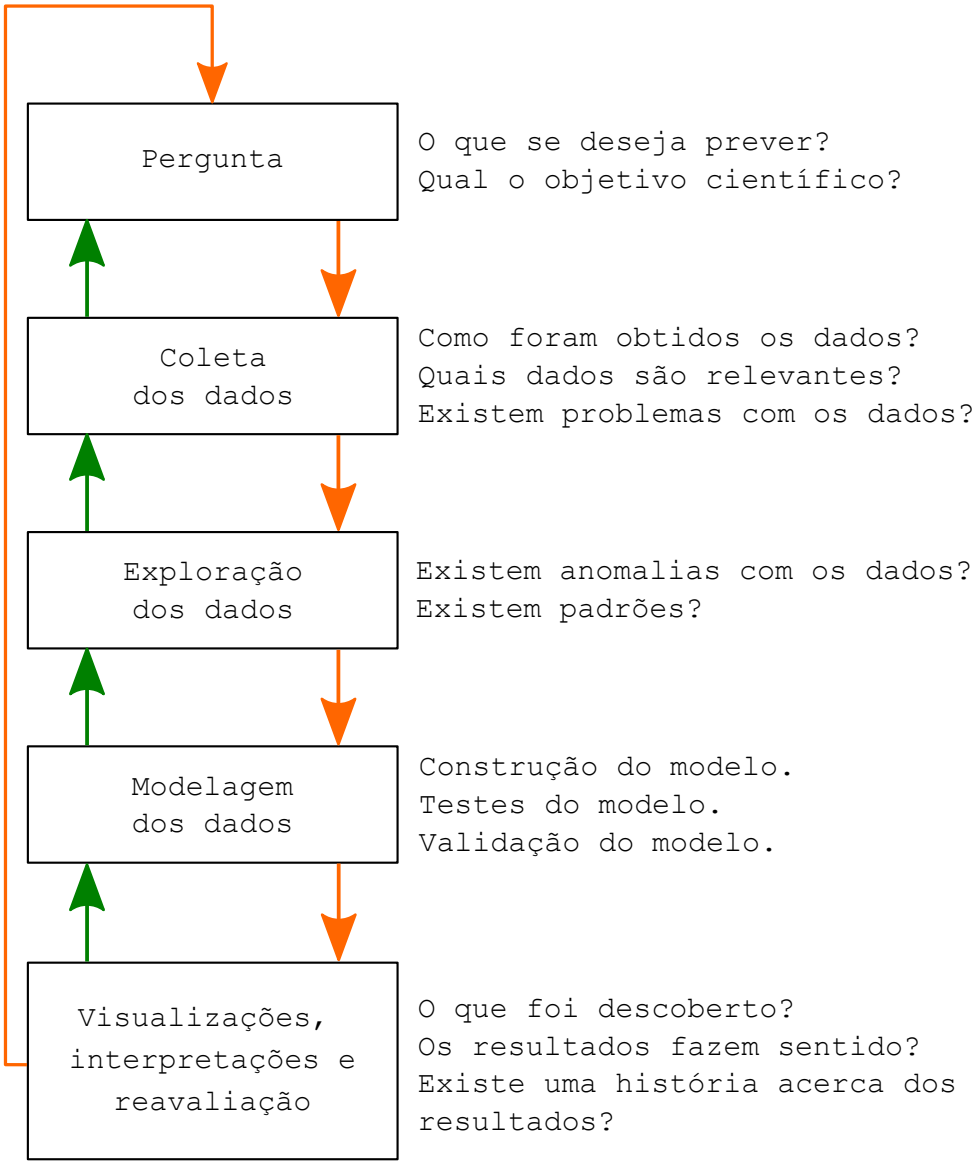
Usuários do mundo todo produzem centenas de dados diariamente utilizando a internet em computadores pessoais, sistemas de telefonia móveis, ou ambos, incluindo dispositivos IoT (*Internet of Things* ou Internet das Coisas). Assim, o armazenamento e o processamento de dados aumentaram em larga escala, oferecendo oportunidades sem precedentes (PRIESTLEY; MCGRATH, 2019). Esse imenso volume de dados necessita de ferramentas e técnicas que permitam realizar sua coleta e tratamento, permitindo analisá-los em busca da descoberta do conhecimento (PORTO; ZIVIANI, 2014).

O uso de ciência de dados para resolução de problemas normalmente ocorre em seis etapas, as quais têm processos bem definidos no fluxo de execução, podendo ocorrer de forma isolada ou simultânea. Essas etapas são conhecidas como *pipeline* de ciência de dados. Cada uma possui sua complexidade e esforço individual, utilizando técnicas e

algoritmos diferentes (FERRAGINA et al., 2015; OLSON et al., 2016; COREY et al., 2018; HALIM; KHAN, 2019). Entretanto, o pré-processamento dos dados é a etapa mais árdua, consumindo horas de trabalho para o tratamento de anomalias nos dados, identificação de parâmetros ausentes ou irrelevantes, valores duplicados ou fora de padrão (BAJPAI; METKEWAR, 2016).

De acordo com Byrne et al. (2017), "Não existe um processo de ciência de dados universalmente acordado.", podendo ocorrer variações nos processos. Essas variações estão relacionadas ao tipo e complexidade do problema a ser resolvido. Na Figura 2.3 pode ser visualizado o fluxo frequentemente utilizado para os processos em ciência de dados.

Figura 2.3 – *Fluxo de processos em ciência de dados.*



Fonte: Adaptado de (BYRNE et al., 2017)

De acordo com a complexidade do objeto de estudo, é necessário realizar subdivisões nos procedimentos para atender aos requisitos. Nas subseções a seguir é descrita cada etapa dos processos.

2.2.2 Formulação da Pergunta

O início do processo ocorre a partir da formulação de uma ou mais perguntas, ou seja, o que se deseja prever e/ou obter. Nessa fase estão presentes os objetivos em se aplicar ciência de dados para a resolução de determinado problema. As perguntas podem estar relacionadas com um problema específico ou um conjunto deles, ou objetivos que podem ser diluídos em várias partes.

2.2.3 Coleta dos Dados

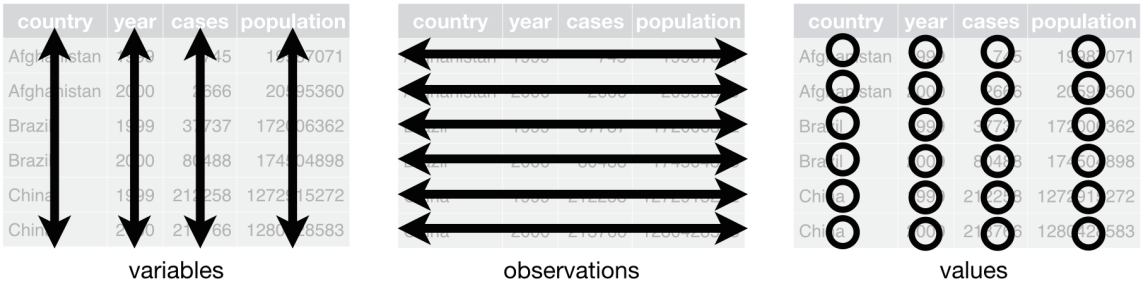
De acordo com Li et al. (2008), o processo de coleta e do tratamento dos dados são as etapas mais complexas e árduas em Ciência de Dados. Coletar dados rotulados é uma tarefa difícil e custosa (BONECHI et al., 2019), uma vez que os mesmos podem se encontrar em formatos diferentes, o que demanda realizar procedimentos para realizar padronização. Normalmente os dados a serem coletados não estão disponíveis em uma só fonte de dados, necessitando realizar o processo repetidas vezes para cada uma delas.

2.2.4 Exploração dos Dados

Após o processo de coleta, é necessário explorá-los, de modo a identificar padrões; realizar sua limpeza para eliminação de anomalias; identificação de erros; registros ausentes, além da tomada de decisões em como lidar com eles; duplicados; corrompidos; padronização das variáveis; entre outros, dependendo do estado em que se encontram. Nesse processo são envolvidas técnicas estatísticas para o tratamento de valores ausentes e *outliers*. Também estão presentes os processos de fusão (*merge*) para o enriquecimento deles, quando necessário. Esses processos permitem a criação de um conjunto de dados ou *dataset*, caracterizado por dispor os dados no formato tabular, Figura 2.4, sendo o principal insumo para a realização dos processos de análise de dados.

Assim, conforme mostrado na Figura 2.4, a primeira linha de um *dataset* é o cabeçalho (*header*), cada coluna uma variável (*feature*), cada linha uma observação (*observation*) e, por fim, cada arquivo/tabela representa um nível de observação (WICKHAM, 2014). Assim, é fundamental definir padrões para manter um elo entre a estrutura dos dados e seus significados, permitindo realizar análises com maior confiabilidade. Essas características são fundamentais para não interferir nos processos e atingir melhores resultados (VELLIDO; MARTÍN-GUERRERO; LISBOA, 2012).

Figura 2.4 – *Características de um dataset.*



Fonte: (SOLTOFF, 2021)

2.2.5 Modelagem dos Dados

Segundo Grus (2019) um modelo é "[...] a especificação de uma relação matemática (ou probabilística) existente entre variáveis diferentes.". Um modelo permite extrair padrões para resolução de problemas por meio do mapeamento do conjunto de dados em relação às variáveis de entrada e à variável de saída, chamados de modelos preditivos. Sua construção é realizada através de técnicas de aprendizagem de máquina que, segundo Mitchell et al. (1997) "[...] é como um programa de computador aprende pela experiência E, com respeito a algum tipo de tarefa T e performance P, se sua performance P nas tarefas em T, na forma medida por P, melhoram com a experiência".

Considerada um ramo da inteligência artificial, tem como finalidade a extração de padrões significativos a partir de exemplos (ERICKSON et al., 2017). Inclui modelos estatísticos, cada qual aplicado de acordo com os dados a serem processados (GRUS, 2019), permitindo a criação de modelos preditivos para a resolução de problemas, de acordo com a natureza e diferentes abordagens (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

Uma vez definido o modelo a ser utilizado, é necessário realizar testes para verificar como está ocorrendo a generalização dos dados, de modo a determinar sua acurácia. Os testes são realizados por validações cruzadas, matrizes de confusão, identificação de falso-positivos e falso-negativos. As técnicas de aprendizagem de máquina são divididas em três categorias: supervisionada, não supervisionada e por reforço (KAVAKIOTIS et al., 2017).

2.2.6 Visualização, Interpretação e Reavaliação do Modelo

Nesta etapa são utilizados gráficos e tabelas com dados estatísticos, de modo a identificar o que foi descoberto. Os valores são utilizados para compreender o comportamento dos dados, além de ilustrar os achados. É a partir das interpretações que as decisões são

tomadas ou então os fenômenos são descritos e explicados (HARDIN et al., 2015).

De acordo com o problema a ser resolvido ou pergunta a ser respondida, faz-se necessário revisitar as etapas anteriores para a reavaliação ou ajuste do modelo de dados desenvolvido. Esse processo é necessário para verificar a ocorrência de erros durante o tratamento dos dados, de modo a garantir sua qualidade (TSAI; CHAN, 2019).

2.3 Autoria e Posicionamento Autoral

O significado do termo autor é amplamente discutido na literatura, visto que existem questões éticas envolvidas na autoria fantasma e honorária (KUMAR, 2018). Segundo Lozano (2014) a maioria dos artigos científicos até o início do século 20 continha apenas um autor. Entretanto, com o avanço da ciência nos últimos 30 anos, a colaboração entre diferentes áreas do conhecimento tem colaborado para o crescimento no número de autores listados nos artigos (ROSENZWEIG et al., 2008; PATIENCE et al., 2019). Assim, o papel do autor em uma publicação científica não se limita apenas a contribuir ativamente para o desenvolvimento dos estudos, seguindo as diretrizes e boas práticas estabelecidas nos diversos campos da ciência, mas também ser responsável pelo seu conteúdo (HILÁRIO et al., 2018).

A identificação do posicionamento autoral em um artigo ocorre por meio da informação dos nomes dos autores em seu início, denominada de *"Author byline"* ou simplesmente *"byline"* (YANG; WOLFRAM; WANG, 2017), estando associada à relevância na publicação. Entretanto, de acordo com a literatura, não há um consenso de como é definido este posicionamento. Assim, estudos sobre a questão do posicionamento autoral têm sido desenvolvidos devido à existência de questões éticas relacionadas à distribuição da autoria para pesquisadores com menor ou nenhuma participação no trabalho (JONES; MCCULLOUGH; RICHMAN, 2005).

Zbar e Frank (2011) estudaram a posição autoral de uma pesquisa para verificar a percepção dos autores sobre a primeira e última posição no *byline*. Eles identificaram que os primeiros autores têm maior probabilidade de desenvolver atividades relacionadas à condução do estudo, redação e contribuições mais relevantes. Por outro lado, os últimos autores estão mais envolvidos com a arrecadação de fundos, mentoria ou chefia do laboratório. No estudo de (MATTSSON; SUNDBERG; LAGET, 2011), foi verificada a relação entre a ordem de autoria e o autor correspondente, o qual tende a ocupar a primeira posição no *byline*, evidenciando que o autor correspondente tem grande participação na questão intelectual no desenvolvimento do artigo, embora o número de autores deva ser considerado para maior precisão na mensuração da contribuição do autor.

No estudo de Igou e Tilburg (2015), realizado em publicações na área da psicologia, foi verificado o posicionamento autoral analisando a quantidade letras no nome do meio dos autores. Os resultados sugerem que as iniciais do nome do meio estão associa-

das às ordens de autoria, indicando que uma quantidade maior delas tende a posicionar um autor a ocupar a primeira posição no *byline*. Segundo os autores, isto pode representar simbolicamente a performance intelectual, em artigos de periódicos acadêmicos em psicologia.

Já no estudo de Russell et al. (2019) foram exploradas as mudanças nas variáveis bibliométricas nos últimos 30 anos em quatro importantes periódicos científicos musculoesqueléticos. Os resultados evidenciaram que a posição do autor correspondente mudou progressivamente da primeira para a última posição no *byline* enquanto o número de autores aumentava. Desta forma, os autores correspondentes são frequentemente considerados aqueles que geraram a ideia de pesquisa ou em cujo laboratório a pesquisa foi conduzida, corroborando com (MATTSSON; SUNDBERG; LAGET, 2011). Assim, os autores considerados mais antigo em comparação com o primeiro autor, tendem a ocupar a última posição. Por fim, a cultura científica sobre posicionamento autoral deve ser considerada em coautorias internacionais por meio da avaliação das colaborações (ARIA; MISURACA; SPANO, 2020).

2.4 Posicionamento Autoral e Contribuições

Um dos critérios para determinar o posicionamento autoral é a contribuição científica, que identifica quais tarefas foram realizadas pelo autor para o desenvolvimento da pesquisa (YANK; RENNIE, 1999; LI et al., 2013). Geralmente, as principais contribuições em artigos científicos ocorrem pelos autores posicionados na primeira e na última posição na ordem de autoria (TSCHARNTKE et al., 2007). A primeira normalmente é ocupada por um aluno de doutorado ou pós-doutorado e a última pelo diretor do laboratório ou responsável pela pesquisa, enquanto os autores do meio colaboram com contribuições menos expressivas (MONGEON et al., 2017). Tradicionalmente o primeiro autor é o que mais contribui no artigo, recebendo a maior parte do crédito (TSCHARNTKE et al., 2007). Apesar da existência de práticas para definir o posicionamento autoral, ainda há desafios em classificar o grau de colaboração e a responsabilidade de cada autor, visto que nem todos colaboram da mesma forma (CHANG, 2019).

Para auxiliar neste processo, está disponível o sistema CRediT *(Contributor Roles Taxonomy)* [1], que é uma taxonomia de alto nível, composta por 14 categorias: *Conceptualization*; *Data curation*; *Formal Analysis*; *Funding acquisition*; *Investigation*; *Methodology*; *Project administration*; *Resources*; *Software*; *Supervision*; *Validation*; *Visualization*; *Writing — original draft*; e *Writing — review & editing*. Essas categorias podem ser usadas para representar as funções normalmente desempenhadas por contribuintes para a produção científica acadêmica, onde cada uma descreve a contribuição específica de cada autor na produção (MCNUTT et al., 2018; DECULLIER; MAISONNEUVE, 2019).

[1] Disponível em: <http://credit.niso.org/>

Principalmente na área de ciências da saúde, a descrição das contribuições de cada autor segue as diretrizes recomendadas pelo ICMJE (SAUERMANN; HAEUSSLER, 2017) para as subáreas da: medicina, enfermagem, veterinária, odontologia e áreas afins (BATES et al., 2004; KUMAR, 2018; ICMJE, 2020), classificadas entre as 27 áreas temáticas do SCImago (SCIMAGO, 2020). Embora as contribuições sejam categorizadas nos periódicos desta área, a nomenclatura das categorias utilizadas por eles divergem em relação às propostas pelo CRediT, gerando despadronização. Publicações realizadas em periódicos que seguem as diretrizes do ICMJE solicitam informações sobre as contribuições dos autores dentro de quatro critérios, descritos no Quadro 2.1.

Quadro 2.1 – *Critérios de contribuição recomendados pelo ICMJE.*

#	**Descrição**
Critério 1	Contribuir substancialmente para a concepção ou planejamento do trabalho, ou aquisição, análise ou interpretação de dados para o trabalho.
Critério 2	Redigir o trabalho ou revisá-lo criticamente para apresentar conteúdo intelectual importante.
Critério 3	Aprovar a versão final para publicação.
Critério 4	Confirmar ser o responsável por todos os aspectos do trabalho, garantindo que as questões relacionadas à precisão ou integridade de qualquer parte do trabalho sejam investigadas e resolvidas de forma adequada.

Fonte: Autor

Apesar da existência do sistema CRediT e das recomendações do ICMJE, para a área das ciências da saúde, ainda existe uma ampla discussão em como determinar o posicionamento autoral. Neste sentido, diversos estudos foram conduzidos com o objetivo de compreender e sugerir métodos para a realização do posicionamento autoral relacionado com as contribuições científicas em publicações. Assim, Tscharntke et al. (2007) estudaram a posição autoral em relação às contribuições em publicações com diversos autores, sugerindo que o posicionamento deveria ocorrer em uma escala de importância decrescente em relação à contribuição, estimulando a pesquisa colaborativa (LAKE; DIEGO, 2010). Por exemplo, o primeiro autor deve receber crédito por todo o impacto, o segundo a metade e assim por diante até a décima posição. A partir daí, a contribuição de cada autor deve ser atribuída em 5%.

No estudo de Costas e Bordons (2011), foi constatado que a posição autoral nas áreas de Biologia, Biomedicina, Ciência dos Materiais e Recursos Naturais é influenciada pela idade e posição profissional. Além disso, pesquisadores mais jovens tendem a aparecer na primeira posição com contribuições mais significativas, enquanto pesquisadores mais

experientes têm maior inclinação a estarem presentes na última posição, corroborando com (MONGEON et al., 2017). Com outra abordagem, Liu e Fang (2012) propuseram modificar o índice "H" para atribuir créditos e posicionar o autor de acordo com um índice obtido pelas contribuições no trabalho. Neste estudo, foram comparadas várias formas de posicionamento com base nas contribuições e no número de autores listados no artigo.

Por outro lado, Abramo, D'Angelo e Rosati (2013) destacaram a importância do número de coautores e a real contribuição de cada um para as publicações científicas em pesquisas institucionais na área das ciências da vida. O estudo constatou que o uso de indicadores bibliométricos não é aconselhável para definir a posição autoral, pois o fracionamento de autoria apresentou menor efeito quando aplicado às citações. Portanto, segundo os autores, a determinação da posição autoral deve considerar a produtividade em pesquisas institucionais com outras organizações internacionais. Essa prática, aliada à divisão igualitária da participação entre os autores, minimiza os créditos para quem efetivamente desenvolveu a pesquisa, superestimando quem pouco colaborou (HAGEN, 2014). Ainda, de acordo com Abramo e D'Angelo (2014), "Medidas de produtividade precisam considerar as contribuições fracionárias de unidades individuais para os resultados." A ponderação dessas medidas pode desempenhar a função de classificação no *byline*, mas elas precisam ser explicitamente definidas. Para Jian e Xiaoli (2013), contagens harmônicas podem distinguir entre cientistas que são frequentemente listados como principais contribuintes e aqueles regularmente listados como coautores, desencorajando práticas de publicação antiéticas, como autoria fantasma e autoria atribuída sem contribuições (*gift authorship*).

Orientações formais para a composição da ordem autoral foram propostas por Kozma et al. (2014), aplicando um sistema de ponderação de contribuição. O estudo mostrou que periódicos de alto impacto aplicam mais rigor para definir os critérios de autoria e contribuição. No entanto, não possuem uma definição ou recomendação formal para o posicionamento dos autores. Assim, a ordem dos autores é influenciada por aspectos relacionados a área de pesquisa, como na das ciências sociais, comportamentais e biomédicas (WALTERS, 2016). Além das áreas, há também questões relacionadas a contextos nacionais, institucionais e histórico (PONTILLE, 2003).

No trabalho de Larivière et al. (2016), os resultados evidenciaram a presença das maiores contribuições para os primeiros e últimos autores, mostrando que a divisão do trabalho na produção de conhecimento é um fator importante e crescente. Para mensurar o nível de contribuição, Rahman et al. (2017) propuseram a criação de uma lista de contribuições para apoio logístico, relacionadas à obra para indicar a posição do autor, com o objetivo de medir o esforço relativo de cada autor nas atividades desempenhadas para produção do artigo. De forma semelhante, Yang, Wolfram e Wang (2017) sugeriram recomendações para a implementação de listas estruturadas de contribuições cruzadas, incluindo critérios de contribuição para autores. Por fim, no estudo de McNutt et al.

(2018) são propostas mudanças nas políticas e procedimentos de autoria de periódicos para fornecer informações sobre qual autor é responsável por quais contribuições, justificando a obtenção sobre o crédito de autoria.

Embora os trabalhos apresentem soluções diversas, ainda se encontra em aberto como definir o posicionamento autoral em relação às contribuições em publicações científicas.

3 IDENTIFICAÇÃO DAS FONTES DE DADOS

Resumo do capítulo

Neste capítulo são descritos os processos para identificação das fontes de dados de periódicos (SCImago), artigos (Scopus), autores (Scopus) e de periódicos (Ann. Intern. Med., JAMA e PLoS Med.) utilizadas neste trabalho. Ademais, descreve suas características e importância para realização desta pesquisa.

3.1 Visão Geral

Em Ciência de Dados, a primeira etapa a ser realizada após a determinação do objetivo (pergunta) é a da identificação das fontes de dados, conforme apresentado no Capítulo 2. Esta etapa pode ocorrer de diferentes formas, dependendo do tipo dos dados ou origem deles. Segundo Aalst (2016), os dados podem ser originados de única uma fonte ou de múltiplas origens.

Desta forma, como o objetivo desta pesquisa é estudar a relação entre posição autoral e as contribuições em publicações científicas, estão envolvidos diferentes objetos (periódicos, artigos, autores e contribuições). Para identificar os diferentes objetos e fontes de dados desse estudo sobre a relação da posição autoral e a lista de contribuições, foi selecionada inicialmente a fonte de dados SCImago, porque nela estão presentes os principais periódicos do mundo (D'UGGENTO; RICCI; TOMA, 2016).

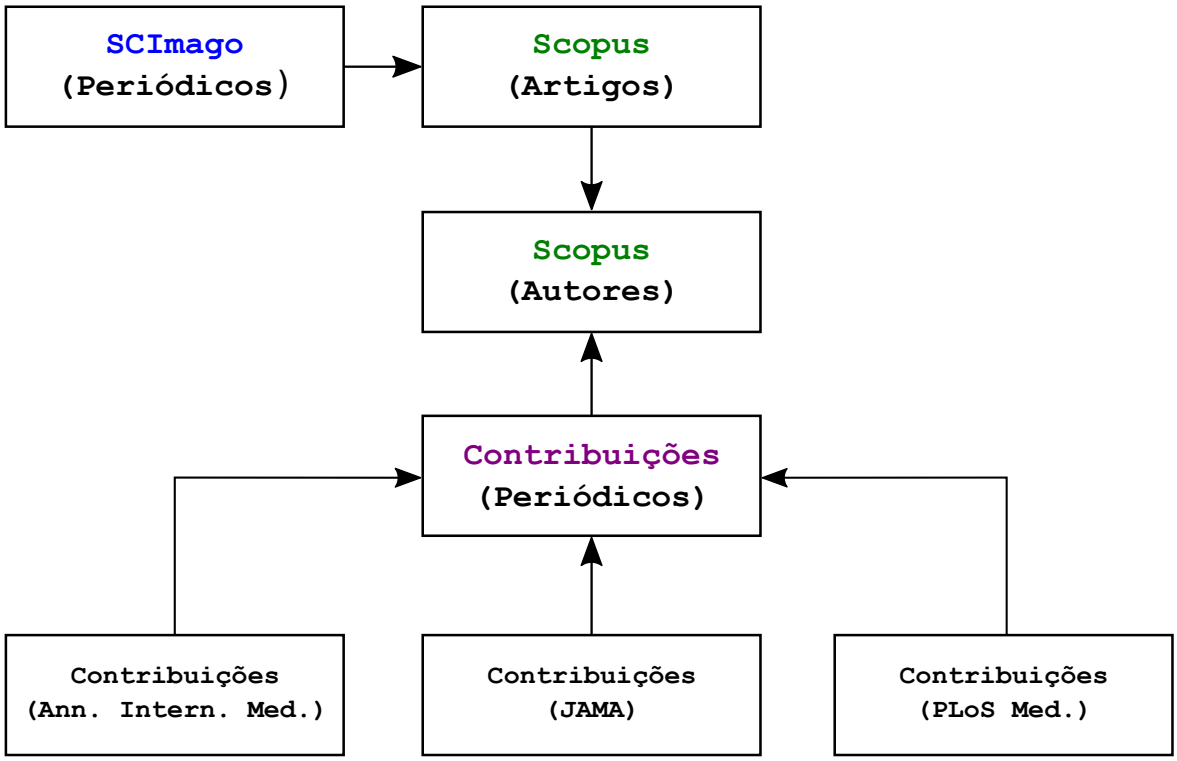
O SCImago disponibiliza dados sobre as 27 áreas temáticas do conhecimento, a partir das quais é possível identificar dados sobre periódicos, onde foi selecionada a área de ciências biológicas e medicina (*Medicine*), contendo 6.824 periódicos (dados disponíveis do ano de 2.020), pois nesta área são oferecidas informações sobre as contribuições dos autores.

Com o objetivo de identificar periódicos na área de ciências biológicas e medicina, foi realizada uma ampla pesquisa na literatura, que apontou para três proeminentes periódicos de primeira linha: Ann. Intern. Med., JAMA e PLoS Med. Esses periódicos foram citados ou utilizados em estudos investigativos sobre posicionamento, ou contribuição autoral (TEIXIERA DA SILVA; DOBRÁNSZKI, 2016; PERNEGER et al., 2017; YANG; WOLFRAM; WANG, 2017; RASMUSSEN et al., 2018; PATIENCE et al., 2019).

Os artigos e os dados dos seus autores, dos periódicos citados, podem ser acessados na fonte de dados Scopus, que é a maior base de dados bibliográfica sobre publicações científicas, disponibilizando resumos, citações e textos completos de periódicos acadêmicos com revisão por pares (MAÑANA-RODRÍGUEZ, 2015). Os dados do Scopus são utilizados pelo SCImago para composição do seu *ranking*. Assim, através de uma busca, realizada em novembro/2019, foram localizados 12.106 artigos publicados no período entre os anos 2000 e 2019, mostrando condições e viabilidade para coleta dos exemplares.

Na Figura 3.1 é apresentado o diagrama da relação entre as fontes de dados identificadas para periódicos (SCImago), artigos e autores (Scopus) e as contribuições (periódicos Ann. Intern. Med., JAMA e PLoS Med.).

Figura 3.1 – *Relação entre as fontes de dados identificadas para coleta de dados.*



Fonte: Autor

Nas seções a seguir são detalhados os procedimentos para identificação das fontes de dados citadas, visto que suas características e os processos de aquisição diferem entre elas.

3.2 Fonte de dados SCImago Journal & Country Rank

O SCImago é um portal que fornece indicadores de produções científicas contidas no banco de dados do Scopus (VILLASEÑOR-ALMARAZ et al., 2019), sobre os principais periódicos do mundo (D'UGGENTO; RICCI; TOMA, 2016). Os indicadores são extraídos de mais de 42.000 títulos agrupados em 27 áreas temáticas, subdivididas em 313 categorias de assuntos específicos, com mais de 5.000 editores internacionais e métricas de desempenho de 239 países.

Dado que o objeto de estudo neste trabalho é compreender a relação entre o posicionamento autoral e as contribuições na área de ciências biológicas e medicina (*Medicine*),

o número de títulos disponíveis no SCImago nesta área (6.824), em 2020, mostrou ser promissor para a realização do estudo. Na Tabela 3.1 é mostrado o total de títulos em cada uma das 27 áreas do SCImago, com destaque para *Medicine*, foco deste trabalho.

Tabela 3.1 – *Quantidade de títulos em cada uma das 27 áreas do SCImago.*

Descrição	**Quantidade**
Medicine	**6.824**
Social Sciences	6.694
Arts and Humanities	4.208
Engineering	2.675
Agricultural and Biological Sciences	2.164
Biochemistry, Genetics and Molecular Biology	2.011
Computer Science	1.661
Mathematics	1.533
Environmental Science	1.467
Business, Management and Accounting	1.427
Psychology	1.270
Materials Science	1.188
Earth and Planetary Sciences	1.138
Economics, Econometrics and Finance	1.111
Physics and Astronomy	1.094
Chemistry	843
Pharmacology, Toxicology and Pharmaceutics	673
Nursing	625
Chemical Engineering	600
Neuroscience	568
Health Professions	568
Immunology and Microbiology	555
Energy	457
Decision Sciences	427
Veterinary	252
Dentistry	201
Multidisciplinary	133
Total de periódicos	**42.367**

Fonte: Autor

Para aquisição da base de dados, foi aplicado um filtro no site do SCImago[1], apenas para publicações classificadas como *journal*. A estrutura do arquivo obtido, que relaciona dados referentes aos periódicos, dispunha de 18 variáveis, descritas no Quadro 3.1.

[1]Recurso disponível através do endereço: <https://www.scimagojr.com/>

Quadro 3.1 – *Variáveis disponíveis no arquivo que relacionam dados referentes aos periódicos.*

Seq.	**Variável**	**Descrição**
1	*Rank*	Posição do periódico no *SCImago*
2	*Sourceid*	ID referente ao periódico na *Scopus*
3	*Title*	Título do periódico
4	*Type*	Tipo de publicação
5	*ISSN*	Número do ISSN do periódico
6	*SJR*	Indicador bibliométrico de influência
7	*SJR Best Quartile*	Classificação do quartil
8	*H index*	Índice H do periódico
9	*Total Docs. (2017)*	Número de documentos indexados em 2017
10	*Total Docs. (3years)*	Total de documentos nos últimos três anos
11	*Total Refs.*	Total de referências
12	*Total Cites (3years)*	Total de citações nos últimos três anos
13	*Citable Docs. (3years)*	Número de documentos citados nos últimos três anos
14	*Cites / Doc. (2years)*	Citações pelo número de documentos nos últimos dois anos
15	*Ref. / Doc.*	Referências pelo número de documentos
16	*Country*	País de origem da publicação
17	*Publisher*	Editora
18	*Categories*	Categorias de publicação

Fonte: Autor

Na seção seguinte, são descritas as fontes de dados sobre artigos e autores (Scopus).

