AtoZ
novas práticas em informação e conhecimento

ISSN: 2237-826X
ARTIGO | PAPER

Identificação de especialistas em câncer no Brasil em plataforma de big data: um estudo de caso minerando a Plataforma Lattes com auxílio de softwares de prospecção

Identification of cancer specialists in Brazil on a big data platform: a case study mining the Lattes Platform with the aid of data prospection softwares

Henrique Koch Chaves [1], Alessandra Moreira de Oliveira [2], Suzanne de Oliveira Rodrigues Schumacher [3], Rafael Cavalcante dos Santos [4], Adelaide Maria de Souza Antunes [5], Jorge Lima de Magalhães [6]

[1] *Fundação Oswaldo Cruz/FIOCRUZ – Instituto de Tecnologia em Fármacos/Farmanguinhos, Rio de Janeiro – RJ, Brasil. ORCID: https://orcid.org/0000-0003-3035-6799*

[2] *Universidade Federal do Rio de Janeiro, Sistema de Informação sobre a Indústria Química (SIQUIM), Escola de Química, Rio de Janeiro – RJ, Brasil. ORCID: https://orcid.org/0000-0003-4203-6637*

[3] *Universidade Federal do Rio de Janeiro, Sistema de Informação sobre a Indústria Química (SIQUIM), Escola de Química, Rio de Janeiro – RJ, Brasil. ORCID: https://orcid.org/0000-0001-8210-5663*

[4] *Universidade Federal do Rio de Janeiro, Programa de Pós-Graduação em Engenharia de Processos Químicos e Bioquímicos, Escola de Química, Rio de Janeiro – RJ, Brasil. ORCID: https://orcid.org/0000-0001-7161-4796*

[5] *Instituto Nacional da Propriedade Industrial, Rio de Janeiro – RJ, Brasil; Universidade Federal do Rio de Janeiro, Escola de Química, Rio de Janeiro – RJ, Brasil. ORCID: https://orcid.org/0000-0002-2245-7517*

[6] *Global Health and Tropical Medicine (GHTM), Instituto de Higiene e Medicina Tropical (IHMT), Universidade NOVA de Lisboa – UNL, Lisboa – Estremadura, Portugal; International Platform for Science, Technology and Innovation in Health (PCTIS), University of Aveiro, Aveiro – Beira Litoral, Portugal; Fundação Oswaldo Cruz/FIOCRUZ – Instituto de Tecnologia em Fármacos/Farmanguinhos, Rio de Janeiro – RJ, Brasil. ORCID: https://orcid.org/0000-0003-2219-5446*

Autor para correspondência/*Mail to:* Henrique Koch Chaves, henrique.chaves@far.fiocruz.br

Recebido/*Submitted*: 26 de julho de 2023; **Aceito/*Approved*:** 15 de agosto de 2024

Resumo

Introdução: Dados de diversas fontes são gerados a cada segundo, exigindo, assim, novas soluções para processá-los e gerenciá-los de maneira rápida. Muitas organizações públicas e privadas têm utilizado a análise de Big Data como estratégia de gestão. No campo da Oncologia, a análise de Big Data é capaz de fornecer subsídios valiosos à tomadores de decisão, seja para o desenho de políticas públicas, a alocação de recursos para pesquisa. Neste trabalho propõe-se uma metodologia para prospecção e análise de dados de CV Lattes visando identificar pesquisadores especialistas no campo da oncologia. **Método**: Utilizou-se, neste estudo, a ferramenta computacional ScriptLattes em combinação com o software KNIME para extração e análise de dados. Foram obtidas informações essenciais dos pesquisadores que atuam em oncologia, na subárea Medicina. A metodologia envolveu a identificação de especialistas, principais produções, distribuição geográfica e redes de colaboração. **Resultados**: A Plataforma Lattes revelou 198 pesquisadores aderentes aos critérios e filtros aplicados na estratégia de busca, dos quais 134 especialistas foram identificados com graduação em medicina e produções na área de oncologia. Artigos científicos figuram como principal produção entre os especialistas de maior destaque. Observou-se maior concentração desses especialistas na Região Sudeste do país e a presença de uma rede de colaborações envolvendo a maior parte dos especialistas mais produtivos. **Conclusão**: As estratégias e metodologias apresentadas neste estudo permitem a prospecção de informações e construção do cenário de especialistas brasileiros em oncologia, sendo promissoras para subsidiar gestores de CTI na tomada de decisões.

Palavras-chave: Câncer; Oncológicos; Plataforma Lattes; ScriptLattes; Medicina; Especialistas.

Abstract

Introduction: *Data from different sources are generated every second, thus requiring new solutions to process and manage them quickly. Many public and private organizations have used Big Data analysis as a management strategy. In the field of Oncology, Big Data analysis is capable of providing valuable subsidies to decision makers, whether for the design of public policies or the allocation of resources for research. This work proposes a methodology for prospecting and analyzing data from CV Lattes in order to identify specialist researchers in the field of oncology.* ***Method****: In this study, the computational tool ScriptLattes was used in combination with the KNIME software for data extraction and analysis. Essential information was obtained from researchers who work in oncology, in the Medicine subarea. The methodology involved the identification of specialists, main productions, geographic distribution and collaboration networks.* ***Results****: The Lattes platform revealed 198 researchers adhering to the criteria and filters applied in the search strategy, of which 134 specialists were identified with a degree in medicine and productions in the field of oncology. Scientific articles are the main production among the most prominent specialists. There was a greater concentration of these specialists in the southeastern region of the country and the presence of a network of collaborations involving most of the most productive specialists.* ***Conclusion****: The strategies and methodologies presented in this study allow the prospection of information and construction of the scenario of Brazilian specialists in oncology, being promising to support ST&I managers in decision-making.*

Keywords: *Cancer, Oncology; Lattes Platform; ScriptLattes; Medicine; Specialists.*

INTRODUÇÃO

Big Data em Saúde

Em tempos hodiernos, quintilhões de dados são disponibilizados diariamente na web em praticamente todos os lugares, como universidades, empresas e residências, no qual esses dados são oriundos das mais diversas fontes, tais como redes sociais, internet das coisas, dispositivos móveis, transações bancárias e comerciais, satélites, sistemas de monitoramento, sensores (dados de localização e dados meteorológicos), registro de softwares etc. Martino, Aversa, Cretella, Esposito, e Kołodziej (2014). Não obstante, agregam-se a este Big Data, o conhecimento científico (artigos) e tecnológico (patentes). Como resultado, novas formas computacionais têm sido desenvolvidas e adotadas visando recuperar e extrair valores desses dados. Seja pelo advento da inteligência artificial ou pelo efeito da Ciência da Informação sobre a Ciência Computacional, novas estruturas ou arquiteturas informacionais surgiram, favorecendo os processos decisórios, políticas e estratégia nas organizações Magalhães, Martins, e Hartz (2014).