3.3 Fonte de dados Scopus

O Scopus é a maior base de dados bibliográfica sobre publicações científicas, disponibilizando resumos, citações e textos completos de periódicos acadêmicos com revisão por pares (MAÑANA-RODRÍGUEZ, 2015). Oferece um panorama abrangente da produção científica mundial nas áreas de ciências, medicina, tecnologia, artes e humanidades. Disponibiliza ferramentas para buscas, permitindo visualizar, analisar e monitorar dados sobre publicações. Por ser uma plataforma independente, o acesso aos dados das publicações são fornecidos parcialmente para visitantes, enquanto para as Instituições de Ensino Superior (IES) e demais assinantes comerciais, integralmente.

As informações sobre buscas são apresentadas em um formato não estruturado, através da listagem em uma página *web*, mostrando variáveis como nome do artigo, periódico e data da publicação. Outra forma de visualização é por exportação manual dos dados, que permite flexibilidade para manipulação e visualização posterior. Para isto, foram realizadas buscas e os resultados exportados manualmente, onde se verificou a existência de variáveis descrevendo periódicos, artigos e autores, insumo para este trabalho. Portanto,

o Scopus foi considerado como uma fonte de dados para este trabalho.

O Scopus se relaciona com o SCImago por meio do ISSN dos periódicos, os quais são informados como um dos critérios para busca de artigos. Desta forma, para realizar a aquisição dos dados de artigos e dos autores, são necessários diversos procedimentos, os quais são detalhados no Capítulo 4.

3.4 Fonte de dados dos Periódicos Ann. Intern. Med., JAMA e PLoS Med.

Os periódicos Ann. Intern. Med., JAMA e PLoS Med. adotam os critérios recomendados pelo ICMJE para declaração das contribuições. Dentro de cada critério existem categorias onde são declaradas as contribuições. Entretanto, cada periódico adota um sistema de nomenclatura para descrevê-las. Como exemplo, no Critério 1 (Quadro 2.1), os periódicos denominam a categoria referente aos dados como: *"Collection and assembly of data"* no Ann. Intern. Med.; *"Acquisition data"* e *"Acquisition, analysis, or interpretation data"* no JAMA e *"Data curation"* no PLoS Med. Apesar da palavra *"data"* estar presente em ambas as denominações, as descrições não seguem um padrão, tornando o processo de comparação ou avaliação entre os periódicos complexas, principalmente quando realizado em grande número de exemplares.

Nas subseções a seguir é apresentada uma visão dos periódicos e sobre como cada um dispõe suas informações sobre contribuições nas categorias, descritas originalmente em língua inglesa.

3.4.1 Periódico Annals of Internal Medicine (Ann. Intern. Med.)

Fundado em 1927 pelo *American College of Physicians* (ACP), o periódico Ann. Intern. Med. tem como missão promover a excelência na medicina, permitindo que médicos e outros profissionais da área da saúde estejam sempre bem informados. É uma das principais revistas médicas que publica estudos e revisões com mérito científico e relevância clínica (FAGGION; BAKAS; WASIAK, 2017).

Em 2019, segundo o próprio site, o periódico possuía 159.000 membros do ACP, incluindo profissionais de saúde e pesquisadores em todo o mundo. Segundo a Clarivate Analytics (2018), em 2019 teve seu fator de impacto (FI) avaliado em $21,317$ com 58.033 citações, sendo o periódico de medicina interna geral mais citado e um dos mais influentes do mundo.

Suas publicações passam por um rigoroso processo de revisão por pares, além de seguir as diretrizes do ICMJE. Assim, recomenda basear as contribuições de autoria em dez categorias, mostradas no Quadro 3.2.

Quadro 3.2 – *Categorias de contribuição utilizadas pelo periódico Ann. Intern. Med.*

#	**Denominação das Categorias**
1	*Conception and design*
2	*Analysis and interpretation of the data*
3	*Drafting of the article*
4	*Critical revision for important intellectual content*
5	*Final approval of the article*
6	*Statistical expertise*
7	*Obtaining of funding*
8	*Administrative technical or logistic support*
9	*Collection and assembly of data*
10	*Provision of study materials or patients*

Fonte: (ANNALS, 2021)

3.4.2 Journal of the American Medical Association (JAMA)

O JAMA é um periódico internacional de medicina geral fundado em 1883, revisado por pares e que promove a ciência, medicina e a melhoria da saúde pública, abordando amplas especialidades para muitos leitores (FIGER et al., 2018).

Segundo o próprio site, é membro do *JAMA Network*, um consórcio de publicações médicas gerais e especializadas, revisadas por pares. Em 2019 foi citado 158.632 vezes, segundo o (Clarivate Analytics, 2018) com FI avaliado em $45,540$. Com 7.9 milhões de visitas por ano e mais de 9.8 milhões de artigos baixados, segue as diretrizes do ICMJE para declaração das contribuições, que são divididas em 10 categorias, mostradas no Quadro 3.3.

Quadro 3.3 – *Categorias de contribuição utilizadas pelo periódico JAMA.*

#	**Denominação das Categorias**
1	*Study concept and design*
2	*Critical Revision the manuscript for important intellectual content*
3	*Drafting of the manuscript*
4	*Statistical analysis*
5	*Study supervision*
6	*Administrative, technical, or material support*
7	*Obtained funding*
8	*Acquisition, analysis, or interpretation data*
9	*Analysis and interpretation data*
10	*Acquisition of data*

Fonte: (JAMA, 2021)

3.4.3 Periódico PLoS Medicine (PLoS Med.)

Criado em 2004, o PLoS Med. é um periódico que publica artigos de interesse geral sobre determinantes biomédicos, ambientais, sociais e políticos da saúde, enfatizando o trabalho que promove a prática clínica, a política de saúde ou a compreensão fisiopatológica para beneficiar a saúde em uma variedade de contextos. Seus critérios para publicação são altamente seletivos, sendo que apenas 10% dos artigos submetidos evoluem para publicação. Para submissões que necessitam de aprofundamento no assunto, o periódico aplica rigorosas revisões editoriais e por pares, supervisionadas por uma equipe de editores profissionais e, no caso de assuntos específicos, por um editor acadêmico independente especialista na área. É uma das principais revista médicas gerais, exigindo expressamente o compartilhamento de dados como condição para publicação de ensaios clínicos randomizados (ECR) (RASMUSSEN et al., 2018).

Segundo o site, adota o sistema de revisão por pares em suas publicações, visando manter os mais altos padrões éticos nas publicações médicas. Inclui ainda o gerenciamento e a divulgação sobre conflitos de interesses em relatórios, revisões e publicações; transparência no processo de revisão e publicação; compartilhamento de dados e reutilização do trabalho; retenção de direitos autorais pelos autores; publicação de acesso aberto sem restrições de disponibilidade e disseminação; e evitar os conflitos de interesse apresentados pela publicidade de medicamentos e dispositivos médicos e venda exclusiva de reimpressões. No ano de 2019 foi avaliado com o fator JCR em $10,5$; sendo citado 32.312 vezes (Clarivate Analytics, 2018). Este periódico segue as diretrizes do ICMJE para declaração das contribuições, que são divididas em 13 categorias, mostradas no Quadro 3.4.

Quadro 3.4 – *Categorias de contribuição utilizadas pelo periódico PLoS Med.*

#	**Denominação das Categorias**
1	*Conceptualization*
2	*Data curation*
3	*Formal analysis*
4	*Funding acquisition*
5	*Investigation*
6	*Methodology*
7	*Project administration*
8	*Resources*
9	*Software*
10	*Supervision*
11	*Validation*
12	*Writing – original draft*
13	*Writing – review & editing*

Fonte: Autor

Portanto, com base na literatura e segundo os indicadores bibliométricos, mostrados na Tabela 3.2, os periódicos Ann. Intern. Med., JAMA e PLoS Med. foram selecionados como fonte de dados para a realização desta pesquisa.

Tabela 3.2 – *Periódicos selecionados para este estudo com os índices do Journal Citation Reports (JCR) e do SCImago Journal & Country Rank (SJR), referentes ao ano de* 2018.

Periódico	**JCR**	**SJR**	**ISSN**
Ann. Intern. Med.	19,31	7,338	1539-3704
JAMA	20,80	7,477	0098-7484
PLoS Med.	11,04	6,626	1549-1277

Fonte: Autor

De acordo com avaliação sobre os dados do SCImago, Scopus e dos periódicos Ann. Intern. Med., JAMA e PLoS Med., essas fontes de dados dispõe das informações necessárias para o desenvolvimento desta pesquisa. No Capítulo 4, a seguir, são descritos em detalhes os métodos para coleta dos dados nas referidas fontes.

4 COLETA DE DADOS DE ARTIGOS, AUTORES E CONTRIBUIÇÕES

Resumo do capítulo

Neste capítulo são apresentados os processos de coleta, preparação e limpeza dos dados de artigos (Scopus), autores (Scopus) e contribuições (Periódicos Ann. Intern. Med., JAMA e PLoS Med.), além da construção de robôs e algoritmos auxiliares para realização dessas tarefas. Tais procedimentos visam responder às seguintes perguntas de pesquisa: (1) "Como automatizar o processo de coleta de dados (Data Gathering) em bases científicas indexadas através de robôs (bots) usando técnicas de raspagem de dados (Data Scraping)?" e (2) "Como melhorar a preparação dos dados coletados (Data Clean) para eliminação de anomalias, identificação de erros e duplicações de forma automática?", referentes a lacuna de pesquisa 2 - "Identificar padrões para os processos de coleta, limpeza e preparação dos dados de bases científicas indexadas", descrita na Seção 1.2.

4.1 Visão Geral

De acordo com Li et al. (2008), o processo de coleta e do tratamento dos dados são as etapas mais complexas e árduas em ciência de dados. Uma vez identificadas a partir do SCImago as fontes de dados do Scopus e os periódicos Ann. Intern. Med., JAMA e PLoS Med., Capítulo 3, foram realizadas análises para verificar como realizar a aquisição dos dados dos autores e contribuições dos artigos nas fontes citadas.

Os dados sobre artigos foram adquiridos da fonte de dados do Scopus utilizando *scripts* de busca e exportação de arquivos da própria plataforma. Esses continham informações sobre 12.106 artigos, representados por 4.118 do Ann. Intern. Med., 5.522 do JAMA e 2.466 do PLoS Med. Após o processo de limpeza dos dados (descrito adiante), o *dataset* criado contabilizou 2.024 artigos com 20.098 autores. Nos arquivos estão presentes algumas variáveis que descrevem os dados como: nome dos autores (*"Authors"*), ID dos autores no Scopus (*"Author(s) ID"*), número de citações (*"Cited by"*), ano de publicação (*"Year"*) e "doi".

A partir desses dados foi realizado um processamento para a individualização dos autores, de modo que cada linha registrou informações do artigo para esta variável de forma compactada. Este formato de armazenamento dispõe todos os autores em uma mesma variável, separados por ponto e vírgula, que requereu o tratamento para sua separação. Para isto, foi criado um algoritmo escrito na linguagem `Python`, executado sobre os 2.024 artigos, utilizando a variável *"Author(s) ID"* como referência. O resultado do processamento gerou um novo *dataset* com 20.098 registros.

A partir das características do *dataset*, foi desenvolvido um robô, escrito na linguagem `Python`, para a extração dos registros dos autores da fonte de dados do Scopus. Segundo

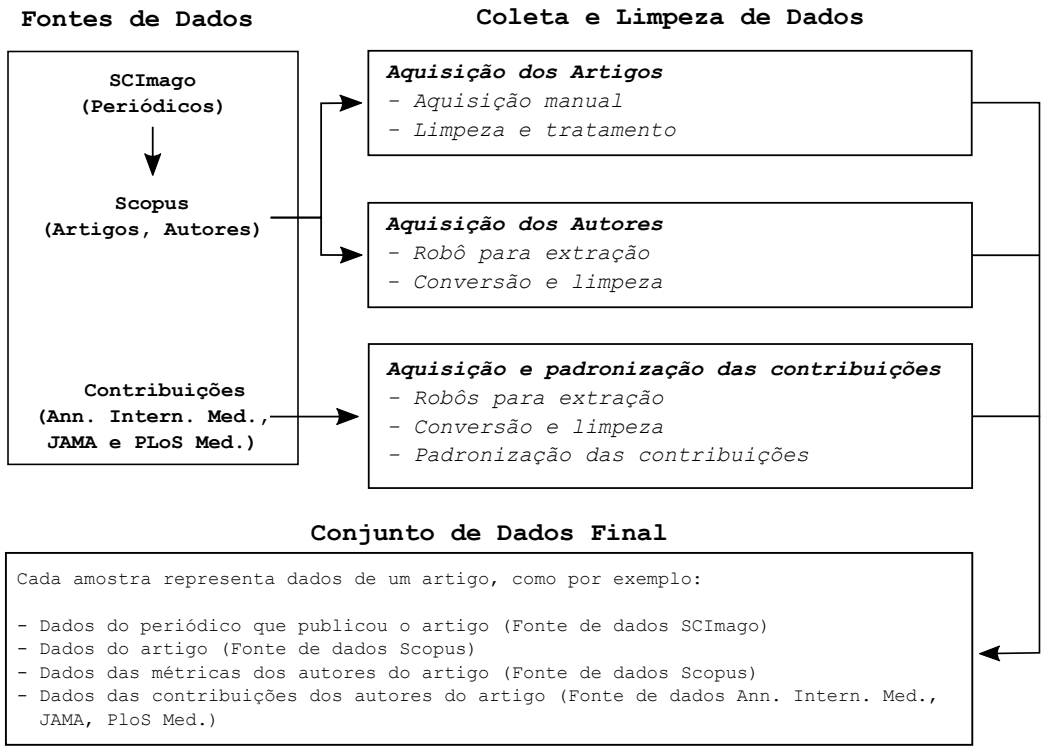
Woolley (2016), robôs são um tipo específico de agente de *software* automatizado escrito para coletar informações, tomar decisões e interagir, imitando usuários reais.

Os dados disponíveis estavam no formato semiestruturado *Extensible Markup Language* (XML), que permite a definição de regras para codificar diferentes categorias de documentos de forma semiestruturada (TEKLI, 2016), relacionando dados dos autores em 19 variáveis. Posteriormente foi criada uma ferramenta para automatizar o processo de conversão dos dados do formato XML (*parser*) para o formato tabular *Comma-separated values* (CSV), formato que armazena os dados separados por vírgula.

Por fim, foram desenvolvidos três robôs para realizar a extração dos dados das contribuições autorais das fontes de dados dos periódicos Ann. Intern. Med., JAMA e PLoS Med. Ao final do processo, foram extraídos 17.385 arquivos. Cada periódico possui um número diferente de variáveis sobre as contribuições dos autores, conforme citado na Subseção 3.4, sendo 10 variáveis para o periódico Ann. Intern. Med., 10 para o JAMA e 13 para o PLoS Med.

Uma vez concluídas as extrações e processamentos necessários, foi realizada uma organização dos dados de todas as fontes de dados, que originou um *dataset* consolidado. Na Figura 4.1 é mostrada uma visão geral das etapas para a coleta e limpeza dos dados até a composição do *dataset* final para a realização do estudo.

Figura 4.1 – *Visão geral das etapas para coleta dos dados até a composição do dataset final.*



Fonte: Autor

Nas seções a seguir são descritos os processos para a aquisição e processamento dos dados de artigos, autores e, por fim, das contribuições, incluindo os algoritmos dos robôs desenvolvidos para este fim.

4.2 Coleta de Dados dos Artigos no Scopus

Os dados brutos disponibilizados pelo Scopus sobre artigos e autores não estão acessíveis de forma direta, além de não estarem estruturados de forma tabular, sendo necessária a obtenção por etapas e a conversão para o formato tabular.

A relação entre os dados do SCImago e do Scopus ocorre por meio do ISSN dos periódicos, os quais são informados como um dos critérios para busca de artigos. Durante a avaliação dos registros do SCImago, foi identificada que a variável ISSN dos periódicos Ann. Intern. Med. e PLoS Med. apresentavam um segundo valor, que corresponde ao seu eISSN (*Electronic International Standard Serial Number*), ambos separados por vírgula. Esta ocorrência exigiu a realização de um processamento para a separação e tratamento desses dados, de modo a adequá-los à formatação de oito dígitos separados por hífen. Este ajuste foi necessário, já que o Scopus armazena os dados destas variáveis neste formato.

Após a execução dos procedimentos, os valores originais foram separados em duas variáveis independentes, ISSN e eISSN, descritas no Quadro 4.1.

Quadro 4.1 – *Separação do campo ISSN com os valores originais em variáveis independentes.*

ISSN Original	ISSN	eISSN
15454487, 00664170	1545-4487	0066-4170

Fonte: Autor

Na primeira etapa de aquisição, foram realizadas buscas para a seleção dos artigos publicados nos periódicos Ann. Intern. Med., JAMA e PLoS Med., utilizando um padrão de *script* proprietário do Scopus, disponível em seu sistema de "busca avançada". Os critérios aplicados para filtragem foram:

- ISSN.
- eISSN.
- Registros classificados apenas como periódico.
- Publicações classificadas apenas como artigo.
- Publicações marcadas como "estágio final de publicação".
- Área da publicação apenas como médica.

- Intervalo de publicação entre os anos de 2000 e de 2019.

Os arquivos resultantes das buscas foram adquiridos manualmente por *download* e contabilizaram 12.106 artigos, representados por 4.118 do Ann. Intern. Med., 5.522 do JAMA e 2.466 do PLoS Med., ambos com uma estrutura contendo 19 variáveis descrevendo seus dados, mostradas no Quadro 4.2.

Quadro 4.2 – *Conjunto de* 19 *variáveis disponíveis nos arquivos obtidos do Scopus com informações sobre artigos.*

#	**Variável**	**Descrição**
1	*Authors*	Nome dos autores do artigo
2	*Author(s) ID*	Identificador do autor no Scopus
3	*Title*	Título do artigo
4	*Year*	Ano de publicação
5	*Source title*	Nome do periódico
6	*Volume*	Número do volume da publicação
7	*Issue*	Edição
8	*Art. No.*	Número do artigo
9	*Page start*	Página inicial
10	*Page end*	Página final
11	*Page count*	Número de páginas
12	*Cited by*	Número de citações
13	*DOI*	Identificador eletrônico
14	*Link*	Endereço para acesso do artigo
15	*Document Type*	Tipo do documento (artigo, *review*, etc.)
16	*Publication Stage*	Estágio da publicação (Final ou *Press Release*)
17	*Access Type*	Tipo de acesso (*Open Access* ou privado)
18	*Source*	Origem do documento (Scopus)
19	*EID*	Identificador do artigo no Scopus

Fonte: Autor

Posteriormente foi realizada a limpeza e tratamento dos dados, visando excluir registros irrelevantes e anômalos, além da padronização dos valores contidos nas variáveis, através de *scripts* `Python` [1] automatizados. Em alguns casos houve necessidade de intervenção manual quando não observados padrões, que impediu o processamento dos dados. Os procedimentos adotados são descritos a seguir:

[1] Disponível em: <https://github.com/EdsonMSouza/PhD/tree/main/Dataclean>

1. Padronização da nomenclatura do nome do periódico quando identificado dois ISSNs diferentes. Esse tratamento foi realizado comparando o nome descrito no arquivo originado do SCImago para a área correspondente e o *"Source Title"* do arquivo de artigos do Scopus.

2. Filtragem do campo *"Author(s) ID"* para verificar a ausência de dados.

3. Exclusão dos registros que não continham valores ou estavam marcados como referência igual a "*No author name available*".

4. Exclusão de artigos que:

   a. continham menos de quatro autores, visto que a distribuição das contribuições para números menores pode prejudicar no modelo de análise dos dados devido ao acúmulo de contribuições dos autores que podem indicá-las em todas as categorias.

   b. não pertenciam exclusivamente à área da medicina, visto que alguns registros recuperados do SCImago estavam relacionados a outras áreas no Scopus.

   c. divergiam entre o número de autores informados nos registros do Scopus e os arquivos obtidos.

   d. não possuíam "doi", visto que a coleta dos dados sobre os autores que publicaram artigos nos periódicos Ann. Intern. Med., JAMA e PLoS Med. utiliza este valor como identificador no site dos periódicos.

5. Tratamento das variáveis *"Page start"*, *"Page end"*, *"Page count"* e *"Cited by"* para padronização do formato numérico, preenchimento de valores ausentes com o valor zero (0), além da correção dos valores que apresentavam caracteres alfanuméricos e outras divergências.

6. O campo *"Page count"* de cada registro foi preenchido com a diferença entre o *"Page end"* e o campo *"Page start"*, quando possível, devido à existência de registros com valores ausentes referentes às páginas iniciais e/ou finais, ou ambas.

7. O campo *"Cited by"* foi preenchido com o valor zero (0) quando não apresentava valor.

Após a limpeza dos dados, foi criado um *dataset* denominado *"dataset_artigos"* com dados de autores dos três periódicos, totalizando 2.024 artigos com 20.098 autores, sendo 17.385 únicos. Da amostra, a diferença de 2.713 indica que esses publicaram em mais de um periódico no período analisado. Na Tabela 4.1 são apresentadas as amostras referentes aos periódicos, artigos e autores.

Tabela 4.1 – *Total de artigos, autores e autores únicos remanescentes dos periódicos Ann. Intern. Med., JAMA e PLoS Med. após a limpeza e tratamento dos dados.*

Periódico	Artigos	Autores	Autores Únicos
Ann. Intern. Med.	901	8.191	6.965
JAMA	790	7.780	6.683
PLoS Med.	333	4.127	3.737
Total	2.024	20.098	17.385

Fonte: Autor

Nas subseções a seguir são discutidos os números referentes a quantidade de publicações selecionadas para o estudo referente aos periódicos Ann. Intern. Med., JAMA e PLoS Med. no intervalo compreendido entre os anos de 2000 e de 2019.

4.2.1 Distribuição dos Artigos por Periódico entre os anos 2000 e 2019

A distribuição das publicações ao longo dos anos difere entre os periódicos, dado que cada um possui regras individuais de publicação, ficando os registros disponíveis conforme indexação do Scopus. Também deve-se considerar que a aquisição dos dados ocorreu em intervalos distintos, podendo apresentar divergências nos registros conforme as atualizações ocorrem na base de dados do Scopus.

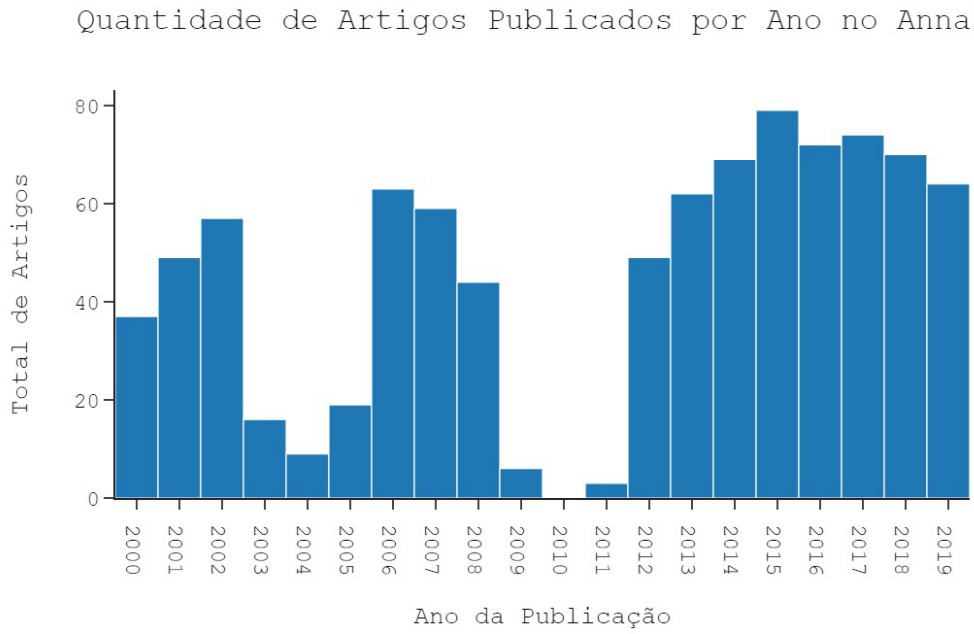
Durante o processo de limpeza dos dados ocorreram exclusões de dados, conforme citado, que reduziram significativamente o número de artigos obtidos inicialmente. A ausência de dados relacionados ao período de publicação não apresenta impacto na realização deste estudo, pois o objetivo é estudar a relação entre a posição autoral e as contribuições, onde o número de artigos e autores são suficientes para sua realização.

4.2.1.1 Periódico Ann. Intern. Med.

Na Figura 4.2 são mostrados os 901 artigos coletados do periódico Ann. Intern. Med. contendo 8.191 autores, onde 6.965 são únicos, indicando que alguns participaram de mais de uma publicação.

Os artigos são divididos em dois intervalos de publicação, onde o primeiro compreende os anos de 2000 e 2009; o segundo entre os anos de 2014 e 2018, concentrando o maior volume de artigos no segundo período. O ano de 2010, que separa os intervalos, não apresenta publicações devido ao processo de limpeza dos dados e aos critérios de seleção sobre o número de autores. Também é possível visualizar uma redução no número de publicações a partir do ano de 2018. A redução recorrente no ano de 2019 está relacionada com o período de disponibilidade das publicações no Scopus no momento da aquisição dos dados, além dos excluídos durante o processo de limpeza.

Figura 4.2 – *Distribuição das publicações no periódico Ann. Intern. Med. entre os anos* 2000 *e* 2019.



Fonte: Autor

4.2.1.2 Periódico JAMA

Na Figura 4.3 são mostrados os 790 artigos coletados do periódico JAMA com 7.780 autores, sendo 6.683 únicos, indicando que alguns participaram de mais de uma publicação.

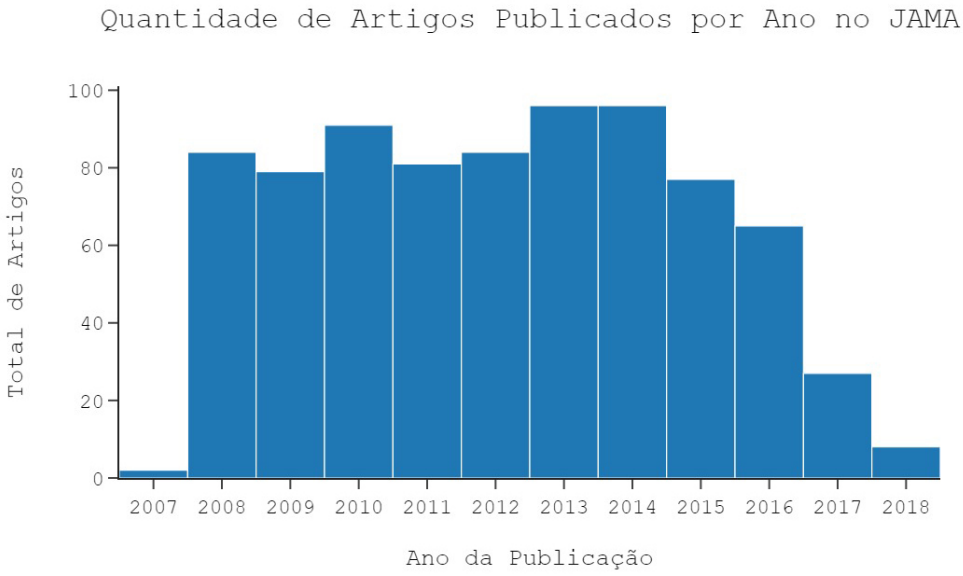
Os 790 artigos estão distribuídos entre os anos de 2007 e 2018. Embora o filtro inicial tenha sido aplicado para o período que compreende os anos entre 2000 e 2019, o processo de limpeza dos dados reduziu exemplares neste este intervalo. A concentração das publicações selecionadas se encontra entre os anos de 2008 e 2016. A redução no número de artigos obtidos a partir do ano de 2016 está relacionada com o processo de limpeza dos dados.

4.2.1.3 Periódico PLoS Med.

Na Figura 4.4 são mostrados os 333 artigos coletados do periódico PLoS Med. entre os anos de 2017 e 2019 com 4.127 autores, sendo 3.737 únicos, indicando que alguns participaram de mais de uma publicação.

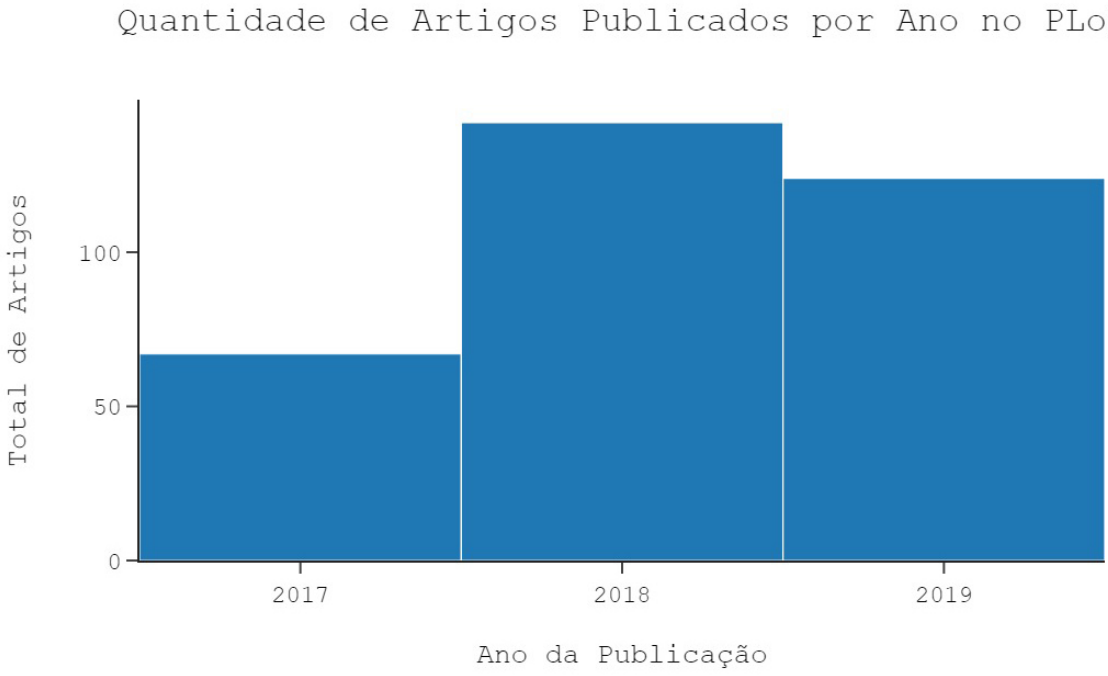
O motivo do intervalo entre os anos ser menor que dos periódicos Ann. Intern. Med. e JAMA é que ao serem coletados seus dados no Scopus, a maioria dos artigos não estavam indicados como "*Medicine*", pertencendo a áreas correlatas. Apesar do conjunto de artigos representar um volume inferior ao dos outros dois periódicos (Ann. Intern. Med. e JAMA),

Figura 4.3 – *Distribuição das publicações no periódico JAMA entre os anos* 2007 *e* 2018.



Fonte: Autor

Figura 4.4 – *Distribuição das publicações no periódico PLoS Med. entre os anos* 2017 *e* 2019.



Fonte: Autor

o número de autores é representativo para o estudo.