Na área da saúde, soluções inovadoras são constantemente exigidas para superar os diversos desafios pelos quais os clínicos se deparam, seja pelo diagnóstico tardio, a eficácia da intervenção ou ainda o fato das patologias serem tratadas geralmente uma de cada vez. Assim, os dados devem constituir uma das bases nos sistemas nacionais de saúde, uma vez que os mesmos auxiliarão no treinamento e desenvolvimento dos algoritmos avançados de inteligência artificial (IA) para a extração de informações essenciais do Big Data **?**.

Um exemplo de Big Data brasileiro é a Plataforma Lattes do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). Esta base de dados integra os currículos (CV) acadêmicos de profissionais de todas as áreas do conhecimento, tanto de brasileiros quanto de pesquisadores estrangeiros residindo em território nacional ou, ainda que em algum momento desenvolveram projetos em parceria com brasileiros. Uma das limitações da base é que o sistema de extração das informações dos CV está baseado na busca individual por nome da pessoa cadastrada. Assim, visando extrair os diversos dados simultaneamente dos CV cadastrados, foi desenvolvida a ferramenta computacional ScriptLattes (Brito, Quoniam, e Mena-Chalco (2016); Magalhães, Hir, Quoniam, Hartz, e Oliveira (2020); Mena-Chalco e Cesar Junior (2009)).

O Scriptlattes é um programa de código aberto e o seu funcionamento baseia-se na execução sequencial de alguns módulos computacionais tendo como base uma lista de nomes criada manualmente (CV específicos de indivíduos conhecidos), ou automatizada (ao criar lista utilizando termos de busca no próprio sistema de procura da Plataforma Lattes e no Diretório de Grupos de Pesquisa do CNPq). O primeiro módulo extrai diretamente da Plataforma Lattes os CV que se deseja analisar (Ferraz, Barnabé, Quoniam, Santos, e Mariosa (2018); Ferraz e Quoniam (2013); Mena-Chalco e Cesar Junior (2013) As informações obtidas com essa ferramenta possibilitam às agências governamentais, sociedades científicas e empresas, analisarem a distribuição da ciência (especialistas em oncologia) nos estados e regiões do país, bem como avaliarem em quais ainda precisam ser desenvolvidas.

Big Data em Oncologia

Segundo o Instituto Nacional do Câncer José Alencar Gomes da Silva - INCA, o "câncer é um termo que abrange mais de 100 diferentes tipos de doenças malignas que têm em comum o crescimento desordenado de células, que podem invadir tecidos adjacentes ou órgãos à distância" INCA (2022b).

Para o triênio 2023-2025 são esperados 704 mil novos casos de câncer, sendo que destes, aproximadamente 70% dos casos se concentram nas regiões Sul e Sudeste do país. Entre os tumores malignos mais incidentes no país estão o de pele não melanoma (31,3% do total de casos), mama feminina (10,5%), próstata (10,2%), cólon e reto (6,5%), pulmão (4,6%) e estômago (3,1%). A estimativa no triênio fornecida pelo INCA fornece dados importantes para a definição de políticas públicas, além de ser a principal ferramenta de gestão e planejamento na área oncológica no Brasil Ministério da Saúde (2023)).

De acordo com o *National Cancer Institute* (EUA) o "Big data fornecerá uma oportunidade sem precedentes para entender o câncer em todos os níveis, desde assinaturas moleculares até estatísticas nacionais" NIH (2022). Fundamental para o tratamento, pois sistemas de saúde fortes com cuidados de sobrevivência, tratamento de qualidade e detecção precoce acessível têm altas taxas de sobrevivência para muitos tipos de câncer WHO (2022).

Estrutura Conceitual em Câncer

A instalação do câncer nos órgãos é normalmente silenciosa e assintomática, dificultando o seu diagnóstico e tratamento em tempo hábil, uma vez que vários casos a sua descoberta ocorrem quando a doença já está em estágio avançado, já acometendo outros órgãos INCA (2022a) . Nesse cenário, a relevância dessa escolha na área da saúde, se fundamenta pelo grande desafio do diagnóstico e da quimioterapia em entregar seletivamente os fármacos às células tumorais com interação mínima com tecidos saudáveis Webster, Parks, Titov, e Beasley (2014) .

Estudos apontam que os gastos com câncer no Brasil aumentam exponencialmente, mesmo ainda estando abaixo do atendimento das necessidades. Somente no ano de 2020, as despesas somente do INCA foi em torno de 300 milhões de reais INCA (2020). Dentre as terapias anticâncer disponíveis incluem-se os agentes quimioterápicos, biológicos, terapia do alvo molecular, radioterapia, cirurgia e oncologia intervencionista Ramos, Rito, e Vieira (2021); Sag, Selcukbiricik, e Mandel (2016). Além da terapia com os medicamentos que agem diretamente no combate nos tumores, outros fármacos são utilizados em conjunto para minimizar as toxicidades causadas pelos medicamentos anticâncer: antieméticos, protetores urinários, corticoides e hidratação venosa, incluindo, quando indicado, transfusões de hemácias e plaquetas, antibióticos e fatores de crescimento Matz e Hsieh (2017); Oun, Moussa, e Wheate (2018).

Desta forma, considerando a importância dos oncológicos para a saúde pública brasileira e mundial, este trabalho objetivou identificar os especialistas mais relevantes em território brasileiro atuando no tema, que tenham CV cadastrado na plataforma. Para tanto, utilizou-se centrar a pesquisa neste artigo em oncológicos, a fim de que pudesse ser dado uma melhor dimensão dessa aplicação na grande área Ciências da Saúde, com ênfase na subárea Medicina.

METODOLOGIA

Busca e Extração de dados de Pesquisadores No que se refere ao resgate dos CV dos pesquisadores na Plataforma lattes, utilizou-se software gratuito ScriptLattes em ambiente Linux para extração dos dados abertos na plataforma Lattes brasileira. Foram selecionados os filtros abaixo listados a fim de identificar os especialistas com maior experiência acadêmica na área[1] **a**. nível (doutorado); **b**. nacionalidade (brasileira e estrangeira); c. bolsistas de produtividade CNPq (todas as categorias); d. presença nos diretórios de Grupos de Pesquisa; e. atuação profissional (grande área: Ciências da Saúde e subárea: Medicina na Plataforma Lattes do CNPq).

No sentido de identificar, preliminarmente, o número de CV disponíveis em cada estratégia de pesquisa, em setembro de 2022, foram realizadas buscas no campo "avançado" da Plataforma Lattes utilizando termos tanto em inglês quanto em português retirados do site do National Institute of Cancer (USA) (NIH, 2022) (ver Figura 1).

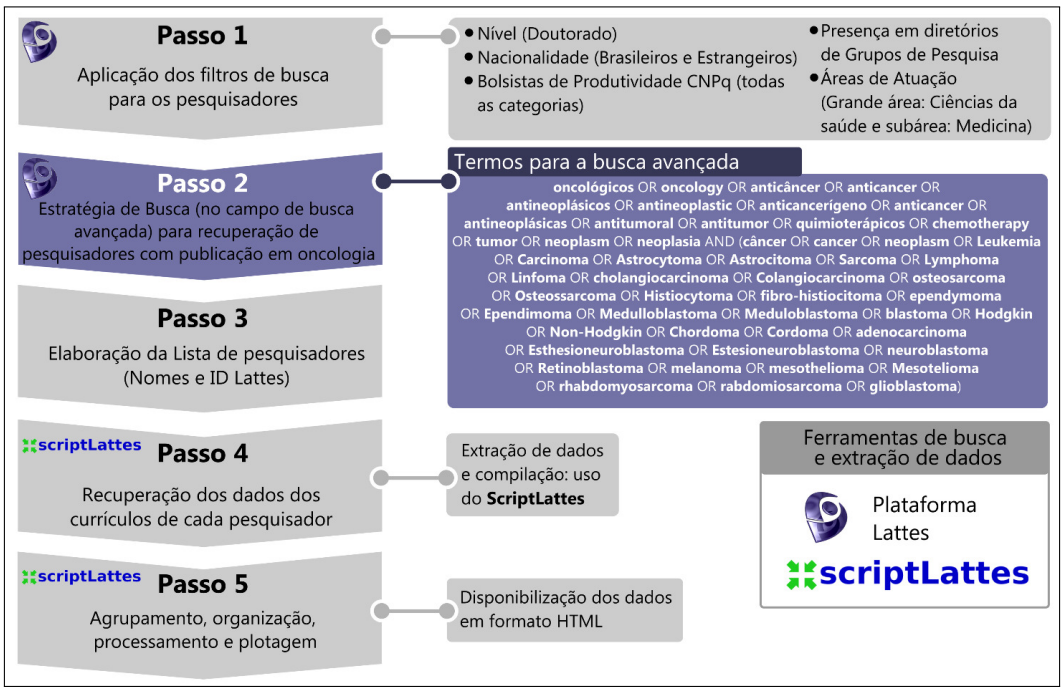

Figura 1 – Metodologia e estratégia de busca... **Fonte:** Elaborado pelos autores, (2023).

[1]As áreas designadas nesta fase do estudo não são as áreas de formação do pesquisador, mas aquelas indexadas pelo pesquisador como área e subárea na Plataforma Lattes.

[2]A análise dos tipos de câncer disponíveis na lista do NIH foi avaliada e comparada com a lista fornecida pelo Instituto Nacional de Câncer José Alencar Gomes da Silva (INCA, 2019b).

[3]Através dos operadores booleanos, a Plataforma Lattes permite obter informações para buscas simples. Buscas relacionais mais complexas ainda não são possíveis na Plataforma Lattes, uma vez que o motor de busca disponível da plataforma aceita somente alguns operadores booleanos como `and`, `or`, `not` e `near`. Já os caracteres `()`, `_`, `&`, `/`, `!`, `~`, `*`, `?` e `$` não são reconhecidos pela Plataforma Lattes Brito et al. (2016). Por exemplo, ao colocar aspas em termos de busca com mais de uma palavra, a plataforma não compreende o operador e as palavras aparecem separadas nos currículos dos pesquisadores. Assim, optou-se por utilizar apenas termos de busca com uma palavra.

Em relação a extração e compilação dos dados na Plataforma Lattes utilizando o ScriptLattes, resgatou-se os CV pelo ID Lattes (*Identity Lattes*) e nome de cada especialista. Os dados foram então, agrupados, organizados, processados e, posteriormente disponibilizados em ambiente na web em formato HTML (**Figura 1**).

Identificação de especialistas e pesquisadores de destaque em oncologia

No sentido de identificar os especialistas em Oncologia com formação na área de medicina, o foco da presente pesquisa está nos pesquisadores mais produtivos) e com formação (graduação) na área de medicina. Os CV foram minerados, buscando-se identificar e separar tais pesquisadores para análise posterior de sua produção. O software livre e gratuito KNIME *Analytic Platform* v.4.6.1 foi aplicado na identificação dos pesquisadores com formação em medicina e que possuam, pelo menos, um artigo, livro ou capítulo de livro publicado, no campo da oncologia. A identificação dessas produções no campo da oncologia foi realizada pelos termos de busca (os mesmos aplicados na estratégia de busca dos CV na Plataforma Lattes) nos títulos das produções mencionadas (**Figura 2.a**).

No que tangue ao TOP 10 pesquisadores de destaque em artigos, livros e capítulos de livros, foi aplicado o software KNIME na detecção e ordenação dos especialistas de maior destaque, segundo seus níveis de produção autodeclarada, em termos da quantidade de artigos completos publicados, livros e capítulos de livros exclusivamente relacionados ao tema de Oncologia. Um algoritmo de mineração de texto **(Figura 2.b)** foi elaborado para identificação contagem e ordenação das produções relacionadas somente a oncologia, utilizando os termos de busca apresentados anteriormente.

Em relação a frequência relativa dos termos de busca, ela foi computada a partir da quantificação da incidência das mesmas (palavras-chave) nos títulos dos artigos, livros e capítulos de livro publicados pelos pesquisadores. A frequência relativa de cada termo foi calculada a partir dos valores totais de termos identificados. Tal quantificação foi realizada por meio de um workflow gerado em KNIME, o qual é mostrado de forma simplificada na **Figura 2.c.**

Referente à distribuição geográfica dos especialistas em oncologia, um mapa apresentou a distribuição geográfica dos TOP 10 pesquisadores de maior destaque, segundo suas produções em artigos, livros e capítulos de livros foi plotado com auxílio do software gratuito QGIS v.3.24.0 (**Figura 2.d**). A distribuição geográfica dos pesquisadores foi computada a partir dos dados de latitude e longitude das instituições as quais os pesquisadores declaram possuir vínculo, extraídos pelo ScriptLattes.

Na **Figura 2.e**, observam-se as análises das redes de coautoria entre os especialistas em oncologia de maior destaque nas produções de artigos, livros e capítulos de livros foram realizadas utilizando o software livre e gratuito Gephi. A rede de colaboração foi plotada com base no número de artigos em comum entre os pares de pesquisadores. Foram identificados os nomes, instituições vinculadas e estados brasileiros (localização das instituições). A metodologia para geração da rede é presentada de forma simplificada.