Na próxima seção é descrito o procedimento para a realização do processamento dos artigos adquiridos.

4.3 Individualização dos Autores

O *"dataset_artigos"*, criado anteriormente, é constituído por 2.024 artigos com 19 variáveis, descritas no Quadro 4.2. Na variável "*authors_id*" estão armazenados os identificadores dos autores separados por ponto e vírgula em cada artigo. No Quadro 4.3 é mostrado um fragmento do *dataset* original, destacando-se a variável "*authors_id*".

Quadro 4.3 – *Fragmento do conjunto de artigos obtidos do Scopus mostrando um artigo com quatro autores, identificados pelos valores contidos na variável authors_id.*

doi	**authors_id**	...
10.7326/M17-2605	**8971597300;57201956561;57127649700;47661495600**	...

Fonte: Autor

Como pode se observar no Quadro 4.3, a variável *"authors_id"* possui quatro valores, identificando cada autor, separados por ponto e vírgula. Visto que o estudo relaciona dados de autores, foi necessário realizar um procedimento para individualizá-los, de modo que cada autor seja armazenado em um único registro no *"dataset_autores"*, ou seja, uma única observação. Para isso foi desenvolvido um algoritmo, implementado com a linguagem `Python`, para processar os dados dos artigos.

4.3.1 Algoritmo de Individualização dos Autores

O algoritmo desenvolvido para a individualização dos autores recebe como parâmetro de entrada o *"dataset_artigos"* (Artigos), processando cada registro até alcançar seu final. Seu funcionamento é apresentado no Algoritmo 1.

Algoritmo 1: Individualização dos autores de cada artigo, onde cada um é inserido em uma nova linha no novo conjunto de dados, totalizando 19 variáveis e seus dados para cada autor, contidas em Artigos.

```
    Entrada: Artigos
    Saída: Dataset
1   Dataset ← ∅
2   para cada artigo i ∈ Artigos faça
3       autor ← ∅
4       para cada autor k ∈ artigo faça
5           autor[k] ← Extrair as informações do artigo i para o autor k ;
6       fim
7       Dataset ← Adicionar os dados do autor ;
8   fim
```

O Algoritmo 1 recebe como parâmetro de entrada "Artigos", contendo uma lista com 2.024 artigos, identificados por 19 variáveis. A cada iteração do algoritmo os dados

de cada artigo, referentes ao autor, são extraídos e adicionados à variável *"autor"*. Após o processamento do artigo, os dados do autor são adicionados à variável *"Dataset"*. Ao término da execução da lista de artigos é retornado o *"Dataset"* (*"dataset_autores"*) atualizado com 20.098 linhas (observações).

Os códigos para a programação dos procedimentos estão disponíveis em forma de código livre no Github [2].

4.3.1.1 Execução da Individualização dos Autores

Para realizar o procedimento de individualização dos autores, os dados contidos no *"dataset_autores"* foram processados com o Algoritmo 1. Ao término do processamento foi gerado um novo *dataset* contendo 20.098 registros e 19 variáveis. O novo *dataset* substituiu o anterior, mantendo-se o nome.

No Quadro 4.4 é apresentado um fragmento do *"dataset_autores"* tratado, mostrando um artigo com seus quatro autores ocupando individualmente uma linha.

Quadro 4.4 – *Fragmento do dataset gerado após o processamento para individualização dos autores, mostrando o resultado de um artigo com quatro autores.*

doi	**author_id**	**author**	...
10.7326/M17-2605	**8971597300**	Abigail M. Judge	...
10.7326/M17-2605	**57201956561**	Jennifer A. Murphy	...
10.7326/M17-2605	**57127649700**	Jose Hidalgo	...
10.7326/M17-2605	**47661495600**	Wendy Macias-Konstantopoulos	...

Fonte: Autor

Na próxima seção são descritos os procedimentos para a coleta dos dados dos autores na fonte de dados do Scopus.

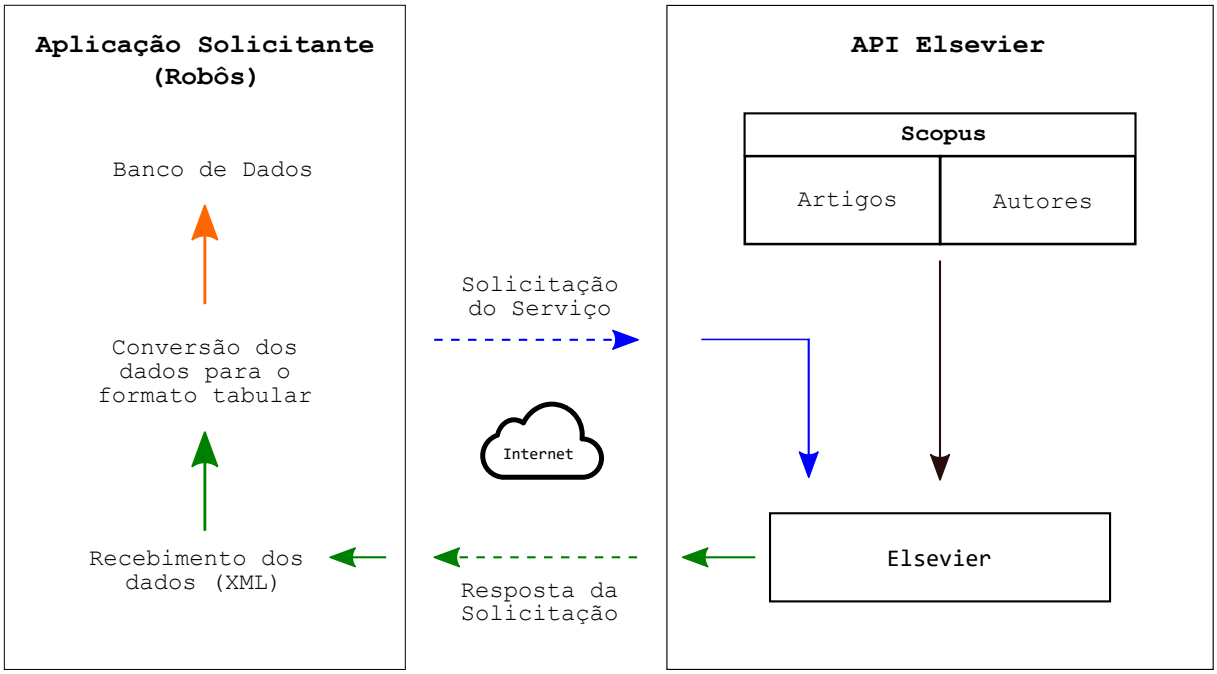
4.4 Coleta de Dados dos Autores na Fonte de Dados do Scopus

O Scopus, por meio da Elsevier, fornece acesso a dados de artigos e autores, descrito no Capítulo 3, mediante uma Interface de Programação de Aplicações, do inglês *Application Programming Interface* (API). Uma API é um conjunto de rotinas e padrões que fornecem acesso a dados, permitindo realizar conexão entre aplicações (VERA; COMENDADOR, 2016). As APIs disponibilizadas pela Elsevier podem ser acessadas via IES através de autenticação, ou então por assinatura comercial. Uma vez autenticado, pode-se explorar os recursos utilizando uma linguagem de programação que interaja via HTTP (*Hypertext Transfer Protocol*) ou Protocolo de Comunicação de Hipermídia, pois as APIs

[2]Recurso disponível em: <https://github.com/EdsonMSouza/PhD/tree/main/Division>

disponibilizadas trabalham sob este protocolo. Na Figura 4.5 é mostrada uma representação para o acesso e recuperação de dados através das APIs da Elsevier, incluindo o processo de conversão dos dados do formato XML para o tabular e armazenamento em banco de dados.

Figura 4.5 – *Representação do processo para recuperação de dados através das APIs da Elsevier, mostrando as etapas para conversão e armazenamento dos dados.*



Fonte: Autor

Na Figura 4.5 é mostrado o diagrama interno da aplicação solicitante (robôs), indicando o fluxo dos processos, que se inicia com a solicitação do serviço à API (cliente) e, ao receber os dados, realiza o processamento. Por outro lado, a API disponibiliza recursos que permitem o acesso aos dados de artigos e autores no Scopus. Seus recursos necessitam de "*TOKENS*" de acesso, que devem ser criados na área do usuário na Elsevier previamente para utilização dos recursos.

Os dados disponibilizados, referentes aos autores em publicações, estão no formato XML. Na Figura 4.6 é mostrado um fragmento de arquivo obtido via API contendo dados de um autor, no qual podem ser identificados dados referentes ao autor como: *"h-index"*, *"cited-by-count"*, *"coauthor-count"*, *"subject-area"* e *"indexed-name"*. Para cada autor coletado na fonte de dados Scopus é gerado um arquivo XML.

As variáveis contidas nos arquivos diferem para cada autor, embora algumas possam ser iguais, visto que são fornecidas de acordo com os dados individuais. Em outras palavras, um arquivo coletado pode ter muitas variáveis, por exemplo, *"journal_history_issn"*, se o autor publicou em mais de um periódico desde a inclusão do seu registro no Scopus. Nestes casos, são gerados registros (linhas) para cada publicação na variável correspondente.

Figura 4.6 – *Fragmento de um arquivo XML extraído do Scopus com dados sobre um autor.*

```
<?xml version="1.0" encoding="UTF-8" ?>
<author-retrieval-response
      xmlns:ait="http://www.elsevier.com/xml/ani/ait"
      xmlns:ce="http://www.elsevier.com/xml/ani/common"
      xmlns:cto="http://www.elsevier.com/xml/cto/dtd"
      xmlns:dc="http://purl.org/dc/elements/1.1/"
      status="found">
 <coredata>
   <dc:identifier>AUTHOR_ID:6504422924</dc:identifier>
   <eid>9-s2.0-6504422924</eid>
   <document-count>1</document-count>
   <cited-by-count>925</cited-by-count>
   <citation-count>925</citation-count>
 </coredata>
 <h-index>1</h-index>
 <coauthor-count>6</coauthor-count>
 <author-profile>
   <preferred-name>
     <initials>A.</initials>
     <indexed-name>Paddags A.</indexed-name>
     <surname>Paddags</surname>
     <given-name>Anne</given-name>
   </preferred-name>
   <affiliation-history>
     <affiliation id="60002975" />
     <affiliation id="60027272" />
     <affiliation id="60001422" />
     <affiliation id="60013278" />
  </affiliation-history>
  <subject-areas>
     <subject-area
       abbrev="BIOC" code="1300">Biochemistry, Genetics and Molecular
           Biology (all)
     </subject-area>
     <subject-area
       abbrev="IMMU" code="2403">Immunology
     </subject-area>
     <subject-area
       abbrev="MEDI" code="2705">Cardiology and Cardiovascular Medicine
     </subject-area>
     <subject-area
       abbrev="MEDI" code="2720">Hematology
     </subject-area>
  </subject-areas>
  <author-profile>
```

Fonte: Elsevier

Para aquisição dos dados de autores, via API, foi desenvolvido um robô em linguagem `Python` para automatizar o processo de extração dos dados do Scopus, descrito a seguir.

4.4.1 Robô Coletor de Dados dos Autores (RCDA)

O algoritmo do robô coletor de dados de autores ou RCDA foi desenvolvido a partir da análise da estrutura dos dados fornecidos pelo Scopus no formato XML. Para executá-lo é necessário realizar uma conexão ao servidor da Elsevier através do endereço HTTP da API [3] a cada solicitação. O seu funcionamento é mostrado no Algoritmo 2.

Algoritmo 2: Robô Coletor de Dados de Autores (RCDA) - Realiza a extração de dados de autores na base de dados do Scopus, utilizando o identificador único (*author_id*).

```
    Entrada: authors_id
    Saída: dataset_autores
1   dataset_autores = {} ;
2   dados_autor = {} ;
3   para cada author_id ∈ authors_id faça
4       arquivo_xml ← Coleta dados a partir do author_id da base Scopus ;
5       dados_autor[author_id] ← author_id ;
6       para cada variavel_xml ∈ arquivo_xml faça
7           dados_autor[variavel_xml] ← conteúdo da variavel_xml ;
8       fim
9       dataset_autores[dados_autor] ← dados_autor ;
10  fim
```

O Algoritmo 2 recebe como parâmetro de entrada um (*dataset*) contendo uma lista de autores com o identificador único de cada autor (*"author_id"*). A lista é processada de forma recorrente até que atinga o seu final.

A cada iteração do algoritmo é realizada a extração dos dados do autor pelo seu *"author_id"* e armazenados na variável *"arquivo_xml"*. As variáveis disponíveis em *"arquivo_xml"* são processadas e seus valores armazenados em um vetor (*"dados_autor"*). O resultado do processamento de cada autor é então armazenado em um outro vetor (*"dataset_autores"*). Ao final da execução da lista *"authors_id"*, o algoritmo retorna um *dataset* (*"dataset_autores"*) contendo os dados coletados de todos os autores.

Os códigos implementados para a programação do robô estão disponíveis em forma de código livre no Github [4].

4.4.1.1 Execução da Coleta com o RCDA

O RCDA, descrito no Algoritmo 2, foi executado para o conjunto de dados criado anteriormente no processo de individualização dos autores, Subseção 4.3.1. Ao final da execução foram obtidos 20.098 registros, referentes aos dados dos autores, dispondo de 22

[3] Recurso disponível em Elsevier Developer Portal: <https://dev.elsevier.com/>

[4] Recurso disponível em: <https://github.com/EdsonMSouza/PhD/tree/main/Bots>

variáveis, descritas no Quadro 4.5.

Quadro 4.5 – *Variáveis extraídas do Scocus com a execução do RCDA referentes aos autores.*

Seq.	**Variável**	**Descrição**
1	*author_id*	Identificador do autor
2	*author_eid*	Identificador secundário do autor
3	*author_name*	Nome do autor
4	*affl_source*	Afiliação
5	*document_count*	Número de publicações
6	*cited_by_count*	Total de citações recebidas
7	*citation_count*	Média simples de citações
8	*h_index*	Índice H do autor
9	*coauthor_count*	Número de coautores em publicações
10	*asjc*	Classificações dos periódicos em que publicou
11	*asjc_frequency*	Frequência de publicações (áreas)
13	*classification_type*	Classificação do periódico (áreas)
14	*journal_history_issn*	ISSN de todos periódicos em que publicou
15	*affl_id*	Id da instituição que o autor é afiliado
16	*affl_name*	Nome da instituição de afiliação
17	*affl_country*	País da instituição
18	*affl_acronym*	Abreviação da instituição
19	*affl_state*	Estado da localização da instituição
20	*affl_city*	Cidade de localização da instituição
21	*affl_history_id*	Id de todas as instituições em que foi afiliado
22	*affl_history_country_count*	Total de países em que o autor trabalhou

Fonte: Autor

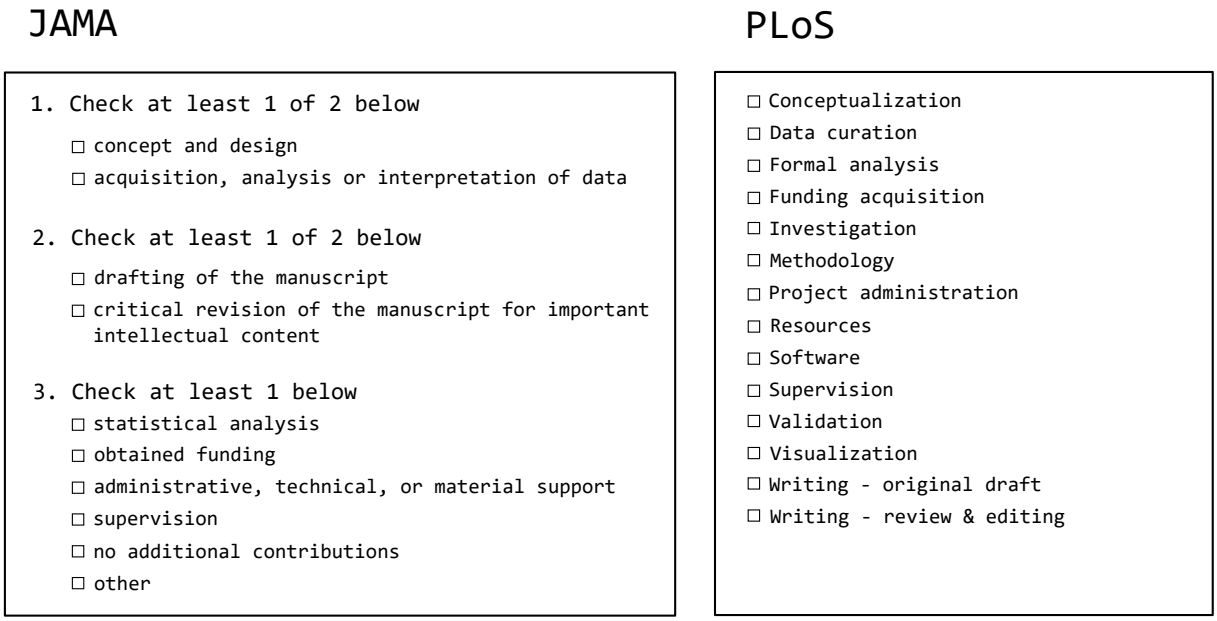
As variáveis e os dados referentes aos autores, obtidos pela execução do RCDA, foram incorporados ao *"dataset_autores"*, processado na etapa anterior de individualização dos autores, Subseção 4.3.1, fornecendo mais indicadores para serem utilizados em análises.

A seguir são apresentados os procedimentos para a extração das contribuições dos autores nos periódicos Ann. Intern. Med., JAMA e PLoS Med.

4.5 Coleta de Dados das Contribuições nos Periódicos

Os periódicos Ann. Intern. Med., JAMA e PLoS Med. disponibilizam em seus respectivos sites os artigos publicados contendo dados sobre as contribuições dos autores no formato semiestruturado *Hypertext Markup Language* (HTML), ou seja, em uma página *web*. As contribuições dos autores são indicadas através de formulários que oferecem uma lista de opções relativas às categorias, conforme mostrado na Figura 4.7. Assim, cada periódico disponibiliza os dados de forma e com nomenclaturas diferentes.

Figura 4.7 – *Fragmento dos formulários de indicação das contribuições dos periódicos JAMA e PLoS Med.*



Fonte: Autor

O acesso aos dados de cada publicação ocorre de forma individual através de um *link* específico para cada artigo, utilizando-se o "doi" como chave de acesso. Cada periódico utiliza uma estrutura diferente de apresentação dos artigos na internet, conforme é mostrado nas subseções a seguir.

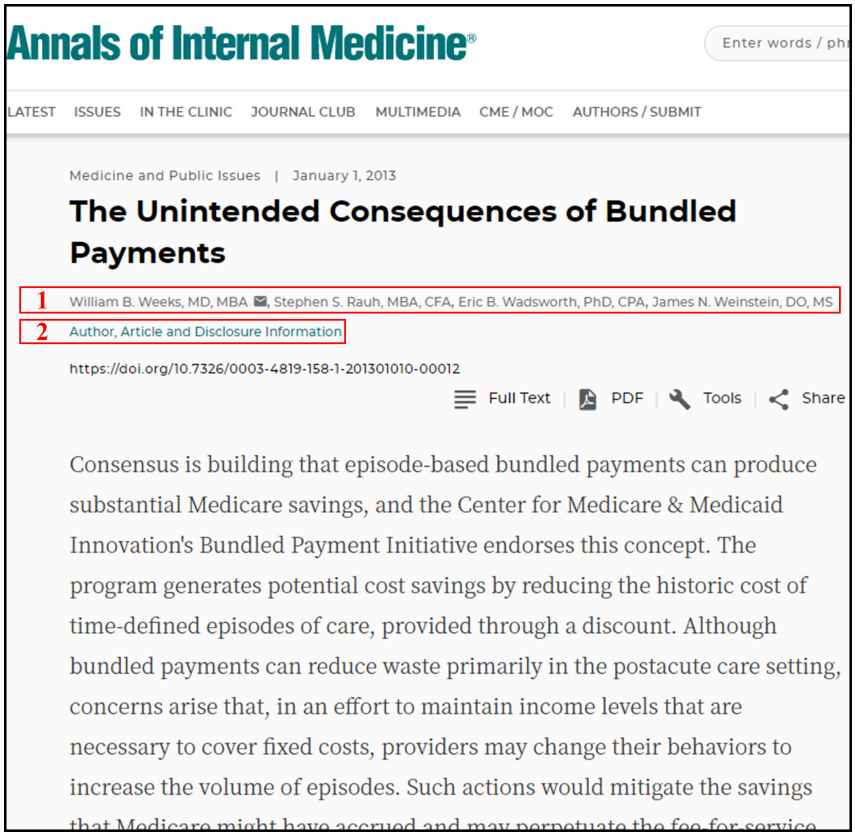
4.5.1 Estrutura de um Artigo no Periódico Ann. Intern. Med.

A estrutura dos artigos fornecidos pelo Ann. Intern. Med. possui no início do documento informações como título, autores, endereço de acesso pelo DOI, além de opções para visualização do texto completo, descarregamento no formato PDF, entre outras. Na Figura 4.8 é apresentado um fragmento inicial do artigo *"The Unintended Consequences of Bundled Payments"*.

Na Figura 4.8, o *"byline"* apresenta o nome dos autores utilizando abreviação no nome do meio, além da titulação acadêmica, destaque 1 na Figura 4.8. Quando a lista de autores atinge o limite da largura da página, é inserido um *link* com a descrição "et. al" em substituição ao nome dos demais autores que extrapolarem o limite da página.

Os dados de contribuição dos autores não possuem acesso direto, sendo necessário acessá-los através de um *link*, destaque 2 na Figura 4.8, necessitando de mais uma etapa para obter a informação. Uma vez acionado o *link*, é aberta uma nova página contendo as informações de contribuição, conforme destaque na Figura 4.9.

Figura 4.8 – *Fragmento da página inicial do artigo "The Unintended Consequences of Bundled Payments" publicado no periódico Ann. Intern. Med., destacando-se o "byline" com o nome dos autores (1) e informações adicionais, que incluem as contribuições no artigo (2).*



Fonte: *Print screen* do site *Annals of Internal Medicine* (ANNALS, 2021)

Na Figura 4.9 são mostradas informações sobre a publicação como o nome do autor correspondente, endereço dos autores e as contribuições de cada um (*"Author Contributions"*). A identificação das contribuições ocorre pela localização do nome da categoria, onde são listados os autores que nela contribuíram.

A ausência de padronização na informação dos nomes dificulta a identificação dos autores quando comparados com os listados no *"byline"* da página inicial do artigo, Figura 4.8, visto que nela os nomes estão escritos no formato (primeiro nome, abreviação do nome do meio e sobrenome). Nas contribuições, apresenta o formato (abreviação do primeiro nome, abreviação do nome do meio e sobrenome).

Figura 4.9 – *Informações adicionais sobre o artigo "The Unintended Consequences of Bundled Payments", incluindo detalhes sobre as contribuições informadas pelos autores, em destaque.*

Fonte: *Print screen* do site *Annals of Internal Medicine* (ANNALS, 2021)

No Quadro 4.6 são mostrados os formatos de apresentação das informações para os dois primeiros autores do artigo, Figura 4.9.

Quadro 4.6 – *Disposição dos nomes dos primeiros dois autores no "byline" e na lista de contribuições do periódico Ann. Intern. Med., mostrando a diferença entre eles.*

Byline	**Contribuições**
Willian B. Weeks, MD, MBA	W.B Weeks
Stephen S. Rauh, MBA, CFA	S.S. Rauh

Fonte: Autor

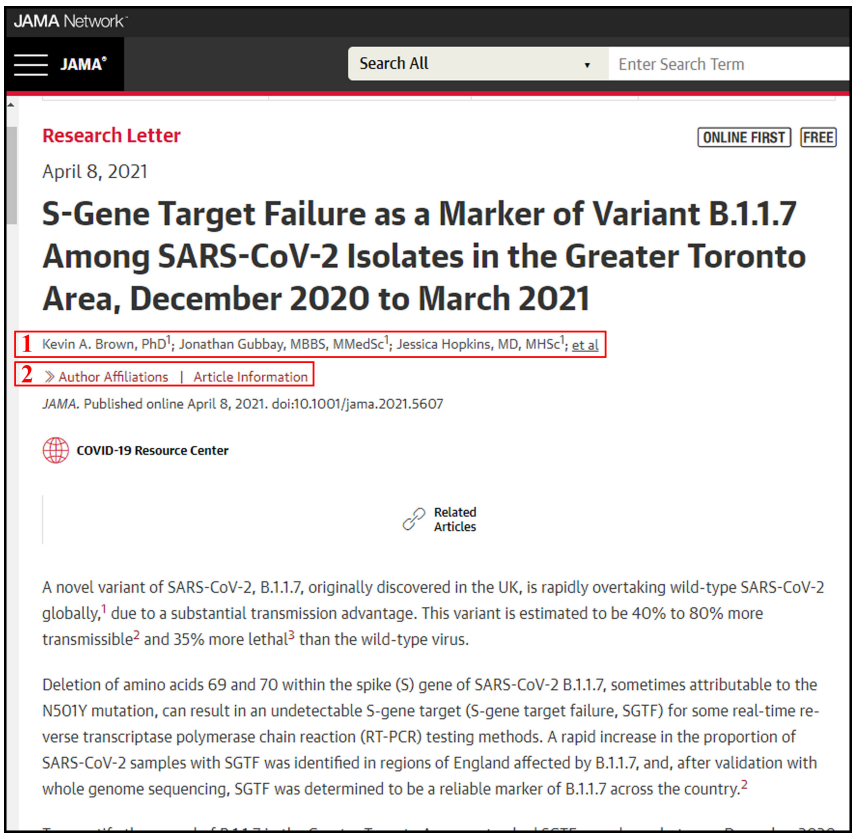
Para extrair os dados de contribuição, é necessário realizar um procedimento para identificar inicialmente na página o nome das categorias e posteriormente fazer uma varredura nos dados para localizar o nome dos autores. Além disso, deve ser feita uma comparação tanto dos nomes informados no *byline*, quanto nas contribuições, com os nomes obtidos dos autores no Scopus para confirmar que se trata do mesmo autor, visto que pode haver homônimos e erros ortográficos.

4.5.2 Estrutura de um Artigo no Periódico JAMA

O periódico JAMA disponibiliza os artigos em um formato que apresenta no início do documento informações como data de publicação, título, autores, endereço de acesso pelo DOI, além de opções para visualização do texto completo, descarregamento no formato PDF, artigos relacionados, entre outras.

Na Figura 4.10 é apresentado um fragmento inicial do artigo *"S-Gene Target Failure as a Marker of Variant B.1.1.7 Among SARS-CoV-2 Isolates in the Greater Toronto Area, December 2020 to March 2021"*. Na figura, destaque 1, é mostrado o *byline*, apresentando o nome dos autores com abreviação no nome do meio, além da titulação acadêmica. Quando a lista de autores atinge o limite da largura da página, é inserido um *link* com a descrição "et. al" em substituição ao nome dos demais autores.

Figura 4.10 – *Fragmento da página inicial do artigo "S-Gene Target Failure as a Marker of Variant B.1.1.7 Among SARS-CoV-2 Isolates in the Greater Toronto Area, December 2020 to March 2021" publicado no periódico JAMA, com destaque para o nome dos autores (1) e acesso às contribuições (2).*



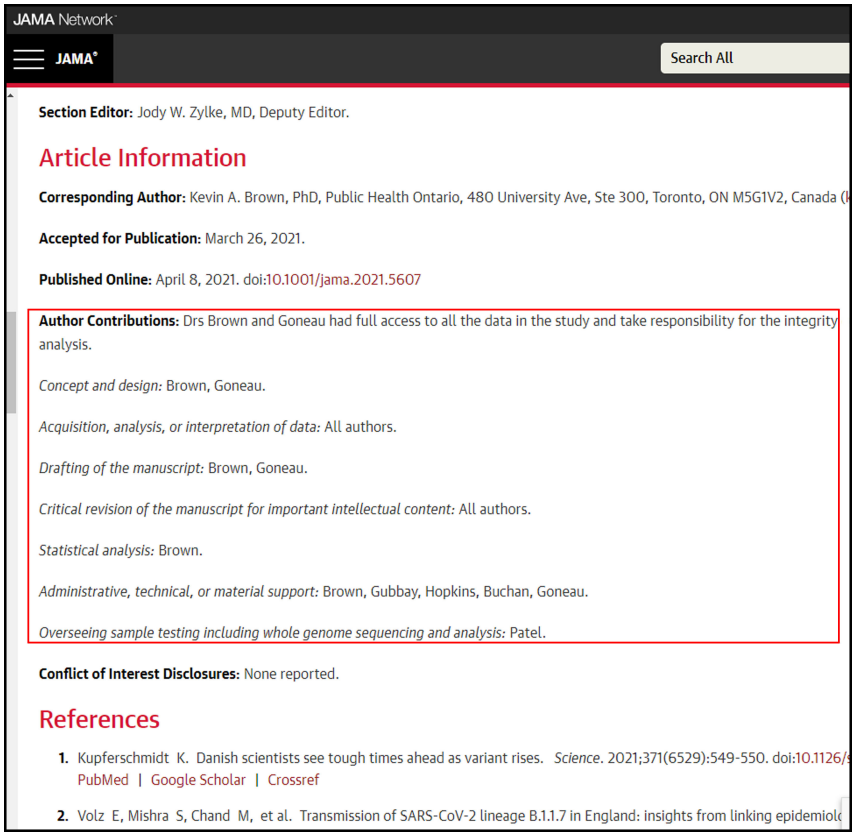
Fonte: *Print screen* do site *Journal of the American Medical Association* (JAMA, 2021)

Ainda, na Figura 4.10, destaque 2, os dados de contribuição podem ser acessados através de um *link*, que direciona para uma região localizada próximo ao rodapé da página, mostrada na Figura 4.11, contendo, além das contribuições, outras informações sobre os

autores.

Na Figura 4.11, são apresentados os dados sobre a publicação como o nome do autor correspondente, data de aceite e da publicação *online* e as contribuições de cada autor (*Author Contributions*). As contribuições são exibidas nas categorias por meio do sobrenome do autor, divergindo dos listados no *byline* da página inicial do artigo, que são informados com o formato (primeiro nome, abreviação do nome do meio e sobrenome), Figura 4.10. Além disso, quando todos os autores contribuíram em uma determinada categoria, é informado apenas como (*"All authors"*). Este formato de disposição dos nomes dificulta a identificação dos autores, assim como sua obtenção.

Figura 4.11 – *Informações adicionais sobre o artigo "S-Gene Target Failure as a Marker of Variant B.1.1.7 Among SARS-CoV-2 Isolates in the Greater Toronto Area, December 2020 to March 2021", incluindo detalhes sobre as contribuições informadas pelos autores, em destaque.*



Fonte: *Print screen* do site *Journal of the American Medical Association* (JAMA, 2021)

No Quadro 4.7 são mostrados os formatos de apresentação das informações para os dois primeiros autores referentes ao artigo apresentado na Figura 4.10.

Para extrair os dados das contribuições, Figura 4.11, é faz necessário identificar primeiro na página o nome das categorias e posteriormente fazer uma varredura nos dados para localizar o nome dos autores. Além disso, deve ser feita uma comparação tanto dos nomes informados no *byline*, quanto nas contribuições, com os nomes obtidos dos autores

Quadro 4.7 – *Disposição dos nomes dos primeiros dois autores no "byline" e na lista de contribuições do periódico JAMA, mostrando a diferença entre eles.*

Byline	Contribuições
Kevin A. Brown, PhD	Brown
Jonathan Gubbay, MBBS, MMedSc	Gubbay

Fonte: Autor

no Scopus para confirmar que se trata do mesmo autor, visto que pode haver homônimos, erros de escrita e ortográficos.

4.5.3 Estrutura de um Artigo no Periódico PLoS Med.

O periódico PLoS Med. fornece acesso aos artigos com uma estrutura bem elaborada, apresentando os dados divididos em blocos. No topo da página são mostrados o título do artigo, nome dos autores, data de publicação e endereço DOI. A divisão em abas facilita a localização das informações, embora o acesso deva ser realizado através de *links*. A primeira aba apresenta o artigo completo e é a primeira a ser carregada no momento do acesso. Na segunda, estão disponíveis informações adicionais, incluindo as contribuições, mostradas na Figura 4.12.

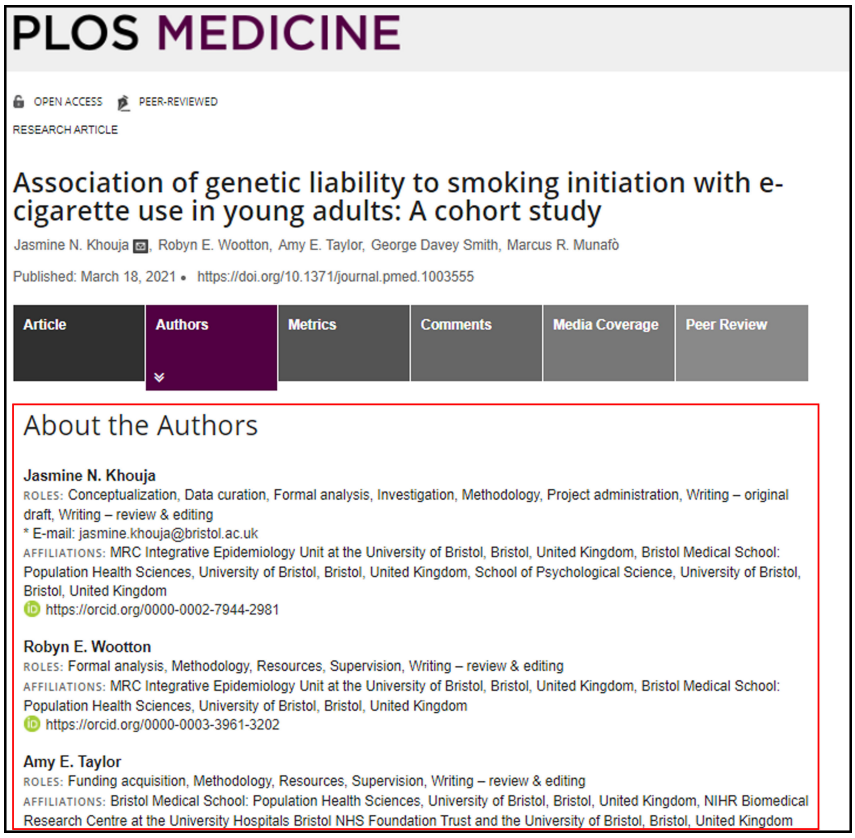
Figura 4.12 – *Fragmento da página inicial do artigo "Association of genetic liability to smoking initiation with e-cigarette use in young adults: A cohort study" publicado no periódico PLoS Med. com destaque para o nome dos autores (1) e a aba "Authors' (2), que fornece acesso a página de contribuições.*

Fonte: *Print screen* do site *PLoS Medicine* (PLOS, 2021)

Na Figura 4.12 é mostrado o fragmento inicial do artigo *"Association of genetic liability to smoking initiation with e-cigarette use in young adults: A cohort study"* e a estrutura da página. com o nome dos autores utilizando abreviação no nome do meio, destaque 1 da Figura 4.12. Quando a lista de autores atinge o limite da largura da página, é inserido um *link* com a descrição "et. al" em substituição ao nome dos demais autores que extrapolam o limite da página.

Os dados de contribuição não estão disponíveis diretamente, sendo necessário acessá-los através de um *link*, destaque 2 na Figura 4.12, que direciona para uma nova página. A Figura 4.13 mostra detalhes da página de contribuições.

Figura 4.13 – *Informações adicionais sobre o artigo "Association of genetic liability to smoking initiation with e-cigarette use in young adults: A cohort study", incluindo detalhes sobre as contribuições dos autores, em destaque.*



Fonte: *Print screen* do site *PLoS Medicine* (PLOS, 2021)

A Figura 4.13 mostra, além das informações sobre as contribuições, os mesmos dados da página inicial, o que facilita a localização do nome dos autores. As contribuições são indicadas pelo nome do autor no formato (primeiro nome, abreviação do nome do meio e sobrenome) e em quais categorias contribuiu.

Conforme descrito, cada periódico adota um sistema diferente de disponibilizar os dados sobre as contribuições, exigindo grande esforço para obtê-los. Assim, como o número de artigos utilizados neste trabalho (2.024) é elevado para obtenção de forma manual, foi

desenvolvido um robô escrito com a linguagem `Python`, aplicando técnicas de raspagem de dados ou *Scraping* para coletar os dados das contribuições dos autores em cada artigo.

Os procedimentos para coleta das contribuições dos autores nos artigos publicados nos três periódicos e apresentado a seguir.

4.5.4 Coleta das Contribuições

O processo de extração de dados das contribuições nos periódicos necessitou de um estudo nas estruturas de suas páginas, visto que os artigos são disponibilizados neste ambiente. O formato HTML, que gera as páginas web com o artigo, possui *tags* de marcação, as quais mantém os valores a serem extraídos. Como cada periódico possui nomenclaturas diferentes, é necessário identificar cada uma delas, na estrutura do documento. Para isso, foram criados três arquivos, um para cada periódico, de modo que as variáveis correspondam às existentes nos dados de extração, incluindo o identificador DOI e o nome dos autores.

Para aquisição dos dados nos três periódicos, foi desenvolvido um algoritmo para um robô coletor de dados de contribuição escrito com a linguagem `Python`, com três variações na implementação, utilizando técnicas de *scraping* ou raspagem de dados. A variações estão relacionadas à disposição dos dados em cada periódico, que requer estratégias diferentes para acessá-las na estrutura do documento.

Para otimizar o processamento e minimizar problemas como perda de conexão ou corrupção do arquivo, o algoritmo realiza o acesso ao documento e faz seu *download* para o computador local. Na próxima etapa realiza o processo de extração dos dados contidos no arquivo, armazenando-os nas variáveis previamente criadas e, por fim, realiza a gravação dos dados no arquivo correspondente ao periódico. A seguir é descrito o algoritmo desenvolvido.

4.5.4.1 Robô coletor de Contribuições RCCA

O algoritmo do robô coletor dos dados de contribuição dos autores ou RCCA foi desenvolvido a partir da análise dos artigos disponibilizados pelos periódicos no formato HTML. Para executá-lo, é necessário realizar o acesso via HTTP a cada artigo individualmente em cada um dos periódicos, utilizando o identificador do artigo (DOI).

O seu funcionamento é mostrado no Algoritmo 3, que recebe como parâmetro de entrada um *dataset* contendo os dados dos autores, incluindo o identificador "doi" do periódico, utilizado para recuperar o artigo no site do periódico. A lista de artigos é processada até que atinja seu final.

Algoritmo 3: Robô Coletor de Contribuições de Autores (RCCA)

```
    Entrada: dataset_autores
    Saída: dataset_contribuicoes
1   dataset_contribuicoes = {} ;
2   categorias_descricao = {descrição de cada categoria do periódico} ;
3   contribuicoes = {} ;
4   autores_contribuicoes = {} ;
5   autores = {} ;
6   function ProcessaContribuicao(arquivo_html, categoria):
7       dados = {} ;
8       lista_categorias = {} ;
9       tag = "tag HTML correspondente a cada periódico" ;
10      para cada categorias ∈ procurar em arquivo_html a tag_html faça
11          para cada categoria ∈ procurar todas as {tags "p"} faça
12              dados[categoria] ← coletar o conteúdo da categoria ;
13          fim
14          para cada categoria ∈ dados faça
15              se encontrou a categoria em categorias_descricao[categoria] então
16                  lista_categorias[categoria] ← Adiciona o nome do autor ;
17              fim
18          fim
19          contribuicoes ← lista_categorias
20      fim
21      retorna contribuicoes ;
22  para cada doi ∈ dataset_autores faça
23      arquivo_html ← Coleta dados a partir do doi ;
24      para cada categoria ∈ categorias_descricao faça
25          autores_contribuicoes[categoria] ←
                ProcessaContribuicao(arquivo_html, categoria) ;
26      fim
27      autores[doi] ← doi ;
28      autores[nomes] ← Coleta os dados do byline ;
29      autores[contribuicao] ← autores_contribuicoes ;
30      dataset_contribuicoes ← autores ;
31  fim
```

A cada iteração, o algoritmo recebe os dados do arquivo a ser processado e armazena na variável *"arquivo_html"*. No próximo passo, faz uma iteração sobre a variável *"categorias_descricao"*, que mantém a lista de categorias do periódico de onde os dados são obtidos. A variável *"autores_contribuicoes_html"* recebe os dados processados através da função "ProcessaContribuicao", que recebe os parâmetros do conteúdo do arquivo e da categoria a ser extraída. Por fim, armazena os resultados nas variáveis correspondentes e retorna um *"dataset"* com os dados extraídos.

A função "ProcessaContribuicao" realiza, a cada iteração, o processamento para a extração das contribuições, utilizando os parâmetros contidos na variável *"tag"*. Ela

armazena uma sequência de caracteres que identifica o posicionamento do elemento HTML dentro do documento. Os valores dessa variável são utilizados como parâmetro para busca de valores na variável *"arquivo_html"*. Uma vez localizado, faz uma iteração em busca de uma *"tag"* "p" do HTML e armazena o resultado na variável (*"dados"*). A variável *"dados"* então é percorrida, iterando sobre cada categoria localizada e armazenando o resultado na variável *"lista_categorias"*. Ao final do processamento de todas as categorias, retorna uma lista contida na variável (*"contribuicoes"*), contendo todas as categorias extraídas.

Os códigos implementados para a programação do robô estão disponíveis em forma de código livre no Github [5].

4.5.4.2 Execução da Coleta com o RCCA

O RCCA, descrito no Algoritmo 3, foi executado para extrair as contribuições dos autores nos periódicos Ann. Intern. Med., JAMA e PLoS. Ao final do processamento, foram obtidos três arquivos contendo as contribuições dos autores. Por fim, os dados desses arquivos foram adicionados ao *"dataset_autores"*.

Na seção a seguir, são apresentados os procedimentos para o processamento das contribuições de cada periódico.

4.6 Processamento das Contribuições

Como citado na Seção 4.5, nos três periódicos os nomes dos autores estão dispostos de formas diferentes nas variáveis que os descrevem. Para resolver esse problema, foram desenvolvidos *scripts* [6] com a linguagem `Python` para filtrar, higienizar e mesclar dados dos autores e contribuições. Em casos especiais, os nomes dos autores diferiam dos indicados *byline* e foram examinados manualmente. Como resultado do procedimento realizado, foram criadas três bases de dados com os dados obtidos para cada autor, contendo as informações originais adquiridas nas coletas anteriores pelos robôs.

Para realização do estudo foram criadas variáveis para cada *dataset* dos periódicos, descrevendo as contribuições, conforme mostrado no Quadro 4.8.

A variável "doi" representa o identificador do artigo; *"authors"* descreve o nome de todos os autores do artigo. Essas variáveis são comuns nos três periódicos e representam a contribuição dos autores, identificadas com o prefixo "*ac_*".

[5] Recurso disponível em: <https://github.com/EdsonMSouza/PhD/tree/main/Bots>
[6] Recurso disponível em: <https://github.com/EdsonMSouza/PhD/tree/main/Contribution>

Quadro 4.8 – *Variáveis que descrevem as contribuições em cada periódico.*

Ann. Intern. Med.	**JAMA**	**PLoS Med.**
doi	doi	doi
authors	authors	authors
ac_administrative_technical_or_logistic_support	ac_acquisition_analysis_interpretation_data	ac_conceptualization
ac_analysis_and_interpretation_of_the_data	ac_acquisition_data	ac_data_curation
ac_collection_and_assembly_of_data	ac_administrative_support	ac_formal_analysis
ac_conception_and_design	ac_analysis_interpretation_data	ac_funding_acquisition
ac_critical_revision_for_important_intellectual_content	ac_critical_revision	ac_investigation
ac_drafting_of_the_article	ac_drafting_manuscript	ac_methodology
ac_final_approval_of_the_article	ac_obtained_funding	ac_project_administration
ac_obtaining_of_funding	ac_statistical_analysis	ac_resources
ac_provision_of_study_materials_or_patients	ac_study_concept_and_design	ac_software
ac_statistical_expertise	ac_study_supervision	ac_supervision
		ac_validation
		ac_writing_original_draft
		ac_writing_review_and_editing

Fonte: Autor

Assim, um artigo extraído de um periódico contém k *in* $\mathbb{N}$ autores e produzirá k amostras para nosso *dataset*, conforme a Equação 4.1.

$$\begin{aligned} x_i^1 &= x_i^{\text{doi}} \circ x_i^{\text{authors}_1} \circ x_i^{\text{ac}_{1,1}} \circ \cdots \circ x_i^{\text{ac}_{n,1}} \\ x_i^2 &= x_i^{\text{doi}} \circ x_i^{\text{authors}_2} \circ x_i^{\text{ac}_{1,2}} \circ \cdots \circ x_i^{\text{ac}_{n,2}} \\ &\vdots \quad\quad \vdots \quad\quad\quad \vdots \quad\quad\quad \vdots \quad\quad \vdots \quad\quad \vdots \\ x_i^{\text{k}} &= x_i^{\text{doi}} \circ x_i^{\text{authors}_k} \circ x_i^{\text{ac}_{1,k}} \circ \cdots \circ x_i^{\text{ac}_{n,k}}, \end{aligned} \tag{4.1}$$

onde x_i^{doi} é o identificador do i-ésimo artigo; $x_i^{\text{authors}_1}, x_i^{\text{authors}_2} \ldots x_i^{\text{authors}_k}$ são os k autores do artigo, e $x_i^{\text{ac}_{1,1}} \ldots x_i^{\text{ac}_{n,k}}$ são valores booleanos que indicam as n contribuições dos k autores do artigo.

4.6.1 Construção do Escore das Contribuições

As contribuições do autor informadas pelos três periódicos não evidenciam ou mostram a existência de pesos ou qualquer índice que permita ou defina uma classificação. No entanto, a ordem de apresentação das contribuições em cada periódico foi mantida para construir a lista de contribuições do autor no banco de dados. O *dataset* inicial dos dados coletados apresenta um artigo identificado pela variável "doi" em cada linha, e a variável "authors" contém todos os autores do artigo separados por ponto e vírgula. O Quadro 4.9 apresenta uma linha da amostra.

Quadro 4.9 – *Variável "authors" contendo quatro autores de um artigo.*

doi	authors
10.7326/0003-4819-136-1-200201010-00009	Andrea Obernosterer; Manuela Aschauer; Wolfgang Schnedl; Rainer W. Lipp

Fonte: Autor

Para quantificar as contribuições individuais do autor, foi aplicada a Equação 4.1 para construir um novo *dataset* criando variáveis binárias para cada contribuição, definindo-as com 0 para "não contribuiu" e 1 para "contribuiu", mostradas no Quadro 4.10.

Quadro 4.10 – *Contribuições dos autores processadas para um artigo, mostrando a individualização dos autores e o preenchimento das variáveis binárias.*

Index	doi	author	ac_conception_and_design	...	ac_statistical_expertise
x_i^1	10.7326/0003-4819-136-1-200201010-00009	Andrea Obernosterer	1	...	0
x_i^2	10.7326/0003-4819-136-1-200201010-00009	Manuela Aschauer	0	...	0
x_i^3	10.7326/0003-4819-136-1-200201010-00009	Wolfgang Schnedl	0	...	1
x_i^4	10.7326/0003-4819-136-1-200201010-00009	Rainer W. Lipp	1	...	0

Embora esse valor permita classificar o grau de contribuição de um autor no artigo, o número de autores por artigo gera uma amostra muito homogênea, dificultando a análise. Assim, foi criado um procedimento para alocar uma taxa de probabilidade de contribuição por autor. Essa taxa pode ser obtida dividindo-se 1 pelo número de autores que contribuíram na categoria, de acordo com o Equação. 4.2.

$$f_{prob}(x_i^{ac}, l, a) = \frac{1}{\sum_{j=1}^{k} x_i^{ac_{l,j}}} \times x_i^{ac_{l,a}}, \tag{4.2}$$

onde f_{prob} é uma função que representa a probabilidade da a-ésima contribuição do l-ésimo autor no artigo x_i. A Equação 4.3 descreve como obter uma matriz de contribuição aplicando a função f_{prob} definida pela Equação 4.2.

$$\begin{array}{rcccccccccc}
x_i^1 & = & x_i^{\text{doi}} & \circ & x_i^{\text{authors}_1} & \circ & f_{prob}(x_i^{\text{ac}}, 1, 1) & \circ & f_{prob}(x_i^{\text{ac}}, 2, 1) & \circ \cdots \circ & f_{prob}(x_i^{\text{ac}}, n, 1) \\
x_i^2 & = & x_i^{\text{doi}} & \circ & x_i^{\text{authors}_2} & \circ & f_{prob}(x_i^{\text{ac}}, 1, 2) & \circ & f_{prob}(x_i^{\text{ac}}, 2, 2) & \circ \cdots \circ & f_{prob}(x_i^{\text{ac}}, n, 2) \\
\vdots & & \vdots & & \vdots & & \vdots & & \vdots & & \vdots \\
x_i^{\text{k}} & = & x_i^{\text{doi}} & \circ & x_i^{\text{authors}_k} & \circ & f_{prob}(x_i^{\text{ac}}, 1, k) & \circ & f_{prob}(x_i^{\text{ac}}, 2, k) & \circ \cdots \circ & f_{prob}(x_i^{\text{ac}}, n, k\ )
\end{array} \tag{4.3}$$

No Quadro 4.11 são mostradas as contribuições com as taxas atualizadas pela Equação 4.3 sobre a amostra anterior.

Quadro 4.11 – *Taxa atualizada de contribuições após a aplicação da equação de probabilidade.*

doi	authors	ac_conception_and_design	...	ac_statistical_expertise
10.7326/0003-4819-136-1-200201010-00009	Andrea Obernosterer	0,5	...	0
10.7326/0003-4819-136-1-200201010-00009	Manuela Aschauer	0	...	0
10.7326/0003-4819-136-1-200201010-00009	Wolfgang Schnedl	0	...	1
10.7326/0003-4819-136-1-200201010-00009	Rainer W. Lipp	0,5	...	0

Após a aplicação da Equação 4.2 o valor atribuído aos autores colocados na primeira e última linhas é $0,5$, diferente do valor 1 atribuído a ambos, Quadro 4.10. Este método

permite verificar de fato quanto um autor contribuiu em uma determinada categoria, Quadro 4.8.

Para cada artigo disponível na amostra, a soma das contribuições do autor em cada categoria é igual a $1,0$, representando 100% das contribuições. Como exemplo, na amostra anterior, a soma das contribuições de "ac_conception_and_design" é 1.0 ($0,5 + 0 + 0 + 0,5$), assim como "ac_statistical_expertise" é 1 ($0 + 0 + 1 + 0$).

4.6.2 *Dataset* Consolidado

Após os procedimentos realizados nas seções anteriores, foram obtidos três conjuntos de dados com variáveis relativas aos artigos e aos autores, incluindo suas contribuições.

Na Tabela 4.2 é apresentada a amostra contendo 2.024 artigos, abrangendo entre os anos 2000 e 2019, o que equivale a 901 do Ann. Intern. Med., 790 do JAMA e 333 do PLoS Med.'; além dos indicadores SCImago H-*index*, *SCImago Journal Rankings* (SJR) e JCR, referentes a periódicos para o ano 2019.

Tabela 4.2 – *Total de artigos selecionados dos periódicos Ann. Intern. Med., JAMA e PLoS Med., com os indicadores do SCImago (H-index), SJR and JCR referentes ao ano de* 2019.

Periódicos	**Artigos**	**SCImago H-index**	**SJR**	**JCR**
Ann. Intern. Med.	901	376	4,738	21,317
JAMA	790	654	5,913	45,540
PLoS Med.	333	215	5,616	10,500
Total	2.024	-	-	-

Fonte: Autor

Na Tabela 4.3 são mostrados os dados dos autores pertencentes à amostra, o número total de artigos e de autores exclusivos de cada periódico.

Tabela 4.3 – *Total de artigos, autores e autores únicos na amostra.*

Periódicos	**Artigos**	**Autores**	**Autores Únicos**
Ann. Intern. Med.	901	8.191	6.965
JAMA	790	7.780	6.683
PLoS Med.	333	4.127	3.737
Total	2.024	20.098	17.385

Fonte: Autor

Ann. Intern. Med. é o periódico que apresenta a maioria dos artigos, autores e autores únicos da amostra, seguido do JAMA. PLoS Med. é o último a ser apresentado. A amostra é bem balanceada, pois nenhum dos três periódicos apresenta mais do que 50% do tamanho

da amostra. Observa-se também que a diferença entre autores totais e únicos é menor que o esperado.

Os artigos utilizados têm 20.098 autores, dos quais 17.385 são únicos, pois alguns participaram em mais de uma publicação no período, descrito na Tabela 4.3. Este número varia entre 4 e 113 autores, onde a média é $14,61$ com desvio padrão $7,78$.

Ao longo deste capítulo foram apresentados os processos para coleta, preparação e limpeza dos dados de artigos, autores e contribuições, além da construção de robôs e algoritmos auxiliares para realização dessas tarefas.

O resultado desses processos permitiram responder às seguintes perguntas de pesquisa referentes a **lacuna de pesquisa 2 - "Identificar padrões para os processos de coleta, limpeza e preparação dos dados de bases científicas indexadas"**, descrita na Seção 1.2: (1) Como automatizar o processo de coleta de dados em bases científicas indexadas através de robôs usando técnicas de raspagem de dados? e (2) Como melhorar a preparação dos dados coletados para eliminação de anomalias, identificação de erros e duplicações de forma automática?

No próximo capítulo é apresentada a análise sobre as contribuições dos dados coletados.

5 ANÁLISE DE CONTRIBUIÇÕES

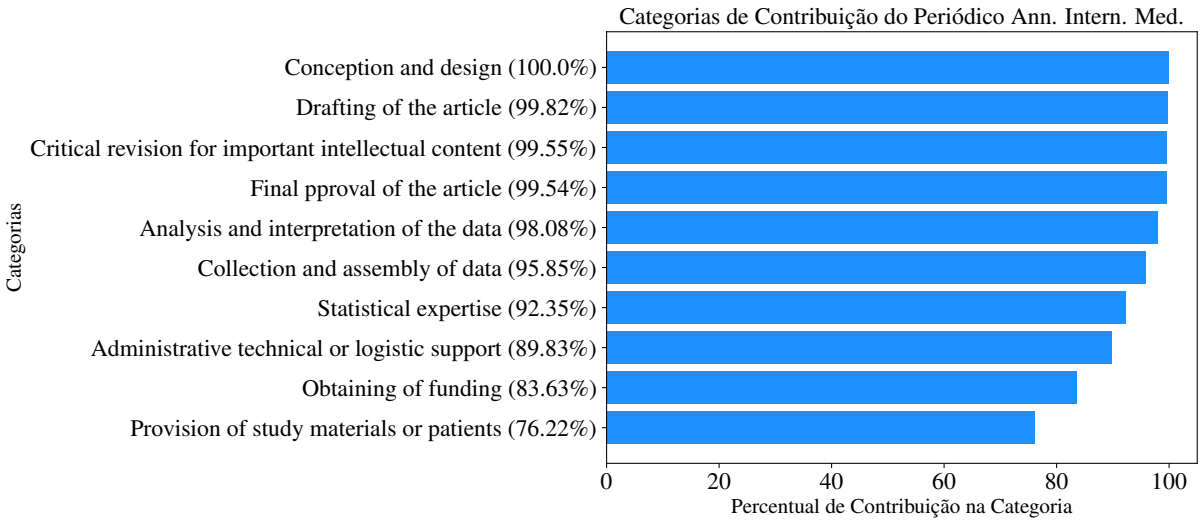
Resumo do capítulo

Neste capítulo é apresentado um estudo sobre as contribuições em publicações científicas utilizando análise fatorial, além de uma tabela de equivalência para categorias de contribuição entre os três periódicos analisados neste estudo. Por fim, apresenta uma proposta para o agrupamento de categorias de contribuições na área de ciências biológicas e medicina que responde parcialmente a seguinte pergunta de pesquisa: "Quais grupos emergem ao mapear a contribuição científica utilizando técnicas estatísticas em conjunto com a Ciência de Dados?", referente a Lacuna 2 desta pesquisa - "Identificar padrões de posição autoral e contribuições na autoria científica", descrita na Seção 1.2.

5.1 Visão Geral

Os periódicos Ann. Intern. Med., JAMA e PLoS Med. foram estudados individualmente de modo a verificar como as contribuições são distribuídas com os dados coletados e processados. Como o ICMJE não determina as nomenclaturas das contribuições, mas sim os critérios de contribuição, a diferença na declaração das contribuições entre os periódicos é significativa. Assim, fazer comparações diretas torna-se uma tarefa árdua e com resultados imprecisos. Na Figura 5.1 é mostrada que a distribuição das contribuições tem baixa variação no Ann. Intern. Med., divididas em dez categorias. A categoria "*Conception and design*" tem 100% de participação dos autores, seguindo recomendações do ICMJE quanto a contribuições substanciais. A menor contribuição "*Provision of study materials or patients*" mostra que nem todas as publicações contêm ensaios clínicos.

Figura 5.1 – *Distribuição do percentual de contribuições nas categorias do Ann. Intern. Med.*

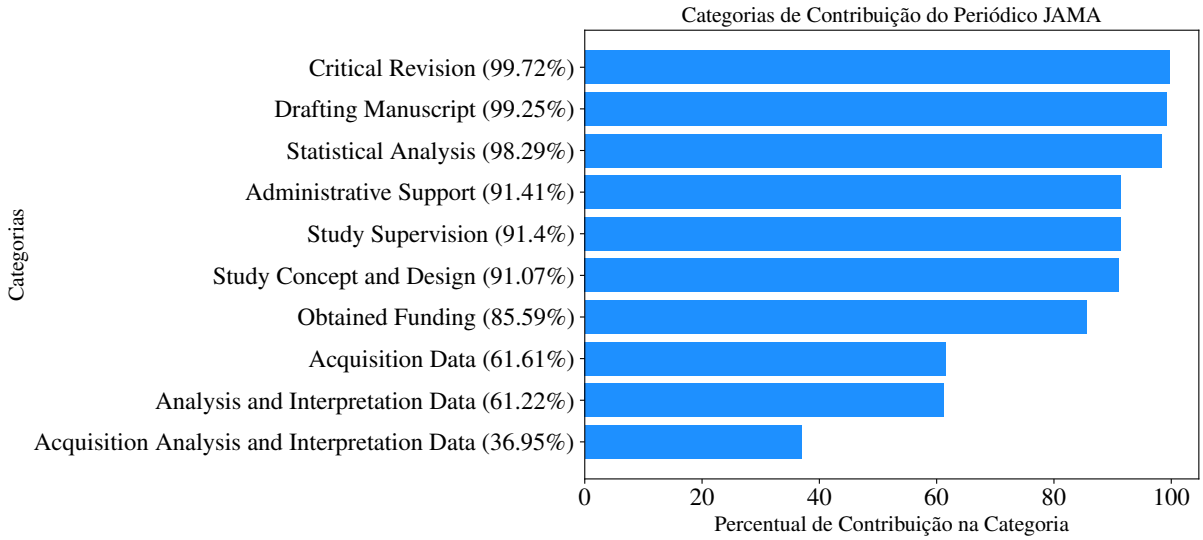


Fonte: Autor

As categorias "*Analysis and interpretation of the data*" e "*Statistical expertise*" possuem uma porcentagem semelhante, porém a segunda parece estar mais relacionada ao conhecimento do que efetivamente a análise dos dados.

Na Figura 5.2 é mostrada que a distribuição das contribuições no JAMA apresenta grande variação, estando divididas em dez categorias. A contribuição "*Critical Revision*" tem $99,72\%$ de participação, mostrando o forte envolvimento dos autores nesta categoria. "*Statistical Analysis*" tem $98,29\%$, indicando a participação da maioria dos autores nesta categoria. "*Analysis and Interpretation Data*" tem $61,22\%$, enquanto "*Acquisition Analysis and Interpretation Data*" $36,95\%$, mostrando que a separação em duas categorias que envolvem dados e são similares, pode causar dúvidas aos autores no momento de declará-las, influenciando a declaração em ambas.

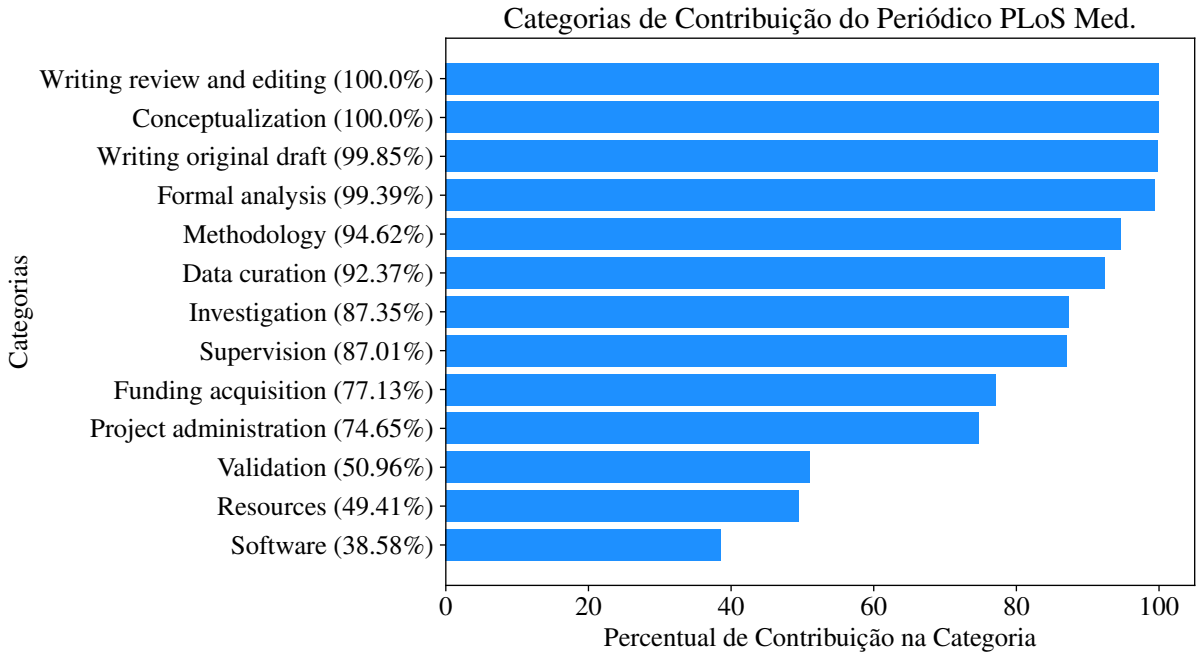
Figura 5.2 – *Distribuição do percentual de contribuições nas categorias do JAMA.*



Fonte: Autor

Por fim, na Figura 5.3 são mostradas as categorias de contribuição para do PLoS Med., que divide as contribuições em 13 categorias, não apresentando homogeneidade. As contribuições "*Writing review and editing*" e "*Conceptualization*" apresentam 100% de participação dos autores, seguindo as recomendações do ICMJE. Por ter suas categorias mais segmentadas em relação aos periódicos Ann. Intern. Med. e JAMA, permite ao autor ser mais pontual na indicação das contribuições. Entretanto, algumas categorias como, por exemplo, "*Software*" (um recurso), pode ser entendida também como "*Resources*" (recursos). Ademais, um número maior de categorias causa maior dispersão nos dados, dificultando a análise das contribuições. As demais categorias apresentam contribuições equivalentes.