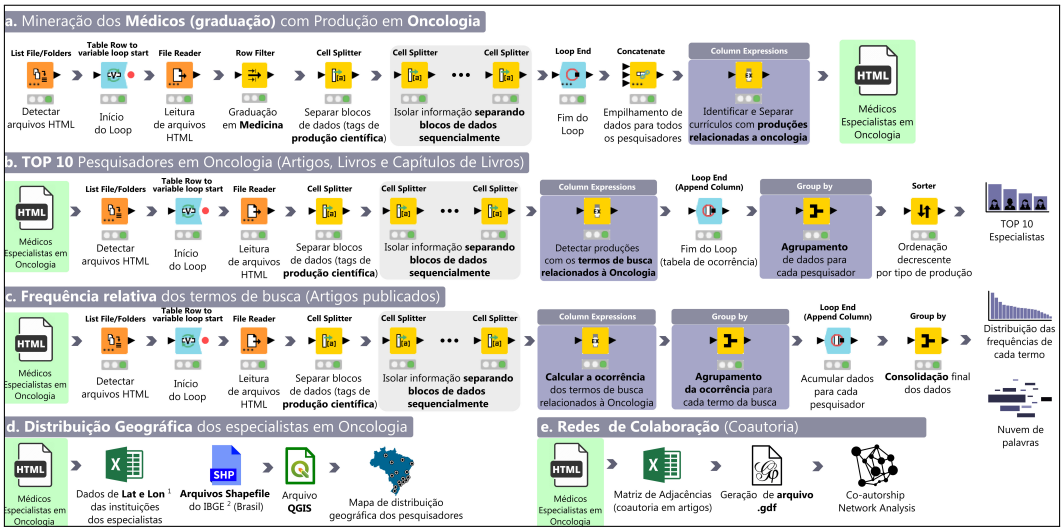

Figura 2 – Workflows para identificação de especialistas em oncologia a partir dos dados do ScriptLattes (**a**) e subsequente quantificação e ranqueamento de especialistas mais destacados (**b**), frequência dos termos de busca (**c**), análise da distribuição geográfica dos especialistas em oncologia (**d**) e redes de colaboração (**e**). As etapas destacadas enfatizam o cálculo mais importante nos fluxos de trabalho baseados em KNIME[a] **Fonte:** Elaborado pelos autores, (2023).

RESULTADOS E DISCUSSÃO

Resgate e Extração de CV da Plataforma Lattes

A plataforma Lattes possui um total de 15.302 CV referentes a pesquisadores brasileiros ou estrangeiros com doutorado. Esse valor decresce para 1.435 CV ao aplicar o filtro "Bolsistas de Produtividade do CNPq", representando uma queda de mais de 1/10 daquela encontrada inicialmente. Acrescentando-se o filtro "Presença no Diretório de Grupos de pesquisa (GP)", a quantidade de CV permanece praticamente constante em todos os termos pesquisados, resultando em 1.340 CV.

Os filtros "Atuação profissional" com grande área em "Ciências da Saúde" e subárea em "Medicina" refina a amostra para 610 e 298 pesquisadores respectivamente. Ressalta-se que o filtro "Atuação profissional" utilizado refere-se à área na qual o pesquisador em questão desenvolve suas atividades científicas. Este é um filtro que permite captar pesquisadores com diferentes formações (graduação, mestrado, doutorado, especialização em diferentes cursos), com atuação no campo de interesse.

Os CV do 298 pesquisadores retornados pela busca utilizando-se todos os termos na Plataforma Lattes foram recuperados utilizando-se o ScriptLattes. A partir dos dados gerados por este software, prosseguiu-se com o necessário refinamento para identificação dos especialistas em oncologia com formação em medicina.

Identificação e Separação dos Especialistas em Oncologia com formação em medicina

Uma avaliação da amostra dos 298 CV mostrou que muitos pesquisadores declaram algum tipo de produção ou atividade profissional relacionada a oncologia, a qual é detectada pela estratégia de busca a partir das palavras-chave características do campo de estudo. Entretanto, vários pesquisadores podem possuir baixa produção neste campo de estudo ou declaram atividades de reduzida contribuição aos campos da oncologia. Dessa forma, filtros adicionais como, a titulação em nível de graduação no curso de Medicina e, pelo menos, uma produção relacionada a oncologia, em termos de artigos, livros e capítulos de livros publicados, reduzem ainda mais a amostra relevante para o estudo, para um total de 134 pesquisadores. Cabe ressaltar que esta filtração adicional dos dados foi realizada com auxílio do software KNIME dado que a Plataforma Lattes não oferece opções adequadas para obtenção deste grau de refinamento. Tal amostra de 134 pesquisadores contempla os especialistas nos diversos campos da oncologia, cujas produções são analisadas a seguir. O número de currículos resgatados referentes à cada conjunto de filtros aplicados pode ser visualizado na **Figura S1** do material suplementar.

TOP 10 pesquisadores de destaque em oncologia

A **Figura 3** apresenta os resultados para a mineração das produções relacionadas a oncologia. A distribuição das produções relacionadas a oncologia, entre as três principais métricas de produção científica analisadas, artigos publicados em periódicos, capítulos de livros e livros completos, pode ser observada na **Figura 3.a**. Da mesma forma, as produções relacionadas estritamente ao campo da oncologia dos pesquisadores mais destacados se concentram nos artigos completos publicados em periódicos, totalizando 5.125 trabalhos (90%), e em capítulos de livros publicados, totalizando 525 trabalhos (9%). Menos de 2% (49 trabalhos) das produções em oncologia são relacionadas à livros completos publicados pelos cientistas analisados.

Nota-se que a produção em termos de capítulos de livros é significativamente maior que a produção em termos de livros completos. Esse resultado pode estar relacionado ao fato de haver menor esforço a ser empreendido na elaboração de capítulos de livros, uma atividade normalmente colaborativa e distribuída, do que na elaboração de Livros completos, usualmente congregando menor número de autores ou apenas um único autor.

A partir desta avaliação global, pode-se inferir que as métricas artigos completos publicados e capítulos de livros abrange 99% da produção em oncologia dos pesquisadores mais destacados, sendo métricas candidatas inclusive como filtros para refinamentos ou busca avançada com enfoque em temas específicos no campo da oncologia.

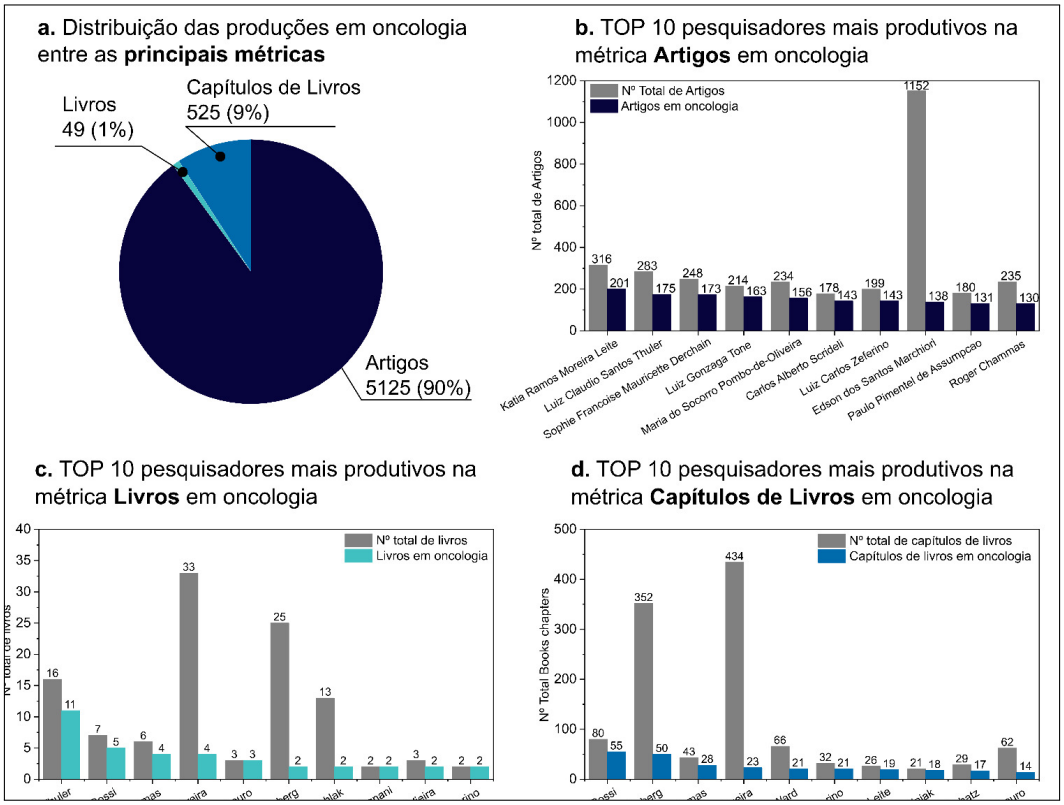

Figura 3 – Produções relacionadas à oncologia... **Fonte:** Elaborado pelos autores, (2023).

Os dados apresentados nas **Figuras 3.b, 3.c, e 3.d** revelam que as produções em oncologia, contabilizadas em cada uma das métricas, podem ser consideradas como as principais produções da maioria dos pesquisadores analisados. Na métrica artigos publicados em periódicos, por exemplo, observa-se que em média, os referidos artigos correspondem a aproximadamente 63% do total publicado pelos pesquisadores, com máximo de 80% e mínimo de 55%.

Em relação aos capítulos de livros publicados, e considerando somente os pesquisadores com produção em oncologia acima de 50%, o valor médio da produção em oncologia é de 69%, com máximo de 86% e mínimo de 59%. A métrica livros completos publicados apresenta os menores valores de produção entre os pesquisadores, entretanto todos os TOP 10 pesquisadores tem pelos menos uma produção em oncologia nesta métrica. O percentual médio, novamente considerando somente os pesquisadores com produção em oncologia acima de 50% do total de sua produção, é de 82% com variação ampla entre um máximo de 100% e um mínimo de 66%.

Observa-se que os níveis de produção em oncologia dos pesquisadores mais destacados acompanham o nível de produção total dos mesmos, no caso da métrica artigos publicados. Para as métricas capítulos de livros e livros completos, os níveis apresentam um comportamento similar aquele dos artigos, com a exceção de alguns pesquisadores, que apresentam produção total extraordinariamente alta, mas a produção relacionada a oncologia compõe menos de 20% de seu acervo. Tais excepcionalidades podem ser atribuídas aos outros campos de investigação científica nos quais esses pesquisadores atuam e possuem produção relevante.

Distribuição Geográfica dos especialistas em oncologia

A distribuição geográfica dos TOP 10 especialistas em oncologia, considerando as três métricas avaliadas, foi plotada na **Figura 4**, de modo similar ao mostrado anteriormente. Conforme pode-se notar, os pesquisadores mais destacados nas três métricas, com atuação principal em Medicina, e que efetivamente possuem produções relacionadas aos temas da oncologia, estão em sua maioria concentrados na região sudeste do território brasileiro com exceção do pesquisador Dr. Paulo P. Assumpção, destaque na produção de artigos em oncologia, vinculado à Universidade Federal do Pará. Entre os estados da Região Sudeste , destaca-se o estado de São Paulo, com 18 pesquisadores atuantes (Dr. Rene A. A. Vieira, Hospital de Câncer de Barretos; Dr. Jose H. T. G. Fregnani, Hospital de Câncer de Barretos;; Dr. Luiz Gonzaga Tone, USP; Dr. Carlos A. Scrideli, USP; Dra. Katia R. M. Leite, USP; Dr. Manoel J. Teixeira, USP; Dr. Nelson Hamerschlak, Hospital Israelita Albert Einstein; Dr. Dan L. Waitzberg, USP; Dra. Lydia M. Ferreira, UNIFESP; Dra. Maria Isabel Achatz, Hospital Sírio-Libanês; Dr. Benedito M. Rossi, Clínica de Cirurgia e Oncologia; Dr. Rubens Chojniak, A. C. Camargo Center; Dr. Roger Chammas, USP; Dra. Sophie F. M. Derchain, UNICAMP; Dra. Laura S. Ward, UNICAMP e Dr. Luiz C. Zeferino, UNICAMP), o estado do Rio de Janeiro, com 3 pesquisadores atuantes (Dr. Luiz C. S. Thuler e Dra. Maria S. Pombo-de-Oliveira, ambos do Instituto Nacional do Câncer; Dr. Edson S. Marchiori, UFRJ), e o

estado do Espírito Santo com apenas um pesquisador de destaque (Dr. Iuri Drumond Louro, UFES).