Figura 5.3 – *Distribuição do percentual de contribuições nas categorias do PLoS Med.*



Fonte: Autor

Os valores apresentados evidenciam a dificuldade em realizar comparações entre as categorias listadas nos periódicos. Portanto, a seguir é descrito o estudo realizado utilizando análise fatorial para compreender a relação das categorias de contribuição entre os periódicos.

5.1.1 Proposta de Padronização das Contribuições de Autores

Como exposto anteriormente, as categorias de contribuição são diferentes entre os periódicos, onde Ann. Intern. Med. e JAMA possuem dez categorias e o PLoS Med. 13 categorias. Diante esse contexto, é proposta uma padronização para as contribuições em publicações científicas na área de ciências biológicas e medicina, divididas em sete categorias, mostradas no Quadro 5.1.

A padronização foi realizada a partir de uma análise sobre o significado de cada uma das categorias de contribuição disponíveis nos três periódicos. Embora ambos sigam as orientações do ICMJE quanto aos critérios de declaração para as contribuições, a lista de declarações não segue um critério rigoroso, diferindo significativamente (YANG; WOLFRAM; WANG, 2017). Isso ocorre devido o ICMJE orientar contribuições apenas no que diz respeito aos critérios e não à descrição. Assim, a política dos periódicos para declarações influencia os nomes e a expressão da categoria de contribuição. Desta forma, o agrupamento proposto no Quadro 5.1, coluna (*Proposal*), reúne categorias que possuem

Quadro 5.1 – *Proposta de padronização das contribuições autorais evidenciando o agrupamento das categorias disponibilizado pelos periódicos Ann. Intern. Med., JAMA e PLoS.*

Categorias	**Periódicos**		
Proposal	Ann. Intern. Med.	JAMA	PLoS Med.
Data Collection	Collection and assembly of data	(1) Acquisition data, (2) Acquisition, analysis, or interpretation data	Data curation
Study Conceptualization	Conception and design	Study concept and design	(1) Conceptualization (2) Investigation (3) Methodology
Study Supervision	Final approval of the paper	Study Supervision	(1) Supervision (2) Validation (3) Project Administration
Critical Revision	Critical revision for important intellectual content	Critical Revision the manuscript for important intellectual content	Writing – review editing
Original Draftting	Drafting of the article	Drafting of the manuscript	Writing – original draft
Statistical Analysis	(1) Statistical expertise (2) Analysis and interpretation of the data	(1) Statistical analysis (2) Analysis and interpretation data	Formal Analysis
Funding and/or Support	(1) Obtaining of funding (2) Administrative technical or logistics support (3) Provision of study materials or patients	(1) Obtained funding (2) Administrative, technical or material support	(1) Funding acquisition (2) Resources (3) Software

Fonte: Autor

argumentos semelhantes quanto ao significado das contribuições.

Enquanto no periódico PLoS Med. existem 13 categorias, no Ann. Intern. Med. e no JAMA são apenas dez. Essa diferença pode impactar positivamente as declarações no PLoS Med., visto que permite indicar com mais assertividade uma contribuição. Por exemplo, *"Supervision"*, *"Validation"* e *"Project Administration"* são contribuições que tendem a receber contribuições de um mesmo autor, pois estão relacionadas a supervisão e condução do estudo (ZBAR; FRANK, 2011). No mesmo contexto de significância, no JAMA só há uma categoria para essa indicação *"Study Supervision"*, diferenciando-se da anterior apenas pelo prefixo *"Study"*. Já no Ann. Intern. Med., a categoria relativa mais aproximada é *"Final approval of the paper"*.

Em termos de nomenclatura, ambos os periódicos possuem categorias que apresentam, de alguma forma, semelhanças em suas descrições. Quando comparadas à taxonomia CRediT (LARIVIÈRE; PONTILLE; SUGIMOTO, 2021), composta pelas categorias: (*Conceptualization*; *Data curation*; *Formal Analysis*; *Funding acquisition*; *Investigation*; *Methodology*; *Project administration*; *Resources*; *Software*; *Supervision*; *Validation*; *Visualization*; *Writing — original draft*; e *Writing — review & editing*), percebe-se a existência de relações entre as descrições. Isso mostra a dificuldade em indicar corretamente uma contribuição na ausência de padronização (HAGEN, 2014).

As categorias são usadas para representar as funções normalmente desempenhadas por autores para a produção científica acadêmica, onde cada uma descreve a contribuição específica de cada autor na produção (MCNUTT et al., 2018; DECULLIER; MAISONNEUVE, 2019). Em publicações com muitos autores, espera-se que as contribuições sejam indicadas de forma distribuída, ou seja, um trabalho que conta com pesquisadores especializados nos assuntos pertinentes à pesquisa. Entretanto, é possível encontrar diversas publicações onde todos os autores indicam participação em todas as categorias (RIVERA,

2018), não permitindo verificar a importância da contribuição de cada um.

Nesse aspecto, o perfil do autor, em relação as suas contribuições, deve ser considerado para a indicação ocorra de forma correta, refletindo sua real contribuição para o trabalho. Porém, as categorias existentes nos periódicos podem ter o mesmo significado, mas com nomenclaturas diferentes. Isso pode influenciar o autor a marcar contribuições diferentes, mas com o mesmo significado. Assim, a padronização das categorias propostas no Quadro 5.1 busca mostrar uma forma de guiar os autores a identificarem suas contribuições de acordo com o seu perfil.

A partir do modelo teórico proposto, foi realizada uma agregação dos dados dos três periódicos. Para isso, as categorias referentes em cada periódico, dentro do agrupamento, foram alocadas em novas variáveis a partir do cálculo da mediana entre elas.

O novo conjunto de dados é composto por 20.098 observações distribuídas em sete variáveis: "*Data Collection*" 51,49% (10.348), "*Study Conceptualization*" 48,83% (9.813), "*Study Supervision*" 51,80% (10.410), "*Critical Revision*" 84,24% (16.931), "*Original Drafting*" 36,64% (7.364), "*Statistical Analysis*" 56,55% (11.366), e "*Funding and/or Support*" 30,36% (6.102). Na Tabela 5.1 são mostrados os resultados das estatísticas sobre o agrupamento das categorias de contribuição entre os três periódicos.

Tabela 5.1 – *Estatísticas sobre os dados agrupados das categorias de contribuição dos periódicos Ann. Intern. Med., JAMA e PLoS Med.*

Estatísticas	**Média**	**Desv. Pad.**	**Min**	**Pctl(25)**	**Pctl(75)**	**Max**
Authors	14,613	13,050	4	8	16	113
Critical Revision	0,100	0,078	0,000	0,045	0,143	1,000
Original Drafting	0,100	0,180	0,000	0,000	0,167	1,000
Study Conceptualization	0,093	0,123	0,000	0,000	0,167	1,000
Statistical Analysis	0,091	0,133	0,000	0,000	0,125	1,000
Study Supervision	0,088	0,133	0,000	0,000	0,143	1,000
Data Collection	0,076	0,137	0,000	0,000	0,100	1,000
Funding and/or Support	0,060	0,136	0,000	0,000	0,111	1,000

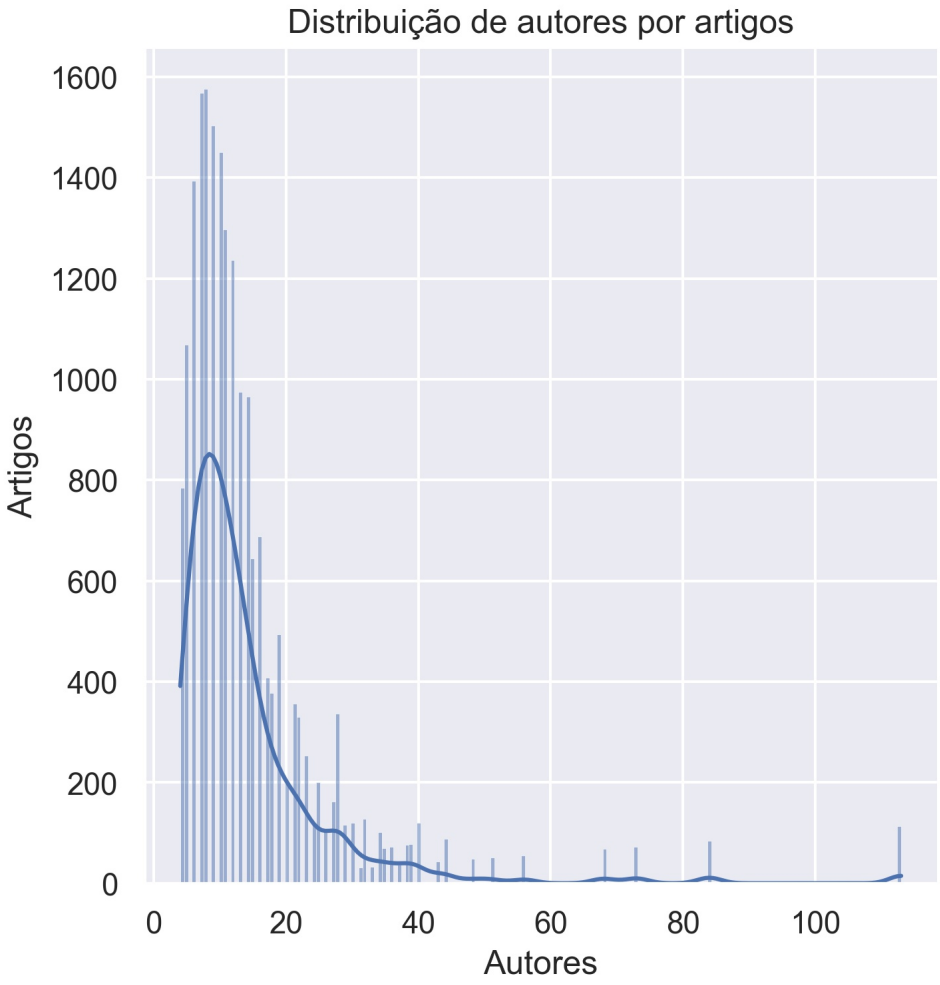
$n = 20.098$

Fonte: Autor

A amostra com 20.098 autores tem média de 14,613 autores por artigo, variando entre 4 e 113, confirmando o crescimento do número de autores em artigos nos últimos anos (ROSENZWEIG et al., 2008; PATIENCE et al., 2019). A Figura 5.4 mostra a distribuição dos autores nos artigos. A média das categorias mostra que os autores contribuem em muitas categorias, embora alguns contribuam mais especificamente. Os menores desvios estão relacionados às categorias "*Critical Revision*" (0,078), "*Study Conceptualization*" (0,123), "*Study Supervision*" (0,133) e "*Statistical Analysis*" (0,133), enquanto "*Funding*

and/or Support" (0, 136), "*Data Collection*" (0, 137) e "*Original Drafting*" (0, 180) com o maior.

Figura 5.4 – *Distribuição de autores por artigo.*



Fonte: Autor

5.1.2 Análise das Contribuições nos Periódicos

O estudo realizado para compreender o relacionamento das categorias de contribuição entre os periódicos foi realizado com a utilização de análise fatorial, que é uma técnica de análise multivariada que trata das relações de um conjunto inicial de variáveis, substituindo-as por outro conjunto menor, chamado de "fatores", os quais explicam a maior parte da variância do conjunto original (ARTES, 1998). Assim, responde questões relacionadas ao exame da possibilidade de algumas variáveis latentes serem responsáveis pela variação de muitas variáveis individuais, chamadas de "itens" (SCHREIBER, 2021). Na análise fatorial as variáveis latentes (fatores) são a "causa" das variáveis observáveis (itens), portanto a redução de dimensão por análise fatorial parte do pressuposto que há

erro de mensuração, pois as variáveis latentes (classificadas como causas) não são observadas. Já no método de análise de componentes principais (*principal components analysis, PCA*) há a mesma redução de fatores que a análise fatorial, porém, a noção de causalidade é reversa, pois na PCA os itens "causam" as variáveis latentes, portanto não há o pressuposto de erros mensuração já que todas as causas são observadas e a variável latente é a consequência dos itens.

Como o objetivo da fatorial é redução de variáveis, pode-se utilizar diversos métodos para estimar os fatores. Dentre esses: Eixos Principais Fatoriais (*Principal Axis Factoring, PAF*), Máxima Verossimilhança (*Maximum Likelihood, ML*); Componente Principal (CP); Mínimos Quadrados Generalizados; Mínimos Quadrados Não Ponderados e Fatoração Alfa (HONGYU, 2018). De forma geral, dentre esses métodos, o de componente principal (CP) e máxima verossimilhança(ML) são os mais recomendados quando as amostras apresentam distribuição não-normal e normal, respectivamente. (JOHNSON; WICHERN et al., 2002; COSTELLO; OSBORNE, 2005).

Além dos métodos de extração, utiliza o conceito de rotação dos fatores que se refere ao método matemático que rotaciona os eixos no espaço geométrico. Esse conceito visa tornar o resultado empírico encontrado mais facilmente interpretável, conservando as suas propriedades estatísticas (FIGUEIREDO FILHO; SILVA JÚNIOR, 2010). Em relação às rotações temos as ortogonais (a mais utilizada é a "varimax") que mantém os fatores não-correlacionados e as rotações oblíquas (a mais utilizada é a "oblimin") que permitem com que os fatores se correlacionem (HONGYU, 2018). Assim, neste trabalho, foi utilizada análise fatorial para estudar individualmente cada periódico, de modo a verificar como as contribuições se relacionam, utilizando o pacote "*psych*" do *software* `R`.

Antes da realização de uma análise fatorial é necessário aplicar alguns testes para verificar a normalidade dos dados, adequação e verificação de correlações. Desta forma, para verificar a normalidade dos dados dos periódicos Ann. Intern. Med., JAMA e PLoS Med., foi aplicado o teste de Shapiro-Wilk (ABRAMO; D'ANGELO, 2017) que resultou um p-valor inferior a $0,05$; mostrando que a amostra não é oriunda de uma população normalmente distribuída. Os testes de Barlett e Kaiser-Meyer-Olkin (KMO) foram aplicados para verificar a suposição de adequação e esfericidade da amostragem, o que confirmou que a análise fatorial era adequada para o conjunto de dados. Os valores obtidos no KMO indicaram $0,82$ para o Ann. Intern. Med., $0,64$ para o JAMA e $0,80$ para o PLoS Med. De acordo com Hutcheson e Sofroniou (1999), valores entre $0,5$ e 1 para o KMO são aceitáveis. O teste de Barlett mostrou que nos três periódicos o p-valor é menor que $0,005$; valor que permite a aplicação de análise fatorial, indicando que a matriz de intercorrelação não é uma matriz identidade, confirmando que existem combinações lineares e variáveis individuais (WATSON, 2017).

Para obter o número de fatores a serem extraídos, foi realizada uma análise paralela com 10.000 iterações com o conjunto de dados dos periódicos usando a rotação "varimax"

(BARTHOLOMEW et al., 2008; PERNEGER et al., 2017), e o método *Principal Axis Factoring (PAF)* (SCHREIBER, 2021). A análise paralela (HORN, 1965) é similar à técnica de *bootstrap*. São geradas N matrizes de correlação da mesma dimensão da matriz de correlação dos dados da amostra com elementos aleatórios e computa os autovalores dessas matrizes e retorna a densidade probabilística completa de cada um dos autovalores das N matrizes. Como os autovalores de uma matriz de correlação seguem uma distribuição decrescente monotônica, conseguimos usar o percentil 95% dos valores dos autovalores das N matrizes como critério de retenção e seleção dos números de fatores. A lógica por trás da análise paralela é que você deve manter apenas fatores com autovalores maiores que 95% de fatores gerados por matrizes aleatórias. Além disso, para a avaliação de confiabilidade de uma estrutura fatorial é utilizado um indicador denominado de Alpha (α) *Cronbah*, que avalia o grau em que os itens de uma matriz de dados estão correlacionados entre si (HONGYU, 2018). De acordo com Ekolu e Quainoo (2019), uma interpretação comum do coeficiente α *Cronbach's* é $\alpha < 0,5$ para confiabilidade baixa; $0,5 < \alpha < 0,8$ para confiabilidade moderada (aceitável); $\alpha > 0,8$ para alta confiabilidade.

Uma vez obtidos os fatores, deve-se avaliar a significância das cargas fatoriais carregadas em cada um. As cargas fatoriais são as estimativas dos fatores (contribuições de cada variável aos fatores) (HAIR et al., 2009). Seus valores devem ser superiores a $0,30$ (HINKIN, 1995) ou $0,40$ (HINKIN, 1998) para que tenham significância ou sejam minimamente aceitáveis Hair et al. (2009). Por fim, Maskey, Fei e Nguyen (2018) defendem valores maiores que $0,50$ para que sejam considerados para significância prática, visto que cargas fatoriais baixas e cargas cruzadas são os principais motivos usados por muitos autores para excluir uma variável. Desta forma, não há consenso sobre o que constitui uma carga fatorial "alta" ou "baixa" (PETERSON, 2000).

Assim, a partir dos dados agrupados, foram realizados testes para verificar a normalidade, adequação e verificação de correlações. Desta forma, foi aplicado o teste de Shapiro-Wilk (ABRAMO; D'ANGELO, 2017), que resultou um p-valor inferior a $0,05$; mostrando que os dados não seguem uma distribuição normal. O teste KMO apresentou o valor de $0,76$, e um p-valor menor que $0,005$ para o teste de Barlett, permitindo continuidade com as análises.

Em seguida foi realizada uma análise fatorial com o parâmetro de rotação "varimax" e com o método PAF para três fatores. Os resultados obtidos mostraram valores fora do intervalo aceitável, apresentando muitas cargas fatoriais cruzadas, além do *Cronbach's* α inferior a $0,50$, mostrando baixa confiabilidade (EKOLU; QUAINOO, 2019). Assim, uma nova análise fatorial foi realizada para dois fatores, mantendo os parâmetros. Os resultados obtidos mostraram cargas fatoriais satisfatórias, apresentadas na Tabela 5.2.

Tabela 5.2 – *Cargas fatoriais para os dados agrupados pela mediana das categorias dos periódicos Ann. Intern. Med., JAMA e PLoS Med.*

Categorias	**Fator 1**	**Fator 2**
Study Conceptualization	0,665	0,291
Study Supervision	0,605	0,122
Funding and/or Support	0,484	0,095
Critical Revision	0,435	0,237
Data Collection	0,265	0,247
Statistical Analysis	0,100	0,804
Original Drafting	0,367	0,469
Cronbach's α	*0.637*	*0.567*

Fonte: Autor

A Tabela 5.3 mostra as estatísticas sobre os fatores apresentados na Tabela 5.2.

Tabela 5.3 – *Estatísticas sobre os fatores do agrupamento dos periódicos.*

Descrição	**Fator 1**	**Fator 2**
Soma das cargas fatoriais^2	1,43	1,09
Proporção da variância explicada	0,20	0,16
Variância acumulada	0,20	0,36

Fonte: Autor

Conforme mostrado na Tabela 5.2, o valor do *Cronbach's* α dos fatores se apresentou dento dos limites Ekolu e Quainoo (2019), permitindo a classificação das categorias propostas. Entre as sete variáveis analisadas, três apresentaram cargas fatoriais acima de $0,50$ (*"Study Conceptualization"*, *"Study Supervision"* e *"Statistical Analysis"*); três superiores a $0,40$ (*"Funding and/or Support"*, *"Critical Revision"* e *"Original Drafting"*) e uma muito abaixo de $0,40$ (*"Data Collection"*) com $0,265$ que ainda apresenta carga cruzada entre os fatores.

Apesar de apresentar uma carga fatorial muito baixa, a variável *"Data Collection"* não pode ser agregada com outras variáveis e nem removida, visto que descreve isoladamente uma categoria de contribuição. Assim, de acordo com (MATOS; RODRIGUES, 2019), apesar de apresentar carga fatorial abaixo dos limites aceitáveis, deve-se considerar sua importância relativa à contribuição para o fator. Segundo Watkins (2018), uma forma de verificar o valor da contribuição para o fator é elevando a carga fatorial ao quadrado para saber a proporção de variância explicada. Neste caso, a variável *"Data Collection"* tem uma carga fatorial de $0,265$, resultando em $0,07$ (7%) de variância explicada. Este valor pode ser considerado significativo de acordo com o número de variáveis existentes.

Portanto, a variável *"Data Collection"* foi mantida.

Embora a carga fatorial da variável *"Data Collection"* (coleta, preparação e disponibilização de dados) esteja mais associada ao fator um, a característica desta categoria de contribuição foi associada ao valor do fator dois, visto que seu conceito está mais relacionado a "*Original Drafting*" (elaboração, criação e/ou apresentação do trabalho publicado, especificamente escrevendo o rascunho inicial (incluindo tradução substantiva) e "*Statistical Analysis*" (aplicação de técnicas estatísticas, matemáticas, computacionais ou outras técnicas formais para analisar ou sintetizar dados do estudo). Estas categorias representam funções metodológicas e/ou logísticas para o desenvolvimento de um artigo.

A correlação entre as categorias de contribuição permitiu um agrupamento bem definido das categorias, considerando a proposta mostrada na tabela 5.1.

Na Tabela 5.4 são apresentadas as estatísticas das contribuições do periódico Ann. Intern. Med. dentro do agrupamento pela mediana, contendo 8.191 autores. Nela pode-se visualizar que a média das contribuições mostra um comportamento bem equilibrado, exceto para "*Funding and/or Support*" $(0,055)$, indicando que poucos autores são responsáveis por esta contribuição. As demais evidenciam que os autores contribuem em muitas categorias.

Tabela 5.4 – *Estatísticas sobre os dados individuais das categorias de contribuição do periódico Ann. Intern. Med., calculados pela mediana dentro do agrupamento das contribuições.*

Estatísticas	**Média**	**Desv. Pad.**	**Min**	**Pctl(25)**	**Pctl(75)**	**Max**
Authors	12,161	7,528	4	7	15	48
Original Drafting	0,110	0,170	0,000	0,000	0,200	1,000
Study Conceptualization	0,110	0,129	0,000	0,000	0,200	1,000
Critical Revision	0,109	0,093	0,000	0,000	0,167	1,000
Study Supervision	0,109	0,081	0,000	0,056	0,143	1,000
Data Collection	0,103	0,169	0,000	0,000	0,167	1,000
Statistical Analysis	0,103	0,130	0,000	0,000	0,125	1,000
Funding and/or Support	0,055	0,148	0,000	0,000	0,000	1,000

$n = 8.191$

Fonte: Autor

O intervalo entre os índices das categorias mostra a existência de quatro grupos de contribuição: *"Funding and/or Support"* $(0,055)$; *"Data Collection"* e *"Statistical Analysis"* $(0,103)$; *"Study Supervision"* e *"Critical Revision"* $(0,109)$; *"Study Conceptualization"* e *"Original Drafting"* $(0,110)$. Entretanto, as categorias *"Study Conceptualization"*, *"Study Supervision"*, *"Critical Revision"* e *"Original Drafting"* possuem valores aproximados.

Na Tabela 5.5 são mostradas as estatísticas de contribuição do periódico JAMA calculados pela mediana dentro do agrupamento das contribuições de 7.780 autores.

Tabela 5.5 – *Estatísticas sobre os dados individuais das categorias de contribuição do periódico JAMA, calculados pela mediana dentro do agrupamento das contribuições.*

Estatísticas	**Média**	**Desv. Pad.**	**Min**	**Pctl(25)**	**Pctl(75)**	**Max**
Authors	13,221	8,198	4	8	16	51
Critical Revision	0,101	0,068	0,000	0,059	0,143	0,500
Original Drafting	0,101	0,169	0,000	0,000	0,167	1,000
Study Conceptualization	0,094	0,124	0,000	0,000	0,167	1,000
Study Supervision	0,094	0,174	0,000	0,000	0,143	1,000
Funding and/or Support	0,087	0,137	0,000	0,000	0,125	1,000
Statistical Analysis	0,084	0,109	0,000	0,000	0,125	1,000
Data Collection	0,050	0,066	0,000	0,000	0,071	0,500

$n = 7.780$

Fonte: Autor

Na Tabela 5.5, a categoria "*Data Collection*" $(0,050)$ é a que apresenta menor média, indicando que esta contribuição é restrita a poucos autores. As categorias *"Funding and/or Support"* $(0,087)$ e *"Statistical Analysis"* $(0,084)$ apresentam pequena variação entre si. Já as categorias *"Critical Revision"* e *"Original Drafting"* $(0,101)$ possuem o mesmo índice de contribuição, o que ocorre também com *"Study Conceptualization"* e *"Study Supervision"* $(0,094)$. O intervalo entre os índices das categorias mostra a existência de três grupos de contribuição. Destaca-se que o índice de *"Data Collection"* $(0,050)$ é $53,48\%$ menor que os das demais categorias $(0,094)$.

Por fim, na Tabela 5.6 são apresentadas as estatísticas das contribuições do periódico PLoS Med. calculados pela mediana dentro do agrupamento das contribuições.

Tabela 5.6 – *Estatísticas sobre os dados individuais das categorias de contribuição do periódico PLoS Med., calculados pela mediana dentro do agrupamento das contribuições.*

Estatísticas	**Média**	**Desv. Pad.**	**Min**	**Pctl(25)**	**Pctl(75)**	**Max**
Authors	22,106	22,773	4	9	26	113
Critical Revision	0,081	0,059	0,000	0,036	0,111	0,500
Original Drafting	0,080	0,215	0,000	0,000	0,000	1,000
Statistical Analysis	0,080	0,174	0,000	0,000	0,100	1,000
Data Collection	0,072	0,152	0,000	0,000	0,091	1,000
Study Conceptualization	0,058	0,098	0,000	0,009	0,108	1,000
Study Supervision	0,033	0,107	0,000	0,000	0,000	1,000
Funding and/or Support	0,020	0,094	0,000	0,000	0,000	1,000

$n = 4.127$

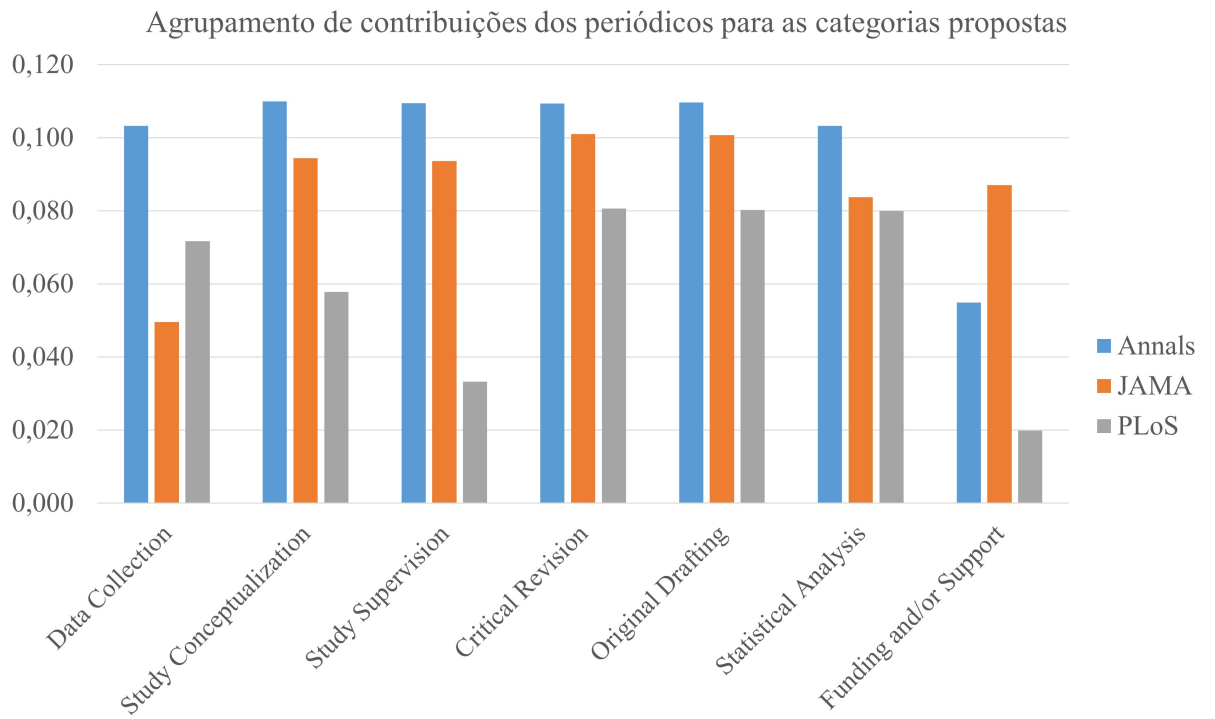
Fonte: Autor

A amostra do PLoS Med. com 4.127 autores, Tabela 5.6, mostra que as contribuições têm grande variação entre as categorias, indicando que os autores contribuem em muitas categorias, embora alguns sejam muito específicos na indicação da contribuição. Essa afirmação pode ser constatada verificando a média das categorias *"Study Conceptualization"* $(0,058)$, *"Study Supervision"* $(0,033)$ e *"Funding and/or Support"* $(0,020)$. Assim, as categorias *"Critical Revision"* $(0,081)$, *"Original Drafting"* $(0,080)$, *"Statistical Analysis"* $(0,080)$ e *"Data Collection"* $(0,072)$ podem ser agrupadas, visto que os valores indicam que os autores contribuem de forma semelhante.

Entre os periódicos, PLoS Med. tem o maior número médio de autores (22.106), seguido por JAMA (13.221) e Ann. Intern. Med. (12.161) que apresenta a maior média de contribuições $(0,100)$, seguido por JAMA $(0,087)$ e pelo PLoS Med. $(0,061)$ com desvio padrão $(0,061)$. O periódico Ann. Intern. Med. é o que apresenta as maiores contribuições e menor variação na distribuição, utilizando nomenclaturas explícitas para as categorias (PERNEGER et al., 2017), o que pode facilitar o entendimento para a indicação das contribuições, seguido por JAMA e PLoS Med. O periódico JAMA se destaca pela categoria "*Data Collection*" apresentar menos contribuições $(0,050)$ em comparação com o Ann. Intern. Med. $(0,103)$ e PLoS Med. $(0,072)$. Uma indicação é a existência de duas categorias relacionadas à coleta de dados *"Acquisition Data"* e *"Acquisition Analysis and Interpretation Data"*, causando uma distorção na indicação correta da contribuição, que pode ser realizada em ambas. No periódico PLoS Med., as contribuições são mais dispersas, indicando que os autores não compartilham muitas contribuições, exceto para as categorias *"Original Drafting"* $(0,080)$, *"Statistical Analysis"* $(0,080)$ e *"Critical Revision"* $(0,081)$ que possuem média de contribuições $(0,080)$, representando $(56,95\%)$ superior as demais que têm média $(0,0458)$. Uma evidência sobre esse achado pode estar relacionada com a adoção do periódico PLoS Med. à taxonomia CRediT em 2016 e que até o final de 2018 já mantinha mais de 30.000 artigos empregando essa nova taxonomia (LARIVIÈRE; PONTILLE; SUGIMOTO, 2021), visto que os artigos utilizados neste estudo estão compreendidos no intervalo entre os anos de 2000 e 2019. Por fim, as contribuições relacionadas a *"Funding and/or Support"* diferem em ambos os periódicos. Na Figura 5.5 é mostrada a média das contribuições de cada periódico dentro do agrupamento pela mediana.

Na Figura 5.5 é mostrado que as categorias do Ann. Intern. Med. e do JAMA mantém um comportamento equilibrado, variando para *"Data Collection"*, que também ocorre quando comparado com o PLoS Med.; e *"Funding and/or Support"* que varia entre os três periódicos. Ainda, as categorias *"Critical Revision"*, *"Original Drafting"* e *"Statistical Analysis"* têm comportamento similar, evidenciando que em ambos os periódicos esta categoria recebe contribuição de muitos autores. A categoria *"Study Conceptualization"*, no PLoS Med., tem um índice menor de contribuição em relação ao Ann. Intern. Med. e JAMA. Quando comparada com as demais categorias, mostra que as contribuições em *"Study Conceptualization"* são mais significativas. Ademais, os dados permitem verificar

Figura 5.5 – *Contribuições médias categorizadas entre os periódicos.*



Fonte: Autor

que o periódico Ann. Intern. Med. tem maior participação de autores que contribuem em muitas categorias, dividindo as responsabilidades. No JAMA e no PLoS Med. as contribuições têm participação de menos autores, apesar de manterem uma relação direta, exceto para *"Data Collection"* que é menor no JAMA.

Os resultados confirmaram que os periódicos apresentam características diferentes com relação à indicação na lista de contribuições (YANG; WOLFRAM; WANG, 2017), apesar de possuírem um perfil parecido no que diz respeito ao conteúdo das publicações, visto que são periódicos renomados de medicina geral (WINGO et al., 2019). Além da diferença na nomenclatura das categorias, alguns fatores podem influenciar na declaração das contribuições como práticas institucionais para a medição precisa da produtividade, considerando a contribuição real da equipe de pesquisa para a produção produzida em colaboração com outras organizações (ABRAMO; D'ANGELO; ROSATI, 2013) e consideração de princípios institucionais, éticos e de boas práticas (FERREIRA; SERRA, 2015).

5.1.3 Agrupamento das Categorias

A partir da proposta do agrupamento das contribuições, Tabela 5.1 e dos resultados obtidos com aplicação de análise fatorial, foi realiza uma interpretação dos fatores para criação dos grupos de contribuição. As categorias, em ambos os periódicos, traduzem conceitos relacionados a metodologia, logística, teoria e supervisão. Assim, as categorias de contribuição (fatores) foram relacionadas em dois grupos: teoria ("*Theory*") para o fator um e metodológicas e/ou logísticas ("*Methodology/Logistics*") para o fator dois. Os valores referentes as cargas fatoriais, estatísticas sobre os dados, e a significância da descrição das categorias direcionaram as decisões para divisão em dois grupos.

De forma similar, Perneger et al. (2017) realizaram um estudo sobre categorias de contribuição do periódico Ann. Intern. Med. para artigos publicados no ano de 2015 e encontraram a existência de três grupos de contribuição com base nos dados de contribuição, titulação dos pesquisadores, afiliação, entre outras, descrevendo-as como *"Thinker"*, *"Soldier"* e *"Scribe"*. Segundo os autores, o estudo foi realizado apenas com o periódico Ann. Intern. Med., visto uma pesquisa piloto revelou que os periódicos proeminentes da área de ciências biológicas e medicina usam diferentes sistemas para relatar as contribuições, impedindo uma análise agrupada. Os grupos propostos pelos autores são semelhantes aos encontrados neste estudo para o mesmo periódico, embora aqui tenham sido utilizados apenas os dados de contribuição, mostrando que a metodologia e a técnica utilizada para o objetivo de agrupar categorias foi apropriada.

Assim, o grupo "*Theory*" representa contribuições teóricas, que são contribuições que se relacionam de alguma forma com a argumentação teórica do artigo. Abrange contribuições relacionadas ao conceito ("*Study Concept*") e supervisão do estudo ("*Study Supervision*"), revisão crítica ("*Critical Revision*") e financiamento e/ou suporte ("*Funding and/or Support*"). Este último, embora não esteja diretamente relacionado ao argumento teórico, é geralmente realizado pelo chefe do laboratório ou mentor, ocupando uma posição sênior vinculada à supervisão do estudo (ZBAR; FRANK, 2011; KOZMA et al., 2014). De acordo com (JABBEHDARI; WALSH, 2017), autores que atuarem especificamente na aquisição de financiamento não devem informar contribuições nesta categoria.

O grupo "*Methodology/Logistic*" representa contribuições metodológicas e logísticas, tais como preparação, criação ou apresentação ("*Original Draft*"), análises estatísticas ("*Statistical Analysis*"), coleta e manipulação de dados ("*Data Collection*"). Este grupo difere de *"Theory"* no que tange as ações praticadas no desenvolvimento do artigo.

De acordo com os dados analisados (20.098 observações) foram calculados os percentuais de contribuição dos autores nas categorias após o agrupamento dos dados, conforme mostrado no Quadro 5.2.

Quadro 5.2 – *Porcentagem de contribuições dos autores nas categorias após o agrupamento proposto referentes a amostra com* 20.098 *autores.*

Grupo	**Categorias**	**Autores**	**Contribuições**
Theory	Critical Revision	16.931	84,24%
	Study Supervision	10.410	51,80%
	Study Concept	9.813	48,83%
	Funding and/or Support	6.102	30,36%
Methodology/Logistic	Statistical Analysis	11.366	56,55%
	Data Collection	10.348	51,49%
	Original Draft	7.364	36,64%

Fonte: Autor

A análise dos dados mostra que a proposta de agrupamento das categorias de contribuição pode oferecer condições para autores analisarem o nível de importância de cada uma das contribuições em artigos científicos da área de ciências biológicas e medicina, visto que ainda falta orientação para isto (YANG; WOLFRAM; WANG, 2017). A representação das contribuições no grupo teórico *"Theory"* têm média de $53,81\%$ de contribuição com expressivos $(84,24\%)$ para *"Critical Revision"*, *"Study Supervision"* $(51,80\%)$, *"Study Concept"* $(48,83)$, e *"Funding and Support"* $(30,36\%)$, mostrando maior importância sobre o argumento teórico.

No grupo *"Methodology/Logistic"* os valores das categorias *"Statistical Analysis"* $(56,55\%)$, *"Data Collection"* $(51,49\%)$, e *"Original Draft''* $(36,64\%)$ mostram a importância no desenvolvimento do artigo no que tange sua operacionalização.

A ordem das categorias agrupadas seguem um critério de classificação baseada sobre os resultados numéricos obtidos, embora possam ser encontrados na literatura alguns fatores que podem tendenciar a indicação de contribuições como, por exemplo, em publicações com colaboração de pesquisadores de países diferentes (MATTSSON; SUNDBERG; LAGET, 2011). Outro fator é o crescente número de publicações na área de ciências biológicas e medicina onde dois ou mais autores tem as mesmas contribuições indicadas (DOTSON, 2013).

Neste estudo foram analisadas as informações sobre a contribuição autoral com base apenas nos dados relatados em publicações científicas da área de ciências biológicas e medicina, utilizando análise estatística para mostrar quais contribuições são mais relevantes e como se relacionam. É importante destacar que não há padronização na descrição das contribuições entre os periódicos e que elas são informadas diretamente pelos autores, o que pode ser superestimado por alguns (HAGEN, 2014), principalmente se as contribuições de todos os autores foram informadas pelo autor que submeteu o artigo.

Em suma, os resultados sobre a análise fatorial permitiram agrupar as categorias de contribuição em dois grupos que podem apoiar os autores de acordo com suas habilidades para contribuir em artigos científicos. Os resultados mostraram que as maiores contribuições se encontram no grupo teórico ("*Theory*") (ZBAR; FRANK, 2011), sinalizando que a experiência acadêmica dos autores é um fator preponderante. No entanto, contribuições metodológicas e logísticas (*"Methodology/Logistic"*) são essenciais para a operacionalização e aplicação dos conceitos teóricos, representando contribuições substanciais (PERNEGER et al., 2017; RAHMAN et al., 2017).

Assim, este capítulo responde parcialmente a seguinte pergunta de pesquisa: "Quais grupos emergem ao mapear a contribuição científica utilizando técnicas estatísticas em conjunto com a Ciência de Dados?", referente a **Lacuna 2 desta pesquisa — "Identificar padrões de posição autoral e contribuições na autoria científica"**, descrita na Seção 1.2.

No capítulo seguinte é apresentado o estudo sobre o posicionamento autoral e contribuições.

6 POSIÇÃO AUTORAL E CONTRIBUIÇÕES

Resumo do capítulo

Neste capítulo são apresentadas as hipóteses, o agrupamento das contribuições e o desenho dos modelos utilizados para estudar a relação entre o posicionamento autoral e as contribuições em publicações científicas. Apresenta ainda uma abordagem numérica para responder às perguntas de pesquisa referentes a Lacuna 2 desta pesquisa "Identificar padrões de posição autoral na autoria científica", descrita na Seção 1.2.

6.1 Visão Geral

No capítulo anterior, Capítulo 5, foi analisado o relacionamento entre as categorias de contribuição dos periódicos Ann. Intern. Med., JAMA e PLoS Med. exclusivamente sobre os dados de contribuição, excluindo-se quaisquer relações com os outros índices referentes aos autores. Os resultados permitiram o agrupamento das categorias de contribuição em dois grupos: (*"Theory"* e *"Methodology/Logistic"*), representando contribuições teóricas e, metodológicas e logísticas, respectivamente.

Neste capítulo é apresentado o estudo realizado para analisar a relação entre o posicionamento autoral e as contribuições científicas em artigos, objetivo principal deste trabalho, expandindo-se o agrupamento das categorias de duas para três.

Os dados utilizados foram os mesmos do estudo anterior, ampliando-se o número de variáveis, as quais relacionam dados dos autores e de suas contribuições. Para isto, foi adicionada uma métrica denominada *Byline* como variável independente, calculada sobre a posição de um autor na lista de autores, além da utilização da variável H-*index* para controlar efeitos que as hipóteses não abrangem.

O estudo foi realizado aplicando-se a técnica de Regressão Linear que, segundo (STOROPOLI; VILS, 2021) "[...] permite com que você use uma ou mais variáveis discretas ou contínuas como variáveis independentes e mensurar o poder de associação com a variável dependente, que deve ser contínua". Assim, a partir de um conjunto de observações, o objetivo é buscar um modelo que melhor explique a relação entre duas ou mais variáveis quantitativas pertencentes a um determinado fenômeno (MAZUCHELI; ACHCAR, 2002).

Na regressão linear a relação entre as variáveis é representada através de um modelo matemático, ou seja, uma equação que associa a uma variável dependente y, cujo comportamento será explicado pela variável independente (x). Essa equação é descrita matematicamente como $a + bx$, ou seja, a equação da reta. Desta forma, o que se busca com a regressão é encontrar valores para a e b, estimando a inclinação da reta, que fornece o efeito em y de uma unidade em x (CHEIN, 2019).

A interpretação geométrica mostra qual é a melhor reta que representa o conjunto de

dados, considerando todas as observações. Do ponto de vista matemático, está relacionada com otimização, visando encontrar a melhor reta entre os pontos que minimiza o erro quadrático médio (*Mean Squared Error* - MSE). Por fim, a interpretação estatística da regressão linear controla os efeitos das diferentes variáveis independentes ao calcular o efeito de uma certa variável independente (DEGROOT, 2012).

Para interpretar os resultados de uma regressão como uma quantidade estatística significativa, precisamos contar com uma série de suposições clássicas (MATLOFF, 2009; STOROPOLI; VILS, 2021): independência e linearidade dos dados; independência e homogeneidade de variância dos erros/resíduos e; ausência de multicolinearidade. Os resultados de uma regressão linear podem ser verificados por meio de um gráfico de dispersão.

Portanto, o objetivo do modelo de regressão linear é, a partir dos valores observados em um conjunto de dados, obter valores $\widehat{\beta_0}$, $\widehat{\beta_1}$ e suas variâncias (CHEIN, 2019).

Com base no exposto, a técnica de regressão linear é aplicada neste estudo para verificar o efeito das variáveis de contribuição científica relacionadas ao posicionamento autoral. Nas seções a seguir são descritos os procedimentos para realização deste estudo.

6.2 Hipóteses

Este estudo se concentra na investigação do efeito das contribuições científicas no posicionamento autoral na área de ciências biológicas e medicina. Assim, são apresentadas as seguintes hipóteses:

H_1: Autores presentes nas posições iniciais do *byline* apresentam maior probabilidade de anotarem contribuições relacionadas à relevância, conceito, redação e condução do estudo (ZBAR; FRANK, 2011). Para Tscharntke et al. (2007) o posicionamento autoral deve ocorrer em uma escala de importância decrescente em relação à contribuição. Outra abordagem seria realizar a alteração no índice "H" para atribuir créditos e posicionar o autor de acordo com um índice obtido pelas contribuições no trabalho (LIU; FANG, 2012). Estudos mais recentes confirmam que os níveis de participação nas listas de contribuição em artigos são maiores para os primeiros autores (PERNEGER et al., 2017; YANG; WOLFRAM; WANG, 2017; PATEL et al., 2019). Assim, contribuições teóricas (*"theoretical contributions"*) estão **negativamente** relacionadas à posição do autor (*"author position"*).

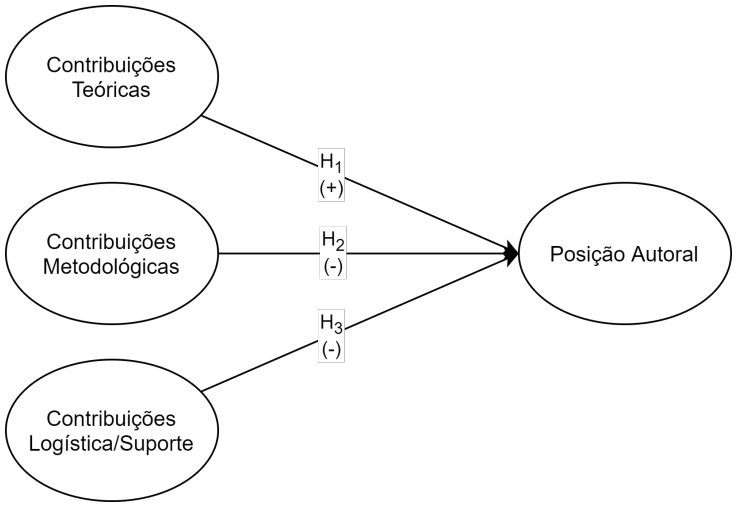
H_2: Contribuições metodológicas agregam informações sobre vários aspectos diferentes, apresentando desafios para extrair informações sobre os tipos específicos de contribuições feitas por um autor como, por exemplo, conceituação *versus* análise de dados (SAUERMANN; HAEUSSLER, 2017). Observa-se que autores indicam em muitos casos, além de contribuições teóricas, outras relacionadas a logística, como coleta de dados (YANG; WOLFRAM; WANG, 2017) e a realização de experimentos (CORRÊA et al., 2017). Assim, contribuições metodológicas (*"methodological contributions"*) estão **nega-**

tivamente relacionadas à posição do autor *"author position"*.

H_3: Contribuições logísticas são aquelas que não contribuem para o desenvolvimento intelectual do artigo e tendem a posicionar o autor como último (ZBAR; FRANK, 2011). Essas contribuições relacionam atividades ligadas a tradução de textos, coleta de dados, aquisição de financiamento (PERNEGER et al., 2017), fornecimento de laboratório e supervisão do estudo (RAHMAN et al., 2017), incluindo aqueles que desempenham funções de mentoria (PATEL et al., 2019). Assim, contribuições logísticas (*Logistics/Support contributions*) estão **positivamente** relacionadas à posição do autor *author position*.

Na Figura 6.1 é mostrada uma representação gráfica das variáveis/constructos desse estudo e suas hipóteses.

Figura 6.1 – *Framework de pesquisa para as variáveis/constructos desse estudo e suas hipóteses.*



Fonte: Autor

6.3 Modelagem das Variáveis

Para o desenvolvimento do estudo, foram selecionados os dados anteriormente coletados e tratados dos três periódicos da área de ciências biológicas e medicina: Ann. Intern. Med., JAMA e PLoS Med., no período compreendido entre os anos de 2000 e 2019, totalizando 20 anos. A amostra é composta por 2.024 artigos.

Na Tabela 6.1 são mostrados os números referentes a amostra, sendo 901 artigos do Ann. Intern. Med., 790 do JAMA e 333 do PLoS Med. Também são apresentados os indicadores SCImago H-*index*, *SCImago Journal Rankings* (SJR) e o JCR, referentes aos três periódicos para o ano de 2019.

Tabela 6.1 – *Total de artigos selecionados dos periódicos Ann. Intern. Med., JAMA e PLoS Med., com os indicadores SCImago (H-index), SJR e JCR referentes ao ano de* 2019.

Periódico	Artigos	SCImago H-index	SJR	JCR
Ann. Intern. Med.	901	376	4,738	21,317
JAMA	790	654	5,913	45,540
PLoS Med.	333	215	5,616	10,500
Total	2.024	-	-	-

Fonte: Autor

Na Tabela 6.2 são apresentados os dados sobre os autores pertencentes a amostra, o total de artigos e o número de autores exclusivos para cada periódico. O periódico Ann. Intern. Med. é o que apresenta a maioria dos artigos, autores e autores únicos na amostra, seguido pelo JAMA. O periódico PLoS Med. é o último a ser apresentado. A amostra é bem equilibrada, uma vez que nenhum dos três periódicos tem mais de 50% do tamanho da amostra. Observa-se também que a diferença entre autores totais e autores únicos é menor que o esperado. Da amostra de 20.095 autores, 2.710 publicaram mais de uma vez nesses periódicos durante o período analisado, justificando a difrença entre o número de autores e autores únicos.

Tabela 6.2 – *Total de artigos, autores e autores únicos na amostra.*

Periódico	Artigos	Autores	Autores Únicos
Ann. Intern. Med.	901	8,191	6,965
JAMA	790	7,779	6,683
PLoS Med.	333	4,125	3,737
Total	2,024	20,095	17,385

Fonte: Autor

6.3.1 Variável Dependente - *Byline*

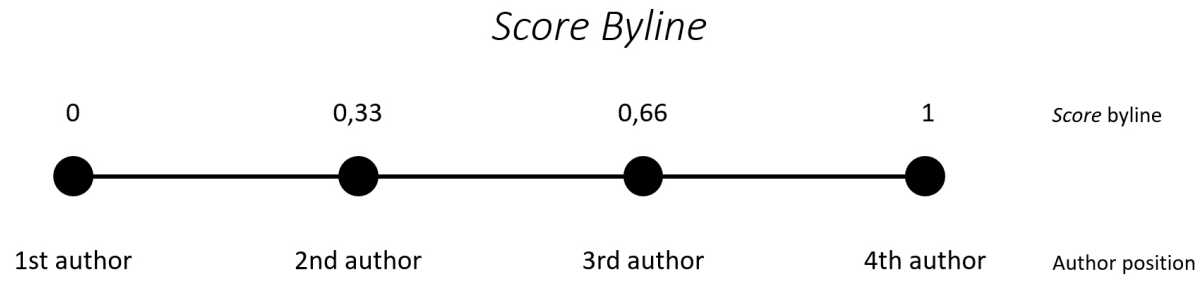
Para identificar o posicionamento autoral de cada autor, foi implementada uma métrica denominada de *Byline*. Esta métrica analisa a posição de um autor na lista de posicionamento autoral, descrita no artigo, a qual varia entre 0 para o primeiro autor e 1 para o último. Esta métrica, descrita na Equação 6.1, permite valorar o posicionamento dos autores de forma linear, onde $n \to \infty$ *byline* $\rightsquigarrow \mathbb{R} \in [0, 1]$

$$\text{Byline}(N) = \begin{cases} 0, & \text{if } \ N = 1, \\ \left(\frac{0}{N-1}, \frac{1}{N-1}, \ldots, \frac{N-1}{N-1}\right), & \text{otherwise.} \end{cases} \tag{6.1}$$

Uma representação visual da métrica *Byline* é exibida na Figura 6.2. No exemplo, é

aplicada para quatro autores e os valores calculados são atribuídos com base na posição dos autores relativos. O primeiro autor tem uma pontuação *Byline* de 0, o segundo autor $0,33$, o terceiro $0,66$ e o último autor tem 1.

Figura 6.2 – *Representação visual da métrica Byline aplicada para quatro autores.*



Fonte: Adaptado de (RIBEIRO; STOROPOLI; CARNEIRO DA CUNHA, 2019)

6.3.2 Variáveis Independentes - Categorias de Contribuição

Apesar de os três periódicos terem contribuições baseadas nas diretrizes do ICMJE, os nomes e descrições das contribuições não possuem padrão entre os periódicos, conforme descrito no Capítulo 5. De fato, eles divergem não apenas na nomenclatura, mas também na quantidade com 10 no Ann. Intern. Med. e no JAMA e 13 no PLoS Med.

A partir dos resultados obtidos é proposta uma expansão do agrupamento para as categorias de contribuição, classificando-as em três grupos (*Group*), descritos no Quadro 6.1.

Quadro 6.1 – *Proposta para nova padronização das contribuições autorais.*

Categories		**Journals**		
Group	Proposal	Ann. Intern. Med.	JAMA	PLoS Med.
Logistic/support	Data Collection	Collection and assembly of data	(1) Acquisition data, (2) Acquisition, analysis, or interpretation data	Data curation
Theory, Methodology	Study Conceptualization	Conception and design	Study concept and design	(1) Conceptualization, (2) Investigation, (3) Methodology
Theory, Methodology, Logistic/support	Study Supervision	Final approval of the paper	Study Supervision	(1) Supervision, (2) Validation, (3) Project Administration
Theory	Critical Revision	Critical revision for important intellectual content	Critical Revision the manuscript for important intellectual content	Writing – review editing
Theory	Original Drafting	Drafting of the article	Drafting of the manuscript	Writing – original draft
Methodology	Statistical Analysis	(1) Statistical expertise, (2) Analysis and interpretation of the data	(1) Statistical analysis, (2) Analysis and interpretation data	Formal Analysis
Logistic/support	Funding and Support	(1) Obtaining of funding, (2) Administrative technical or logistics support, (3) Provision of study materials or patients	(1) Obtained funding, (2) Administrative, technical, or material support	(1) Funding acquisition, (2) Resources, (3) Software,

Fonte: Autor

A proposta mostrada no Quadro 6.1 tem como objetivo classificar os tipos de categorias relacionados às contribuições. Na coluna *Group*, foram agrupadas as contribuições em três categorias: *Theory*, *Methodology* e *Logist/support*. A coluna *Theory* representa contribuições teóricas, que são contribuições que estão relacionadas de alguma forma ao

argumento teórico do artigo. Abrange as contribuições relacionadas ao conceito (*Concept*) e à supervisão (*Supervision*) do estudo, o rascunho original (*Original Draft*) e a revisão crítica (*Critical Revision*). A metodologia (*Methodology*) representa as contribuições metodológicas, que compartilha algumas contribuições teóricas, como o conceito (*Concept*) e a supervisão do estudo (*Study Supervision*), mas também possui uma contribuição única denominada análises estatísticas (*Statistical Analysis*). Finalmente, logística e apoio (*Logistic/Support*) reúne todas as contribuições que não estão relacionadas à teoria (*Theory*) ou metodologia (*Methodology*), mas que são importantes de alguma forma para trazer o artigo para a publicação final. Ela compartilha da contribuição com a teoria (*Theory*) e a metodologia (*Methodology*), que é a supervisão do estudo (*Study Supervision*). As duas contribuições restantes são: coleta de dados (*Data Collection*), financiamento e apoio (*Funding and Support*). Embora sejam contribuições importantes, elas não têm o requisito de conhecimento em teoria (*Theory*) ou metodologia (*Methodology*).

6.3.3 Atribuição das Contribuições para os Autores

Nos dados obtidos referentes às contribuições de todos os autores nos periódicos Ann. Intern. Med., JAMA e PLoS Med., foi realizado um procedimento para todos os autores da amostra. Foi atribuído o valor "0" para autores que não contribuíram e "1" para os que contribuíram nas categorias, visando atribuir a cada autor uma medida de quanto eles contribuem em cada categoria. A soma das contribuições deu origem a um índice de contribuição *IC*. O objetivo foi ponderar igualmente todas as contribuições de uma determinada categoria. Assim, por exemplo, se a teoria (*Theory*) somar quatro contribuições, soma-se todos os valores codificados dessas quatro contribuições teóricas e divide-se pelo número total de categorias, neste caso quatro. Dessa forma, foi calculado o índice de contribuição teórica (*Theory*) dos autores. O mesmo se aplica ao índice de metodologia (*Methodology*) e ao índice *Logistic/Support*, com a ressalva de que ambos têm três contribuições, em vez de quatro índices da *Theory*. As equações para todos os índices das categorias são descritas pelas Equações 6.2, 6.3 e 6.4. Deve-se observar que para cada índice o intervalo de valores possíveis é entre 0 a 1.

$$theory\ index = \frac{\begin{array}{r}Study\ Concept + Study\ Supervision \\ + Critical\ Revision + Original\ Draft\end{array}}{4} \tag{6.2}$$

$$methodology\ index = \frac{\begin{array}{r}Study\ Concept + Study\ Supervision \\ + Statistical\ Analysis\end{array}}{3} \tag{6.3}$$

$$logistic/support\ index = \frac{\begin{array}{r}Data\ Collection + Study\ Supervision \\ + Funding\ and\ Support\end{array}}{3} \tag{6.4}$$

6.3.4 Variável de Controle H-*index*

Para controlar alguns efeitos que as hipóteses não abrangem, foi adicionado o H-*index* do Scopus de cada autor como uma medida aproximada da experiência em pesquisa de autores. Esse dado foi extraído do banco de dados dos autores e representa o H-*index* dos autores a partir do ano de 2019, uma vez que apenas o H-*index* atual estava disponível no Scopus. Isso implica em uma limitação dessa métrica, pois a amostra é desenhada para os anos entre 2000 e 2019, mas ainda é considerada um *proxy* adequado para medir a experiência em pesquisa de autores. Assim, nas análises subsequentes, pode-se ver o impacto das variáveis independentes na posição do autor, mantendo o H-*index* fixo.

A análise dos dados foi realizada através de regressão linear para testar as hipóteses, descritas na Seção 6.2. Especificamente, a variável dependente é o *Byline* e as variáveis independentes são *theory index*, *methodology index* and *logistic/support index* em conjunto com a variável de controle H-*index*. Foi utilizada a função `lm` do pacote de estatísticas `stats` da `linguagem R` na Versão 3.6.2.

6.4 Resultados e Análises

Inicialmente é apresentado um resumo estatístico de todas as variáveis na Tabela 6.3, composta por média, desvio padrão, percentis 25% e 75% e valor mínimo e máximo. A variável dependente *byline* tem média $0,5$, valor mínimo de 0 e valor máximo de 1. As variáveis independentes, compostas pelos índices de contribuição *theory, methodology e logistic/support*, têm, respectivamente os valores $0,60$; $0,69$ e $0,67$. Isso significa que, em média, os autores contidos na amostra têm mais contribuições metodológicas (*Methodology*) do que logísticas/apoio (*Logistic/support*) e teóricas (*Theory*). Mas, como os valores são um pouco semelhantes, pode-se dizer que essa diferença é baixa. Observa-se também que as contribuições têm um desvio padrão alto, inferindo que uma variação alta está presente em todas as contribuições. A variável de controle, H-*index*, tem uma média de $34,14$; o que é esperado, dado o alto prestígio relacionado à publicação em periódicos de alto impacto, como os da amostra utilizada.

Na Tabela 6.4 é mostrada a matriz de correlação de todas as variáveis. Essa é uma matriz diagonal inferior esquerda que exibe a correlação de cada variável pela interseção da coluna e da linha. Como, por exemplo, a correlação entre *Byline* e H-*index* é representada pela interseção da coluna *Byline* com a linha H-*index*, fornecendo o valor $0,184$.

Na Tabela 6.5, é mostrada a regressão para todas as variáveis do modelo, as quais têm um Fator de Inflação de Variância (FIV) apropriado: abaixo de 5. A variável dependente é *Byline* e todos os coeficientes são exibidos com sua média e intervalo de confiança de 95%. O número total de observações é $20,095$, o R^2 é $0,044$. As variáveis significativas no modelo são *Theory index* e H-*index* com $p < 0,01$ e *Logistic/support index* com $p < 0,1$.

Tabela 6.3 – *Resumo das Variáveis.*

Variável	**Média**	**Desv. Pad.**	**Mín.**	**Pctl(25)**	**Pctl(75)**	**Máx.**
Byline	0,50	0,32	0	0,2	0,8	1
Theory index	0,60	0,26	0,00	0,25	0,75	1,00
Methodology index	0,69	0,35	0,00	0,3	1,00	1,00
Logistic/support index	0,67	0,36	0,00	0,33	1,00	1,00
H-index	34,14	32,38	0	10	49	283

$n = 20.095$
Fonte: Autor

Tabela 6.4 – *Matriz de correlação entre as variáveis.*

	Byline	Theory index	Methodology index	Logistic/ support index
Theory index	$-0,065$			
Methodology index	$-0,051$	$0,850$		
Logistic/support index	$-0,010$	$0,694$	$0,681$	
H-index	$0,184$	$0,105$	$0,097$	$0,061$

Fonte: Autor

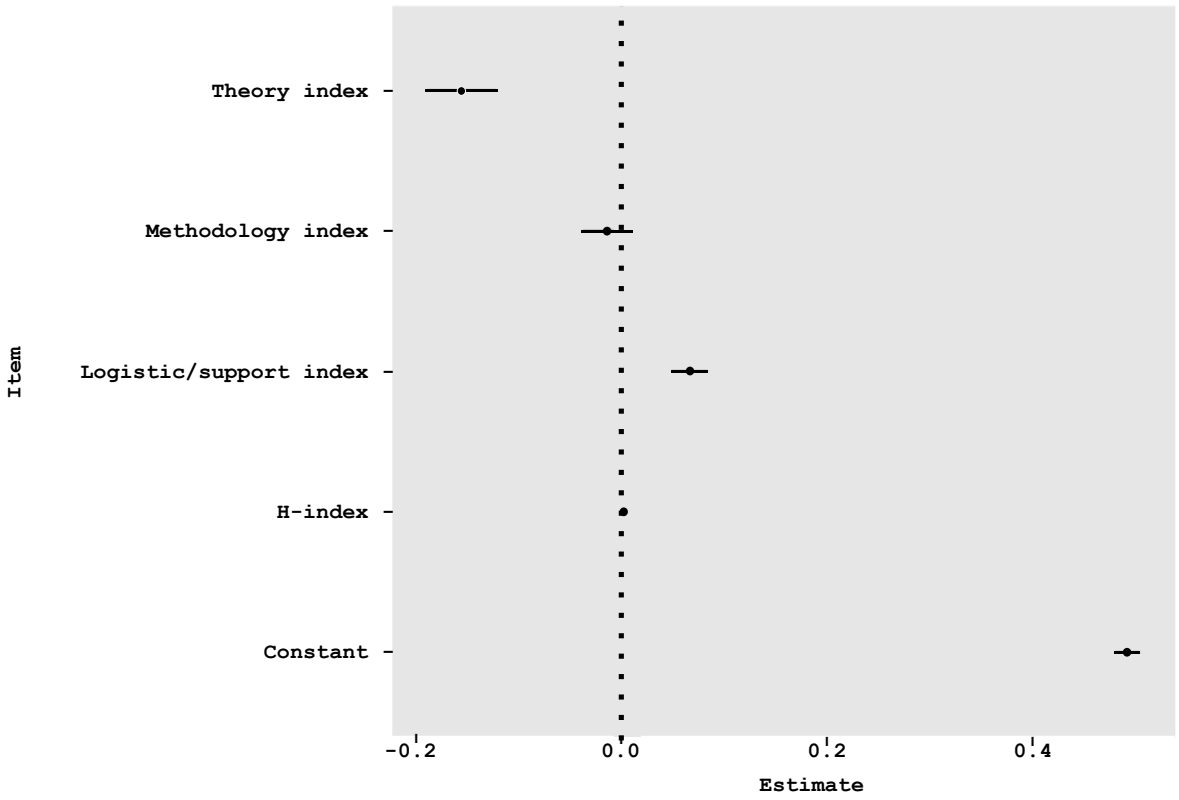
Tabela 6.5 – *Resultados da Regressão.*

	Variável dependente:
	Byline
Theory index	$-0{,}150^{***}$
	$(-0{,}190;\ -0{,}120)$
Methodology index	$-0{,}013$
	$(-0{,}038;\ 0{,}011)$
Logistic/support index	$0{,}067^{*}$
	(0,049, 0,084)
H-index	$0{,}002^{***}$
	(0,002; 0,002)
Constant	0,490
	(0,480; 0,500)
Observations	20.095
R^2	0,044
Adjusted R^2	0,044
Residual Std. Error	0,310 (df = 20090)
F Statistic	231.000 (df = 4; 20090)
Observação:	*p<0,1; **p<0,05; ***p<0,01

Fonte: Autor

Finalmente, na Figura 6.3, pode-se ver uma representação visual dos coeficientes do modelo. Os pontos são a média do coeficiente, a linha em negrito representa o intervalo de confiança de 95% dos valores do coeficiente.