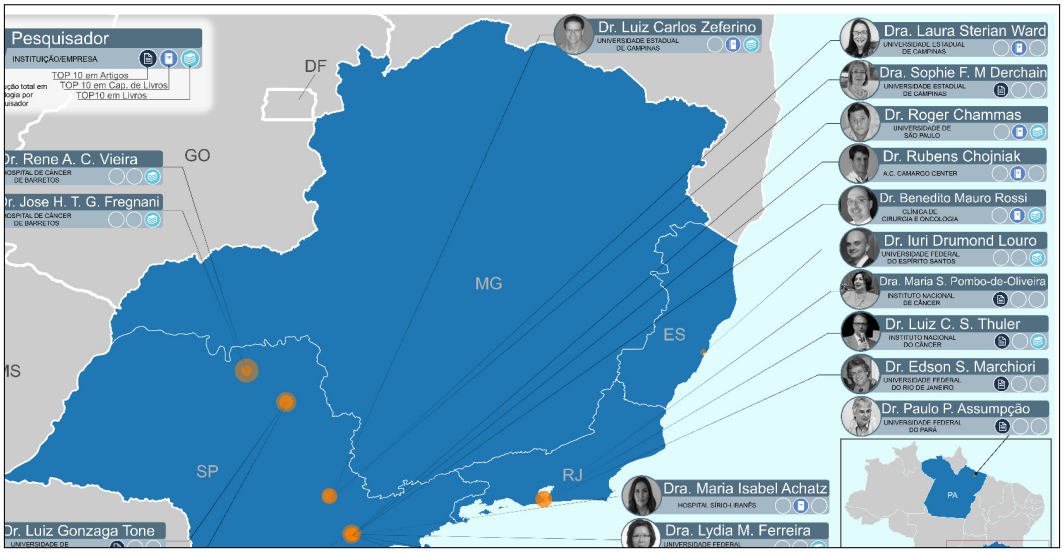

Figura 4 - Localização geográfica dos pesquisadores mais produtivos (instituições vinculadas) nas três principais métricas (vide ícones na legenda) e seus respectivos níveis de produções totais relacionadas à oncologia (artigos + capítulos de livros + livros, círculos em laranja)
Fonte: Elaborado pelos autores, (2023).

Universidades e centros de pesquisa públicos de notório reconhecimento e consagrada contribuição aos diversos campos da ciência, como a Universidade de São Paulo, figuram entre as instituições às quais os pesquisadores de maior destaque estão vinculados.

Análise da frequência dos termos de busca relacionados a oncologia entre as produções

A ocorrência de cada palavra-chave, que compõe a sinonímia utilizada na identificação dos 134 pesquisadores especialistas em oncologia e suas produções, pode ser analisada de forma a se obter algum insight a respeito das temáticas tratadas dentro dos trabalhos em oncologia desenvolvidos pelos pesquisadores. Na **Figura 5** é apresentada a frequência acumulada para as palavras-chave utilizadas. A frequência acumulada é computada a partir da frequência absoluta de cada termo referente ao tema da oncologia (veja **Figura S3** e **S4** do material suplementar). Por simplificação e para melhor visualização da informação, as palavras do dicionário sinonímico, e suas variantes em gênero e número, foram agregadas segundo sua significação.

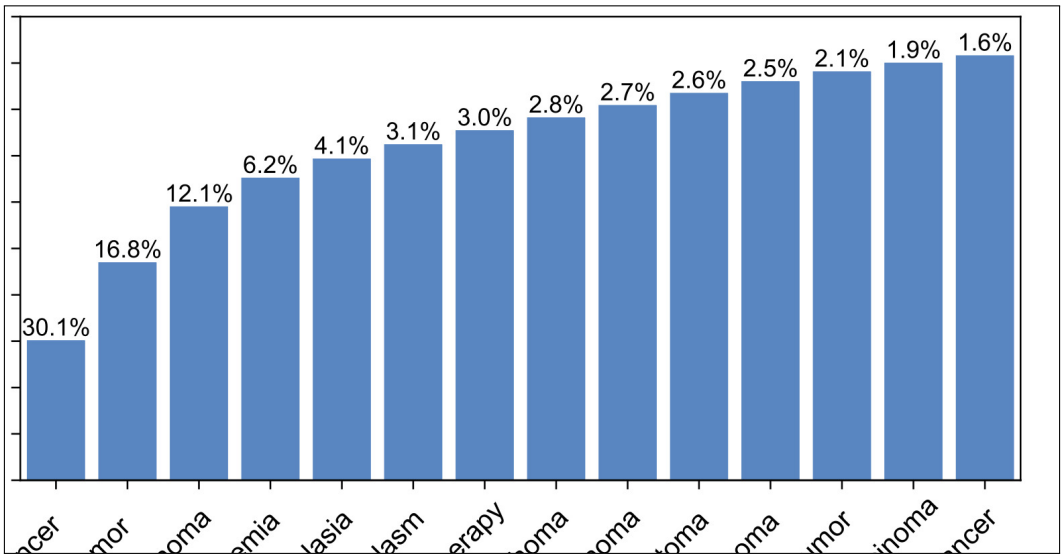

Figura 5 – Frequência cumulativa para o termos de busca cobrindo pelo menos 90% da ocorrência total computada para todas as palavras-chave utilizadas na mineração das produções em oncologia relacionadas aos 134 especialistas identificados
Fonte: Elaborado pelos autores, (2023).

Observa-se que grande parte dos trabalhos (artigos, livros e capítulos de livros) possuem títulos compostos de termos generalistas como “cancer”, “tumor”, “carcinoma” e “neoplasia”. Entre os principais termos que identificam os tipos de câncer mais pesquisados estão “leukemia”, “lymphoma”, “melanoma”, “blastoma”, “sarcoma” e “adenocarcinoma”. O desenvolvimento de pesquisas e a criação de competências focadas nestes tipos de neoplasia

é de interesse da própria administração da saúde pública. Leucemias, linfomas e melanomas são tipos de cânceres de considerável frequência entre as neoplasias diagnosticadas, especialmente na população brasileira INCA (2022b). "chemotherapy", "antitumor" e "anticancer" também são frequentes entre os termos mais utilizados e indexam trabalhos ligados às aplicações terapêuticas em oncologia como ação antitumoral de compostos químicos, por exemplo. A maior parte dos termos referentes a tipos específicos de câncer foram detectados entre os trabalhos analisados o que pode indicar certa diversidade da expertise desenvolvida pelos pesquisadores brasileiros. As inferências apresentadas estão obviamente restritas pelas limitações da metodologia de busca, limitações no dicionário sinonímico utilizado para a mineração de texto e métricas aplicadas. É previsto que as produções utilizando exclusivamente termos técnicos ou jargões específicos no campo da oncologia, para descrever os temas tratados, não serão quantificados pelos algoritmos computacionais como produções em oncologia, apesar de o serem de fato. Por exemplo, títulos de artigos, livros ou capítulos de livros contendo exclusivamente a abreviação "PC-3 cell lines" (linhagem de células de câncer humano de próstata), são qualificados como produções relacionadas ao campo de estudos da oncologia, no entanto, não são quantificados pela metodologia aplicada neste trabalho. Estas restrições não comprometem os resultados, uma vez que a sinonímia utilizada indexa grande parte dos estudos relacionados à câncer, de acordo com as bases do National Institute of Cancer (USA), o que converge com o escopo deste trabalho, o qual visa fornecer uma visão abrangente sobre a produção registrada na plataforma Lattes dos especialistas em oncologia brasileiros com atuação nos mais variados ramos da Oncologia.

Redes de colaboração entre pesquisadores de destaque

A colaboração acadêmica entre os pesquisadores selecionados foi identificada baseada nas publicações realizadas em coautoria entre eles. No presente estudo, o gráfico da Figura 6, apresenta a rede de colaboração entre os pesquisadores de maior destaque em artigos, livros e capítulos de livros, e foi produzido com o auxílio do software gratuito e de código aberto Gephi, que permite a visualização e exploração para todos os tipos de gráficos e redes Bastian, Heymann, e Jacomy (2009).