Figura 6.3 – *Coeficientes da regressão com o intervalo de confiança com as médias e intervalo de confiança de* 95%.



Fonte: Autor

É perceptível que, sem dúvida, as contribuições teóricas (*theoretical contributions*) (ZBAR; FRANK, 2011) produzem uma posição de autor muito menor (com tendência a ser o primeiro autor), com um intervalo de confiança de 95% entre $-0,19$ e $-0,12$, corroborando H_1. Essas contribuições abrangem as categorias agrupadas *Original Drafting* e *Study Conceptualization* que, de acordo com a relação em cada periódico, indica que os primeiros autores são os que mais contribuem (YANG; WOLFRAM; WANG, 2017).

As contribuições metodológicas (*methodological contributions*) não evidenciam afetar a posição do autor, pois o intervalo de confiança de 95%, captures zero ($-0,038$ a $0,011$), não dando suporte a H_2. Isso mostra que as contribuições metodológicas agregam informações sobre vários aspectos diferentes, apresentando desafios para extrair informações sobre os tipos específicos de contribuições feitas por um autor (SAUERMANN; HAEUSSLER, 2017).

As contribuições logísticas (*logistic/Support contributions*) aponta para uma posição de autor mais alta (inclinando-se para ser o último autor) (PERNEGER et al., 2017;

RAHMAN et al., 2017), com um intervalo de confiança de 95% entre $0,049$ e $0,084$, também corroborando H_3.

Por fim, H-*index* não evidencia afetar a posição do autor, pois a média e o intervalo de confiança de 95% estão em torno de zero $(0,002)$. Isto pode estar associado ao fato de que o H-*index* é atribuído de forma igual para todos os autores do artigo (HIRSCH, 2019), independente das contribuições realizadas.

6.5 Discussão Sobre Posição Autoral e Contribuições

Para ter uma melhor compreensão sobre as contribuições teóricas (*theoretical contributions*) (H_1) e contribuições metodológicas (*methodological contributions*) (H_2), foram realizadas duas análises *post-hoc* depois que as hipóteses foram testadas. A primeira diz respeito a quais contribuições teóricas (*theoretical contributions*) (H_1) têm mais influência sobre a *posição do autor*. A segunda, especifica se a análise estatística (*Statistical Analysis*) tem impacto sobre a posição do autor (*author position*). Para ambas as análises, o controle foi realizado pelo H-*index*. Além disso, nenhum FIV foi superior a 5, o que mostra ausência de multicolinearidade.

A tabela 6.6 exibe os coeficientes da regressão para um modelo especificado apenas com as contribuições que pertencem à categoria de teoria (*Theory*).

O *Original Drafting* foi a contribuição teórica mais associada à posição inferior do autor, com um intervalo de confiança de 95% entre $-0,13$ e $-0,11$. As categorias *Study Conceptualization* e *Critical Revision* tiveram intervalos de confiança de 95% semelhantes com associações negativas mais brandas do que *Original Drafting* com a posição do autor. Notavelmente, *Study Supervision* foi a única contribuição teórica com uma associação positiva com a posição do autor (intervalo de confiança de 95% entre $0,053$ e $0,075$).

As contribuições teóricas (*theoretical contributions*) e contribuições metodológicas (*methodological contributions*) compartilham duas categorias: *Study Conceptualization* e *Study supervision*, mostradas no Quadro 6.1. Como contribuições metodológicas (*Methodological contributions*) possuem três categorias, há uma que não é compartilhada com contribuições teóricas (*Theoretical contributions*) ou contribuições logísticas (*Logistic/-support contributions*). A única contribuição metodológica (*Methodological contributions*) é *Statistical Analysis*, objeto da segunda análise *post-hoc*.

A Tabela 6.7 exibe os coeficientes de regressão para um modelo especificado apenas com *Statistical Analysis*. Apesar das contribuições metodológicas (*methodological contributions*) ter um efeito nulo na posição do autor (H_2), descobriu-se que a contribuição específica *Statistical Analysis* tem uma associação negativa leve com a posição do autor (intervalo de confiança de 95% entre $-0,092$ e $-0,073$).

Tabela 6.6 – *Estimativas da regressão para (Theoretical Contributions).*

	Variável dependente:
	Byline
Study Conceptualization	$-0{,}027^{***}$
	$(-0{,}039; -0{,}014)$
Critical Revision	-0.034^{***}
	$(-0{,}044; -0{,}025)$
Original Drafting	-0.120^{***}
	$(-0{,}130; -0{,}110)$
Study Supervision	0.064^{*}
	(0,053; 0,075)
H-index	0.002^{***}
	(0,002; 0,002)
Constant	0,500
	(0,490; 0,510)
Observations	20.095
R^2	0,059
Adjusted R^2	0,059
Residual Std. Error	0,310 (df = 20089)
F Statistic	254.000 (df = 5; 20089)
Observação:	*p<0,1; **p<0,05; ***p<0,01

Fonte: Autor

Tabela 6.7 – *Estimativas da regressão para contribuição (Statistical Analysis).*

	Variável dependente:
	Byline
Statistical Analysis	$-0{,}083^{***}$
	$(-0{,}092; -0{,}073)$
H-index	$0{,}002^{***}$
	(0,002; 0,002)
Constant	0,490
	(0,480; 0,500)
Observations	20.095
R^2	0,048
Adjusted R^2	0,048
Residual Std. Error	0,310 (df = 20092)
F Statistic	509.000 (df = 2; 20092)
Observação:	*p<0.1; **p<0.05; ***p<0.01

Fonte: Autor

O posicionamento autoral no *"byline"* e as contribuições científicas têm grande importância na indicação de crédito e responsabilidades em publicações (YANG; WOLFRAM; WANG, 2017), além das questões éticas (JONES; MCCULLOUGH; RICHMAN, 2005). Os resultados da regressão, obtidos sobre a proposta para o agrupamento das contribuições, confirmaram que autores com contribuições teóricas tendem a ser o primeiro autor (ZBAR; FRANK, 2011), enquanto as contribuições logísticas tendem a posicionar um autor na última posição (RAHMAN et al., 2017). Estudos sobre posicionamento autoral com base em contribuições, de forma geral, focam no primeiro e no último autor (TSCHARNTKE et al., 2007), considerando que os autores posicionados entre eles contribuem de forma menos expressiva (MONGEON et al., 2017). Essa afirmação pode ser interpretada sobre diferentes formas, visto que questões institucionais e de tradição (PONTILLE, 2003) podem influenciar no posicionamento autoral, independente das contribuições realizadas por um autor. Já para as contribuições metodológicas, os resultados não evidenciaram alterar o posicionamento de um autor no *byline* quando considerados o H-*index* e as contribuições realizadas, utilizados também na verificação das contribuições teóricas e logísticas. Uma evidência sobre essa constatação é que contribuições desta natureza agregam informações sobre vários aspectos diferentes, dificultando extrair informações sobre os tipos específicos de contribuição (SAUERMANN; HAEUSSLER, 2017).

Um aspecto a ser considerado é que autores podem indicar contribuições em muitas categorias para aumentar sua participação no artigo, mesmo que não tenham efetivamente contribuído, causando distorções. Por outro lado, um autor pode contribuir em muitas categorias, incluindo as que abrangem os conceitos teóricos e metodológicos. Desta forma, fica evidente a dificuldade em identificar o grau de colaboração de cada autor, uma vez que o ICMJE recomenda critérios de contribuição para as áreas de ciências biológicas e da medicina, mas não indica o grau de importância de cada uma em termos qualitativos, apenas quantitativo (TEIXIERA DA SILVA; DOBRÁNSZKI, 2016), dificultando classificá-las.

Com base nos resultados da regressão, Tabela 6.5, fica evidente que a correlação entre as variáveis de contribuição tem um efeito muito pequeno ao considerar o valor do R^2 ajustado $(0,044)$, ou seja, $4,5\%$ da variação da variável dependente (*Byline* com p-valor $< 0{,}01$) é explicada pelas variáveis independentes (*"theory index"*, *"methodology index"* e *"logistic/support index"*). Esse valor é considerado baixo, partindo da premissa que valores do R^2 para variáveis latentes endógenas devem ser avaliadas como $0,26$ (substancial), $0,13$ (moderado), $0,02$ (fraco) (COHEN, 1988). Assim, apesar do valor do R^2 ser $0,044$, superior a $0,02$, encontra-se muito abaixo dos $0,13$ para ser considerado como moderado.

O pequeno efeito registrado na correlação entre as variáveis de contribuição indica que o posicionamento autoral não é definido pelas contribuições dos autores. Esse achado mostra que, apesar da adoção das recomendações do ICMJE pelos periódicos Ann. Intern. Med., JAMA e PLoS Med. e de seus esforços em exigir a indicação das contribuições nas

publicações para transparência sobre questões éticas e de responsabilidade, as contribuições não surtem efeito para a classificação do posicionamento autoral. Esse achado mostra a necessidade de novas investigações para determinar quais são os aspectos que influenciam no posicionamento autoral nas áreas de ciências biológicas e da medicina, quando consideradas as contribuições.

Assim, neste capítulo foi apresentado um estudo sobre a relação entre o posicionamento autoral e as contribuições em publicações científicas sobre dados coletados de artigos e periódicos da área de ciências biológicas e da medicina. O estudo mostrou que contribuições teóricas produzem uma posição de autor muito menor (com tendência a ser o primeiro autor), corroborando H_1. As contribuições metodológicas não evidenciam afetar a posição do autor, não dando suporte a H_2. Por fim, as contribuições logísticas apontam para uma posição de autor mais alta (inclinando-se para ser o último autor), corroborando H_3. Embora os resultados obtidos mostrem a relação do posicionamento autoral e das contribuições, o pequeno efeito registrado na correlação entre as variáveis de contribuição indica que o posicionamento autoral não é definido pelas contribuições dos autores.

Apresentou também uma proposta de padronização para contribuições autorais, agrupadas por categorias que se relacionam de acordo com o significado dentro de cada uma no que tange as características teórica, metodológica e logística.

Por fim, os resultados obtidos permitiram responder à pergunta de pesquisa referente a **Lacuna 2 desta pesquisa "Identificar padrões de posição autoral na autoria científica"**, descrita na Seção 1.2.

7 CONCLUSÕES

Neste trabalho foram estudadas as categorias de contribuição em publicações científicas, além da identificação de padrões entre o posicionamento autoral e contribuições na autoria científica, utilizando técnicas estatísticas e de ciência de dados, com o foco na área de ciências biológicas e medicina. O foco foi necessário dado que há uma preocupação latente com o discernimento das contribuições de autores na área de ciências biológicas e medicina.

Os dados utilizados nesta pesquisa somaram 2.024 artigos contendo 17.385 autores, originados das fontes de dados do SCImago, Scopus e dos periódicos da área de ciências biológicas e medicina Ann. Intern. Med., JAMA e PLoS Med. Para o processo de obtenção, foram desenvolvidos robôs computacionais para automatizar as tarefas de coleta, preparação e limpeza dos dados de artigos, autores e contribuições, que mostraram diminuição do tempo na realização e a acurácia dos processos. Assim, as técnicas de ciência de dados utilizadas neste estudo mostraram ser adequadas e eficientes para o estudo da ciência.

As categorias de contribuição foram estudas com a utilização de análise fatorial, Capítulo 5. Os resultados evidenciaram a existência de dois grupos de contribuições que podem apoiar os autores de acordo com suas habilidades para contribuir em artigos científicos. Também mostraram que as maiores contribuições se encontram no grupo teórico ("*Theory*"), sinalizando que a experiência acadêmica dos autores é um fator preponderante. No entanto, contribuições metodológicas e logísticas (*Methodology and Logistic*) são essenciais para a aplicação dos conceitos teóricos, representando contribuições substanciais.

O estudo para identificação da relação entre o posicionamento autoral e as contribuições em publicações científicas foi realizado utilizando regressão linear, da ciência de dados, para o desenho dos modelos, Capítulo 6. Os resultados mostraram que autores que publicam na área de ciências biológicas e medicina e que contribuem teoricamente tendem a ser o primeiro autor. Já as contribuições logísticas tendem a posicionar um autor como último. Por fim, contribuições metodológicas não evidenciam impacto no posicionamento autoral. Ademais, apresentou uma proposta de padronização para contribuições autorais, agrupadas por categorias que se relacionam de acordo com o significado dentro de cada uma no que tange as características teórica, metodológica e logística, por meio da construção de uma tabela de equivalência de contribuições.

A declaração das contribuições em publicações científicas é latente na área de ciências biológicas e medicina, promovendo transparência e identificação da colaboração e responsabilidade dos autores. Entretanto, há dificuldades em identificar corretamente as categorias de contribuição disponibilizadas pelos periódicos.

Assim, este trabalho mostrou, além da construção de robôs e algoritmos para automatização dos processos de coleta limpeza e preparação dos dados, duas alternativas

que podem auxiliar autores a indicarem suas contribuições nas categorias que estejam mais alinhadas com sua experiência acadêmica e participação nas publicações. A primeira alternativa, Quadro 5.2, oferece condições de autores analisarem o nível de importância de cada uma das contribuições relacionadas aos conceitos teóricos (*"Theory"*) e metodológicos e logísticos (*"Methodology and Logistic"*), por meio do agrupamento das categorias de contribuições. A segunda apresenta uma proposta de modelo universal de contribuições na área de ciências biológicas e medicina, Quadro 6.1, com três categorias: teórica (*"Theory"*), metodológica (*"Methodology"*) e logística (*"Logistic/support"*), que tem como objetivo classificar os tipos de categorias relacionados às contribuições. Por fim, mostrou que as contribuições na autoria científica não afetam o posicionamento autoral na área de ciências biológicas e medicina.

8 TRABALHOS FUTUROS

Como propostas de trabalhos futuros, pretende-se:

1. Ampliar este estudo utilizando um número maior de periódicos para verificar se o comportamento do posicionamento autoral em relação às contribuições apresenta o mesmo comportamento obtido nesta pesquisa, visto que cada periódico utiliza um sistema de nomenclatura diferente para declaração das contribuições.

2. Criar uma ferramenta *online* utilizando inteligência artificial que possibilite autores verificarem quais categorias de contribuição são mais alinhadas com sua experiência de acordo com as nomenclaturas oferecidas pelos periódicos.

3. Verificar a idiossincrasia de contribuições em publicações científicas entre as 27 áreas do SCImago para periódicos que adotam algum sistema de indicação das contribuições em suas publicações.

4. Expandir a funcionalidade dos robôs coletores de dados para atender à extração de dados para um número maior de periódicos com parâmetros configuráveis.

REFERÊNCIAS BIBLIOGRÁFICAS

AALST, W. van der. ***Process Mining***. [S.l.]: Springer Berlin Heidelberg, 2016. 1–467 p. doi:10.1007/978-3-662-49851-4. ISBN 978-3-662-49850-7. Citado na pág. 36.

ABBASI, A.; ALTMANN, J.; HOSSAIN, L. **Identifying the effects of co-authorship networks on the performance of scholars: A correlation and regression analysis of performance measures and social network analysis measures**. *Journal of Informetrics*, Elsevier Ltd, v. 5, n. 4, p. 594–607, oct 2011. ISSN 17511577. doi:10.1016/j.joi.2011.05.007. Citado na pág. 20.

ABRAMO, G.; D'ANGELO, C. A. **How do you define and measure research productivity?** *Scientometrics*, v. 101, n. 2, p. 1129–1144, 2014. ISSN 15882861. doi:10.1007/s11192-014-1269-8. Citado na pág. 34.

ABRAMO, G.; D'ANGELO, C. A. **Does your surname affect the citability of your publications?** *Journal of Informetrics*, Elsevier Ltd, v. 11, n. 1, p. 121–127, 2017. ISSN 18755879. doi:10.1016/j.joi.2016.12.003. Citado na pág. 78, 79.

ABRAMO, G.; D'ANGELO, C. A.; MURGIA, G. **The combined effects of age and seniority on research performance of full professors**. *Science and Public Policy*, v. 43, n. 3, p. 301–319, 2016. doi:10.1093/scipol/scv037. Citado na pág. 21.

ABRAMO, G.; D'ANGELO, C. A.; ROSATI, F. **Measuring institutional research productivity for the life sciences: The importance of accounting for the order of authors in the byline**. *Scientometrics*, v. 97, n. 3, p. 779–795, 2013. ISSN 01389130. doi:10.1007/s11192-013-1013-9. Citado na pág. 34, 84.

ALLEN, L.; O'CONNELL, A.; KIERMER, V. **How can we ensure visibility and diversity in research contributions? How the Contributor Role Taxonomy (CRediT) is helping the shift from authorship to contributorship**. *Learned Publishing*, v. 32, n. 1, p. 71–74, 2019. ISSN 17414857. doi:10.1002/leap.1210. Citado na pág. 18.

AMADEU, S. **Governo Dos Algoritmos**. *Revista de Políticas Públicas*, v. 21, n. 1, p. 267–281, 2017. doi:10.18764/2178-2865.v21n1p267-281. Citado na pág. 27.

ANNALS. ***Annals of Internal Medicine***. 2021. Disponível em: <https://www.acpjournals.org/doi/10.7326/0003-4819-158-1-201301010-00012/>. Acesso em: 10 abr. 2021. Citado na pág. 41, 59, 60.

ARAUJO, B. S. D.; ALMEIDA, H. L. S. D.; MELLO, F. L. D. **Computational intelligence methods applied to the fraud detection of electric energy consumers**. *IEEE Latin America Transactions*, v. 17, n. 1, p. 71–77, 2019. ISSN 15480992. doi:10.1109/TLA.2019.8826697. Citado na pág. 27.

ARIA, M.; MISURACA, M.; SPANO, M. **Mapping the Evolution of Social Research and Data Science on 30 Years of Social Indicators Research**. *Social Indicators Research*, Springer Netherlands, n. 0123456789, feb 2020. ISSN 0303-8300. doi:10.1007/s11205-020-02281-3. Citado na pág. 32.

ARTES, R. **Aspectos estatísticos da análise fatorial de escalas de avaliação**. *Rev. psiquiatr. clín.(São Paulo)*, p. 223–8, 1998. Citado na pág. 77.

AZAM, A. ***The First Rule of Data Science***. 2016. Disponível em: <http://berkeleysciencereview.com/article/first-rule-data-science/>. Acesso em: 15 abr. 2020. Citado na pág. 26.

BAJPAI, J.; METKEWAR, P. S. **Data Quality Issues and Current Approaches to Data Cleaning Process in Data Warehousing**. *GRD Journals- Global Research and Development Journal for Engineering*, v. 1, n. 10, p. 14–18, 2016. Citado na pág. 19, 28.

BARTHOLOMEW, D. J. et al. ***Analysis of multivariate social sciencia data***. [S.l.]: Taylor and Francis Group, LLC, 2008. 371 p. Citado na pág. 79.

BATES, T. et al. **Authorship Criteria and Disclosure of Contributions**. *Jama*, v. 292, n. 1, p. 86–88, 2004. ISSN 0098-7484. doi:10.1001/jama.292.1.86. Citado na pág. 33.

BAUDISCH, A. R. ***Ciência de Dados é explorar Big Data para fazer perguntas para prever o futuro***. 2016. Disponível em: <https://medium.com/@AlfredBaudisch/o-que-%C3%A9-ci%C3%AAncia-de-dados-data-science-7af5bdac101a>. Acesso em: 18 mai. 2020. Citado na pág. 26.

BISHOP, C. M. ***Pattern recognition and machine learning***. [S.l.]: springer, 2006. Citado na pág. 19.

BONECHI, S. et al. **Confidence Measures for Deep Learning in Domain Adaptation**. *Applied Sciences*, v. 9, n. 11, p. 2192, may 2019. ISSN 2076-3417. doi:10.3390/app9112192. Citado na pág. 29.

BYRNE, C. et al. Development Workflows for Data Scientists. *O'Reilly*, 2017. Disponível em: <http://oreilly.com/safari>. Citado na pág. 28.

CAI, L.; ZHU, Y. **The Challenges of Data Quality and Data Quality Assessment in the Big Data Era**. *Data Science Journal*, v. 14, p. 2, may 2015. ISSN 1683-1470. doi:10.5334/dsj-2015-002. Citado na pág. 19.

CAMPÊLO, R. A. et al. **A brief survey on replica consistency in cloud environments**. *Journal of Internet Services and Applications*, Journal of Internet Services and Applications, v. 11, n. 1, p. 1–13, 2020. ISSN 18690238. doi:10.1186/s13174-020-0122-y. Citado na pág. 27.

CAO, L. **Data Science**. *ACM Computing Surveys*, v. 50, n. 3, p. 1–42, oct 2017. ISSN 0360-0300. doi:10.1145/3076253. Citado na pág. 26.

CHANG, Y.-W. **Definition of authorship in social science journals**. *Scientometrics*, v. 118, n. 2, p. 563–585, feb 2019. ISSN 0138-9130. doi:10.1007/s11192-018-2986-1. Citado na pág. 18, 32.

CHEIN, F. ***Introdução aos modelos de regressão linear: um passo inicial para compreensão da econometria como uma ferramenta de avaliação de políticas públicas***. [S.l.: s.n.], 2019. 76 p. p. ISBN 9788525601155. Citado na pág. 88, 89.

Clarivate Analytics. ***Journal Impact Factor, Journal Citation Reports***. 2018. Disponível em: <http://help.incites.clarivate.com/incitesLiveJCR/JCRGroup/howtoCiteJCR.html>. Acesso em: 19 mai. 2020. Citado na pág. 40, 41, 42.

CLEMENT, T. P. **Authorship Matrix: A Rational Approach to Quantify Individual Contributions and Responsibilities in Multi-Author Scientific Articles**. *Science and Engineering Ethics*, v. 20, n. 2, p. 345–361, jun 2014. ISSN 14715546. doi:10.1007/s11948-013-9454-3. Citado na pág. 21.

COHEN, J. **Statistical power analysis for the behavioral sciences**. *Lawrence Earlbam Associates, Hillsdale, NJ*, 1988. Citado na pág. 99.

COLE, J. M. **A Design-to-Device Pipeline for Data-Driven Materials Discovery**. *Accounts of Chemical Research*, v. 53, n. 3, p. 599–610, mar 2020. ISSN 0001-4842. doi:10.1021/acs.accounts.9b00470. Citado na pág. 19.

COREY, K. M. et al. **Development and validation of machine learning models to identify high-risk surgical patients using automatically curated electronic health record data (Pythia): A retrospective, single-site study**. *PLOS Medicine*, v. 15, n. 11, p. e1002701, nov 2018. ISSN 1549-1676. doi:10.1371/journal.pmed.1002701. Citado na pág. 19, 27, 28.

CORRÊA, E. A. et al. **Patterns of authors contribution in scientific manuscripts**. *Journal of Informetrics*, Elsevier Ltd, v. 11, n. 2, p. 498–510, 2017. ISSN 18755879. doi:10.1016/j.joi.2017.03.003. Citado na pág. 89.

COSTAS, R.; BORDONS, M. **Do age and professional rank influence the order of authorship in scientific publications? Some evidence from a micro-level perspective**. *Scientometrics*, v. 88, n. 1, p. 145–161, 2011. ISSN 01389130. doi:10.1007/s11192-011-0368-z. Citado na pág. 21, 33.

COSTELLO, A. B.; OSBORNE, J. W. **Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis**. *Practical Assessment, Research and Evaluation*, v. 10, n. 7, 2005. ISSN 15317714. doi:10.7275/jyj1-4868. Citado na pág. 78.

DECULLIER, E.; MAISONNEUVE, H. **Have ignorance and abuse of authorship criteria decreased over the past 15 years?** *Journal of Medical Ethics*, p. 255–258, 2019. ISSN 14734257. doi:10.1136/medethics-2019-105737. Citado na pág. 18, 32, 75.

DEGROOT, M. H. ***Probability and statistics***. [S.l.]: Pearson, 2012. Citado na pág. 89.

DOTSON, B. **Equal contributions and credit assigned to authors in pharmacy journals**. *American Journal of Pharmaceutical Education*, v. 77, n. 2, 2013. ISSN 15536467. doi:10.5688/ajpe77239. Citado na pág. 86.

DREW, C. ***The Data Science VENN Diagram***. 2020. Disponível em: <http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram>. Acesso em: 29 mai. 2020. Citado na pág. 27.

D'UGGENTO, A. M.; RICCI, V.; TOMA, E. **An indicator proposal to evaluate research activities based on Scimago institutions ranking (SIR) data:**

An application for italian high education institutions. *Electronic Journal of Applied Statistical Analysis*, v. 9, n. 4, p. 655–674, 2016. ISSN 20705948. doi:10.1285/i20705948v9n4p655. Citado na pág. 36, 37.

EGGHE, L.; LIANG, L.; ROUSSEAU, R. **The byline: Thoughts on the distribution of author ranks in multiauthored papers**. *Mathematical and Computer Modelling*, v. 38, n. 3-4, p. 323–329, 2003. ISSN 08957177. doi:10.1016/S0895-7177(03)90090-2. Citado na pág. 19.

EKOLU, S. O.; QUAINOO, H. **Reliability of assessments in engineering education using Cronbach's alpha, KR and split-half methods**. *Global Journal of Engineering Education*, v. 21, n. 1, p. 24–29, 2019. ISSN 13283154. Citado na pág. 79, 80.

ERICKSON, B. J. et al. **Machine learning for medical imaging**. *Radiographics*, v. 37, n. 2, p. 505–515, 2017. ISSN 15271323. doi:10.1148/rg.2017160130. Citado na pág. 30.

FAGGION, C. M.; BAKAS, N. P.; WASIAK, J. **A survey of prevalence of narrative and systematic reviews in five major medical journals**. *BMC Medical Research Methodology*, BMC Medical Research Methodology, v. 17, n. 1, p. 176, dec 2017. ISSN 1471-2288. doi:10.1186/s12874-017-0453-y. Citado na pág. 40.

FERRAGINA, A. et al. **Bayesian regression models outperform partial least squares methods for predicting milk components and technological properties using infrared spectral data**. *Journal of Dairy Science*, Elsevier, v. 98, n. 11, p. 8133–8151, nov 2015. ISSN 00220302. doi:10.3168/jds.2014-9143. Citado na pág. 28.

FERREIRA, M. A. S. P. V.; SERRA, F. R. **A Coautoria Em Artigos Científicos De Administração: Perspectivas De Pesquisadores Internacionais**. *Administração: Ensino e Pesquisa*, v. 16, n. 4, p. 663, 2015. ISSN 2177-6083. doi:10.13058/raep.2015.v16n4.381. Citado na pág. 84.

FIGER, B. et al. **An evaluation of reporting of consent declines in three high impact factor journals**. *Clinical Ethics*, v. 13, n. 4, p. 189–193, dec 2018. ISSN 1477-7509. doi:10.1177/1477750918802428. Citado na pág. 41.

FIGUEIREDO FILHO, D. B.; SILVA JÚNIOR, J. A. dA. **Visão além do alcance: uma introdução à análise fatorial**. *Opinião Pública*, v. 16, n. 1, p. 160–185, jun 2010. ISSN 0104-6276. doi:10.1590/S0104-62762010000100007. Citado na pág. 78.

FORTUNATO, S. et al. **Science of science**. *Science*, v. 359, n. 6379, 2018. ISSN 10959203. doi:10.1126/science.aao0185. Citado na pág. 19.

GRUS, J. ***Data Science do zero: Primeiras regras com o Python***. [S.l.]: Alta Books, 2019. Citado na pág. 30.

HAGEN, N. T. **Reversing the byline hierarchy: The effect of equalizing bias on the accreditation of primary, secondary and senior authors**. *Journal of Informetrics*, Elsevier Ltd, v. 8, n. 3, p. 618–627, jul 2014. ISSN 17511577. doi:10.1016/j.joi.2014.05.003. Citado na pág. 34, 75, 86.

HAIR, J. F. et al. ***Análise multivariada de dados***. [S.l.]: Bookman editora, 2009. Citado na pág. 79.

HALIM, Z.; KHAN, S. **A data science-based framework to categorize academic journals**. Springer International Publishing, v. 119, n. 1, p. 393–423, 2019. ISSN 15882861. doi:10.1007/s11192-019-03035-w. Citado na pág. 28.

HARDIN, J. et al. **Data Science in Statistics Curricula: Preparing Students to "Think with Data"**. *The American Statistician*, v. 69, n. 4, p. 343–353, oct 2015. ISSN 0003-1305. doi:10.1080/00031305.2015.1077729. Citado na pág. 19, 26, 31.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. ***The elements of statistical learning: data mining, inference, and prediction***. [S.l.]: Springer Science & Business Media, 2009. Citado na pág. 30.

HE, B. et al. **Mining diversity subgraph in multidisciplinary scientific collaboration networks: A meso perspective**. *Journal of Informetrics*, Elsevier Ltd, v. 7, n. 1, p. 117–128, 2013. ISSN 17511577. doi:10.1016/j.joi.2012.09.005. Citado na pág. 19.

HILÁRIO, C. M. et al. **Authorship in science: A critical analysis from a Foucauldian perspective**. *Research Evaluation*, v. 27, n. 2, p. 63–72, apr 2018. ISSN 0958-2029. doi:10.1093/reseval/rvx041. Citado na pág. 18, 31.

HINKIN, T. R. **A review of scale development in the study of behavior in organizations**. *Journal of Management*, v. 21, p. 967–988, 1995. Citado na pág. 79.

HINKIN, T. R. **A brief tutorial on the development of measures for use in survey questionnaires**. *Organizational research methods*, v. 1(1), p. 104–121, 1998. Citado na pág. 79.

HIRSCH, J. E. **h α : An index to quantify an individual's scientific leadership**. *Scientometrics*, Springer International Publishing, v. 118, n. 2, p. 673–686, 2019. ISSN 15882861. doi:10.1007/s11192-018-2994-1. Disponível em: <https://doi.org/10.1007/s11192-018-2994-1>. Citado na pág. 97.

HONGYU, K. **Análise Fatorial Exploratória: resumo teórico, aplicação e interpretação**. *E&S Engineering and Science*, v. 7, n. 4, p. 88–103, dec 2018. ISSN 2358-5390. doi:10.18607/ES201877599. Citado na pág. 78, 79.

HORN, J. L. **A rationale and test for the number of factors in factor analysis**. *Psychometrika*, Springer, v. 30, n. 2, p. 179–185, 1965. Citado na pág. 79.

HUTCHESON, G. D.; SOFRONIOU, N. ***The Multivariate Social Scientist: Introductory Statistics Using Generalized Linear Models***. SAGE Publications, 1999. (Statistics Series). ISBN 9780761952015. Disponível em: <https://books.google.com.br/books?id=NiF4--8lvf0C>. Citado na pág. 78.

ICMJE. ***International Committee of Medical Journal Editors***. 2020. Disponível em: <http://www.icmje.org/recommendations/>. Acesso em: 26 mai. 2020. Citado na pág. 33.

IGOU, E. R.; TILBURG, W. A. van. **Ahead of others in the authorship order: Names with middle initials appear earlier in author lists of academic articles in psychology**. *Frontiers in Psychology*, v. 6, n. MAR, p. 1–9, 2015. ISSN 16641078. doi:10.3389/fpsyg.2015.00469. Citado na pág. 31.

IOANNIDIS, J. P. **Why most published research findings are false**. *PLoS medicine*, Public Library of Science, v. 2, n. 8, p. e124, 2005. Citado na pág. 19.

JABBEHDARI, S.; WALSH, J. P. **Authorship Norms and Project Structures in Science**. *Science Technology and Human Values*, v. 42, n. 5, p. 872–900, 2017. ISSN 15528251. doi:10.1177/0162243917697192. Citado na pág. 85.

JAMA. ***Journal of the American Medical Association***. 2021. Disponível em: <https://jamanetwork.com/journals/jama/fullarticle/2778599/>. Acesso em: 10 abr. 2021. Citado na pág. 41, 61, 62.

JIA, Z. et al. **Equal contributions and credit: an emerging trend in the characterization of authorship in major spine journals during a 10-year period**. *European Spine Journal*, v. 25, n. 3, p. 913–917, 2016. ISSN 14320932. doi:10.1007/s00586-015-4314-2. Citado na pág. 21.

JIAN, D.; XIAOLI, T. **Perceptions of author order versus contribution among researchers with different professional ranks and the potential of harmonic counts for encouraging ethical co-authorship practices**. *Scientometrics*, v. 96, n. 1, p. 277–295, 2013. ISSN 01389130. doi:10.1007/s11192-012-0905-4. Citado na pág. 34.

JOHNSON, R. A.; WICHERN, D. W. et al. ***Applied multivariate statistical analysis***. [S.l.]: Prentice hall Upper Saddle River, NJ, 2002. v. 5. Citado na pág. 78.

JONES, J. W.; MCCULLOUGH, L. B.; RICHMAN, B. W. **The ethics of bylines: Would the real authors please stand up?** *Journal of Vascular Surgery*, v. 42, n. 4, p. 816–818, 2005. ISSN 07415214. doi:10.1016/j.jvs.2005.06.026. Citado na pág. 31, 99.

KATAL, A.; WAZID, M.; GOUDAR, R. H. **Big data: Issues, challenges, tools and Good practices**. *2013 6th International Conference on Contemporary Computing, IC3 2013*, IEEE, p. 404–409, 2013. doi:10.1109/IC3.2013.6612229. Citado na pág. 19.

KAVAKIOTIS, I. et al. **Machine Learning and Data Mining Methods in Diabetes Research**. *Computational and Structural Biotechnology Journal*, The Authors, v. 15, p. 104–116, 2017. ISSN 20010370. doi:10.1016/j.csbj.2016.12.005. Citado na pág. 19, 27, 30.

KIM, J.; KIM, J.; OWEN-SMITH, J. **Generating automatically labeled data for author name disambiguation: an iterative clustering method**. *Scientometrics*, Springer International Publishing, v. 118, n. 1, p. 253–280, 2019. ISSN 15882861. doi:10.1007/s11192-018-2968-3. Citado na pág. 20.

KOZMA, E. et al. **Authorship: how to decide the order of authors on the byline?** *Current Medical Research and Opinion*, v. 30, p. S21–S21, 2014. Citado na pág. 34, 85.

KUMAR, S. **Ethical concerns in the rise of co-authorship and its role as a proxy of research collaborations**. *Publications*, v. 6, n. 3, 2018. ISSN 23046775. doi:10.3390/publications6030037. Citado na pág. 31, 33.

LAKE, D. A.; DIEGO, S. **Trumps the Alphabet**. n. January, p. 43–47, 2010. doi:10.1017/S1049096509990801. Citado na pág. 33.

LARIVIÈRE, V. et al. **Contributorship and division of labor in knowledge production**. *Social Studies of Science*, v. 46, n. 3, p. 417–435, 2016. ISSN 14603659. doi:10.1177/0306312716650046. Citado na pág. 34.

LARIVIÈRE, V.; PONTILLE, D.; SUGIMOTO, C. R. **Investigating the division of scientific labor using the Contributor Roles Taxonomy (CRediT)**. *Quantitative Science Studies*, v. 2, n. 1, p. 111–128, 2021. doi:10.1162/qss_a_00097. Citado na pág. 18, 75, 83.

LI, G. et al. **EASE: an effective 3-in-1 keyword search method for unstructured, semi-structured and structured data**. In: *Proceedings of the 2008 ACM SIGMOD international conference on Management of data - SIGMOD '08*. [S.l.]: ACM Press, 2008. p. 903. ISBN 9781605581026. ISSN 07308078. doi:10.1145/1376616.1376706. Citado na pág. 20, 29, 44.

LI, Z. et al. **Equal Contributions and Credit: An Emerging Trend in the Characterization of Authorship in Major Anaesthesia Journals during a 10-Yr Period**. *PLoS ONE*, v. 8, n. 8, 2013. ISSN 19326203. doi:10.1371/journal.pone.0071430. Citado na pág. 32.

LIU, X. Z.; FANG, H. **Modifying h-index by allocating credit of multi-authored papers whose author names rank based on contribution**. *Journal of Informetrics*, Elsevier Ltd, v. 6, n. 4, p. 557–565, 2012. ISSN 17511577. doi:10.1016/j.joi.2012.05.002. Citado na pág. 34, 89.

LIU, X. Z.; FANG, H. **Scientific group leaders' authorship preferences: An empirical investigation**. *Scientometrics*, v. 98, n. 2, p. 909–925, 2014. ISSN 01389130. doi:10.1007/s11192-013-1083-8. Citado na pág. 18.

LIU, X. Z.; FANG, H. **The impact of publications from mainland China on the trends in alphabetical authorship**. *Scientometrics*, v. 99, n. 3, p. 865–879, 2014. ISSN 01389130. doi:10.1007/s11192-013-1219-x. Citado na pág. 21.

LOWNDES, J. S. S. et al. **Our path to better science in less time using open data science tools**. *Nature Ecology & Evolution*, v. 1, n. 6, p. 0160, jun 2017. ISSN 2397-334X. doi:10.1038/s41559-017-0160. Citado na pág. 19, 20.

LOZANO, G. A. **Ethics of Using Language Editing Services in An Era of Digital Communication and Heavily Multi-Authored Papers**. *Science and Engineering Ethics*, v. 20, n. 2, p. 363–377, 2014. ISSN 14715546. doi:10.1007/s11948-013-9451-6. Citado na pág. 31.

MAÑANA-RODRÍGUEZ, J. **A critical review of SCImago Journal & Country Rank**. *Research Evaluation*, v. 24, n. 4, p. 343–354, oct 2015. ISSN 0958-2029. doi:10.1093/reseval/rvu008. Citado na pág. 36, 39.

MARQUES, F. ***Hierarquia Complexa***. 2011. Disponível em: <http://revistapesquisa.fapesp.br/2011/06/19/hierarquia-complexa/>. Acesso em: 28 mai. 2021. Citado na pág. 18.

MASKEY, R.; FEI, J.; NGUYEN, H. O. **Use of exploratory factor analysis in maritime research**. *Asian Journal of Shipping and Logistics*, Elsevier B.V., v. 34, n. 2, p. 91–111, 2018. ISSN 20925212. doi:10.1016/j.ajsl.2018.06.006. Citado na pág. 79.

MATHESON, A. **How industry uses the ICMJE guidelines to manipulate authorship-and how they should be revised**. *PLoS Medicine*, v. 8, n. 8, p. 1–5, 2011. ISSN 15491277. doi:10.1371/journal.pmed.1001072. Citado na pág. 21.

MATLOFF, N. S. ***From algorithms to Z-Scores: Probabilistic and statistical modeling in computer science***. [S.l.]: University Press of Florida Gainesville, 2009. Citado na pág. 89.

MATOS, D. A. S.; RODRIGUES, E. C. ***Análise Fatorial***. [S.l.]: Enap Fundação Escola Nacional de Administração Pública, Brasília, DF, 2019. v. 1. 74 p. ISBN 9788525601186. Citado na pág. 80.

MATTSSON, P.; SUNDBERG, C. J.; LAGET, P. **Is correspondence reflected in the author position? A bibliometric study of the relation between corresponding author and byline position**. *Scientometrics*, v. 87, n. 1, p. 99–105, 2011. ISSN 01389130. doi:10.1007/s11192-010-0310-9. Citado na pág. 31, 32, 86.

MAZUCHELI, J.; ACHCAR, J. A. **Algumas considerações em regressão não linear**. *Acta Scientiarum. Technology*, v. 24, n. 6, p. 1761–1770, 2002. ISSN 1807-8664. Citado na pág. 88.

MCNUTT, M. K. et al. **Transparency in authors' contributions and responsibilities to promote integrity in scientific publication**. *Proceedings of the National Academy of Sciences*, v. 115, n. 11, p. 2557–2560, mar 2018. ISSN 0027-8424. doi:10.1073/pnas.1715374115. Citado na pág. 32, 35, 75.

MIKALEF, P. et al. **Exploring the relationship between big data analytics capability and competitive performance: The mediating roles of dynamic and operational capabilities**. *Information and Management*, Elsevier, v. 57, n. 2, p. 1–15, 2020. ISSN 03787206. doi:10.1016/j.im.2019.05.004. Citado na pág. 27.

MITCHELL, T. M. et al. ***Machine learning***. [S.l.]: McGraw-hill New York, 1997. 432 p. Citado na pág. 19, 30.

MONGEON, P. et al. **The rise of the middle author: Investigating collaboration and division of labor in biomedical research using partial alphabetical authorship**. *PLOS ONE*, v. 12, n. 9, p. 1–14, sep 2017. ISSN 1932-6203. doi:10.1371/journal.pone.0184601. Citado na pág. 18, 32, 34, 99.

NAUR, P. ***Computing: A Human Activity - Selected Writings From 1951 To 1990***. New York: ACM Press/Addison-Wesley, 1992. 630 p. ISBN 0-201-58069-1. Citado na pág. 26.

OLSON, R. S. et al. **Evaluation of a Tree-based Pipeline Optimization Tool for Automating Data Science**. *GECCO 2016 - Proceedings of the 2016 Genetic and Evolutionary Computation Conference*, p. 485–492, mar 2016. doi:10.1145/2908812.2908918. Citado na pág. 28.

PATEL, V. M. et al. **Collaborative patterns, authorship practices and scientific success in biomedical research: a network analysis**. *Journal of the Royal Society of Medicine*, v. 112, n. 6, p. 245–257, 2019. ISSN 01410768. doi:10.1177/0141076819851666. Citado na pág. 89, 90.

PATIENCE, G. S. et al. **Intellectual contributions meriting authorship: Survey results from the top cited authors across all science categories**. *PLoS ONE*, v. 14, n. 1, p. 1–20, 2019. ISSN 19326203. doi:10.1371/journal.pone.0198117. Citado na pág. 18, 31, 36, 76.

PEIDU, C. **Can authors' position in the ascription be a measure of dominance?** *Scientometrics*, Springer International Publishing, v. 121, n. 3, p. 1527–1547, 2019. ISSN 0138-9130. doi:10.1007/s11192-019-03254-1. Citado na pág. 18.

PERNEGER, T. V. et al. **Thinker, Soldier, Scribe: Cross-sectional study of researchers' roles and author order in the Annals of Internal Medicine**. *BMJ Open*, v. 7, n. 6, 2017. ISSN 20446055. doi:10.1136/bmjopen-2016-013898. Citado na pág. 24, 36, 79, 83, 85, 87, 89, 90, 96, 97.

PETERSON, R. A. **A meta-analysis of variance accounted for and factor loadings in exploratory factor analysis**. *Marketing letters*, Springer, v. 11, n. 3, p. 261–275, 2000. Citado na pág. 79.

PLOS. ***PLOS Medicine***. 2021. Disponível em: <https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1003555/>. Acesso em: 10 abr. 2021. Citado na pág. 63, 64.

PONTILLE, D. **Authorship practices and institutional contexts in sociology: Elements for a comparison of the United States and France**. *Science, Technology, & Human Values*, Sage Publications, v. 28, n. 2, p. 217–243, 2003. Citado na pág. 34, 99.

PORTO, F. A. M.; ZIVIANI, A. **Ciência de Dados**. In: *Anais do III Seminário de Grandes Desafios da Computação no Brasil*. [s.n.], 2014. Disponível em: <https://www.lncc.br/~ziviani/papers/III-Desafios-SBC2014-CiD.pdf>. Acesso em: 17 abr. 2020. Citado na pág. 19, 26, 27.

PRIESTLEY, J.; MCGRATH, R. J. **The Evolution of Data Science**. *International Journal of Knowledge Management*, v. 15, n. 2, p. 97–109, apr 2019. ISSN 1548-0666. doi:10.4018/IJKM.201904010. Citado na pág. 18, 27.

RAHMAN, M. T. et al. **The need to quantify authors' relative intellectual contributions in a multi-author paper**. *Journal of Informetrics*, Elsevier Ltd, v. 11, n. 1, p. 275–281, feb 2017. ISSN 18755879. doi:10.1016/j.joi.2017.01.002. Citado na pág. 34, 87, 90, 96, 97, 99.

RASMUSSEN, K. et al. **Collaboration between academics and industry in clinical trials: Cross sectional study of publications and survey of lead academic authors**. *BMJ (Online)*, v. 363, 2018. ISSN 17561833. doi:10.1136/bmj.k3654. Citado na pág. 36, 42.

RIBEIRO, L. D. R.; STOROPOLI, J. E.; CARNEIRO DA CUNHA, J. A. **Coautoria em Administração: uma análise da posição autoral**. *Desenvolvimento em Questão*, v. 17, n. 48, p. 52–70, aug 2019. ISSN 2237-6453. doi:10.21527/2237-6453.2019.48.52-70. Disponível em: <https://www.revistas.unijui.edu.br/index.php/desenvolvimentoemquestao/article/view/8354>. Citado na pág. 92.

RIVERA, H. **Inappropriate Authorship**. *J Korean Med Sci*, v. 33, n. 13, p. 105, 2018. doi:10.3346/jkms.2018.33.e105. Citado na pág. 76.

ROSENZWEIG, J. S. et al. **Authorship, collaboration, and predictors of extramural funding in the emergency medicine literature**. *The American Journal of Emergency Medicine*, v. 26, n. 1, p. 5–9, jan 2008. ISSN 07356757. doi:10.1016/j.ajem.2007.02.028. Citado na pág. 18, 20, 31, 76.

RUSSELL, A. F. et al. **A Bibliometric Study of Authorship and Collaboration Trends Over the Past 30 Years in Four Major Musculoskeletal Science Journals**. *Calcified Tissue International*, Springer US, v. 104, n. 3, p. 239–250, 2019. ISSN 14320827. doi:10.1007/s00223-018-0492-3. Disponível em: <http://dx.doi.org/10.1007/s00223-018-0492-3>. Citado na pág. 32.

SAUERMANN, H.; HAEUSSLER, C. **Authorship and contribution disclosures**. *Science Advances*, v. 3, n. 11, p. 1–13, nov 2017. ISSN 2375-2548. doi:10.1126/sciadv.1700404. Citado na pág. 18, 33, 89, 96, 99.

SCHOENHERR, T.; SPEIER-PERO, C. **Data Science, Predictive Analytics, and Big Data in Supply Chain Management: Current State and Future Potential**. *Journal of Business Logistics*, v. 36, n. 1, p. 120–132, mar 2015. ISSN 07353766. doi:10.1111/jbl.12082. Citado na pág. 19.

SCHREIBER, J. B. **Issues and recommendations for exploratory factor analysis and principal component analysis**. *Research in Social and Administrative Pharmacy*, Elsevier Inc., v. 17, n. 5, p. 1004–1011, may 2021. ISSN 15517411. doi:10.1016/j.sapharm.2020.07.027. Citado na pág. 77, 79.

SCIMAGO. ***SCImago Journal & Country Rank***. 2020. Disponível em: <http://www.scimagojr.com>. Acesso em: 17 abr. 2020. Citado na pág. 18, 33.

SIVERTSEN, G.; ROUSSEAU, R.; ZHANG, L. **Measuring scientific contributions with modified fractional counting**. *Journal of Informetrics*, Elsevier Ltd, v. 13, n. 2, p. 679–694, 2019. ISSN 18755879. doi:10.1016/j.joi.2019.03.010. Citado na pág. 21.

SOLTOFF, B. ***Tidy data***. Computing for the Social Sciences, 2021. Disponível em: <https://cfss.uchicago.edu/notes/tidy-data/>. Acesso em: 15 fev. 2021. Citado na pág. 30.

STOROPOLI, J.; VILS, L. ***Estatística com R: Regressão Linear***. 2021. Disponível em: <https://storopoli.io/Estatistica/6-Regressao_Linear.html>. Citado na pág. 88, 89.

TEIXIERA DA SILVA, J. A.; DOBRÁNSZKI, J. **Multiple Authorship in Scientific Manuscripts: Ethical Challenges, Ghost and Guest/Gift Authorship, and the Cultural/Disciplinary Perspective**. *Science and Engineering Ethics*, v. 22, n. 5, p. 1457–1472, 2016. ISSN 14715546. doi:10.1007/s11948-015-9716-3. Citado na pág. 36, 99.

TEKLI, J. **An Overview on XML Semantic Disambiguation from Unstructured Text to Semi-Structured Data: Background, Applications, and Ongoing Challenges**. *IEEE Transactions on Knowledge and Data Engineering*, IEEE, v. 28, n. 6, p. 1383–1407, 2016. ISSN 10414347. doi:10.1109/TKDE.2016.2525768. Citado na pág. 45.

TSAI, W.-L.; CHAN, Y.-C. **Designing a Framework for Data Quality Validation of Meteorological Data System**. *IEICE Transactions on Information and Systems*, E102.D, n. 4, p. 800–809, apr 2019. ISSN 0916-8532. doi:10.1587/transinf.2018DAP0021. Citado na pág. 31.

TSCHARNTKE, T. et al. **Author sequence and credit for contributions in multiauthored publications**. *PLoS Biology*, v. 5, n. 1, p. 0013–0014, 2007. ISSN 15449173. doi:10.1371/journal.pbio.0050018. Citado na pág. 32, 33, 89, 99.

VELLIDO, A.; MARTÍN-GUERRERO, J. D.; LISBOA, P. J. **Making machine learning models interpretable**. *ESANN 2012 proceedings, 20th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, n. April, p. 163–172, 2012. Disponível em: <https://pdfs.semanticscholar.org/ce0b/8b6fca7dc089548cc2e9aaac3bae82bb19da.pdf>. Acesso em: 22 mar. 2020. Citado na pág. 19, 29.

VERA, M. C. S.; COMENDADOR, B. E. V. **A Web-Based Student Support Services System Integrating Short Message Service Application Programming Interface**. *International Journal of Future Computer and Communication*, v. 5, n. 2, p. 77–82, 2016. ISSN 20103751. doi:10.18178/ijfcc.2016.5.2.448. Citado na pág. 53.

VILLASEÑOR-ALMARAZ, M. et al. **Impact factor correlations with Scimago Journal Rank, Source Normalized Impact per Paper, Eigenfactor Score, and the CiteScore in Radiology, Nuclear Medicine & Medical Imaging journals**. *La radiologia medica*, Springer Milan, v. 124, n. 6, p. 495–504, jun 2019. ISSN 0033-8362. doi:10.1007/s11547-019-00996-z. Citado na pág. 37.

WALTERS, G. D. **Adding authorship order to the quantity and quality dimensions of scholarly productivity: evidence from group- and individual-level analyses**. *Scientometrics*, Springer Netherlands, v. 106, n. 2, p. 769–785, 2016. ISSN 15882861. doi:10.1007/s11192-015-1803-3. Citado na pág. 21, 34.

WATKINS, M. W. **Exploratory Factor Analysis: A Guide to Best Practice**. *Journal of Black Psychology*, v. 44, n. 3, p. 219–246, 2018. ISSN 15524558. doi:10.1177/0095798418771807. Citado na pág. 80.

WATSON, J. C. **Establishing evidence for internal structure using exploratory factor analysis**. *Measurement and Evaluation in Counseling and Development*, Taylor & Francis, v. 50, n. 4, p. 232–238, 2017. ISSN 19476302. doi:10.1080/07481756.2017.1336931. Citado na pág. 78.

WICKHAM, H. **Tidy Data**. *Journal of Statistical Software*, v. 59, n. 10, p. 1–23, 2014. ISSN 1548-7660. doi:10.18637/jss.v059.i10. Disponível em: <http://www.jstatsoft.org/v59/i10/>. Citado na pág. 29.

WILSON, R.; LADEIRA, S. **Automação de Diagnóstico para Manutenção Preditiva Baseada em Análise de Fluídos de Equipamentos com Machine Learning ***. In: *21° Seminário de Automação e TI. Anais do Seminário de Automação e TI.* [S.l.]: ABM Wek, 2017. v. 21, p. 268–279. Citado na pág. 26.

WINGO, M. T. et al. **Update in Outpatient General Internal Medicine: Practice-Changing Evidence Published in 2018**. *American Journal of Medicine*, Elsevier

Inc., v. 132, n. 8, p. 926–930, 2019. ISSN 15557162. doi:10.1016/j.amjmed.2019.02.023. Citado na pág. 84.

WOOLLEY, S. C. **Automating power: Social bot interference in global politics**. *First Monday*, First Monday, v. 4, p. 21, 2016. doi:10.5210/fm.v21i4.6161. Citado na pág. 45.

WRIGHT, A. **Big data meets big science**. *Communications of the ACM*, v. 57, n. 7, p. 13–15, jul 2014. ISSN 0001-0782. doi:10.1145/2617660. Citado na pág. 26.

YAMAKAWA, E. K. et al. **Comparativo dos softwares de gerenciamento de referências bibliográficas: Mendeley, EndNote e Zotero**. *Transinformação*, v. 26, n. 2, p. 167–176, 2014. ISSN 01033786. doi:10.1590/0103-37862014000200006. Citado na pág. 25.

YANG, S.; WOLFRAM, D.; WANG, F. **The relationship between the author byline and contribution lists: a comparison of three general medical journals**. *Scientometrics*, Springer Netherlands, v. 110, n. 3, p. 1273–1296, mar 2017. ISSN 0138-9130. doi:10.1007/s11192-016-2239-0. Citado na pág. 24, 31, 34, 36, 74, 84, 86, 89, 96, 99.

YANK, V.; RENNIE, D. **Disclosure of researcher contributions: A study of original research articles in the Lancet**. *Annals of Internal Medicine*, v. 130, n. 8, p. 661–670, 1999. ISSN 00034819. doi:10.7326/0003-4819-130-8-199904200-00013. Citado na pág. 32.

ZBAR, A.; FRANK, E. **Significance of authorship position: An open-ended international assessment**. *American Journal of the Medical Sciences*, Elsevier Masson SAS, v. 341, n. 2, p. 106–109, 2011. ISSN 00029629. doi:10.1097/MAJ.0b013e3181f683a1. Citado na pág. 31, 75, 85, 87, 89, 90, 96, 99.

APÊNDICES

A : Expressões de busca para pesquisa de artigos no Scopus

Autoria, Posicionamento Autoral e Contribuições

Quadro 1 – *Script de consulta do Scopus para pesquisa sobre autoria.*

TITLE-ABS-KEY("author byline"OR "authorship byline"OR "author order"OR "authorship order"OR "author contribution"OR "authorship contribution"OR "author responsibilities"OR "authorship responsibilities"OR "author position"OR "authorship position"OR "author impact"OR "authorship impact"OR "author prestigious"OR "authorship prestigious"OR "byline contribution"OR "byline order"OR "byline position"OR "byline impact"OR "coauthorship") AND DOCTYPE(ar) AND (LIMIT-TO (PUBSTAGE,"final")) AND (LIMIT-TO (PUBYEAR,2019) OR LIMIT-TO (PUBYEAR,2018) OR LIMIT-TO (PUBYEAR,2017) OR LIMIT-TO (PUBYEAR,2016) OR LIMIT-TO (PUBYEAR,2015) OR LIMIT-TO (PUBYEAR,2014) OR LIMIT-TO (PUBYEAR,2013)) AND (LIMIT-TO (LANGUAGE,"English")) Filtros aplicados: Intervalo (2013-2019), tipo (Artigos), estágio de publicação (Final), idioma (Inglês) Total: 661

Fonte: Autor

Ciência de Dados

Quadro 2 – *Script de consulta do Scopus para pesquisa sobre data science.*

TITLE-ABS-KEY("data science"OR "data science pipeline"OR "data gathering"OR "data cleaning"OR "data scraping"OR "datas cience pipeline") AND DOCTYPE(ar) AND (LIMIT-TO (PUBSTAGE,"final")) AND (LIMIT-TO (PUBYEAR,2019) OR LIMIT-TO (PUBYEAR,2018) OR LIMIT-TO (PUBYEAR,2017) OR LIMIT-TO (PUBYEAR,2016) OR LIMIT-TO (PUBYEAR,2015) OR LIMIT-TO (PUBYEAR,2014) OR LIMIT-TO (PUBYEAR,2013)) AND (LIMIT-TO (LANGUAGE,"English")) AND (LIMIT-TO (SUBJAREA,"COMP") OR LIMIT-TO (SUBJAREA,"MEDI") OR LIMIT-TO (SUBJAREA,"DECI")) Filtros aplicados: Intervalo (2013-2019), tipo (Artigos), estágio de publicação (Final), áreas (computação, medicina, decisão), idioma (Inglês) Total: 511

Fonte: Autor

B : Expressões de busca para pesquisa de artigos na Web of Science

Autoria, Posicionamento Autoral e Contribuições

Quadro 3 – *Script de consulta da Web of Science para pesquisa sobre autoria.*

ALL=("author byline"OR "authorship byline"OR "author order"OR "authorship order"OR "author contribution"OR "authorship contribution"OR "author responsibilities"OR "authorship responsibilities"OR "author position"OR "authorship position"OR "author impact"OR "authorship impact"OR "author prestigious"OR "authorship prestigious"OR "byline contribution"OR "byline order"OR "byline position"OR "byline impact"OR "coauthorship") Filtros aplicados: Intervalo (2013-2019), tipo (Artigos), idioma (Inglês) Total: 721

Fonte: Autor

Ciência de Dados

Quadro 4 – *Script de consulta da Web of Science para pesquisa sobre data science.*

ALL=("data science"OR "data science pipeline"OR "data gathering"OR "data cleaning"OR "data scraping"OR "datas cience pipeline") Filtros aplicados: Intervalo (2013-2019), tipo (Artigos, Proceedings Papers), idioma (Inglês) Total: 148

Fonte: Autor

C : Expressões de busca para pesquisa de artigos no Scielo

Autoria, Posicionamento Autoral e Contribuições

Quadro 5 – *Script de consulta do Scielo para pesquisa sobre autoria.*

"author byline"OR "authorship byline"OR "author order"OR "authorship order"OR "author contribution"OR "authorship contribution"OR "author responsibilities"OR "authorship responsibilities"OR "author position"OR "authorship position"OR "author impact"OR "authorship impact"OR "author prestigious"OR "authorship prestigious"OR "byline contribution"OR "byline order"OR "byline position"OR "byline impact"OR "co-authorship" Filtros aplicados: (Ano de publicação: 2013) (Ano de publicação: 2014) (Ano de publicação: 2016) (Ano de publicação: 2019) (Ano de publicação: 2015) (Ano de publicação: 2017) (Ano de publicação: 2018) Total: 13

Fonte: Autor

Ciência de Dados

Quadro 6 – *Script de consulta do Scielo para pesquisa sobre data science.*

"data science"OR "data science pipeline"OR "data gathering"OR "data cleaning"OR "data scraping"OR "datas cience pipeline" Filtros aplicados: (Idioma: Inglês) (Ano de publicação: 2018) (Ano de publicação: 2016) (Ano de publicação: 2014) (Ano de publicação: 2017) (Ano de publicação: 2019) (Ano de publicação: 2013) (Ano de publicação: 2015) (Citáveis e não citáveis: Citável) (Tipo de literatura: Artigo) Total: 30

Fonte: Autor

D : Trabalhos Resultantes da Tese

Publicação

1. DE SOUZA, E. M.; STOROPOLI, J. E. ; ALVES, W. A. L. **FERRAMENTA DE EXTRAÇÃO DE DADOS PARA A WEB OF SCIENCE**. **In**: SETII - Seminário em Tecnologia da Informação Inteligente, 2019, São Paulo. Universidade Nove de Julho.

Sob Avaliação

1. DE SOUZA, E. M.; ALVES, W. A. L.; STOROPOLI, J. E. **Scientific Contribution List Categories Investigation**. *Estrato A1 (CAPES)*, 2021.

Em Desenvolvimento

1. DE SOUZA, E. M.; ALVES, W. A. L.; VILS, L. STOROPOLI, J. E. **What type of contributions makes an author first or last? (Título provisório)**. *Estrato A1 (CAPES)*, 2021. Previsão de submissão em Jul. 2021.

Produção de Software

1. DE SOUZA, E. M. **Desenvolvimento de robôs para coleta e tratamento de dados sobre artigos científico**s, 2020. Disponível em forma de *software* livre em: <https://github.com/EdsonMSouza/PhD>.

2. DE SOUZA, E. M. **SJMetrics**. Ferramenta *online* para auxiliar no processo inicial de seleção de periódicos com base em índices bibliométricos extraídos das plataformas: Web of Science, SCImago, Google Schollar e Researchgate, 2020. Disponível online em: <https://www.edsonmelo.com.br/sjmetrics/>.

3. DE SOUZA, E. M.; ALVES, W. A. L.; STOROPOLI, J. E. **Web of Science Extractor**, 2019. Programa para extração de índices bibliométricos sobre artigos mantidos pela indexadora de periódicos *Web of Science*. Disponível em forma de *software* livre em: <https://github.com/EdsonMSouza/web_of_science_metrics>.