Na **Figura 6** é apresentada a rede de colaboração, no qual os vértices (21) representam os pesquisadores e as arestas (23) representam as ligações entre os vértices. Nesse gráfico, a espessura das arestas (linhas que concetam os vértices) é proporcional ao número de publicações em conjunto, entre os dois autores, as cores dos vértices identificam os estados do Brasil, nos quais localizam-se as instituições de pesquisa dos autores. Uma métrica muito utilizada nas análises desse tipo de gráfico é a centralidade de grau (centrality degree), que corresponde ao número de arestas incidentes ou ao número de vértices adjacentes a ele (GIORDANO, BRUNING, BORDIN, 2015). Dessa maneira, a centralidade de grau irá indicar quais são os autores que mais colaboraram, publicando conjuntamente com outros autores, levando em consideração o número de coautores que colaboraram com um determinado autor, juntamente com o número de publicações que estes fizerem em parcerias (BORDIN, GONÇALVES, TODESCO, 2014).

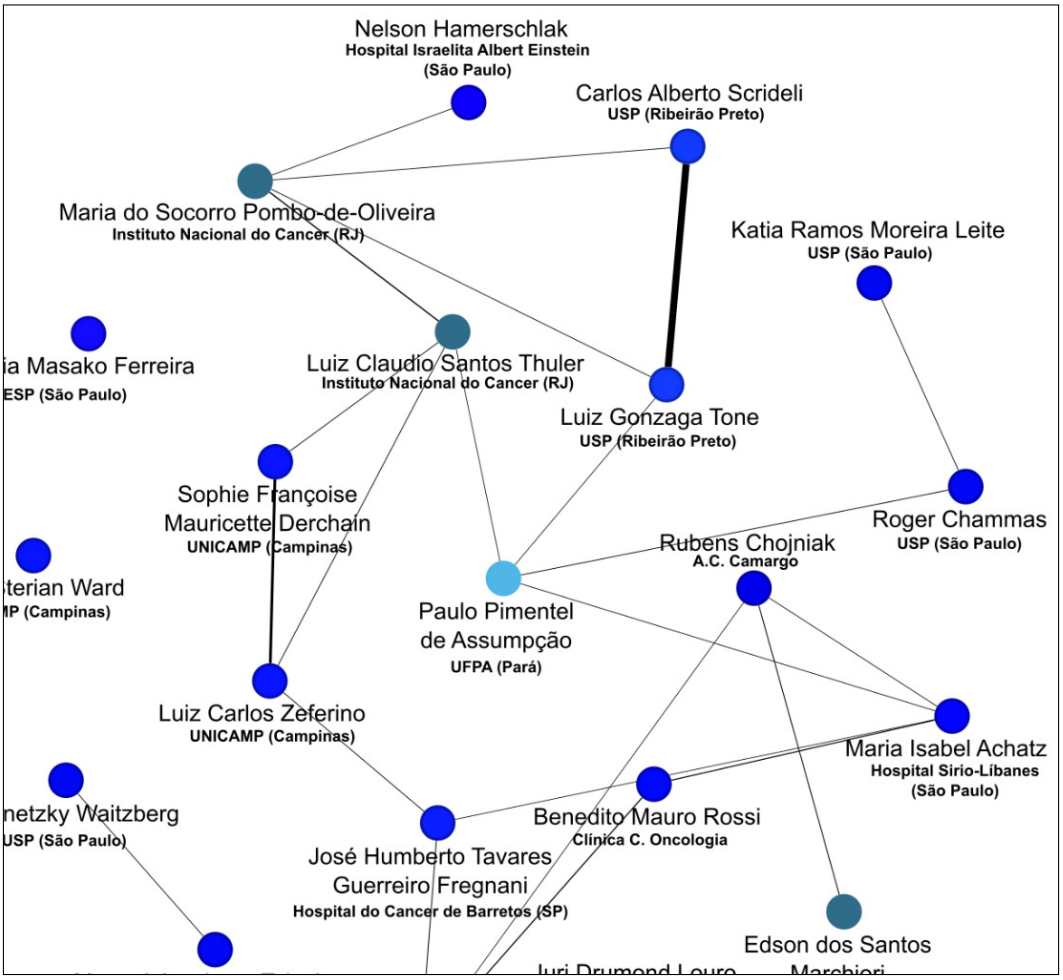

Figura 6 – Rede de colaborações entre os pesquisadores mais destacados na produção de artigos, livros e capítulos de livros no campo da oncologia
Fonte: Elaborado pelos autores, (2023).

Cinco pesquisadores de destaque no campo da oncologia, possuem centralidade de grau igual a 4, colaborando assim, com quatro outros especialistas. Estes especialistas são a Dra. Maria do Socorro Pombo-de-Oliveira, Dr. Luiz Claudio Santos Thuler, Dr. Paulo Pimentel de Assunção, Dra. Maria Isabel Achatz e Dr. José Roberto Tavares Guerreiro Fregnani. Convém destacar que estes pesquisadores estão em uma mesma rede de colaboração, havendo ligações entre eles de forma direta (coautoria direta) e indireta (cadeia de pesquisadores). A rede a qual pertence os pesquisadores com maior centralidade de grau também é a rede principal de colaboradores, abarcando a maioria dos especialistas analisados entre os de maior destaque. Os demais pesquisadores podem estar conectados a esta rede principal por meio de outros especialistas cuja posição no ranking, segundo a metodologia utilizada neste estudo, seja superior à décima posição.

CONCLUSÕES

Os resultados obtidos nesse estudo demonstraram a eficácia do software ScriptLattes em relação às análises de informações propostas, bem como à disponibilização das mesmas por meio de páginas de fácil acesso em formato HTML. Outro ponto fonte do programa é o fato de trazer as informações dos pesquisadores analisados de maneira organizada, uma vez que essas informações eram obtidas de maneira fragmentada, disponíveis apenas individualmente nos CV desses pesquisadores. Foram analisados os CV dos 134 dos pesquisadores, identificados como especialistas em oncologia, com formação de graduação e atuação profissional na subárea Medicina da Plataforma Lattes, um dos ramos da grande área Ciências da Saúde. O estudo também providenciou o ranqueamento dos pesquisadores de maior destaque no campo da oncologia. Artigos completos e Capítulos de Livros compõem mais de 90% do acervo técnico direcionado aos estudos de oncologia. Palavras-chaves generalistas, como “câncer” e “carcinoma” são as mais utilizadas na titulação dos trabalhos, além disso, as palavras leukemia”, “lymphoma”, “melanoma”, “blastoma”, “sarcoma” e “adenocarcinoma” foram as mais frequentes entre os termos ligados à tipos específicos de cânceres. A maior parte da expertise relacionada aos temas de oncologia concentra-se geograficamente na região sudeste do Brasil. A sistemática utilizada na criação de expressão de busca fornece uma nova metodologia para identificação do corpo de conhecimento, representado na figura dos pesquisadores, em uma área específica. O ScripLattes é uma ferramenta estratégica para recuperar informações acadêmicas na Plataforma Lattes em diversas áreas do conhecimento, inclusive na pesquisa sobre

o câncer. A posterior aplicação de ferramentas de análise de dados, como a plataforma KNIME *Analytic*, na base de dados criada a partir dos outputs dos ScriptLattes, permitiu a obtenção de insights profundos sobre a identificação de especialistas e sua produção no campo da oncologia.

REFERÊNCIAS

Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi: An open source software for exploring and manipulating networks. In *Proceedings of the international aaai conference on web and social media* (v. 3, p. Artigo 1).

Brito, A. G. C. d., Quoniam, L., & Mena-Chalco, J. P. (2016). Exploração da plataforma lattes por assunto: Proposta de metodologia. *Transinformação*, *28*(1), 77–86. doi: 10.1590/2318-08892016002800006

Ferraz, R. R. N., Barnabé, A. S., Quoniam, L., Santos, A. M. d., & Mariosa, D. F. (2018). Aspectos históricos da criação dos grupos de pesquisa em dengue no brasil com a utilização da ferramenta computacional scriptgp. *Ciência & Saúde Coletiva*, *23*, 837–848. doi: 10.1590/1413-81232018233.00862016

Ferraz, R. R. N., & Quoniam, L. M. (2013). A utilização da ferramenta computacional scriptlattes para avaliação das competências em pesquisa no brasil. *PRISMA.COM*, *21*, Artigo 21.

INCA. (2020). *Receitas e despesas.* https://www.gov.br/inca/pt-br/acesso-a-informacao/receitas-e-despesas.

INCA. (2022a). *Estatísticas de câncer.* https://www.gov.br/inca/pt-br/assuntos/cancer/numeros/estatisticas-de-cancer.

INCA. (2022b). *O que é câncer?* https://www.gov.br/inca/pt-br/assuntos/cancer/o-que-e-cancer/o-que-e-cancer.

Magalhães, J., Hir, M., Quoniam, L., Hartz, Z., & Oliveira, D. A. d. (2020). A management tool to aid in the tropical outbreak of the 21st century: Senior scientists and their knowledge of the triple threat dengue, zika and chikungunya. *Problems of Management in the 21st Century*, *15*(1), 40–55. doi: 10.33225/10.33225/pmc/20.15.40

Magalhães, J., Martins, M. d. R. O., & Hartz, Z. (2014). Big data em medicina tropical: Um panorama do conhecimento científico e tecnológico em malária no mundo e a contribuição de portugal. *Anais do Instituto de Higiene e Medicina Tropical*, *13*, 47–58. doi: 10.25761/anaisihmt.171

Martino, B. D., Aversa, R., Cretella, G., Esposito, A., & Kołodziej, J. (2014). Big data (lost) in the cloud. *International Journal of Big Data Intelligence*, *1*(1/2), 3. doi: 10.1504/IJBDI.2014.063840

Matz, E. L., & Hsieh, M. H. (2017). Review of advances in uroprotective agents for cyclophosphamide- and ifosfamide-induced hemorrhagic cystitis. *Urology*, *100*, 16–19. doi: 10.1016/j.urology.2016.07.030

Mena-Chalco, J., & Cesar Junior, R. (2009). Scriptlattes: An open-source knowledge extraction system from the lattes platform. *J. Braz. Comp. Soc.*, *15*, 31–39. doi: 10.1007/BF03194511

Mena-Chalco, J., & Cesar Junior, R. (2013). *Prospecção de dados acadêmicos de currículos lattes através de scriptlattes.* doi: 10.13140/RG.2.1.5183.8561

Ministério da Saúde. (2023). *Inca lança a estimativa 2023 – incidência de câncer no brasil.* https://bvsms.saude.gov.br/inca-lanca-a-estimativa-2023-incidencia-de-cancer-no-brasil/.

NIH. (2022). *A to z list of cancer types—nci (ncigloba,ncienterprise).* https://www.cancer.gov/types.

Oun, R., Moussa, Y. E., & Wheate, N. J. (2018). The side effects of platinum-based chemotherapy drugs: A review for chemists. *Dalton Transactions*, *47*(19), 6645–6653. doi: 10.1039/C8DT00838H

Ramos, M. J., Rito, P. d. N., & Vieira, V. V. (2021). Monitoramento ambiental na manipulação de medicamentos oncológicos injetáveis à luz das normativas vigentes. *Vigil Sanit Debate*, *9*(2), Artigo 2. doi: 10.22239/2317-269X.01811

Sag, A. A., Selcukbiricik, F., & Mandel, N. M. (2016). Evidence-based medical oncology and interventional radiology paradigms for liver-dominant colorectal cancer metastases. *World Journal of Gastroenterology*, *22*(11), 3127–3149. doi: 10.3748/wjg.v22.i11.3127

Webster, W. D., Parks, G. T., Titov, D., & Beasley, P. (2014). The production of radionuclides for nuclear medicine from a compact, low-energy accelerator system. *Nuclear Medicine and Biology*, *41*, e7–e15. doi: 10.1016/j.nucmedbio.2013.11.007

WHO. (2022). *Cancer.* https://www.who.int/health-topics/cancer.

Como citar este artigo (APA):
Chaves, H. K., Oliveira, A. M., Schumacher, S. O. R., Santos, R. C., Antunes, A. M. S. & Magalhães, J. L. (2025). Identificação de especialistas em câncer no Brasil em plataforma de big data: um estudo de caso minerando a Plataforma Lattes com auxílio de softwares de prospecção . *AtoZ: novas práticas em informação e conhecimento, 14*, 1 – 12. Recuperado de: http://dx.doi.org/10.5380/atoz.v14.91962

NOTAS DA OBRA E CONFORMIDADE COM A CIÊNCIA ABERTA

CONTRIBUIÇÃO DE AUTORIA

Papéis e contribuições	Henrique K. Chaves	Alessandra M. de Oliveira	Suzanne O. R. Schumacher	Rafael C. dos Santos	Adelaide M. de S. Antunes	Jorge L. de Magalhães
Concepção do manuscrito	X	X	X	X	X	X
Escrita do manuscrito	X	X	X	X	X	X
Metodologia	X	X	X	X		
Curadoria dos dados	X	X	X			
Discussão dos resultados	X	X	X	X	X	
Análise dos dados	X	X	X	X		X

FINANCIAMENTO

O(s) autor(es) declara(m) que esta pesquisa recebeu financiamento conforme dados indicados a seguir e o documento comprobatório foi anexado como documento suplementar: **Desenvolvimento Científico e Tecnológico em Saúde (FIOTEC)**